What if we do not have multiple videos of the same action? — Video Action Localization Using Web Images

Waqas Sultani, Mubarak Shah
Center for Research in Computer Vision (CRCV), University of Central Florida (UCF)
waqassultani@knights.ucf.edu, shah@crcv.ucf.edu

Abstract

This paper tackles the problem of spatio-temporal action localization in a video, without assuming the availability of multiple videos or any prior annotations. Action is localized by employing images downloaded from internet using action name. Given web images, we first dampen image noise using random walk and evade distracting backgrounds within images using image action proposals. Then, given a video, we generate multiple spatio-temporal action proposals. We suppress camera and background generated proposals by exploiting optical flow gradients within proposals. To obtain the most action representative proposals, we propose to reconstruct action proposals in the video by leveraging the action proposals in images. Moreover, we preserve the temporal smoothness of the video and reconstruct all proposal bounding boxes jointly using the constraints that push the coefficients for each bounding box toward a common consensus, thus enforcing the coefficient similarity across multiple frames. We solve this optimization problem using variant of two-metric projection algorithm. Finally, the video proposal that has the lowest reconstruction cost and is motion salient is used to localize the action. Our method is not only applicable to the trimmed videos, but it can also be used for action localization in untrimmed videos, which is a very challenging problem. We present extensive experiments on trimmed as well as untrimmed datasets to validate the effectiveness of the proposed approach.

1. Introduction

Bounding box annotations have played a crucial role in development of several computer vision applications, such as: object/action recognition, detection, tracking and segmentation [36, 33, 6]. However, these annotations are cumbersome to obtain, require hundreds of hours and are subject to human biases.

To mitigate this annotation challenge, several weakly-supervised approaches have been introduced recently [22, 5, 21, 29, 3, 9], particularly in the object domain. In general, all these approaches assume presence of dominant centered objects in multiple images. For instance, the method proposed in [9] annotate objects from previously annotated images, [21] obtain bounding boxes by involving human eye tracking, and [29, 3] achieve object annotation using multiple images, where most of these images contain the object of interest.

As compared to object annotation, spatio-temporal action annotations in videos are far more challenging and, therefore, it is not surprising that most of recent action

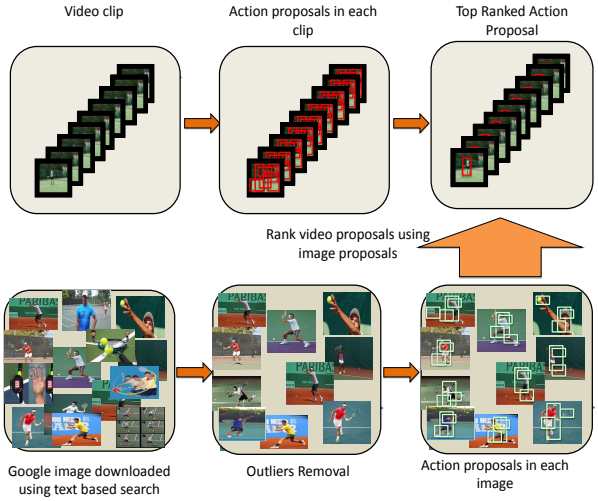

Figure 1: This figure illustrates our key idea of action localization in a video using images. We first download images of an action of interest from internet. After removing noisy images, we co-localize all the images jointly to obtain action proposals in each of the image. Then, given the candidate action locations in a video, we leverage image proposals to discover the most action representative proposal in a video.

datasets [28, 14] contain only a few or no spatio-temporal annotations. The straightforward approach to obtain spatio-temporal action annotations in a video would be to extend any of the previously mentioned methods from image domain to video domain. However, temporal extension has many challenges due to large search space and critical differences between spatial and temporal dimensions [33]. More importantly, what if we do not have available multiple videos of the same action?

To tackle the challenge of action localization in a single video, we propose to leverage images downloaded from the internet using text-based queries. In contrast to previous works in object annotations, we neither assume availability of bounding box annotations nor the presence of multiple videos of the same class. Furthermore, we do not assume the availability of clean images either.

$Images$ are usually taken to capture key poses, descriptive viewpoints and important instances of an action or event [16, 17]. Our key idea is to exploit this useful information to obtain precise spatio-temporal action localization in $videos$. To operationalize our intuition (see Fig. 1) , we first download several images of the action of interest using the action label as a query from Google. These images contain human performing actions in different locations (not necessarily at the center), backgrounds and include many irrelevant and noisy images. To circumvent these issues, we remove irrelevant noisy images using random walk. To handle the challenge of variable locations and backgrounds, we co-localize the action in multiple images using a recently proposed unsupervised localization method [3]. The output of these steps is the candidate action localization in the images.

Our ultimate goal is to obtain spatio-temporal annotations in a video. Therefore, given a video clip, we first obtain action proposals [20]. These proposals represent candidate spatio-temporal action locations in the video. However, not all proposals are truly action representative as many are due to camera motion and cluttered backgrounds. Therefore, we remove highly overlapping action proposals using non-maximal suppression by exploiting optical flow gradient within the proposals. To obtain the most action representative proposal, we propose to reconstruct action proposals in the video by leveraging the action proposals in images. Furthermore, we preserve the temporal smoothness of the video by introducing consensus regularization. Consensus regularization enforces consistency among coefficients vectors of multiple frames within the proposal. The proposal with the lowest reconstruction error and a high motion saliency is selected as a final action localization.

