

Online Localization and Prediction of Actions and Interactions

Khurram Soomro, *Member, IEEE*, Haroon Idrees, *Member, IEEE*, and Mubarak Shah, *Fellow, IEEE*

Abstract—This paper proposes a person-centric and online approach to the challenging problem of localization and prediction of actions and interactions in videos. Typically, localization or recognition is performed in an offline manner where all the frames in the video are processed together. This prevents timely localization and prediction of actions and interactions - an important consideration for many tasks including surveillance and human-machine interaction.

In our approach, we estimate human poses at each frame and train discriminative appearance models using the superpixels inside the pose bounding boxes. Since the pose estimation per frame is inherently noisy, the conditional probability of pose hypotheses at current time-step (frame) is computed using pose estimations in the current frame and their consistency with poses in the previous frames. Next, both the superpixel and pose-based foreground likelihoods are used to infer the location of actors at each time through a Conditional Random Field enforcing spatio-temporal smoothness in color, optical flow, motion boundaries and edges among superpixels. The issue of visual drift is handled by updating the appearance models, and refining poses using motion smoothness on joint locations, in an online manner. For online prediction of action (interaction) confidences, we propose an approach based on Structural SVM that operates on short video segments, and is trained with the objective that confidence of an action or interaction increases as time progresses. Lastly, we quantify the performance of both detection and prediction together, and analyze how the prediction accuracy varies as a time function of observed action (interaction) at different levels of detection performance. Our experiments on several datasets suggest that despite using only a few frames to localize actions (interactions) at each time instant, we are able to obtain competitive results to state-of-the-art *offline* methods.

Index Terms—Action Localization, Action Prediction, Interactions, Dynamic Programming, Structural SVM

1 INTRODUCTION

Predicting *what* and *where* an action or interaction will occur is an important and challenging computer vision problem for automatic video analysis [1], [2], [3], [4]. It involves the use of limited motion information in partially observed videos for frame-by-frame localization and label prediction, and has varied applications in many areas. For human-computer or human-robot interaction, it allows the computer to automatically localize and recognize actions and gestures as they occur, or predict the intention of actors, thereafter creating appropriate responses for them. It is especially relevant to the monitoring of elderly, where detection of certain actions, e.g. *falling*, must trigger an immediate automated response and alert the care giver or a staff member. Moreover, this allows their interactions with other people to be monitored and quantified for overall well-being. In visual surveillance, online localization and prediction can be used for detecting abnormal actions such as assault or interactions of criminal nature, e.g., drug exchange and alert the human monitors in a timely manner. In automated robot navigation or autonomous driving, the timely detection of human actions in the environment will lead to requisite alteration in path or speed, e.g., a child jumping in front of the car. In this paper, we address the very problem of *Online Action and Interaction Localization*, which

aims at localizing actions (interactions) and predicting their class labels in a streaming video (see Fig. 1).

In this work, for online action (interaction) localization and prediction, 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 actor from the background. The superpixel-based model incorporates visual appearance using color and motion features, whereas the pose-based model captures the structural cues through joint locations. Using both the foreground and pose models we generate a confidence map, that is later used to locate the action segments by inferring on a Conditional Random Field in an online manner. Since the appearance of an actor changes due to articulation and camera motion, we retrain foreground model as well as impose spatio-temporal smoothness constraints on poses to maintain representation that is both robust and adaptive. As soon as the human actors are localized at the current frame, we proceed to recognize and predict the label of the action (interaction). There can be multiple approaches to perform online prediction, since the windows over which the visual features are accumulated can be defined in various ways. In [5], we used a hybrid of binary SVM and dynamic programming on short intervals to predict the class labels in an online manner. However, this requires multiple classifiers to be trained for each sub-action or segment of an action. In this paper, we present an alternate approach that uses Structural SVM, trained with the objective that the score of the action (interaction) over positive instances should increase as time progresses.

• K. Soomro, H. Idrees and M. Shah are with the Center for Research in Computer Vision (CRCV), University of Central Florida, Orlando, FL 32816. E-mail: {ksoomro,haroon,shah}@eecs.ucf.edu

Online Action Localization = Action Prediction + Detection

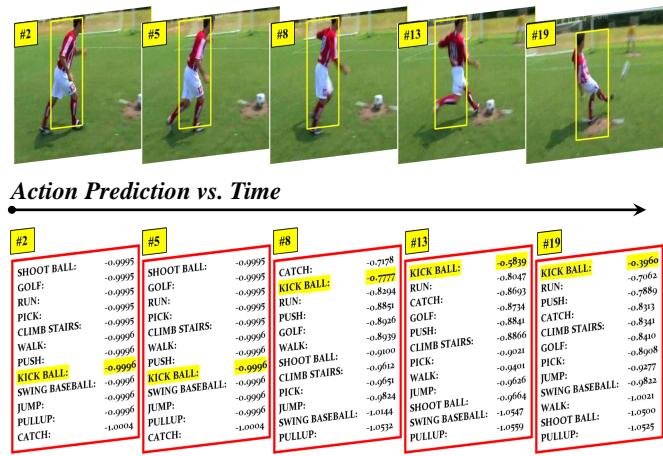


Fig. 1. This figure illustrates the problem of *Online Action Localization* that we address in this paper. The top row shows *kick ball* action being performed by a soccer player with frame number shown in top-left of each frame. The goal is to localize the actor (shown with yellow rectangles in top row) and predict the class label of the action (shown in red boxes in second row) as the video is streamed. As can be seen in bottom row, the confidence of *kick ball* action increases and comes to the top as more action is observed across time. This problem contrasts with *offline* action localization where action classification and detection is performed after the action or video clip has been observed in its entirety.

Finally, we perform rigorous experiments on four action and two interaction datasets, and introduce measures for consistent evaluation across both actions and interactions.

Existing *offline* action localization methods [1], [2], [3], [6], [7], [8] classify and localize actions after completely observing an entire video sequence. The goal is to localize an action by finding the volume that encompasses an entire action. Some approaches are based on sliding-windows [6], [9], while others segment the video into supervoxels which are merged into action proposals [7], [10], [8]. The action proposals from either methods are then labeled using a classifier. Essentially, an action segment is classified *after* the entire action volume has been localized. Similarly, the videos are processed for classification [11], [12], retrieval [13], [14] or localization [15] in an offline manner for the case of interactions. Since offline methods have entire video and action segments at their disposal, they can take advantage of observing entire motion of action instances, and for practical purposes do not provide action detection in a timely manner. Similarly, there have been recent efforts to predict activities by early recognition [16], [17], [18], [19]. However, these methods only attempt to predict the label of the action without any localization. Thus, the important question about *where* an action is being performed remains unanswered, which we tackle in this work.

In summary, our contributions in this paper can be summarized as follows: 1) We address the problem of *Online Action and Interaction Localization* in streaming videos, 2) by using high-level pose estimation to learn mid-level superpixel-based foreground models at each time instant. 3) We employ spatio-temporal smoothness constraints on joint locations in human poses to obtain stable and robust action segments in an online manner. 4) The label and confi-

dences for action (interactions) segments are *predicted* using Structural SVM trained on partial action clips, which enforces the constraint that the confidence of positive samples increases monotonically over time. Finally, 5) we introduce an evaluation measure to quantify performance of action (interaction) prediction and online localization and perform experiments on six action and interaction datasets with a consistent evaluation framework.

Compared to our CVPR 2016 paper [5], we extend our approach to interactions and perform experiments on three additional datasets. Moreover, in contrast to the Binary-SVM and dynamic programming hybrid for online prediction, we employ Structural SVM formulation which requires one classifier per action and is computationally efficient. Furthermore, we also introduce a unified framework of evaluation for actions and interactions. The rest of the paper is organized as follows. In Sec. 2 we review literature relevant, whereas Sections 3 and 4 cover the technical details of our approach. We report results in Sec. 5 and conclude with suggestions for future work in Sec. 6.

2 RELATED WORK

Online 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 and their confidences. Li and Fu [16] predict human activities by mining sequence patterns, and modeling causal relationships between them. Zhao *et al.* [20] 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 [21] simulate the sequential arrival of data while training, and train detectors to recognize incomplete events. Similarly, Lan *et al.* [17] propose hierarchical ‘movemes’ to describe human movements and develop a max-margin learning framework for future action prediction.

Ryoo [18] proposed integral and dynamic bag-of-words for activity prediction, and divide the training and testing videos into small segments and match the segments sequentially. In follow-up work, Ryoo and Aggarwal [15] treat interacting people as a group and recognize interactions in continuous videos by computing group motion similarities. Similarly, Kong *et al.* [19] 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.* [22] 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.* [23] 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. For the case of interactions, Huang and Katani [24] predict the reaction in a two-person setting by modeling it as an optimal control problem. Recently, there have been works on online temporal detection [25], [26] without localization. In contrast to these works, we perform both action prediction and localization in an online manner.

