View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Universidad Carlos III de Madrid e-Archivo

This is a postprint version of the following published document:

Murtaza, F., Yousaf, M.H., Velastin, S.A. (2018). PMHI: Proposals from Motion History Images for Temporal Segmentation of Long Uncut Videos. *IEEE Signal Processing Letters*, 25(2), pp. 179-183.

DOI: 10.1109/LSP.2017.2778190

© 2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

PMHI: Proposals from Motion History Images for Temporal Segmentation of Long Uncut Videos

Fiza Murtaza, Muhammad Haroon Yousaf, Sergio A. Velastin SMIEEE

***Abstract*—This letter proposes a method for the generation of temporal action proposals for the segmentation of long uncut video sequences. The presence of consecutive multiple actions in video sequences makes the temporal segmentation a challenging problem due to the unconstrained nature of actions in space and time. To address this issue, we exploit the non-action segments present between the actual human actions in uncut videos. From the long uncut video, we compute the energy of consecutive non-overlapping Motion History Images (MHIs) which provides spatiotemporal information of motion. Our Proposals from MHIs (PMHI) are based on clustering the MHIs into actions and non-action segments by detecting minima from the energy of MHIs. PMHI efficiently segments the long uncut videos into a small number of non-overlapping temporal action proposals. The strength of PMHI is that it is unsupervised, which alleviates the requirement for any training data. Our temporal action proposal method outperforms existing proposal methods on the MuHAVi-uncut and CVPR 2012 Change Detection datasets with an average recall rate of 86.1% and 86.0% respectively.**

***Index Terms*—Motion History Images (MHIs), temporal segmentation, uncut videos, MuHAVi-uncut, Change Detection CVPR-2012, action proposals.**

I. Introduction

WITH the advent of digital cameras, and smartphones, multimedia collections in the form of videos are increasing day by day. Videos are captured and analyzed for various purposes i.e. sharing over the internet, surveillance, content-based search and retrieval, sports analysis, wildlife monitoring, etc. These application domains can benefit significantly from automatic recognition of desired actions in long uncut videos containing multiple actions. To achieve this objective, temporal segmentation of the videos encompassing multiple consecutive actions is required [1]. To address this issue, there is a need for computer vision algorithms to automatically segment the long uncut videos in a meaningful manner as shown in Fig. 1.

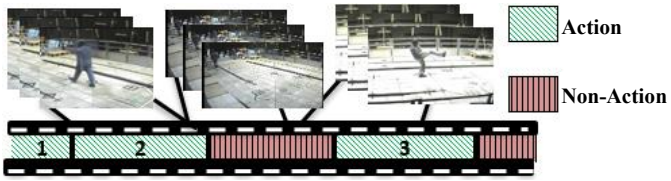


Fig. 1. Visualization of temporal video segments of an uncut video having three action and 2 non-action regions. Our method retrieves the locations of the action and non-action regions in an unsupervised way.

Most existing methods exhaustively apply an action classifier at every frame in a sliding window fashion for video segmentation [2-6]. These approaches are computationally expensive for the analysis of large-scale videos. In [7-10] researchers used training data comprised of manually segmented videos to learn key-instances in uncut videos. Temporal localization of actions was then performed through supervised learning. Such solutions are not attractive as they require manually segmented videos for training purpose.

The exhaustive computation of video classifiers (sliding window) has been avoided in [11-16]. In those methods, first a number of candidate segments containing human actions, known as *action proposals*, are produced. Then an action classifier is applied for action recognition. Recent methods produce either spatiotemporal proposals using tube-let [14], the action-ness measurement [17], action tubes [15], segments based upon dense trajectories [12] or temporal proposals using e.g. fast activity proposals [18], the bag of fragments [13] and Gaussian process regression (GPR) [19] based methods. For proposal generation these methods either use hierarchical grouping methods or the dense trajectories of motion fields, which are computationally expensive for large-scale videos [18].