Our experimental results reveal that it is possible to automatically annotate an action in a video by employing web images of the same action through mitigating the effect of distracting backgrounds within images and by preserving the temporal structure of video during reconstruction.

Most of the previous works demonstrate action localization accuracy either on trimmed videos or carefully staged clean untrimmed videos. However, these videos do not represent the real-world videos, which are long, have variable scenes and backgrounds and contain multiple or no instance of the action of interest. Since proposed approach does not require multiple videos and prior annotations, it can easily be applied to more realistic untrimmed videos. We have evaluated our approach on trimmed [23, 13] as well as on the part of untrimmed [14] datasets and have obtained encouraging results.

In summary, 1) We demonstrate the feasibility of using images to achieve spatio-temporal action localization in videos, 2) By utilizing video proposal sparse reconstruction error with motion saliency, we achieve impressive localization results on popular trimmed action datasets, 3) We are the first to report spatio-temporal action localization results on (the part of) challenging untrimmed action dataset [14]. Furthermore, we will release spatio-temporal annotations of 35, 000 frames of [14] to facilitate further research in this direction.

2. Related work

With the exponential increase in the size of object/action datasets, obtaining annotations is becoming increasingly daunting task. Moreover, it is subject to human biases in terms of start and end of the activity and the sizes of the exact spatial boxes around an actor.

One way to avoid these time consuming annotations altogether is to use weakly supervised object/action detector approaches such as [25]. This type of approaches only use image/video level labels and learn the object/action detector without requiring bounding box annotations. Although impressive, their accuracy is still far behind that of detectors trained on hundreds of bounding box annotations. Another interesting area of research relates weakly supervised annotations methods [30, 3, 15]. Tang $et\ al.$ [30] introduced co-localization method where the objective is to obtain bounding boxes around common objects among multiple images. Their joint image and box formulation can also handle the presence of noisy images to some extent. Joulin $et\ al.$ [15] extended [30] to videos and co-localize objects in several frames using multiple videos. Both methods require image or video level labels only. Recently, Cho $et\ al.$ [3] introduced part based matching approach to localize common objects across multiple images, without requiring images level labels. Given several images of different object classes, this method efficiently localize objects which are common in multiple images. Although encouraging results have been obtained, these methods require multiple images of the object of interest and cannot localize the objects if multiple images containing the same object are not

available.

Recently, [32] and [26] respectively introduced weakly supervised methods to obtain object segmentation and bounding box annotations in a single video. Similar to our approach, these methods do not require multiple images of object/action of interest. However, in contrast to our method, they rely on negative data (the videos that do not contain the object of interest). These methods robustly segment and locate object/action in a single video. We compare our approach with both of these methods and show superior performance.

There is an increasing interest in leveraging images and videos to improve the performance of either domain or both. The method in [22] used YouTube videos to gather more examples for training object detectors. Chen *et al.* [2] used unlabeled video to learn action detector for images. Kevin *et al.* [31] proposed to adapt object detector from images to video. The approach presented in [16, 27] use images to produce effective video summarization. The authors in [17] presented an efficient framework to produce a joint summary of a video and Flicker images. Recently, Jain *et al.* [12] demonstrates that object classifiers can be used to improve action recognition accuracy.

However, we are not aware of any previous work that uses images to localize an action in a video. In what follows, we first describe our approach in detail for trimmed videos and then present its extension to untrimmed videos (Section 6).

3. Weakly Supervised Action localization in Images

The first step of our approach is to obtain candidate action proposals in downloaded images. For this purpose, we download images from internet and obtain candidate action locations in each image. The details of each step are given below.

3.1. Web Image Collection

Using the action name such as tennis swing, golf swing etc., as a text query we download images from Google Image search engine. Although, Google image search quality has been improved significantly over last few years, the retrieved images still contain outliers and irrelevant images due to in-accurate query text and polysemy.

We perform random walk over these images to get rid of image noise. The key benefit of using random walk is that it can discover both small cluster of outliers as well as the images far away from all other images (in feature space) [19]. We define a fully connected graph $\mathbf{Z}(\mathbf{N}, \mathbf{E})$, where $\mathbf{N}$ is the set of all images and $\mathbf{E}$ represents set of edges between them. The weight between any two nodes i and j on the graph is measured by Euclidean distance between

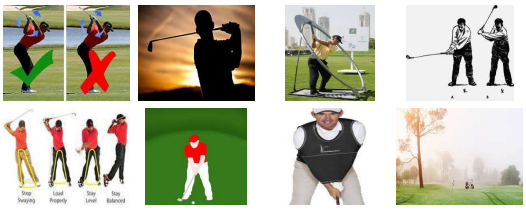

Figure 2: Noisy golf swing images removed by random walk. These images include cartoons, people in unusal backgrounds and clipart. Last image (bottom right) represents the failure case, which random walk is unable to remove (perhaps due to its similarity to golf swing in the feature space).