Offline Localization has received significant attention in the past few years, both for actions [27], [28], [29], [30] as well as interactions [31], [32]. For actions, the first category of approaches uses either rectangular tubes or cuboid-based representations. Lan *et al.* [1] treated the human position as a latent variable, which is inferred simultaneously while localizing an action. Yuan *et al.* [33] used branch-and-bound with dynamic programming, while Zhou *et al.* [34] used a split-and-merge algorithm to obtain action segments that are then classified with LatentSVM [35]. Oneata *et al.* [9] presented an approximation to Fisher Vectors for tractable action localization. Tran and Yuan [36] used Structural SVM to localize actions with inference performed using Max-Path search method. Ma *et al.* [37] automatically discovered spatio-temporal root and part filters, whereas Tian *et al.* [6] developed Spatio-temporal Deformable Parts Model [35] to detect actions in videos and can handle deformities in parts, both in space and time. Recently, Yu and Yuan [38] 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 [10], [7]. Jain *et al.* [7] recently proposed a method that extends selective search approach [39] to videos. They merge supervoxels using appearance and motion costs and produce multiple layers of segmentation for each video. Gkioxari and Malik [40] use selective search [39] 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 [41], [38], which yield fewer regions of interest in videos. For interaction recognition in videos, Kong *et al.* [12], [42] learn high-level descriptions called interactive phrases to express binary semantic motion relationships between interacting people. A hierarchical model is used to encode interactive phrases based on latent SVM framework where interactive phrases are treated as latent variables. Wu *et al.* [43] also decompose interaction video segments into spatial cells and learn relationship between them. Similar to these methods, our approach can delineate contours of actions and interactions, but with the goal of performing prediction and localization in a streaming fashion.

There are recent works for temporal detection using neural networks with reinforcement learning [44] or multi-stage CNNs [45] which only localize the actions temporally. Others use sparse representation to find temporal action proposals [46], or statistical language models [47] to temporally localize actions in videos. In contrast, we localize actions both temporally and spatially.

Pose for Recognition. Low-level motion features, both hand-crafted [48] and deep learned [49], have imparted significant gains to the performance of action recognition and localization algorithms. However, human actions inherently consists of articulation which low-level features cannot model explicitly. The compact and low-dimensional nature of high-level representations such as human poses might make them sensitive and unstable for the task of action localization and recognition. Nonetheless, human

pose estimation has been successfully employed for action recognition in several works. For instance, Majiwa *et al.* [50] implicitly capture poses through ‘poselet activation vector’ and later use them for action recognition in static images. Xu *et al.* [51] detect poses through [52] and couple them with independently computed local motion features around the joints for action recognition. Wang *et al.* [53] also extended [52] to videos and represented videos in terms of spatio-temporal configurations of joints to perform action recognition. Raptis and Cigal [54] recognize and detect interactions from videos by modeling poselets as latent variables in a structural SVM formulation. Joint recognition of action and pose estimation in videos was recently proposed by Xiaohan *et al.* [55]. They divide the action into poses and their spatio-temporal parts, and model their inter-relationships through And-Or graphs. Pirsiavash *et al.* [56] 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.* [2], who detect actions using a unified approach that discovers action parts using dynamical poselets, and the relations between them.

Similarly, several works model and determine head orientation and upper body pose for recognition and localization of interactions. Patron-Perez *et al.* [13] developed a per-person descriptor which incorporates head orientation and the local spatio-temporal context around each person to detect interactions. Vahdat *et al.* [57] represented each individual by a set of key poses and formulated spatio-temporal relationships among them in their model. The frame-wise interaction model in Patron-Perez *et al.* [14] combines local and global descriptors and incorporates visual attention of people by modeling their head orientations. Although Hoai and Zisserman [11] do not detect poses per se, they develop a technique to detect different upper body configurations each consisting of multiple parts. In contrast to these methods, we use pose in conjunction with low-level features and mid-level superpixels to predict and localize actions (interactions) in an online manner. Our work is at the cross roads of both online prediction and offline localization, in a unified framework for both actions and interactions operable in partially observed videos.

3 ONLINE LOCALIZATION OF ACTIONS AND INTERACTIONS

The pipeline of our approach for localization is shown in Fig. 2. Given a testing video, we initialize the localization algorithm with several pose estimations in individual frames and refine the poses using multiple spatio-temporal constraints from previous frames. Next, we segment the testing video frames into superpixels. The features computed within each superpixel are used to learn a superpixel-based appearance model, which distinguishes the foreground from the background by training a discriminative classifier with superpixels within each pose bounding box as foreground and the rest of superpixels as background. Simultaneously, the conditional probability of pose hypotheses at current time-step (frame) is computed using pose confidences and consistency with poses estimated

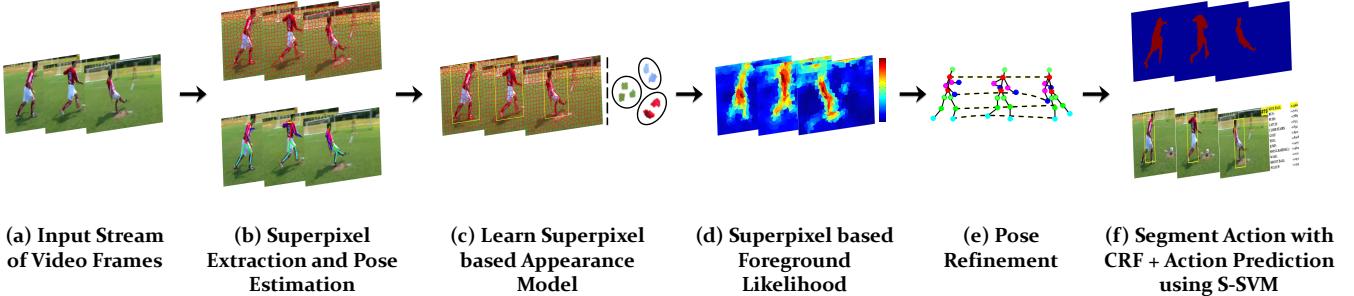


Fig. 2. This figure shows the framework of the approach proposed in this paper. (a) Given an input video, (b) we over-segment each frame into superpixels and detect poses using an off-the-shelf method [58]. (c) An appearance model is learned using all the superpixels inside a pose bounding box as positive, 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 at each frame, on which actions (interactions) are predicted through Structural SVM.

in previous frames. The superpixel and pose-based foreground probability is used to infer the location of actors at each frame through a Conditional Random Field enforcing spatio-temporal smoothness in color, optical flow, motion boundaries and edges among superpixels. After localizing actions (interactions) at each time-step (frame), we refine poses by imposing consistency in locations and appearance of joints as well as scale of poses. Once the pose has been estimated and refined at current time-step, the superpixel-based appearance model is updated to avoid visual drift. This process is repeated for every frame in an online manner (see Fig. 2) and gives human localization at every frame. After localization, the spatio-temporal tubes are then used for prediction and recognition of labels at each frame, discussed later in Sec. 4. Thus, the pose estimation not only provides initialization for the proposed discriminative appearance models, as it is more robust compared to human detection in action (interaction) videos due to articulation, it also allows computation of pose features which we use during label prediction (Sec. 4). Note that the pose estimations can consist of any or multiple body configurations such as upper or full body, as well as multiple humans interacting or performing actions. To simplify the treatment in this section, we assume we are dealing with a single actor or action, without loss of generality.

Let s_t represent a superpixel by its centroid in frame t and p_t represent one of the poses in frame t . Since our goal is to localize the actor 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 at that time instant. 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:

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

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 streaming

mode. The state which maximizes the posterior (MAP) estimate in Eq. 1 is selected as the new state. An implication of Eq. 1 is that the state or localization cannot be altered in the past frames, which makes online localization different from the existing offline methods. 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 (interactions) from the background. Given foreground and background superpixels in the previous frames $t - \delta, \dots, t - 1$, we group them into $k = 1 \dots K$ clusters. Furthermore, let ζ_k define the ratio of foreground to background superpixels for the k th cluster through k-means. Then, the appearance-based foreground score using color, ϕ_{color} , and flow, ϕ_{flow} , features in the superpixels is given by:

$$H_{fg}(\mathbf{s}_t) = \exp\left(\frac{\|\phi_{color}(\mathbf{s}_t) - \mathbf{q}_k\|}{r_k}\right) \cdot \zeta_k \\ + \exp\left(\frac{\|\phi_{flow}(\mathbf{s}_t) - \boldsymbol{\mu}_k\|}{\rho_k}\right), \quad (2)$$

where \mathbf{q}_k and r_k are the cluster center and radius, respectively, whereas $\boldsymbol{\mu}_k$ and ρ_k represent the mean and variance of optical flow for the k th 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 superpixels within a foreground bounding box are inevitably considered as foreground initially, however the later segmentation through Conditional Random Field serves to alleviate this problem by separating foreground superpixels within the bounding box localization. The ζ_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: $p(\mathbf{S}_t | \mathbf{X}_t) = \alpha_{fg} \cdot H_{fg}(\mathbf{s}_t)$, where α_{fg} is the normalization factor.