In this letter, we propose Proposals from Motion History Images (PMHI) which generate the temporal action proposals in long duration uncut videos in an unsupervised manner. We have the following contributions. First, we propose a clustering algorithm that can segment the Motion History Images (MHIs) into actions and non-action segments. Second, our approach is unsupervised hence it does not require prior training which

Revised manuscript submitted on August 30, 2017. Sergio A Velastin acknowledges funding by the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement nº 600371, el Ministerio de Economía y Competitividad (COFUND2013-51509) and Banco Santander. Authors also acknowledges support from the Directorate of ASR & TD, University of Engineering and Technology Taxila, Pakistan. F. Murtaza and M.H. Yousaf are with Department of Computer Engineering, University of Engineering and Technology Taxila, Pakistan. S.A. Velastin is with University Carlos III Madrid Spain, Avenida Gregorio Peces-Barbas, 22 28270 Colmenarejo, Madrid, SPAIN and with Queen Mary University of London, UK

elevates the need for training data. Third, we experimentally demonstrate that the small number of non-overlapping temporal proposals can segment long uncut videos more accurately than the methods producing a large number of overlapping spatiotemporal proposals [12]. Experiments show that PMHI outperforms the recall rate of recent methods on the MuHAVi-uncut [20] dataset as well as the CVPR 2012 Change Detection dataset [21]. The MuHAVi-uncut is relatively new [22] and thus, to our knowledge, it has not yet been used for temporal segmentation purposes. Therefore, our work can also be used as a baseline for the temporal segmentation of the MuHAVi-uncut dataset.

II. Action Proposals from Motion History Images

Here, the temporal segmentation of long uncut videos is done by generating the temporal action proposals from Motion History Images (MHIs). Fig. 2 shows an overview of the approach. Given a long untrimmed video, we compute multiple MHIs over non-overlapping temporal windows of fixed size. We next cluster those MHIs to find temporal action proposals by finding the Energy Minima between energies of each MHIs. To output these non-overlapping temporal proposals, the long uncut video is efficiently segmented into actions and non-action regions.

The following subsections describe the approach in detail.

A. Generating Motion History Images

We have used readily available silhouettes data which may have noise due to imperfect image segmentation, therefore, pre-processing i.e. noise reduction steps might be needed. Then, we compute consecutive non-overlapping Motion History Images (MHIs) for every τ frames. We used MHIs because they can effectively represent the human motion in spatiotemporal fashion [23]. MHIs encodes how recently motion occurred at a pixel. Let $I(x,y,t)$ be a binary silhouette image, where $I(x,y,t)=1$ denotes that the pixel at location (x,y) contains foreground at time t. The function $M(x,y,t)$ computes the MHI at time t as:

$$M(x,y,t)=\begin{cases}\tau & \text{if } I(x,y,t)=1\\ \max(0,M(x,y,t-1)-1 & \text{otherwise}\end{cases} \quad (1)$$

where τ is the size of the temporal window and $t=1:\tau$. For a long untrimmed video with N frames, there are total $w=N/\tau$ non-overlapping temporal windows. For every k^{th} temporal window, $MHI_k(x,y)$ is calculated using lines 2-5 of Algorithm 1. We store these MHIs in a sequential order, e.g. $MHI_1(x,y)$ is computed for the first τ frames then the window is moved to the next τ frames to compute $MHI_2(x,y)$ and so on. This sequence is necessary because in this way we can have information for the starting and ending frames for each MHI.

B. Clustering of MHIs into Action Proposals

To generate a set of action proposals for an uncut video, we propose a clustering algorithm to cluster MHIs into actions and non-action proposals. For clustering, first we project the

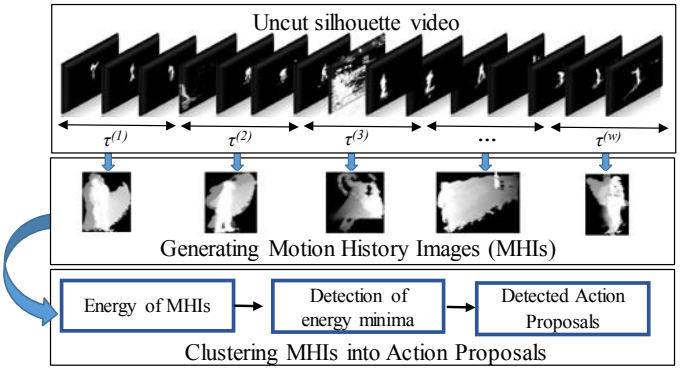


Fig. 2. Block diagram of our proposed method

Algorithm 1: Finding $MHI_k(x,y)$

Input: Silhouette frames $I(x,y,t)$ of uncut videos and τ
Output: $MHI_k(x,y)$ for all temporal windows
Procedure:
1: **for** $k=1:w$ do % w is the total temporal windows
2: **for** $t=1:\tau$ **do**
3: Find $M(x,y,t)$ using (1)
4: **end for**
5: $MHI_k(x,y)=M(x,y,\tau)$ % after above loop $t=\tau$
6: **end for**