their features $\phi(i)$ and $\phi(j)$, where ϕ represents deep learning features [35] computed over the whole image. Finally, the transition probability between any two nodes i and j is given by

$$p(i,j) = \frac{e^{-\gamma\|\phi(i)-\phi(j)\|_2}}{\sum_{m=1}^{k} e^{-\gamma\|\phi(i)-\phi(m)\|_2}}. \quad (1)$$

The random walk over the graph is then formulated as:

$$r_k(j) = \beta \sum_i r_{k-1}(i) p_{ij} + (1-\beta) v_j, \quad (2)$$

where $r_k(j)$ represents relevance score of the image j at k^{th} iteration, v_j is its initial probabilistic score and β controls the contribution of both terms to the final score. Due to the absence of any prior knowledge about images, we assign the same initial probabilistic score to all the images. The relevance score $r_k(j)$ is iteratively updated for all nodes until fixed number of iterations are achieved. The images with low relevance score can be considered as outliers and subsequently removed. In our experiments, we removed 30% of the originally downloaded images. When removing images more than 30%, we start losing good quality images. In experiments, we use β=0.99. Some of the typical images removed by random walk are shown in Figure. 2.

3.2. Action Proposals in Images

Although images downloaded using the text query belong to the same overall concept; they are mostly captured in different scenes and contain distracting backgrounds. Using these images naively is detrimental to video proposals ranking (see Table 1). Therefore, to get rid of unnecessary backgrounds, we propose to localize the action in images.

To localize the action in the downloaded images, we use recently proposed state-of-art unsupervised localization method [3]. We use this method because of its excellent performance on many complex datasets [6].

Following [3], we extract hundreds of candidate action proposals [18] from each image. The objective is to obtain

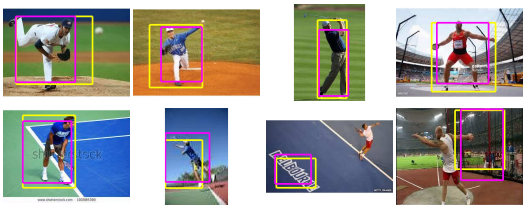

Figure 3: Automatically generated action proposals in images. In bottom row, last two images (from right) show the failure cases due to very small size of actor and cluttered background.

the proposals which represent the most common concept (the action in our case) across all the images. To achieve this, we efficiently match action proposals across all the images using Probabilistic Hough Matching (PHM) [3]. The PHM matching is performed on local regions within proposals by carefully considering their scale and localization variations. The score of a local region in proposal p_m with respect to p'_m is given as:

$$\psi(p) = \max_{r'} c((r, r')|(p_m, p'_m)), \tag{3}$$

where c represents Hough matching confidence of local region r in p_m with respect to p'_m.

The high region score represents the highest matched proposal across images. However, it does not provide the explicit action localization as the background regions can also have good matches. Therefore, we use both the standout score [3] of each proposal and the PHM based region matching to obtain the final action localization in images. In our experiments, we use only top two action proposals from each image. Selecting more than two proposals increases the computational time of next steps, while not always helping the performance. Figure 3 shows automatically generated images' proposals for actions of THUMOS14. The right most images in the second row show the failure cases, where we are unable to localize the action of interest.

4. Action Proposals in videos

Our end goal is to obtain spatio-temporal action localization in a video using image action proposals generated in the previous section. Therefore, we first estimate action locations in the video and try to remove the majority of camera and background generated proposals.

In order to obtain spatio-temporal action localization in a video, we first need to obtain candidate action locations in a video. Traditional ways to achieve this is to use 3D (spatio-temporal volume) sliding window approach. However, this approach has two main limitations. Firstly, it produces extremely large number of candidate locations. Secondly, 3D cuboids contain a large amount of background particularly in case of dynamic actions. To circumvent these problems, recently, action proposals have been presented [11, 20, 34]. Compared to sliding window, these techniques provide far less number of high quality action proposals.

In this work, we employ supervoxel segmentation based approach to generate action proposals [20]. However, our method does not depend on specific action proposal methods and any action proposals method can be used [34, 11]. We compute fixed number of superpixels from each video frame and estimate mean color, color histogram and optical flow histogram within each superpixel. Given n number of superpixels, we build a graph $\mathbf{G}(\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is a set of superpixels and $\mathbf{E}$ represents a set of edges between them. We use discontinuity preserving first and second order spatial edge weights between superpixels m and n, where the first order edge weight is given by:

$$\begin{aligned} e^{nm,s} = \alpha_1 d_1(n,m) + \alpha_2 d_2(n,m) + \alpha_3 d_3(n,m) \\ + \alpha_4 d_4(n,m) + \alpha_5 d_5(n,m), \end{aligned} \tag{4}$$

where d_1 corresponds to distance between color means, d_2 and d_3 represent distance between color and flow histograms and d_4 and d_5 represent geodesic distance between superpixel centroids computed through motion and color boundaries.

In addition to spatial edges, we also build temporal edges given as

$$e^{nm,t} = \alpha_7 d_1(n,m) + \alpha_8 d_2(n,m) + \alpha_9 d_3(n,m), \tag{5}$$

where d_1, d_2 and d_3 are the same as described before and m and n represents temporal neighbors. Hierarchical clustering on this graph results into supervoxels segmentations. Finally, action proposals are built by merging supervoxels using randomized Prim's maximum span tree algorithm [18], extended to videos. During proposals generation, appearance, motion and size similarities of superpixels are taken in account. Typical examples of few action proposals for UCF Sports videos are shown in Figure 4. Although, the above method generates significantly less number of action proposals (approx. 2000 in each video), their number is still huge for our application, since we want to obtain only the most action representative proposal in each video clip.