3.2 Pose-based Foreground Likelihood

We represent each pose p_t graphically with a tree, given by $\mathbf{T} = (\Pi, \Lambda)$. The body joints $\pi \in \Pi$ are based on

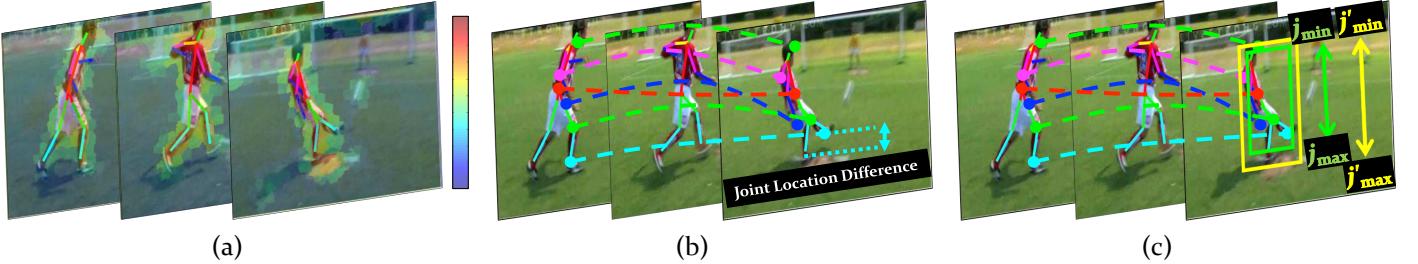


Fig. 3. This figure shows a visualization of the joint smoothness costs used in pose-based foreground likelihood for (a) appearance smoothness of joints (J_{app}), (b) location smoothness of joints (J_{loc}) and (c) scale smoothness of joints (J_{sc}).

appearance connected by $\lambda \in \Lambda$ edges capturing the structure of the pose. The joint j with its location in pose \mathbf{p}_t is represented by π_t^j , consisting of its x and y locations. Then, the raw cost (or negated detection score) for a particular pose \mathbf{p}_t is the sum of appearance and deformation costs:

$$H_{\text{raw}}(\mathbf{p}_t) = \sum_{j \in \Pi_t} \psi(\pi_t^j) + \sum_{(j, j') \in \Lambda_t} \chi(\pi_t^j, \pi_t^{j'}), \quad (3)$$

where ψ and χ are linear functions of appearance features of pose joints, and the relative joint displacements (deformations) w.r.t each other. We use a pre-trained pose detector to obtain pose hypotheses in each frame. In [5], we used Flexible Mixture-of-Parts [52] for pose estimation, which optimizes over latent variables that capture different joint locations and pose configurations. In this paper, we report results using Convolutional Pose Machines (CPM) [58] which uses deep learning. For CPM, the deformation costs are embedded within joint costs in Eq. 3. Since the pose estimation in both methods works on individual frames, it is inherently noisy and does not take into account the temporal information available in videos. We impose the following smoothness constraints (as shown in Fig. 3 (a-c)) in the previous δ frames to re-evaluate poses in Eq. 3 for the current time-step.

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:

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

where \hat{s}_t^j is the enclosing superpixel of the joint π_t^j .

Location Smoothness of Joints: Since human motion is naturally smooth, we ensure that displacements in joint locations over time are small. This is achieved by fitting a 2D spline using piecewise polynomials to each joint j on the past δ frames, γ_t^j . Then the location smoothness cost over all joints is given by:

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

Scale Smoothness of Joints: Let j_{\min} and j_{\max} denote the vertical minimum and maximum for all the splines $\gamma_\tau, \forall \tau \in$

$\{t-\delta, \dots, t\}$, i.e., the y -axis components of the bounding box circumscribing all the splines fitted on joints. Furthermore, let j'_{\min}, j'_{\max} denote minimum and maximum for joints in actual poses $\pi_t \in \Pi_t$. Then, the scale smoothness cost essentially computes the overlap between the two heights:

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

The combined cost of a particular pose is defined as its raw cost plus the smoothness costs across space and time, i.e.,

$$H_{\text{pose}}(\mathbf{p}_t) = H_{\text{raw}}(\mathbf{p}_t) + J_{\text{app}}(\mathbf{p}_t) + J_{\text{loc}}(\mathbf{p}_t) + J_{\text{sc}}(\mathbf{p}_t). \quad (7)$$

The change in pose and appearance of an actor may cause visual drift. Similar to Sec. 3.1, we use a temporal window of past δ frames to refine the pose locations. This helps in better prediction of the highly probable foreground locations in current frame. We propose an iterative approach to select poses in the past $\{t-\delta, \dots, t\}$ frames. Given an initial set of poses, we fit a spline to each joint π_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:

$$(*\mathbf{p}_{t-\delta}, \dots, *\mathbf{p}_t) = \arg \min_{\mathbf{p}_{t-\delta}, \dots, \mathbf{p}_t} \sum_{\tau=t-\delta}^t \left(H_{\text{pose}}(\mathbf{p}_\tau) \right). \quad (8)$$

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, such that the joints are spatially accurate and temporally consistent with the motion of the action. This procedure is summarized in Algorithm 1. Note that the poses in previous frames of the batch are only refined simultaneously, however, the pose at current time step is used by the algorithm. 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 Actor Segmentation using Conditional Random Fields (CRF)

Once we have the superpixel and pose-based foreground likelihoods in Eq. 1, we proceed to infer the action segment and its contour 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 with superpixels as nodes connected through

Algorithm 1 : Algorithm to refine pose locations in a batch of frames in Q iterations.

Input: $\mathbf{P}_{t-\delta}, \dots, \mathbf{P}_t$
Output: ${}^*\mathbf{P}_{t-\delta}, \dots, {}^*\mathbf{P}_t$

```

1: procedure REFINEPOSES()
2:   for  $\tau = t - \delta$  to  $t$  do
3:      ${}^*\mathbf{p}_\tau = \arg \min(H_{\text{raw}}(\mathbf{p}_\tau))$ 
4:   end for
5:   for  $n = 1$  to  $Q$  do
6:     Fit a spline  $\gamma^j$  to each joint using locations
7:      $[{}^*\pi_{t-\delta}^j, \dots, {}^*\pi_t^j]$ 
8:     Compute  $J_{\text{app}}(\mathbf{p}_t)$  using Eq. 4
9:     Compute  $J_{\text{loc}}(\mathbf{p}_t)$  using Eq. 5
10:    Compute  $J_{\text{sc}}(\mathbf{p}_t)$  using Eq. 6
11:    Find  $({}^*\mathbf{p}_{t-\delta}, \dots, {}^*\mathbf{p}_t)$  through Eq. 8.
12:   end for
13: end procedure

```

spatial and *temporal* edges. Let variable a denote the foreground/background label of a superpixel. Then, the objective function of CRF becomes:

$$\begin{aligned}
& -\log(p(a_{t-\delta}, \dots, a_t | \mathbf{s}_{t-\delta}, \dots, \mathbf{s}_t, \mathbf{p}_{t-\delta}, \dots, \mathbf{p}_t)) \\
&= \sum_{\tau=t-\delta}^t \left(\underbrace{\Theta(a_\tau | \mathbf{s}_\tau, \mathbf{p}_\tau)}_{\text{unary potential}} + \underbrace{\Upsilon(a_\tau, a'_\tau | \mathbf{s}_\tau, \mathbf{s}'_\tau)}_{\text{spatial smoothness}} \right) \\
&\quad + \sum_{\tau=t-\delta}^{t-1} \underbrace{\Gamma(a_\tau, a'_{\tau+1} | \mathbf{s}_\tau, \mathbf{s}'_{\tau+1})}_{\text{temporal smoothness}}, \quad (9)
\end{aligned}$$

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

$$\Theta(a_\tau | \mathbf{s}_\tau, \mathbf{p}_\tau) = \alpha_{\text{fg}} H_{\text{fg}}(\mathbf{s}_\tau) + \alpha_{\text{pose}} H_{\text{pose}}(\mathbf{p}_\tau), \quad (10)$$

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

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

and

$$\begin{aligned}
& \Gamma(a_\tau, a'_{\tau-1} | \mathbf{s}_\tau, \mathbf{s}'_{\tau-1}) = \beta_{\text{col}} d_{\text{col}}(\mathbf{s}_\tau, \mathbf{s}'_{\tau-1}) \\
&\quad + \beta_{\text{hof}} d_{\text{hof}}(\mathbf{s}_\tau, \mathbf{s}'_{\tau-1}) + \beta_\mu d_\mu(\mathbf{s}_\tau, \mathbf{s}'_{\tau-1}), \quad (12)
\end{aligned}$$

respectively. In Eqs. 11, and 12, $\beta_{\text{col}} d_{\text{col}}(\cdot)$ is the cost of color features in HSI color space, $\beta_{\text{hof}} d_{\text{hof}}(\cdot)$ and $\beta_\mu d_\mu(\cdot)$ compute compatibility between histogram of optical flow and mean of optical flow magnitude of the two superpixels, respectively. Similarly, $\beta_{\text{mb}} d_{\text{mb}}(\cdot)$ and $\beta_{\text{edge}} d_{\text{edge}}(\cdot)$ quantify incompatibility between superpixels with prominent boundaries.