spatiotemporal information of every k^{th} MHI into only temporal information by finding its energy, E_k:

$$E_k=\sum_{x,y} MHI_k(x,y) \quad (2)$$

where $k=1:w$. The energy from each MHI is concatenated in a vector form as $E=[E_1|...|E_w]$. Each E_k is normalized to E_k' using unity-based normalization given as:

$$E_k'=\frac{E_k-\min(E)}{\max(E)-\min(E)}. \quad (3)$$

Long uncut videos mostly have non-action regions having lower energies as compared to action regions. Therefore, we use Energy Minima to detect the boundaries between actions. Once we calculated the normalized energy E'_k for each MHI, we next cluster the MHIs by finding the energy minima $E_{\min}$ using Algorithm 2. We find the temporal locations of non-action segments, i.e. G, by concatenating in G those values of k for which $E'_k \le E_{min}$ (line 3 of Algorithm 2). The values for G are concatenated (line 3) until the ratio R (given in line 6) between the length of G and total temporal windows w is greater than the threshold r. The *card(.)* in line 6 of Algorithm 2 represents the length of the vector. We will show in Section III B, how the different values of r (line 1) affect the results. We find the temporal locations of the action regions, i.e. A, by taking the complement of non-action locations, i.e. G, with total w possible locations. The length of A will be $m=w-n$ where n is the length of G.

Finally, we cluster all the action locations A in action proposals P_a, based on their locality using:

Algorithm 2: Finding the temporal locations A of action regions

Input: $E_{\min}=0,\ E'_k, r, w,\ G=[\ ]$

Output: $E_{\min}$, A

Procedure:

1: **while** $R > r$ **do** % r is threshold value given in section III B
2: **for** $k=1:w$ **do**
3: $G=\begin{cases}[G \mid k] & if\ E'_k \le E_{\min} \\ G & otherwise\end{cases}$ % G is temporal locations of non-action segments
4: **end for**
5: $E_{\min}=E_{\min}+0.01$ % 0.01 is the step size
6: $R=card(G)/w$ % *card(.)* finds the length
7: **end while**
8: $A=comp(G,[1:w])$ % *comp(.)* finds the complement

$$P_a=\begin{cases}[P_a \mid A(i)] & if\ A(i+1)-A(i)=1 \\ a=a+1 & otherwise\end{cases} \quad (4)$$

Using (4), for all temporal action locations $i=1:m$, we obtain a number of proposals i.e. $[P_1,...,P_a]$, which are temporally non-overlapping. Hence we can directly use these proposals as multiple temporal segments of a long uncut video.

III. Experimental Results

A. Datasets and Evaluation measure

For evaluation purposes, we chose MuHAVi-uncut [20] and the thermal videos of the CVPR 2012 Change Detection (CCD) dataset [21]. MuHAVi-uncut is a dataset of long RGB video recordings (8 cameras) of people doing prescribed actions [22]. The dataset provides a set of silhouettes obtained by a good but not perfect foreground estimation algorithm. As a result, videos contain noise that any action recognition or temporal segmentation algorithm needs to cope with. MuHAVi-uncut has salt and pepper noise, typically of size less than 15x15 pixels which is removed (for all the experiments, including comparison with other methods) using a median filter of size 15x15 [21]. This dataset also provides a ground truth consisting of temporal markers and action labels. It has a large variation in styles of execution, camera viewpoints, and contains background clutter and movement. We chose the CCD dataset as it also contains readily available silhouette videos having consecutive actions and non-action segments similar to MuHAVi-uncut. It has a large variation in object size and intensity contrast. All experiments are performed using MATLAB 2016 with Intel Core i3 at 1.70 GHz, 4GB RAM, in a 64-bit operating system.

We measure the quality of the action proposals by calculating the temporal overlap between each detected action proposal and the available ground truth action regions. To do this, we compute the *temporal Intersection over Union (tIoU)* (similar to [18]) of time intervals of the ground truth segment and predicted action proposal. If the tIoU of a predicted proposal is above a predefined tIoU threshold, the detection is considered as a true positive otherwise a false positive.