Human actions are mainly characterize by motion. We use this important cue for two purposes. First, we use it to discard camera and background generated proposal (as they would have small optical flow gradients). Secondly, we use it to facilitate action proposal ranking. To this end, we use optical flow gradients within each video proposal. We first compute Frobenius norm of optical flow within each proposal, defined as:

$$\|U_X\|_F = \left\| \begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix} \right\|_F, \tag{6}$$

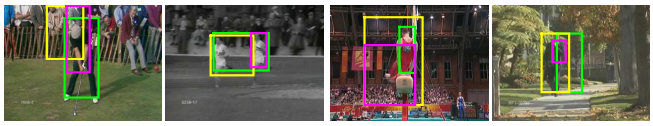

Figure 4: Video action proposals. Colors in the figures are randomly assigned.

where $U = (u, v)$ represents forward optical flow and u_x, v_x, u_y and v_y are optical flow gradients.

The motion score, η_p, of each video proposal, p_v, is then defined as weighted summation of Frobenius norm, namely,

$$\eta_p = G_l(x_c, y_c) \times G_s(h, w) \times \sum \|U_X\|_F \,, \qquad (7)$$

where x_c, y_c, h, w represent center coordinates, height and width of the proposal respectively. Gaussians G_l and G_s encourage proposals that are in the center of the video and are in vertical shapes since humans in these action videos are mostly in the center and are in upright position.

Assuming η_p, as a detection score, we perform Non-Maximal Suppression (NMS) to obtain a few proposals, which have high optical flow gradients and have small overlap with each other. In our experiments, we keep at most fifty proposals from each video. This results in a huge decrease in computation for further steps. Finally, we normalize motion score η_p of all proposal within a video between zero and one. We use this normalize score to represent motion saliency of each proposal.

5. Ranking Video Action Proposals using Image Action Proposals

In this section, we present our key idea of ranking video action proposals, $\mathbf{P_v}$, using image action proposals, $\mathbf{P_m}$. We propose to achieve this by reconstructing video action proposals as a linear combination of image action proposals. The main idea is that video action proposals which can easily be reconstructed using image action proposals (i.e., have low reconstruction error) can be considered to be capturing the key poses and viewpoints of the specific action and therefore represents the action of interest.

Suppose a video contains k number of video action proposals, $\mathbf{P}_v = [p_v^1, p_v^2, \ldots, p_v^k]$. Within each proposal, we extract visual features [35] from each of the key frame (bounding box). Let $\mathbf{\Pi^f} \in \mathbb{R}^{d \times n}$ represents the matrix obtained by vertical concatenation of all key frames features within a proposal, where d is the dimension of visual feature, and n is the number key-frames within proposal.

Similarly, $\mathbf{\Upsilon^f} \in \mathbb{R}^{d \times m}$, represents vertical concatenation of visual features from all image proposals, where m represents the total number of image proposals.

The straightforward approach would be to reconstruct each of the video proposal bounding box independently using image proposals and aggregate the reconstruction error for all the bounding boxes to obtain overall proposal action score. Although appealing, it ignores the underlying temporal structure of the video. Videos are not just the collection of frames but the sequence of frames and hence contain temporal information. Therefore, we propose to reconstruct all proposal bounding boxes jointly using the constraints that push the coefficients for each bounding box towards a common consensus, thus enforcing the coefficient similarity across multiple frames. Moreover, we introduce sparsity constraint to take care of noise in image data. Consensus regularization has been introduced recently for different applications [4, 37].

To achieve above goal, we minimize following the objective function:

$$\mathbf{Z} = \min_{\mathbf{C}} \left\|\mathbf{\Pi^f} - \mathbf{\Upsilon^f C}\right\|_{\mathbf{F}}^{\mathbf{2}} + \lambda_{\mathbf{1}} \left\|\mathbf{C} - \bar{\mathbf{C}}\right\|_{\mathbf{F}}^{\mathbf{2}} + \lambda_{\mathbf{2}} \left\|\mathbf{C}\right\|_{\mathbf{1}}, \qquad (8)$$

where the first term minimizes reconstruction error and second and third term enforce consistency (across columns) and sparsity in coefficient matrix $\mathbf{C}$, respectively. The consensus matrix $\bar{\mathbf{C}}$ is obtained by columns-wise concatenation of mean of coefficient matrix $\mathbf{C}$.