4 ONLINE PREDICTION OF ACTIONS AND INTERACTIONS

For online recognition and class-label prediction of actions (interactions) in streaming videos, the classifier has to be

applied on-the-fly on short temporal intervals. In particular, training videos of an action (interaction) class c are divided into M clips of equal-sized interval Ω . The average length of each segment is saved as prior information, which during testing allows us to compute features in intervals of the desired length. Next, we present a baseline approach using Support Vector Machine and Dynamic Programming hybrid (from our CVPR 2016 paper [5]) which divides videos into short segments, and trains a classifier independently for each segment. The online update of action confidences is achieved through dynamic programming on segment scores. In this paper, we present an alternate approach which makes structured prediction by training a single classifier per action and modeling temporal dependence between action segments. In this section, we present the formulation in terms of linear classifiers for simplicity, however, in practice we used SVM with histogram intersection kernel.

Let m index over temporal segments, i.e., $m \in 1, \dots, M$ and $\mathbf{x}_{i,m}$ denote the m th segment as well as its feature vector in video i . Next, we present the two approaches we use to recognize and predict the class label at time step t of a testing video.

4.1 Binary SVMs with Dynamic Programming Inference (DP-SVM)

First, we present a baseline for online prediction [5] in our localization framework. For training binary SVMs for segments in an action (interaction) class c , we assume availability of N trimmed positive and negative training videos. For linear SVM, we obtain a single weight vector \mathbf{w}_m per segment by optimizing the following objective function,

$$\begin{aligned}
\min & \quad \frac{1}{2} \|\mathbf{w}_m\|^2 + C \sum_{i=1}^N \sum_{m=1}^M \xi_{i,m} \\
\text{s.t.} & \quad \mathbf{y}_{i,m} \langle \mathbf{w}_m, \mathbf{x}_{i,m} \rangle \geq 1 - \xi_{i,m}, \quad \xi_{i,m} \geq 0, \quad \forall i, m \quad (13)
\end{aligned}$$

where C controls the trade-off between regularizer and constraints, and $\mathbf{y}_{i,m} = 1$, for desired m if $i \in c$ and -1 , otherwise. Effectively, the training videos are divided into short intervals and an SVM is trained for each interval m independently. While testing on videos, the classification is performed on features accumulated on interval lengths learned from training videos. To exploit and preserve the sequential information present in videos, this is followed by dynamic programming on the short interval clips. 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 class, and gives the confidence for that class. This shares resemblance to Dynamic Bag-Of-Words [18] which used RBF function to compute score between training and testing segments, and applied it on trimmed videos.

Let $F(t, z)$ be the result of dynamic programming at time t assuming the current interval is z for a particular class. The result of applying classifier on testing video o on features computed between $t - \Omega$ and t is given by $\sigma(\langle \mathbf{w}_m, \mathbf{x}_{o,t} \rangle)$, where σ is the sigmoid function. If the testing video is

trimmed, then $F(t, z)$ is computed using the following recursion:

$$F(t, z) = \max_m F(t - \Omega, z - m) \cdot \sigma(\langle \mathbf{w}_m, \mathbf{x}_{o,t} \rangle). \quad (14)$$

At each time instant, the maximum value at time t gives the desired confidence for the class under consideration.

4.2 Structural SVM (S-SVM)

Ideally the prediction confidence for the correct class should increase as more action (interaction) in the video is observed over time. There is rich structure that can be derived from division of actions into sub-actions, and modeling the spatio-temporal dependence between them. Given testing video segments, we then apply Structural SVM detector to each segment of the test video. For this case, we redefine intervals w.r.t start time of an action (interaction), i.e., the start time of interval m is 0 of the trimmed training video. We set the problem in a Structural Support Vector Machine (S-SVM) with margin re-scaling construction, given by:

$$\begin{aligned} \min \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{N*M} \xi_i \\ \text{s.t.} \quad & \langle \mathbf{w}, \Psi_i(\mathbf{x}_i, \mathbf{y}_i) - \Psi_i(\mathbf{x}_i, \mathbf{y}) \rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i, \\ & \forall \mathbf{y} \in \mathcal{Y} \setminus \mathbf{y}_i, \xi_i \geq 0, \forall i \end{aligned} \quad (15)$$

where the joint feature map for input and output is given by $\Psi(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \text{sign}(\mathbf{y})$, and $\mathcal{Y} = \{-1, 1, \dots, M\}$ is the set of all labels. In Eq. 15, ξ represents the slack variables for the soft-margin SVM, which optimizes over the learned weight vector \mathbf{w} and the slack variables ξ . The constraint with the loss function $\Delta(\mathbf{y}_i, \mathbf{y})$ ensures that the score with the correct label \mathbf{y}_i is greater than alternate labels. Since the number of constraints can be tremendous, only subset of constraints are used during the optimization. For each training sample, the label \mathbf{y} which maximizes $\langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle + \Delta(\mathbf{y}_i, \mathbf{y})$ is found and the constraint which maximizes this loss is added into the subset, known as the most violated constraint. For both actions and interactions, the temporal component of the loss is defined as:

$$\Delta(\mathbf{y}_i, \mathbf{y}_{i'}) = \begin{cases} |\mathbf{y}_i - \mathbf{y}_{i'}|, & i \in c \wedge i' \in c \\ M + \epsilon, & i \in c \wedge i' \notin c \\ \epsilon, & \text{otherwise.} \end{cases} \quad (16)$$

The above loss function ensures that the confidence increases as the action (interaction) happens in the testing video, i.e., the evaluation during a positive test instance, possibly over a long video, yields a unique signature of confidence values that increases over time. This approach results in one S-SVM per class, and can be applied indiscriminately to untrimmed videos. For interactions an additional loss captures the relationship between actors. Once the weight vector \mathbf{w} has been learned, the score for a clip in the testing video is computed using $\arg \max_{\mathbf{y} \in \mathcal{Y}} \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}) \rangle$.

Note that the performance of detection or prediction for action (interaction) localization depends on the quality of localized tubes / cuboids, as the classifiers are only evaluated on such video segments. This is in contrast to previous prediction methods [18], [16], [19], [21] which do not spatially localize the actions (interactions).

5 EXPERIMENTS

We evaluate our *online action localization* approach on six challenging datasets: 1) JHMDB, 2) Sub-JHMDB, 3) MSR-II, 4) UCF Sports, 5) TV Human Interaction and 6) UT Interaction datasets. We provide details for the experimental setup followed by the performance evaluation and analysis of the proposed approach.

Features: For each frame of the testing video we extract superpixels using SLIC [59]. This is followed by extraction of color features (HSI) for each superpixel in the frame, as well as improved Dense Trajectory features (iDFT: HOG, HOF, MBH, Traj) [48] within the streamed volumes of the video. Each superpixel descriptor has a length of 512 and we set $K = 20$. The pose detections are obtained using [58] and pose features using [60]. 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 [1], we show ROC @ 20% overlap. Furthermore, Area Under the Curve (AUC) of ROC at various thresholds gives an overall measure of performance. The proposed evaluation metrics are computed over all action and interaction datasets for consistency. For MSR-II dataset, we also report results using Precision and Recall curves typically used for this dataset.

Inspired from early action recognition and prediction works [18], we also quantify the performance as a function of *Observation Percentage* of actions (interactions). For this evaluation method, the localization and classification for testing videos are sampled at different percentages of observed video/action (0, 0.1, 0.2, ..., 1). The ROC curve is computed at multiple overlap thresholds, and AUC is computed under ROC curves at respective thresholds. In the case of untrimmed videos, we evaluate the prediction accuracy as a function of observation percentage within the temporal boundaries of actions (interactions).

Note that, in online action (interaction) localization, the prediction and localization is performed instantaneously at each frame in a streaming video, therefore once locations are detected and predictions are made, retroactive modifications or changes to results are not possible.

5.1 Datasets

JHMDB Dataset: The JHMDB [60] dataset is a subset of the larger HMDB51 [61] 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 [40]. We use a codebook size of $K = 4000$ to train SVMs using iDFT features.