For evaluation, we also used detection rate η and over segmentation ratio γ [19] given by:

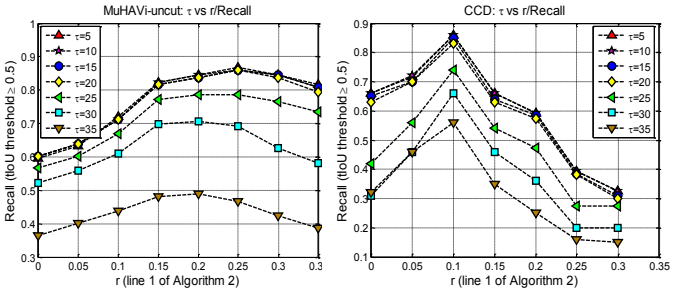
$$\eta=\frac{\text{number of True Positive Propsals}}{\text{number of segments in ground truth}} \quad (5)$$

$$\gamma=\frac{\text{number of False Positive Proposals}}{\text{number of segments in ground truth}}. \quad (6)$$

In (5), the True Positive Proposals means correct detections for a specified tIoU threshold. The value of η ranges between 0 and 1 (1 being best), whereas $\gamma > 0$ indicates that extra segments are detected which are not part of ground truths.

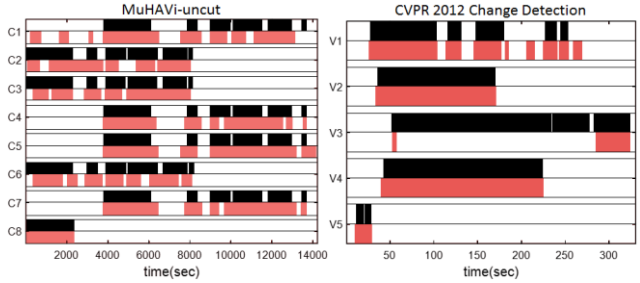
B. Evaluating PMHI Parameters

We measured and plotted in Fig. 3, different recall values for tIoU threshold ≥ 0.5 by iteratively changing the values of r (Algorithm 1) and fixing the value of τ. In Algorithm 2, r is the proportion of non-action frames present in a video while τ is related to the temporal range of a movement. The choices of τ and r are somewhat dependent on video content. The parameter τ needs to be small enough so that it does not tend to encompass both action and non-action and large enough so that it captures what we might call "atomic" actions. In a typical video, this would be around half to one second or 10-25 frames, something that it is not too complicated to observe manually. In fact, we have experimentally observed (Fig. 3 for both datasets), that for a range of τ values from 5 to 20, recall is similar because two successive MHIs are quite similar (but the smaller values would result in longer processing times) and recall worsens with larger values of τ movement. Therefore, from now on we use τ=20 for both datasets. The value of r is related to the proportion of non-action frames present in the data and that could be estimated quite well by manual observations. Experiments confirmed this when finding that



(a) (b)

Fig. 3. Evaluation of r (x-axis) versus recall (y-axis) for different values of τ for the (a) MuHAVi-uncut and (b) CCD (change detection dataset).



(a) (b)

Fig. 4. Final segmentation results for tIoU≥0.5 where the ground truth is shown in black color, segmentation results of our method is shown in red for (a) MuHAVi-uncut and (b) CCD dataset.

TABLE I
TEMPORAL ACTION SEGMENTATION RESULTS FOR MUHAVI-UNCUT AND CCD DATASET USING TIOU THRESHOLD ≥ 0.5

Video Name	Recall (%)	Precision (%)	Recall (%)	Precision (%)
	PMHI (our)		APT [12]	
MuHAVi- uncut dataset				
C1:Camera1	94.1	80.0	50.0	53.0
C2:Camera2	53.0	50.0	100	54.4
C3:Camera3	94.1	94.1	90.0	56.0
C4:Camera4	88.2	62.5	29.4	52.0
C5:Camera5	94.1	67.0	43.1	30.3
C6:Camera6	71.0	67.0	70.0	44.1
C7:Camera7	94.1	76.2	50.0	39.0
C8:Camera8	100	100	50.0	53.0
Average	**86.1**	**74.6**	60.3	48.0
CCD dataset				
V1:corridor	80	66.7	20.0	35.7
V2:diningRoom	100	50.0	100	90.0
V3:lakeSide	100	100	33.3	85.0
V4:library	100	100	100	100
V5:park	50	100	33.3	100
Average	**86.0**	**83.3**	57.3	82.1

r=0.25 for MuHAVi-uncut and r=0.1 for the CCD dataset (values that are consistent with the corresponding proportions of non-action frames), result in good recall rates for a range of values of τ. Therefore, although an adaptive search method would be useful, simple video contents observation gives a good approximation to appropriate values of τ and r. In Fig. 4 we also present how close (in locations) the obtained segments are in comparison to the available ground truth locations of candidate segments.