We solve the optimization mentioned in Eq. 8 using variant of two-metric projection algorithm [7, 24]. We divide the optimization variables c_i into two sets: active set and working set. Active set, $\mathcal{A}$, contains the variables that have positive partial derivative and are close to zero.

$$\mathcal{A} = \{i | c_i < \epsilon, \nabla_i \mathbf{Z}(\mathbf{C})\} \qquad (9)$$

Similarly, variables that have negative partial derivative or that are sufficiently non-zero belong to working set, $\mathcal{W}$. We compute a diagonally-scaled projected pseudo-gradient step for active set variables and a projection of Newton step along working set, namely,

$$\begin{aligned} \mathbf{C}_{\mathcal{W}} &\leftarrow \mathcal{P}[\mathbf{C}_{\mathcal{W}} - \sigma \mathbf{H}_{\mathcal{W}}^{-\mathbf{1}} \nabla_{\mathcal{W}} \mathbf{Z}(\mathbf{C})] \\ \mathbf{C}_{\mathcal{A}} &\leftarrow \mathcal{P}[\mathbf{C}_{\mathcal{A}} - \sigma \mathbf{D}_{\mathcal{A}} \nabla_{\mathcal{A}} \mathbf{Z}(\mathbf{C})], \end{aligned} \qquad (10)$$

where $\mathcal{P}$ is orthant projection and H is Hessian matrix. Note that, given positive diagonal scaling matrix $D_{\mathcal{A}}$, combined gradient direction is descent, unless $\mathbf{C}$ is optimal. We iteratively solve the above equations until we obtain the optimal solution or the maximum number of iterations are met.

We optimize Eq. 8 for every proposal in the video clip and estimate the reconstruction error. We normalize reconstruction errors of all proposals within a video between zero and one. The final action score Λ_p of each proposal, p_v, is simply given as:

$$\Lambda_p = (1 - \mathcal{R}_p) + \eta_p, \qquad (11)$$

where $\mathcal{R}_p$ and η_p represent reconstruction error and motion saliency (calculated in Section 4) of proposal, p_v.

Note that we have experimented with several state-of-art domain adaptation methods such as [8, 10], however either they do not help at all or have diminishing effect on the performance.

6. Action localization in Untrimmed videos

The proposed approach is generic in nature and can be applied to any action dataset including recently introduced extremely challenging untrimmed action datasets such as THUMOS14 [14]. This dataset contains long YouTube sports videos, mostly gathered from news and documentaries. The general trend in these videos is that they contain: newscaster or reporter, clips showing the crowd and stadium, people talking about the specific sport and finally the actual action clips somewhere in between these irrelevant clips.

To use our approach on untrimmed videos, we first divide long videos into shots [1]. We start with the assumption that each shot contains an action. By considering each shot as a $trimmed$ video, we compute top ranked action proposal from each video using exactly the same procedure as described in Section 3, 4 and 5. After computing the most representative action proposal in each shot (Section 5), we compare the action score (Eq. 11) of these top ranked proposals across the shots. Intuitively, the shots that contain an action would have top ranked proposals with high action score as compared to the shots that do not contain action. We max-normalize the reconstruction error of shots across the video. By sweeping the threshold of reconstruction error, we generate ROC curve as shown in Figure 7.

7. Experimental Results

The main goal of our experiments is to quantitatively evaluate the performance of proposed approach, verify that each component contributes to its final accuracy and demonstrate the generality of our approach. To this end, we performed extensive experiments on trimmed as well untrimmed action datasets.

For shot detection, we computed RGB histogram of frames as a feature representation. For all other experiments, we used CNN features [35], computed within image/video proposals bounding boxes. We set the parameters in Equation 8 as λ_1=0.06 and λ_2=0.16.

7.1. Experiments on Trimmed Action Dataset

For experiment on trimmed dataset, we have chosen UCF-Sports [23] and THUMOS13 [13] because of their complexity and that several recent works have used in their experiments [34, 11, 33]. In these datasets, an action spans the complete video clip. These broadcast videos contain large camera motion, cluttered background, variable viewpoints and occlusion. UCF-Sports dataset contains 150 videos and include 10 actions including: diving, golf swing, kicking, lifting, horse riding, running, etc. THUMOS13 is a subset of UCF101 [28] and contains 24 human actions that have spatio-temporal annotations. These actions include: cricket bowling, biking, salsa spin, etc. This dataset has 3207 videos. We used all videos of both datasets for evaluation (except Walk-Front-005 in UCF-Sports since it is actually a running action).

To evaluate localization accuracy, we use the standard intersection over union metric at 20% threshold [11, 33]. The localization accuracy of our complete method for UCF-Sports is given in Table 2. We compare our method with two strong baselines: CRANE [32] and Negative Mining [26]. Similar to the proposed approach, both of these techniques are weakly supervised annotation methods, i.e., they only assume video level labels. The comparison shown in Table 2 indicates the significantly improved localization accuracy of our method. Note that we use the same features [35] for all three methods.

In Figure. 5, we show qualitative examples of localization. We show four frames for a video from each action. It can be seen that our method performs quite well despite large camera motion (diving, kicking), scale changes (walking), cluttered background (horse riding, skateboarding), small actor size (running, golf swing) and abrupt motion (swinging).

Our method contains several components. We evaluate the contribution of each component towards final localization accuracy in Table 1. First row indicates localization accuracy, where we use all of the downloaded images (without removing noisy ones) in our reconstruction framework. Removing noisy images gives 3% improvement in localization accuracy (second row). Reducing the effect of images background noise through proposal, we achieve further 15% improvement. By enforcing consistency and sparsity in coefficient vectors among multiple frames of the proposal, we obtain 5% improvement. Finally, by adding motion score, we achieve further 5% improvement. Our results demonstrate that each component of our approach is necessary and contributes towards final localization accuracy. Moreover, our results reinforce that the web images do have the ability to make a significant impact on action localization in videos.