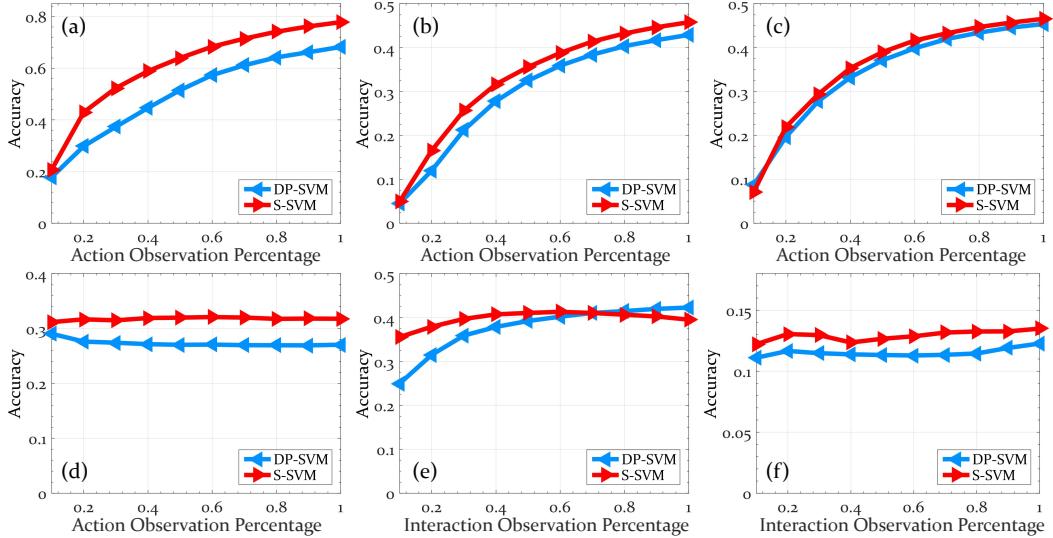


Fig. 4. This figure shows action prediction performance as a function of observed percentage of action or interaction for (a) UCF Sports, (b) JHMDB, (c) sub-JHMDB, (d) MSR-II, (e) TV Human Interaction and (f) UT Interaction datasets. Prediction performance by the baseline Binary SVM with Dynamic Programming approach is shown in blue, and that of Structural SVM with the red curve.

sub-JHMDB Dataset: The sub-JHMDB dataset has all human body joints visible in each frame. It contains a total of 316 videos over 12 action classes: *catch*, *climb stairs*, *golf*, *kick ball*, etc. The presence of the entire human within each frame makes it more challenging to recognize and localize the actions as compared to JHMDB dataset, due to high articulation of human body joints and complex variations in appearance and motion. A codebook size of $K = 4000$ was used for iDFTF, and SVMs were trained with a bag-of-words representation inside the ground truth action volumes.

UCF-Sports Dataset: The UCF Sports [62], [63] dataset is collected from broadcast television channels and consists of 150 videos. It includes a total of 10 action classes: *diving*, *golf swing*, *kicking*, *lifting*, *riding horse*, *skateboarding*, etc. Videos are captured in a realistic setting with intra-class variations, camera motion, background clutter, scale and viewpoint changes. We evaluated our method using the methodology proposed by [1], who use a train-test split with intersection-over-union criterion at an overlap of 20%. To train SVM, we use a codebook size of $K = 1000$ on iDFTFs using all the training videos.

MSR-II Dataset: The MSR-II dataset [33] consists of 54 untrimmed videos and 3 action classes: Boxing, Handclapping and Handwaving. We follow the experimental methodology of [33], having cross-dataset evaluation, where KTH [64] dataset is used for training and testing is performed on MSR-II dataset. A codebook size of $K = 1000$ was used to train SVM on iDFTFs. We show quantitative comparison using Precision-Recall curves with state-of-the-art *offline* methods. However, for uniformity with other datasets we also report results using ROC and AUC curves.

TV Human Interaction (TVHI): The TVHI dataset [13], [14] is collected from 23 different TV shows and is composed of 300 untrimmed videos. It includes 4 interaction classes: *hand shake*, *high five*, *hug* and *kiss*, with 50 videos each. It also contains a negative class with 100 videos, that have none of the listed interactions. The videos have varying number of

actors in each scene, different scales and abrupt changes in camera viewpoint at shot boundaries. For our experiments we only use the 4 interaction classes (excluding negative class) for interaction localization. We use the suggested experimental setup of two train/test splits. The localization performance is reported using ROC and AUC curves.

UT Interaction: The UT Interaction dataset [65], [15] contains untrimmed videos of 6 interaction classes: *handshaking*, *hugging*, *kicking*, *pointing*, *punching*, and *pushing*. Similar to [14], we add *being kicked*, *being punched* and *being pushed* as interactions. The dataset consists of two sets, where each set has 10 video sequences and each sequence having at least one execution per interaction. Videos involve camera jitter with varying background, scale and illumination. We follow the recommended experimental setup by using 10-fold leave-one-out cross validation per set. That is, within each set we leave one sequence for testing and use remaining 9 for training. We report the average localization performance of the proposed approach using ROC @ 20% overlap and AUC curves.

5.2 Results and Analysis

Action (Interaction) Prediction with Time: The prediction accuracy is evaluated with respect to the percentage of action (interaction) observed. Fig. 4 shows the accuracy against time for (a) UCF Sports, (b) JHMDB, (c) sub-JHMDB, (d) MSR-II, (e) TV Human Interaction and (f) UT Interaction datasets. The results show that Structural SVM in general performs better than Binary SVM with Dynamic Programming as it learns to predict higher confidence as more action is observed. 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. However, the curves for MSR-II (Fig. 4(e)) and UT Interaction (Fig. 4(g)) datasets do not reflect noticeable change as more action (interaction) is observed. This is partially due to the reason that both

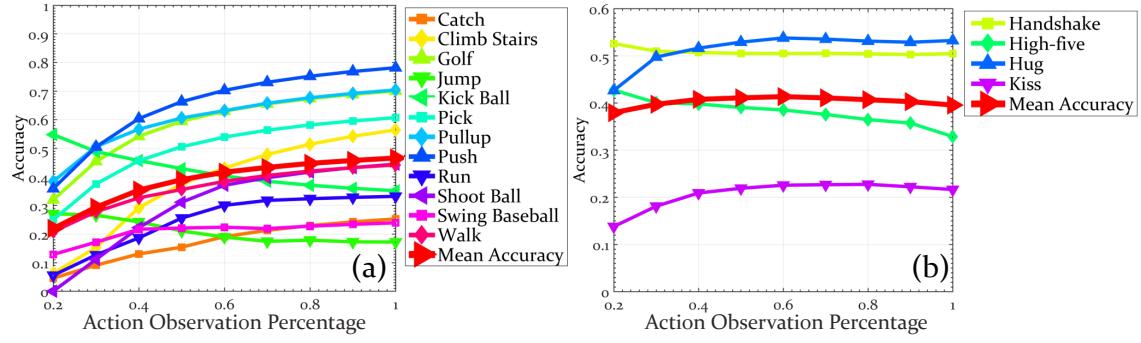


Fig. 5. This figure shows per-action prediction accuracy as a function of observed action (interaction) percentage for (a) sub-JHMDB and (b) TV Human Interaction datasets. The mean accuracy for all actions (interactions) is shown with bold red curve.

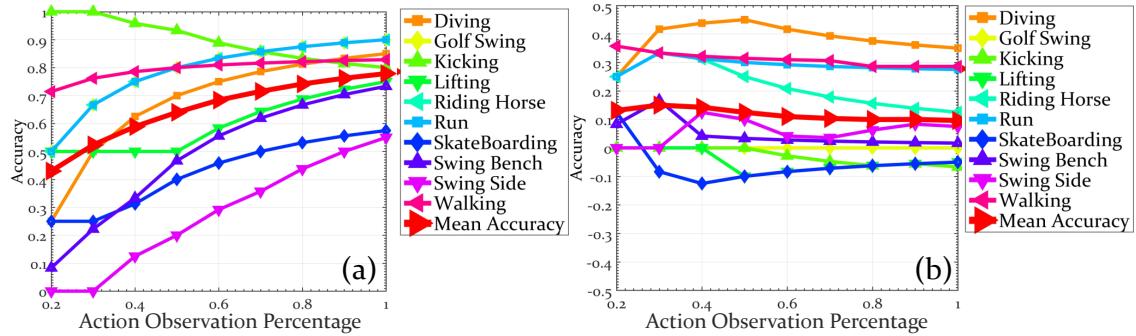


Fig. 6. This figure shows per-action prediction accuracy as a function of observed action percentage for UCF Sports dataset for (a) Structural SVM approach (Sec. 4.2) and (b) and its difference with SVM and Dynamic Programming (Sec. 4.1). On average, S-SVM outperforms DP-SVM.

TABLE 1

This table shows the percentage of video observation required to achieve a prediction accuracy of 30%. Results in the first two rows are from JHMDB, then from sub-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.