C. Action Proposals Quality

We compared the quality of our generated action proposals with Action localization Proposals from dense Trajectories (APT) [12]. APT is an unspuervised method originally aimed to find spatiotemporal proposals, and we have used only the temporal information that it computes for direct comparisons.

A good temporal segmentation method should achieve a high recall rate (but considerably good precision rate too) by finding as many true activity segments in a video as possible [18] and we analyzed the quality of competing temporal segmentation methods using precision and recall measures for tIoU threshold ≥ 0.5. In Table I, we summarize the comparison results of our method with APT for both the MuHAVi-uncut and the CCD dataset. Our method achieves a good average recall of 86.1% for MuHAVi-uncut as compared to APT. Similarly, Table 1 shows that APT produces less average precision rate because APT produces many overlapping false positive proposals. For CCD we achieve an average recall of 86.0% while APT achieves an average recall of 57.3%. In Fig. 5 we plot recall rate of our method in comparison with APT. We observe a better recall behavior of our method against APT for different tIoU thresholds.

In Table II, we summarize the comparison results of our method with [12] and [19] based on detection rate η using (5) and over segmentation ratio γ using (6) on CCD dataset. The method proposed in [19] is unsupervised in nature, and a one-class classification (OCC) technique is used based on Gaussian process regression (GPR) [19]. Table II shows that our method does not detect extra segments ($\gamma = 0$) for all videos except V1. For V1, extra segments are detected due to background variation. For V3 and V5, our method has low detection rate also due to background variation. From Table III, comparison results for MuHAVi-uncut dataset show that our method obtains high average detection rate of 0.86 and less average over-segmentation ratio of 0.37 compared to APT. Results from Table I-III reveal that non-overlapping temporal proposals, generated by our method, can segment long uncut videos more accurately than the APT which produces many overlapping spatiotemporal proposals.

TABLE II
COMPARISON RESULTS FOR CCD DATASET USING TIOU THRESHOLD ≥ 0.5

Video Name	η	γ	η	γ	η	γ
	PMHI (our)		GPR [19]		APT [12]	
V1:corridor	1.00	0.60	0.68	0.49	0.20	1.00
V2:diningRoom	1.00	0.00	0.90	0.48	1.00	0.00
V3:lakeSide	0.33	0.00	0.23	0.05	0.33	0.33
V4:library	1.00	0.00	1.00	1.67	1.00	1.50
V5:park	0.50	0.00	0.00	0.00	1.00	0.00
Average	**0.77**	**0.12**	0.56	0.53	0.71	0.56

TABLE III
COMPARISON RESULTS FOR MUHAVI-UNCUT DATASET USING TIOU THRESHOLD ≥ 0.5

Video Name	η	γ	η	γ
	PMHI (our)		APT [12]	
C1:Camera1	0.94	0.24	0.52	1.31
C2:Camera2	0.53	0.52	0.57	2.18
C3:Camera3	0.94	0.06	0.80	0.90
C4:Camera4	0.88	0.53	1.00	0.69
C5:Camera5	0.94	0.47	0.60	0.64
C6:Camera6	0.71	0.35	0.47	0.17
C7:Camera7	0.94	0.29	1.00	0.19
C8:Camera8	1.00	0.00	0.90	0.82
Average	**0.86**	**0.37**	0.73	0.86

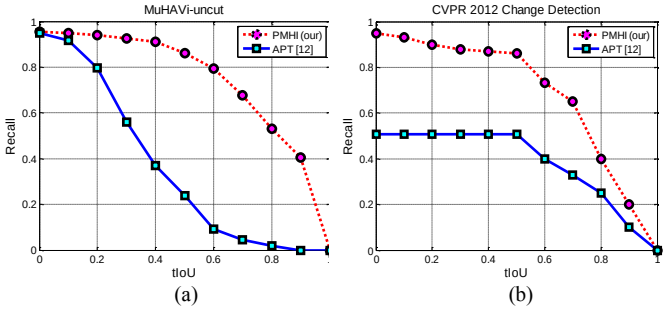


Fig. 5. Recall at different tIoU thresholds for (a) MuHAVi-uncut (b) CCD dataset.