Table 3 shows localization accuracy of top ranked proposals in UCF-Sports and UCF-101 across different overlap thresholds.

7.2. Experiments on Un-Trimmed Action Dataset

To demonstrate the effectiveness of the proposed approach, we have evaluated it on a part of recently released un-trimmed action dataset [14]. This dataset was released in 2014 in THUMOS challenge workshop. In addition to having cluttered background, severe occlusion and huge camera motion, these extremely challenging real-world videos

Method	Diving	Golf Swing	Kicking	Lifting	Riding Horse	Run	Skateboarding	Swing Bench	Swing Sideangle	Walk	Avg
All Images w/o Noise removal	78.57	77.78	60.00	100	75.00	61.54	50.00	80.00	7.69	61.90	65.25
Images w/ Noise removal	78.57	83.33	65.00	100	75.00	61.54	50.00	85.00	30.77	52.38	68.16
ImageProp w/o constraints	85.71	83.33	80.00	100	100	69.23	66.67	100	61.57	90.48	83.70
Reconstruction model (Eq. 8)	92.86	**100**	85.00	100	91.67	**92.31**	**88.33**	95.00	69.23	76.19	88.56
Complete Method (Eq. 11)	**100.00**	94.44	85.00	100	100	84.62	83.33	100	**84.62**	**95.24**	**92.72**

Table 1: Quantitative results for UCF-Sports. Top row shows localization accuracy of reconstructing video proposals using all images (including noisy ones). The second row shows the same after noise removal using random walk. The third row shows localization accuracy of reconstructing video proposals from image proposals without enforcing sparsity and consensus constraints. Localization accuracy of complete reconstruction model (Eq.8) is shown in fourth row. Finally, fifth row shows accuracy of complete method. The results indicate that noise removal, image proposals, regularization and motion saliency; all contribute to overall localization accuracy.

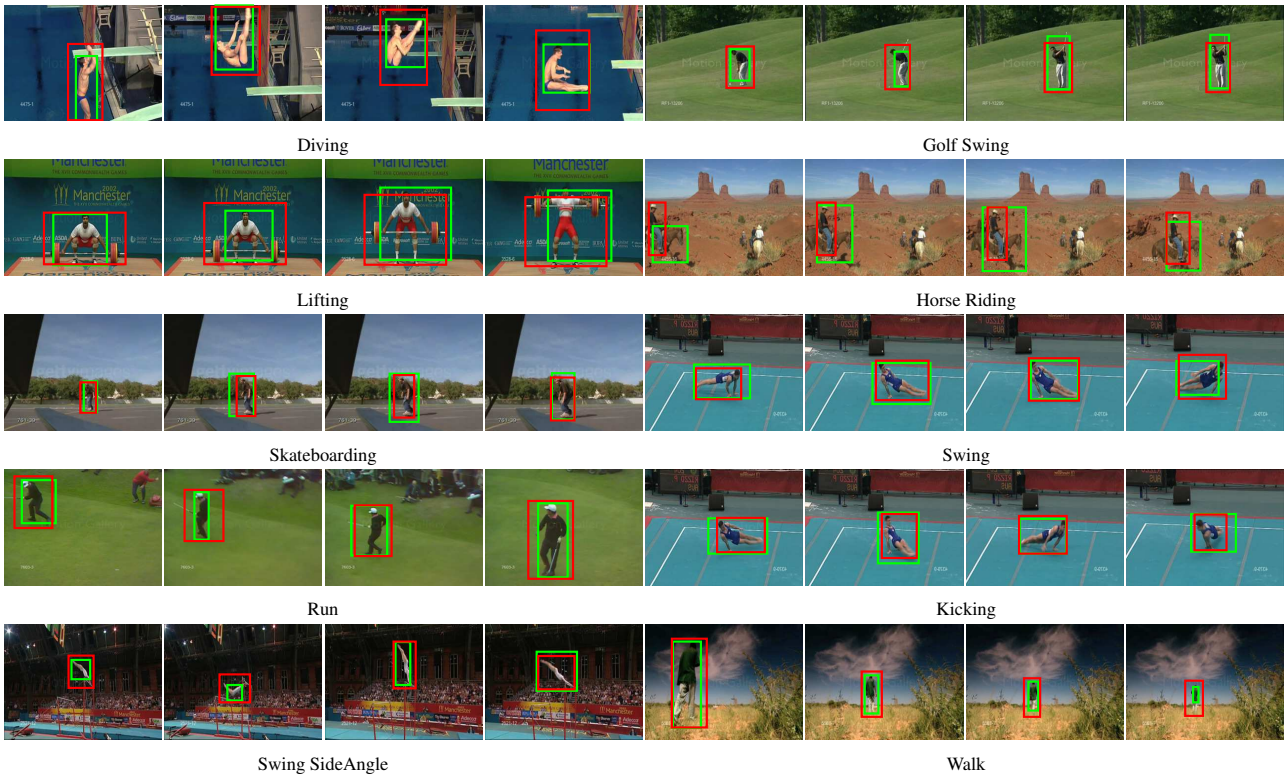

Figure 5: Localization results (Top ranked proposal) from UCF-Sports. We show four frames of each action video. Red box indicates ground truth and green box shows localization results.

contain several irrelevant frames such as non-action frames and multiple instance of the same action.