JHMDB Actions	Shoot Ball	Shoot Gun	Pull up	Golf	Clap	Climb Stairs	Shoot Bow	Brush Hair	Pour	Push	Walk	
Video (%)	1%	1%	16%	19%	25%	26%	28%	32%	32%	36%	36%	
JHMDB Actions	Swing											
Video (%)	40%	40%	48%	60%	Catch	Jump	Pick	Kick Ball	Throw	Wave		
sub-JHMDB Actions	Kick Ball	Pullup	Golf	Push	Walk	Pick	Climb Stairs	Shoot Ball	Run	Catch	Jump	Swing Baseball
Video (%)	1%	17%	18%	18%	20%	24%	41%	48%	60%	-	-	-
UCF Sports Actions	Kicking	Lifting	Walking	Golf	Riding Horse	Run	Diving	Swing Bench	Skate Boarding	Swing Side		
Video (%)	1%	1%	1%	15%	15%	15%	22%	36%	37%	61%		

these datasets have very few classes (3 and 6, respectively), and there is little confusion among classes from the onset of actions. An analysis of prediction accuracy per action class is shown in Fig. 5 for (a) sub-JHMDB and (b) TV Human Interaction datasets. Similarly Fig. 6(a) shows per-action results for UCF Sports. A common theme among the results of all the datasets is that actions which have actors in upright standing position are always easy to predict and localize compared to other actions. This is also visible from the curves of *kicking* (UCF Sports), *kick ball* (sub-JHMDB) and *hand shake*, *high five* (TV Human Interaction) which begin with a high prediction accuracy and drop slightly as

observation time period progresses, thus suggesting strong bias of classifier towards such actions (interactions). For sub-JHMDB, high prediction accuracy actions include *push* and *pull up*, both of which have humans in upright position making pose estimation easy, whereas *jump* is the most difficult action to predict. An inspection of videos for this action reveals that most of the instances were taken from parkour exhibiting high articulation and intra-class variation. For TV Human Interaction dataset, *hug* is easy to predict whereas *kiss* is the most difficult due to its subtle motion and high confusion with *hug*. For UCF Sports, high performing actions are *kicking*, *walking* and *running*, all upright with

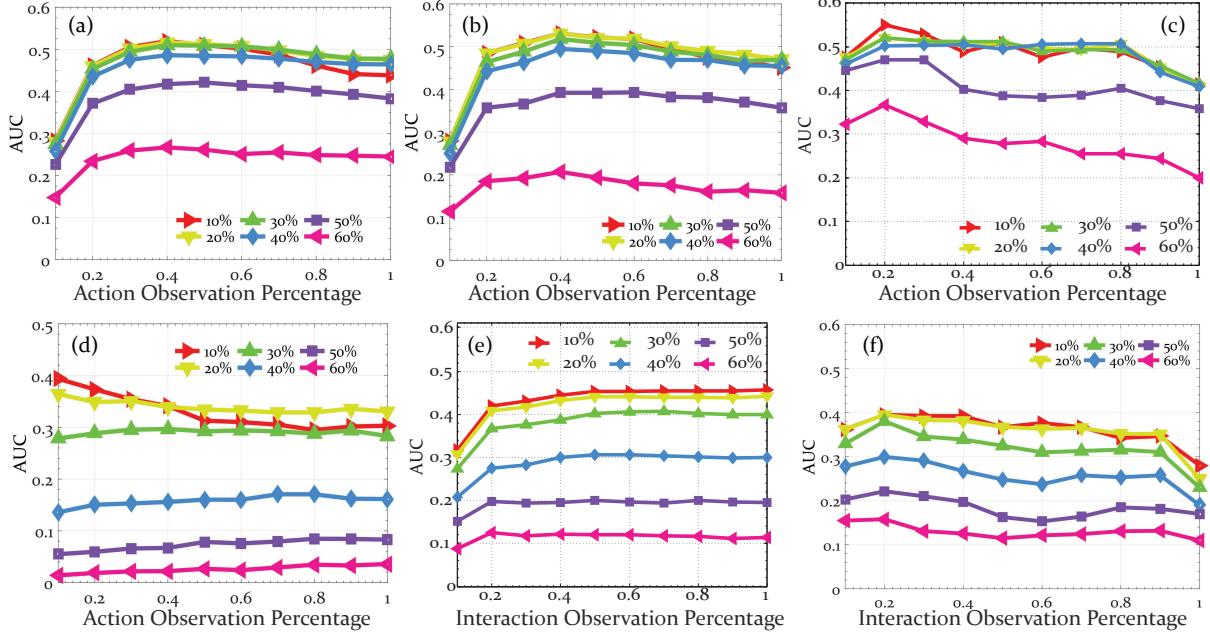


Fig. 7. This figure shows online action (interaction) localization performance as a function of observed action percentage on (a) JHMDB, (b) sub-JHMDB, (c) UCF-Sports, (d) MSR-II, and as a function of observed interaction percentage for (e) TV Human Interaction and (f) UT Interaction datasets. Different curves show evaluations at different overlap thresholds: 10% (red), 30% (green) and 60% (pink).

smooth motion of legs. For this dataset, the most difficult action is *swing side* due to high articulation with most of the instances depicting swinging of the sportsperson from the very first frame with different pose at the beginning of each action instance. Finally, we also analyze the performance of DP-SVM and S-SVM in Fig. 6(b) which shows the difference of prediction accuracy DP-SVM and S-SVM. Longer duration actions such as *diving*, *walking*, *running*, *riding horse* gain significant boost in prediction accuracy, with average performance increasing by about 12% over the baseline DP-SVM for this dataset.

Since each action has its own predictability, we also analyze how early we can predict each action. We arbitrarily set the prediction accuracy to 30% and show the percentage of action observation required for each action of JHMDB, sub-JHMDB and UCF Sports datasets in Table 1. Although we set a reasonable prediction target, certain actions do not reach such prediction accuracy even until the completion of the video. This highlights the challenging nature of online action prediction and localization.

Action (Interaction) Localization with Time: To evaluate online performance, we analyze how the localization performance varies across time by computing prediction accuracy as a function of observed action (interaction) percentage. Fig. 7 shows the AUC against time for different overlap thresholds (10% – 60%) for (a) JHMDB, (b) sub-JHMDB, (c) UCF Sports and (d) MSR-II action datasets. The AUC as a percentage of observed interaction percentage is shown for (e) TV Human Interaction and (f) UT Interaction datasets as well. 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, sub-JHMDB and MSR-II datasets in Fig. 7(a,b,d), 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. 7(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 observation percentage increases, the online algorithm never attains the peak again for many overlap thresholds despite observing the entire action.

Comparison with Offline Methods: We also evaluate the performance of our method against existing *offline* state-of-the-art action localization methods. Fig. 8(a) shows the

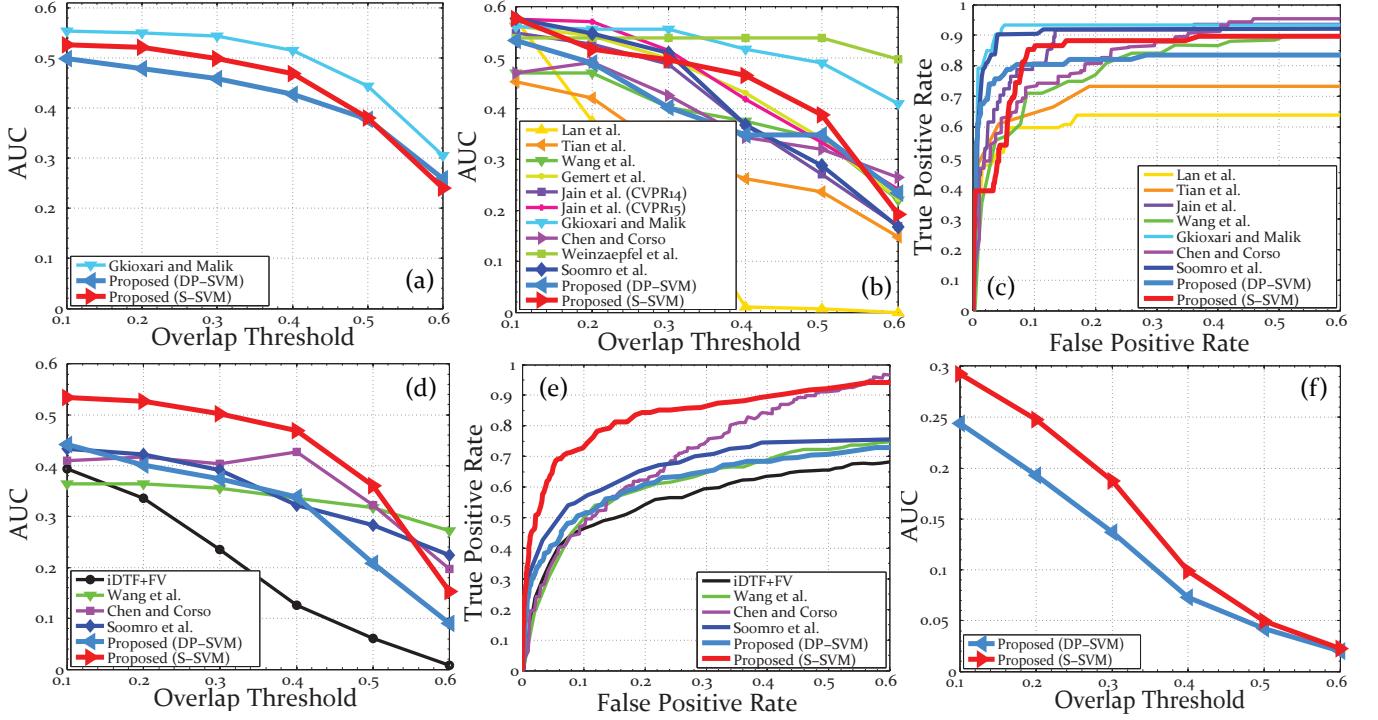


Fig. 8. This figure shows action localization results of the baseline Binary SVM with Dynamic Programming (DP-SVM) and Structural SVM (S-SVM) approaches, along with existing *offline* methods on four action datasets (JHMDB, UCF Sports, sub-JHMDB and MSR-II). (a) shows AUC curves for JHMDB, while (b) and (c) show AUC and ROC @ 20%, respectively, for UCF Sports dataset. AUC and ROC @ 20% overlap are shown in (d) and (e) for sub-JHMDB dataset, finally AUC for MSR-II dataset is shown in (f). The curve for S-SVM method is shown in red and DP-SVM is shown in blue, while other *offline* localization methods including Lan et al. [1], Tian et al. [6], Wang et al. [2], van Gemert et al. [66], Jain et al. [7] [30], Gkioxari and Malik [40], Chen and Corso [67], Weinzaepfel et al. [68] and Soomro et al. [8] are shown with different colors. Despite being online, the proposed approach performs competitively overall compared to existing offline methods.