IV. CONCLUSION

In this letter, we have proposed PMHI for the generation of temporal action proposals by finding the energy minima in MHIs. PMHI produces non-overlapping action proposals which can directly segment the uncut videos having both actions non-action segments. The results obtained on the large and challenging MuHAVi-uncut dataset and also on CCD dataset revealed that detection of Energy minima from the Energy of MHIs can discriminate between actions and non-action regions accurately. The proposed method is unsupervised and hence it saves time for long and complex videos. In future work, we can model the relationship between the segmented regions (actions and non-action regions) to have even better results. We will also plan perform action recognition on the detected segments to classify between different action classes.

REFERENCES

[1] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem, "Daps: Deep action proposals for action understanding," in *European Conference on Computer Vision*, 2016, pp. 768-784.

[2] D. Oneata, J. Verbeek, and C. Schmid, "The LEAR submission at Thumos 2014," 2014.

[3] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 1194-1201.

[4] T. Lan, Y. Wang, and G. Mori, "Discriminative figure-centric models for joint action localization and recognition," in *2011 International Conference on Computer Vision*, 2011, pp. 2003-2010.

[5] S. Karaman, L. Seidenari, and A. Del Bimbo, "Fast saliency based pooling of Fisher encoded dense trajectories," in *ECCV THUMOS Workshop*, 2014, p. 6.

[6] C. Orrite, M. Rodriguez, E. Herrero, G. Rogez, and S. A. Velastin, "Automatic segmentation and recognition of human actions in monocular sequences," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*, 2014, pp. 4218-4223.

[7] K.-T. Lai, F. X. Yu, M.-S. Chen, and S.-F. Chang, "Video event detection by inferring temporal instance labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2243-2250.

[8] K.-T. Lai, D. Liu, M.-S. Chen, and S.-F. Chang, "Recognizing complex events in videos by learning key static-dynamic evidences," in *European Conference on Computer Vision*, 2014, pp. 675-688.

[9] C. Sun, S. Shetty, R. Sukthankar, and R. Nevatia, "Temporal localization of fine-grained actions in videos by domain transfer from web images," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 371-380.

[10] J. Yuan, Z. Liu, and Y. Wu, "Discriminative video pattern search for efficient action detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33, pp. 1728-1743, 2011.

[11] G. Yu and J. Yuan, "Fast action proposals for human action detection and search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1302-1311.

[12] J. van Gemert, M. Jain, E. Gati, and C. Snoek, "APT: Action localization proposals from dense trajectories," in *BMVC*, 2015, p. 4.

[13] P. Mettes, J. C. van Gemert, S. Cappallo, T. Mensink, and C. G. Snoek, "Bag-of-fragments: Selecting and encoding video fragments for event detection and recounting," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 2015, pp. 427-434.

[14] M. Jain, J. Van Gemert, H. Jégou, P. Bouthemy, and C. G. Snoek, "Action localization with tubelets from motion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 740-747.

[15] G. Gkioxari and J. Malik, "Finding action tubes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 759-768.

[16] S. Ma, L. Sigal, and S. Sclaroff, "Learning activity progression in lstms for activity detection and early detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1942-1950.

[17] W. Chen, C. Xiong, R. Xu, and J. J. Corso, "Actionness ranking with lattice conditional ordinal random fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 748-755.

[18] F. C. Heilbron, J. C. Niebles, and B. Ghanem, "Fast Temporal Activity Proposals for Efficient Detection of Human Actions in Untrimmed Videos," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 1914-1923, 2016.

[19] M. V. Krishna, P. Bodesheim, M. Körner, and J. Denzler, "Temporal video segmentation by event detection: A novelty detection approach," *Pattern recognition and image analysis,* vol. 24, p. 243, 2014.

[20] S. Singh, S. A. Velastin, and H. Ragheb, "Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, 2010, pp. 48-55.

[21] N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, "Changedetection. net: A new change detection benchmark dataset," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, 2012, pp. 1-8.

[22] F. Murtaza, M. H. Yousaf, and S. A. Velastin, "Multi-view human action recognition using 2D motion templates based on MHIs and their HOG description," *IET Computer Vision,* 2016.

[23] F. Murtaza, M. H. Yousaf, and S. A. Velastin, "Multi-view Human Action Recognition using Histograms of Oriented Gradients (HOG) Description of Motion History Images (MHIs)," in *2015 13th International Conference on Frontiers of Information Technology (FIT)*, 2015, pp. 297-302.