THUMOS14 test-set contains 20 actions, where only temporal annotations are provided *without any spatial annotations*. To evaluate spatio-temporal localization accuracy of our method on this dataset, we manually annotated four actions: baseball pitch, golf swing, tennis swing and throw discus. Specifically, we annotated around 35, 000 video frames (these annotations will be made publicly available). Baseball pitch, golf swing, tennis swing and throw discus contain 40, 141, 80, and 28 number of action instances, respectively. Given a video, we first divide it into

Method	CRANE[32]	NML [26]	Ours
Localization	65.41	63.01	**92.72**

Table 2: A comparison of our approach with related weakly supervised annotation methods on UCF-Sports

Threshold	0.1	0.2	0.3	0.4	0.5	0.6
UCF-Sports	93.9	92.7	82.1	61.0	40.7	18.5
UCF-101	78.0	62.7	47.8	28.8	13.8	4.6

Table 3: Localization accuracy of UCF-Sports and THUMOS13 (24 classes) at various thresholds.

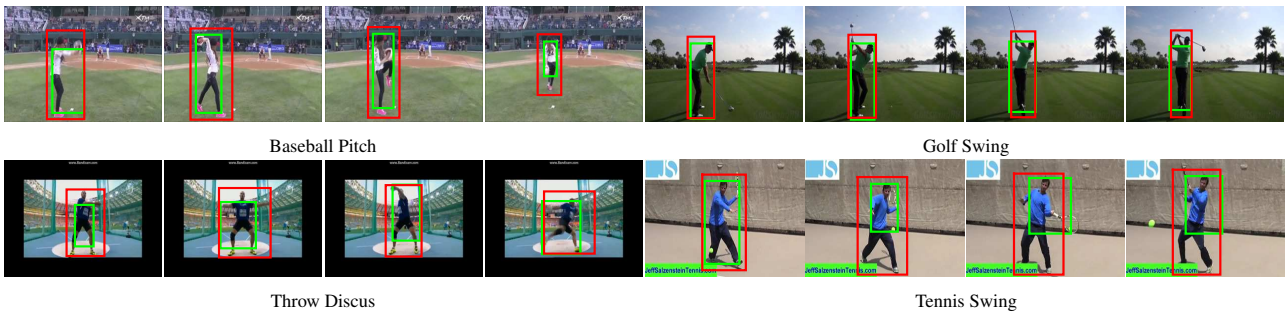

Figure 6: Localization results (Top ranked proposal) from four actions of THUMOS14. We show four frames of each action video. Red box indicates ground truth and green box shows localization results.

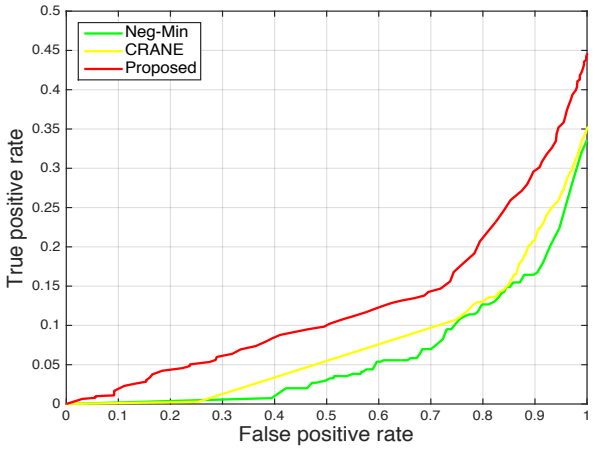

Figure 7: Mean ROC curves for four actions of THUMOS14: Tennis swing, Golf swing, Throw Discus, and Baseball pitch. The results are shown for Negative Mining approach [26] (green), CRANE [32] (yellow) and Proposed method (red).

shots or clips. We, then, compute video action proposals within each clip by assuming that each clip contains the action. We compute the action scores of all proposals within the shot and obtain the most action representative proposal in every shot. We consider the action score as a action detection score and evaluate localization accuracy using intersection over union metric at 10% threshold. The mean ROC curve for all four actions is shown in Figure 7. Again, we compare our results with weakly supervised annotation methods [32, 26] and obtain better results. Improved results as compared to strong baseline methods signify the effectiveness of the proposed approach. We use lower threshold criterion due to extreme difficulty of the dataset. Even though the results of all three methods are lower as compared to state-of-art results on similar actions in trimmed datasets, we consider these results encouraging, due to the complexity of dataset. The qualitative results for all four actions are shown in Figure 6.

Figure. 8 shows some of typical failure cases on THUMOS14 dataset. The figure on the top-left shows a frame from golf swing video. In this video of more than 5000 frames, the complete golf swing action happens only for 500 frames. In the rest of the video, the person is teaching golf swing techniques and performing in-complete golf swing action several times. Although, we achieve good localization over the actor, our method has problem in distinguishing complete action from the in-complete ones. Other failures occurred due to actor's occlusion and blurred video.