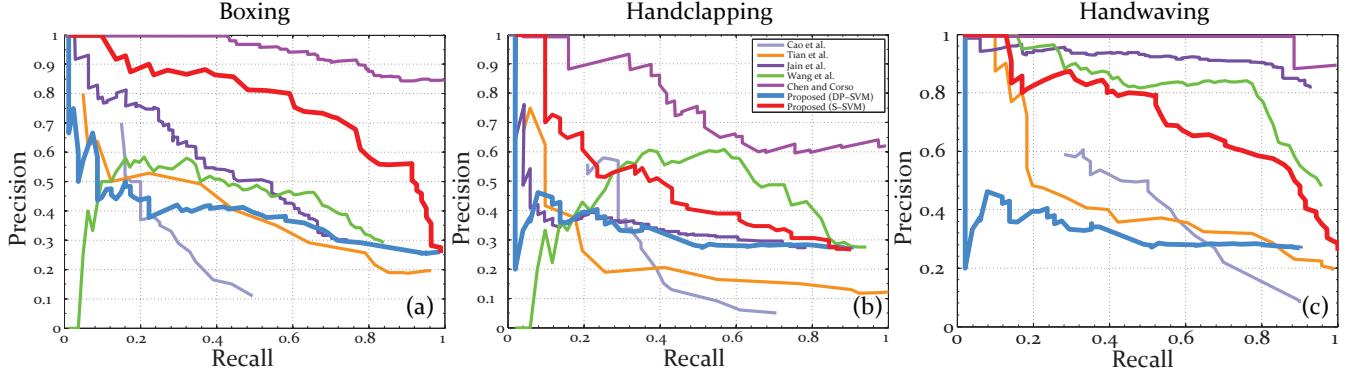


Fig. 9. This figure shows action localization results on MSR-II dataset. The precision/recall curves are drawn for three actions: (a) boxing, (b) Hand clapping and (c) hand waving. We perform competitive to many existing *offline* methods. Red curve shows the proposed S-SVM approach, while blue curve shows the results of the baseline DP-SVM method.

results of the proposed S-SVM method, on JHMDB dataset, in red and the baseline DP-SVM in blue, while that of [40] in cyan. The difference in performance is attributed to the online vs. offline nature of the methods, as well as the use of CNN features by [40]. Quantitative comparison on UCF Sports using AUC and ROC @ 20% is shown in Fig. 8(b) and (c) respectively. Fig. 8 also shows the results of S-SVM and DP-SVM over all datasets where S-SVM outperforms DP-SVM highlighting the importance of structured prediction. The biggest gain in performance is visible in sub-JHMDB dataset, as shown by the AUC and ROC curves in Fig. 8(d) and (e), where despite being online S-SVM outperforms

existing state-of-the-art methods.

For MSR-II dataset, we evaluate action localization and prediction using two separate metrics with: 1) precision/recall curve to draw comparison with existing methods as shown in Fig 9 for the three different actions: (a) boxing, (b) hand clapping and (c) hand waving. 2) AUC performance is also shown in Fig. 8 (f) for consistent evaluation with other datasets. The average precision per action is presented in Table 2.

Generally, interaction datasets have either been used for classification [15], activity prediction [18] or video retrieval [13]. We are the first to evaluate online localization on these

TABLE 2

This table shows the average precision for MSR-II dataset on three different actions: (a) Boxing, (b) Handclapping and (c) Handwaving.

Method	Boxing	Handclapping	Handwaving
Cao et al.[69]	17.5	13.2	26.7
Tian et al.[6]	38.9	23.9	44.4
Jain et al.[7]	46.0	31.4	85.8
Wang et al.[2]	41.7	50.2	80.9
Chen and Corso [67]	94.4	73.0	87.7
Proposed (DP-SVM)	37.3	28.3	42.9
Proposed (S-SVM)	75.3	43.4	71.3

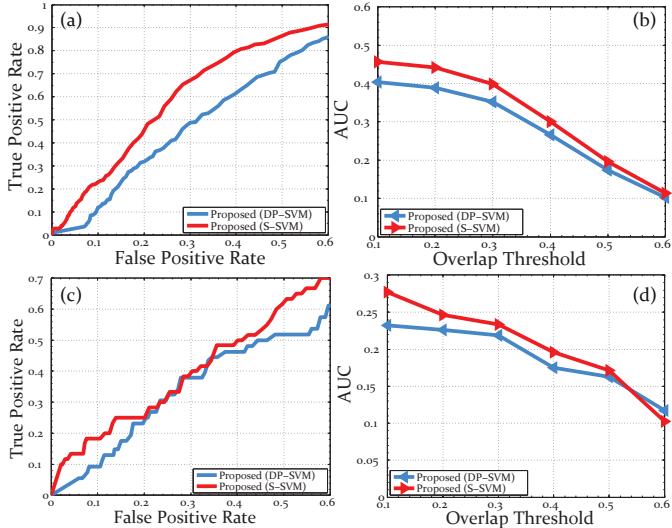


Fig. 10. This figure shows interaction localization results on two interaction datasets. ROC @ 20% overlap and AUC curves for TV Human Interaction dataset are shown in (a) and (b), and for UT Interaction dataset in (c) and (d). In this figure, S-SVM shown in red and DP-SVM (baseline) in blue.

datasets. To keep evaluation metrics uniform, we present our performance on localization and prediction of human interactions in Fig. 10 using ROC and AUC curves for TV Human interaction (a,b) and UT interaction (c,d).

Pose Refinement: Pose-based foreground likelihood refines poses in an iterative manner using spatio-temporal smoothness constraints. Our qualitative results in Fig. 11 show the improvement in pose joint locations on two example videos..

Action Segments: Since we use superpixel segmentation to

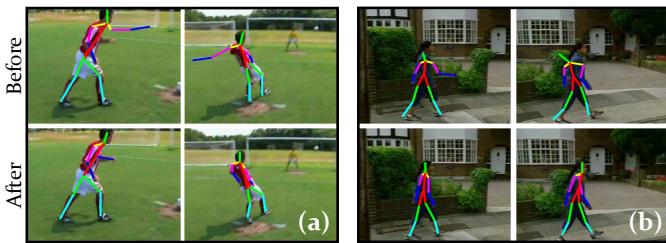


Fig. 11. This figure shows qualitative results for pose refinement. Results show a comparison of raw poses (top row) and refined poses (bottom row) for (a) Kicking and (b) Walking.

represent the foreground actor, our approach outputs action segments. Our qualitative results in Fig. 12 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.

6 CONCLUSION

In this paper, we introduced a new prediction problem of online action and interaction localization, where the goal is to simultaneously localize and predict action (interaction) 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 (interaction) prediction. We also refine pose estimation online using spatio-temporal constraints. The localized tubes are obtained using CRF, and prediction confidences come from the classifier. We showed that the Structural SVM (S-SVM) formulation outperforms the baseline dynamic programming with SVM (DP-SVM) hybrid. The intermediate results and analysis indicate that such an approach is capable of addressing this difficult problem, and performing competitive to some of the recent offline action localization methods.

REFERENCES

- [1] T. Lan, Y. Wang, and G. Mori, "Discriminative figure-centric models for joint action localization and recognition," in *ICCV*, 2011.
- [2] L. Wang, Y. Qiao, and X. Tang, "Video action detection with relational dynamic-poselets," in *ECCV*, 2014.
- [3] 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*, vol. 13, no. 3, 2011.
- [4] A. Dehghan, H. Idrees, and M. Shah, "Improving semantic concept detection through the dictionary of visually-distinct elements," in *CVPR*, 2014.
- [5] K. Soomro, H. Idrees, and M. Shah, "Predicting the where and what of actors and actions through online action localization," in *CVPR*, 2016.
- [6] Y. Tian, R. Sukthankar, and M. Shah, "Spatiotemporal deformable part models for action detection," in *CVPR*, 2013.
- [7] M. Jain, J. Gemert, H. Jegou, P. Bouthemy, and C. Snoek, "Action localization with tubelets from motion," in *CVPR*, 2014.
- [8] K. Soomro, H. Idrees, and M. Shah, "Action localization in videos through context walk," in *ICCV*, 2015.
- [9] D. Oneata, J. Verbeek, and C. Schmid, "Efficient action localization with approximately normalized fisher vectors," in *CVPR*, 2014.
- [10] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid, "Spatio-temporal object detection proposals," in *ECCV*, 2014.
- [11] M. Hoai and A. Zisserman, "Talking heads: Detecting humans and recognizing their interactions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 875–882.
- [12] Y. Kong, Y. Jia, and Y. Fu, "Learning human interaction by interactive phrases," in *European Conference on Computer Vision*. Springer, 2012, pp. 300–313.
- [13] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. D. Reid, "High five: Recognising human interactions in tv shows," in *BMVC*, vol. 1. Citeseer, 2010, p. 2.
- [14] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman, "Structured learning of human interactions in tv shows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2441–2453, 2012.
- [15] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 1593–1600.
- [16] K. Li and Y. Fu, "Prediction of human activity by discovering temporal sequence patterns," *IEEE TPAMI*, vol. 36, no. 8, 2014.