8. Conclusion

We present a new approach to spatio-temporally localize an action in a single video. As compared to previous similar works, we don't assume availability of multiple videos, prior annotations or clean images. Our experimental results show that impressive action localization can be achieved by reconstructing candidate action locations by leveraging freely available internet images. Our framework tackles noisy images through random walk and sparse representation, removes background and camera generated video proposals through optical flow gradients and preserves temporal smoothness of video by enforcing consistency of coefficient vectors across multiple frames. Our extensive experiments on trimmed as well as un-trimmed action datasets validate the effectiveness of proposed ideas and the framework.

Acknowledgments: We thank Boqing Gang for valuation discussions on this project.

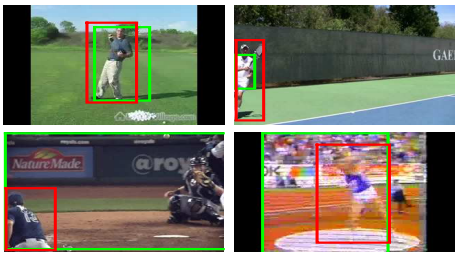

Figure 8: Failure cases on THUMOS14 dataset

References

[1] K. Bleakley and J.-P. Vert. The group fused lasso for multiple change-point detection. In *arXiv preprint arXiv:1106.4199*, 2011. 6

[2] C.-Y. Chen and K. Grauman. Watching unlabeled video helps learn new human actions from very few labeled snapshots. In *CVPR*, 2013. 3

[3] M. Cho, S. Kwak, C. Schmid, and J. Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *CVPR*, 2015. 1, 2, 3, 4

[4] A. Dehghan, H. Idrees, and M. Shah. Improving semantic concept detection through the dictionary of visually-distinct elements. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 5

[5] A. B. Deselaers, T. and V. Ferrari. Weakly supervised localization and learning with generic knowledge. In *IJCV*, 2012. 1

[6] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015. 1, 3

[7] E. Gafni and D. Bertsekas. Two-metric projection methods for constrained optimization. In *SIAM Journal on Control and Optimization*, 1982. 5

[8] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012. 6

[9] M. Guillaumin and V. Ferrari. Large-scale knowledge transfer for object localization in imagenet. In *CVPR*, 2012. 1

[10] H. D. III. Frustratingly easy domain adaptation. In *ACL*, 2007. 6

[11] M. Jain, J. van Gemert, H. Jegou, P. Bouthemy, and C. Snoek. Action localization with tubelets from motion. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 740–747, June 2014. 4, 6

[12] M. Jain, J. C. van Gemert, and C. G. M. Snoek. What do 15,000 object categories tell us about classifying and localizing actions? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 3

[13] Y.-G. Jiang, J. Liu, A. Roshan Zamir, M. Piccardi, I. Laptev, M. Shah, and R. Sukthankar. Thumos challenge: Action recognition with a large number of classes. 2013. 2, 6

[14] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014. 2, 6

[15] A. Joulin, K. Tang, and L. Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *ECCV*, 2014. 2

[16] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *CVPR*, 2013. 2, 3

[17] G. Kim, L. Sigal, and E. Xing. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *CVPR*, 2014. 2, 3

[18] S. Manen, M. Guillaumin, and L. Van Gool. Prime object proposals with randomized prim's algorithm. In *ICCV*, 2013. 3, 4

[19] H. Moonesinghe and P.-N. Tan. Outlier detection using random walks. In *Tools with Artificial Intelligence, 2006. ICTAI '06. 18th IEEE International Conference on*, 2006. 3

[20] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid. Spatio-Temporal Object Detection Proposals. 2014. 2, 4

[21] D. P. Papadopoulos, A. D. F. Clarke, F. Keller, and V. Ferrari. Training object class detectors from eye tracking data. In *ECCV*, 2014. 1

[22] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012. 1, 3

[23] M. Rodriguez, J. Ahmed, and M. Shah. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 2, 6

[24] M. Schmidt. Graphical model structure learning with l1-regularization. In *Ph.D. Thesis*, 2010. 5

[25] N. Shapovalova, A. Vahdat, K. Cannons, T. Lan, and G. Mori. Similarity constrained latent support vector machine: An application to weakly supervised action classification. In *ECCV*, 2012. 2

[26] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In *ECCV*, 2012. 3, 6, 7, 8

[27] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, 2015. 3

[28] K. Soomro, R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. In *ICCV*, 2013. 2, 6

[29] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei. Co-localization in real-world images. In *CVPR*, pages 1464–1471, June 2014. 1

[30] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei. Co-localization in real-world images. In *CVPR*, 2014. 2

[31] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller. Shifting weights: Adapting object detectors from image to video. In *NIPS*, 2012. 3

[32] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei. Discriminative segment annotation in weakly labeled video. In *CVPR*, 2013. 3, 6, 7, 8

[33] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *CVPR*, 2013. 1, 2, 6

[34] J. van Gemert, M. Jain, G. Ella, and C. Snoek. Apt: Action localization proposals from dense trajectories. In *BMVC*, pages 740–747, June 2015. 4, 6

[35] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. 2015. 3, 5, 6

[36] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *CVIU*, 115(2):224–241, 2010. 1

[37] M. Yang, L. Zhang, D. Zhang, and S. Wang. Relaxed collaborative representation for pattern classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012. 5