Fig. 12. This figure shows qualitative results of the proposed approach for all six datasets, where each action segment is shown with yellow contour and ground truth with green bounding box.

- [17] T. Lan, T.-C. Chen, and S. Savarese, "A hierarchical representation for future action prediction," in *ECCV*, 2014.
- [18] M. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *ICCV*, 2011.
- [19] Y. Kong, D. Kit, and Y. Fu, "A discriminative model with multiple temporal scales for action prediction," in *ECCV*, 2014.
- [20] X. Zhao, X. Li, C. Pang, X. Zhu, and Q. Z. Sheng, "Online human gesture recognition from motion data streams," in *ACM MM*, 2013.
- [21] M. Hoai and F. De la Torre, "Max-margin early event detectors," *IJCV*, vol. 107, no. 2, 2014.
- [22] G. Yu, J. Yuan, and Z. Liu, "Predicting human activities using spatio-temporal structure of interest points," in *ACM MM*, 2012.
- [23] 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.
- [24] D.-A. Huang and K. M. Kitani, "Action-reaction: Forecasting the dynamics of human interaction," in *European Conference on Computer Vision*. Springer, 2014.
- [25] R. De Geest, E. Gavves, A. Ghodrati, Z. Li, C. Snoek, and T. Tuytelaars, "Online action detection," in *ECCV*, 2016.
- [26] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, "Online human action detection using joint classification-regression recurrent neural networks," in *ECCV*, 2016.
- [27] 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.
- [28] 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.
- [29] C. Desai and D. Ramanan, "Detecting actions, poses, and objects with relational phraselets," in *ECCV*, 2012.
- [30] 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.
- [31] P. Rota, N. Conci, N. Sebe, and J. M. Rehg, "Real-life violent social interaction detection," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 3456–3460.
- [32] Y. Kong and Y. Fu, "Modeling supporting regions for close human interaction recognition," in *European Conference on Computer Vision*. Springer, 2014, pp. 29–44.
- [33] J. Yuan, Z. Liu, and Y. Wu, "Discriminative video pattern search for efficient action detection," *IEEE TPAMI*, vol. 33, no. 9, 2011.
- [34] Z. Zhou, F. Shi, and W. Wu, "Learning spatial and temporal extents of human actions for action detection," *IEEE Transactions on Multimedia*, vol. 17, no. 4, 2015.
- [35] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE TPAMI*, vol. 32, no. 9, 2010.
- [36] D. Tran and J. Yuan, "Max-margin structured output regression for spatio-temporal action localization," in *NIPS*, 2012.

- [37] S. Ma, J. Zhang, N. Ikizler-Cinbis, and S. Sclaroff, "Action recognition and localization by hierarchical space-time segments," in *ICCV*, 2013.
- [38] G. Yu and J. Yuan, "Fast action proposals for human action detection and search," in *CVPR*, 2015.
- [39] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, 2013.
- [40] G. Ghioxari and J. Malik, "Finding action tubes," in *CVPR*, 2015.
- [41] W. Chen, C. Xiong, R. Xu, and J. J. Corso, "Actionness ranking with lattice conditional ordinal random fields," in *CVPR*, 2014.
- [42] Y. Kong, Y. Jia, and Y. Fu, "Interactive phrases: Semantic descriptions for human interaction recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 9, pp. 1775–1788, 2014.
- [43] J. Wu, F. Chen, and D. Hu, "Human interaction recognition by spatial structure models," in *International Conference on Intelligent Science and Big Data Engineering*. Springer, 2013, pp. 216–222.
- [44] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," in *CVPR*, 2016.
- [45] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *CVPR*, 2016.
- [46] F. C. Heilbron, J. C. Niebles, and B. Ghanem, "Fast temporal activity proposals for efficient detection of human actions in untrimmed videos," in *CVPR*, 2016.
- [47] A. Richard and J. Gall, "Temporal action detection using a statistical language model," in *CVPR*, 2016.
- [48] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013.
- [49] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," *arXiv preprint arXiv:1505.04868*, 2015.
- [50] S. Majiwa, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *CVPR*, 2011.
- [51] 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.
- [52] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR*, 2011.
- [53] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *CVPR*, 2013.
- [54] M. Raptis and L. Sigal, "Poselet key-framing: A model for human activity recognition," in *CVPR*, 2013.
- [55] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *CVPR*, 2015.
- [56] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *ICCV*, 2014.
- [57] A. Vahdat, B. Gao, M. Ranjbar, and G. Mori, "A discriminative key pose sequence model for recognizing human interactions," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1729–1736.
- [58] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.
- [59] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE TPAMI*, vol. 34, no. 11, 2012.
- [60] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *ICCV*, 2013.
- [61] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *ICCV*, 2011.
- [62] M. Rodriguez, A. Javed, and M. Shah, "Action mach: a spatio-temporal maximum average correlation height filter for action recognition," in *CVPR*, 2008.
- [63] K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," in *Computer Vision in Sports*. Springer, 2014, pp. 181–208.
- [64] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *ICPR*, 2004.
- [65] M. S. Ryoo and J. K. Aggarwal, "UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)," http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010.
- [66] J. C. van Gemert, M. Jain, E. Gati, and C. G. Snoek, "Apt: Action localization proposals from dense trajectories," in *BMVC*, vol. 2, 2015, p. 4.
- [67] W. Chen and J. J. Corso, "Action detection by implicit intentional motion clustering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.
- [68] P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Learning to track for spatio-temporal action localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.
- [69] L. Cao, Z. Liu, and T. S. Huang, "Cross-dataset action detection," in *CVPR*, 2010.



Khurram Soomro received his B.Sc and M.Sc degrees in Computer Engineering from Lahore University of Management Sciences, Lahore, Pakistan, in 2007 and 2011, respectively. He joined Center for Research in Computer Vision at University of Central Florida in 2011, where he is currently pursuing his Ph.D degree in Computer Vision. His research interests include Action Recognition and Localization, Human Detection, Visual Surveillance and Tracking, and Sports Analytics. He is a member of the IEEE and Upsilon Pi Epsilon (UPE) Honor Society.



Haroon Idrees is a Postdoctoral Associate at the Center for Research in Computer Vision at University of Central Florida. He has published several papers in conferences and journals such as CVPR, ICCV, ECCV, Journal of Image and Vision Computing, Computer Vision and Image Understanding, and IEEE Transactions on Pattern Analysis and Machine Intelligence. His research interests include crowd analysis, action recognition and localization, object detection, visual tracking, multi-camera and airborne surveillance, and multimedia content analysis. He received the BSc (Hons) degree in Computer Engineering from the Lahore University of Management Sciences, Pakistan in 2007, and the PhD degree in Computer Science from the University of Central Florida in 2014.



Mubarak Shah, the Trustee chair professor of computer science, is the founding director of the Center for Research in Computer Vision at the University of Central Florida (UCF). He is an editor of an international book series on video computing, editor-in-chief of Machine Vision and Applications journal, and an associate editor of ACM Computing Surveys journal. He was the program cochair of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) in 2008, an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence, and a guest editor of the special issue of the International Journal of Computer Vision on Video Computing. His research interests include video surveillance, visual tracking, human activity recognition, visual analysis of crowded scenes, video registration, UAV video analysis, and so on. He is an ACM distinguished speaker. He was an IEEE distinguished visitor speaker for 1997-2000 and received the IEEE Outstanding Engineering Educator Award in 1997. In 2006, he was awarded a Pegasus Professor Award, the highest award at UCF. He received the Harris Corporations Engineering Achievement Award in 1999, TOKTEN awards from UNDP in 1995, 1997, and 2000, Teaching Incentive Program Award in 1995 and 2003, Research Incentive Award in 2003 and 2009, Millionaires Club Awards in 2005 and 2006, University Distinguished Researcher Award in 2007, Honorable mention for the ICCV 2005 Where Am I? Challenge Problem, and was nominated for the Best Paper Award at the ACM Multimedia Conference in 2005. He is a fellow of the IEEE, AAAS, IAPR, and SPIE.