Movement Tracks for the Automatic Detection of Fish Behavior in Videos

Declan McIntosh [1] **Tunai Porto Marques** [1] **Alexandra Branzan Albu** [1] **Rodney Rountree** [2] **Fabio De Leo** [3]

Abstract

Global warming is predicted to profoundly impact ocean ecosystems. Fish behavior is an important indicator of changes in such marine environments. Thus, the automatic identification of key fish behavior in videos represents a much needed tool for marine researchers, enabling them to study climate change-related phenomena. We offer a dataset of sablefish (Anoplopoma fimbria) startle behaviors in underwater videos, and investigate the use of deep learning (DL) methods for behavior detection on it. Our proposed detection system identifies fish instances using DL-based frameworks, determines trajectory tracks, derives novel behavior-specific features, and employs Long Short-Term Memory (LSTM) networks to identify startle behavior in sablefish. Its performance is studied by comparing it with a state-of-the-art DL-based video event detector.

1. Introduction

Among the negative impacts of climate change in marine ecosystems predicted for global warming levels of $1.5°$C to $2°$C (*e.g.* significant global mean sea level rise [1], sea-ice-free Artic Oceans [2], interruption of ocean-based services) are the acidification and temperature rise of waters.

The behavioral disturbance in fish species resulting from climate change can be studied with the use of underwater optical systems, which have become increasingly prevalent over the last six decades [3, 4, 5]. However, advancements in automated video processing methodologies have not kept pace with advances in the video technology itself. The manual interpretation of visual data requires prohibitive amounts of time, highlighting the necessity of semi- and fully-automated methods for the enhancement [6, 7] and annotation of marine imagery.

[1]University of Victoria, BC, Canada [2]Biology Department, University of Victoria, BC, Canada [3]Ocean Networks Canada, BC, Canada. Correspondence to: Alexandra Branzan Albu <aalbu@uvic.ca>.

Tackling Climate Change with Machine Learning Workshop at the 34^{th} Conference on Neural Information Processing Systems (NeurIPS 2020).

As a result, the field of automatic interpretation of underwater imagery for biological purposes has experienced a surge of activity in the last decade [8]. While numerous works propose the automatic detection and counting of specimen [9, 10, 11], ecological applications require more complex insights. Video data provides critical information on fish behavior and interactions such as predation events, aggressive interactions between individuals, activities related to reproduction and startle responses. The ability to detect such behavior represents an important shift in the semantic richness of data and scientific value of computer vision-based analysis of underwater videos: from the focused detection and counting of individual specimens, to the context-aware identification of fish behavior.

Given the heterogeneous visual appearance of diverse behaviors observed in fish, we initially focus our study on a particular target: startle motion patterns observed in sablefish (*Anoplopoma fimbria*). Such behavior is characterized by sudden changes in the speed and trajectory of sablefish movement tracks.

We propose a novel end-to-end behavior detection framework that considers 4-second clips to 1) detect the presence of sablefish using DL-based object detectors [12]; 2) uses the Hungarian algorithm [13] to determine trajectory tracks between subsequent frames; 3) measures four handcrafted and behavior-specific features and 4) employs such features in conjunction with LSTM networks [14] to determine the presence of startle behavior and describe it (*i.e.* travelling direction, speed, and trajectory). The remainder of this article is structured as follows. In Section 2 we discuss works of relevance to the proposed system. Section 3 describes the proposed approach. In Section 4 we present a dataset of sablefish startle behaviors and use it to compare the performance of our system with that of a state-of-the-art event detector [15]. Section 5 draws conclusions and outlines future work.

2. Related Works

Related works to our approach include DL-based methods for object detection in images and events in videos.

Deep learning-based object detection for static images. Krizhevsky *et al.* [16] demonstrated the potential of using Convolutional Neural Networks (CNNs) to extract and

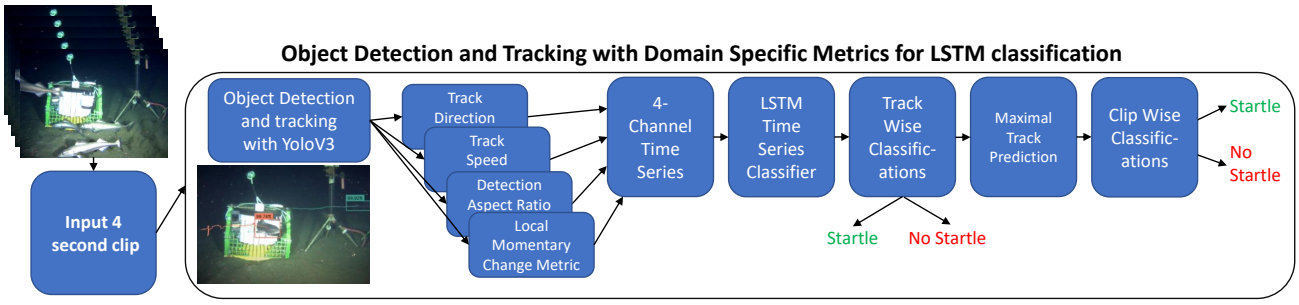

Figure 1. Computational pipeline of the fish behavior detection system proposed. The framework is able to provide clip-wise and movement track-wise classifications alike (see 4.2).

classify visual features from large datasets. Their work motivated the use of CNNs in object detection, where frameworks perform both *localization* and *classification* of regions of interest (RoI). Girshick *et al*. [17] introduced R-CNN, a system that uses a traditional computer vision-based technique (selective search [18]) to derive RoIs where individual classification tasks take place. Processing times are further reduced in Fast R-CNN [19] and Faster R-CNN [20]. A group of frameworks [21, 22, 23] referred to as "one-stage" detectors proposed the use of a single trainable network for both creating RoIs and performing classification. This reduces processing times, but often leads to a drop in accuracy when compared to two-stage detectors. Recent advancements in one-stage object detectors (*e.g*. a loss measure that accounts for extreme class imbalance in training sets [23]) have resulted in frameworks such as YOLOv3 [12] and RetinaNet [23], which offer fast inference times and performances comparable with that of two-stage detectors.

DL-based event detection in videos. Although object detectors such as YOLOv3 [12] can be used in each frame of a video, they often ignore important temporal relationships. Rather than individual images employed by aforementioned methods, recent works [24, 15, 25, 26, 27, 28, 29] used video's inter-frame temporal relationship to detect relevant events. Saha *et al*. [25] use Fast R-CNN [19] to identify motion from RGB and optical flow inputs. The outputs from these networks are fused resulting in action tubes that encompass the temporal length of each action. Kang *et al*. [24] offered a video querying system that trains specialized models out of larger and more general CNNs to be able to efficiently recognize only specific visual targets under constrained view perspectives with claimed processing speed-ups of up to 340×. Coşar *et al*. [28] combined an object-tracking detector [30], trajectory- and pixel- based methods to detect abnormal activities. Ionescu *et al*. [27] offered a system that not only recognizes motion in videos, but also considers context to differentiate between normal (*e.g*. a truck driving on a road) and abnormal (*e.g*. a truck driving on a pedestrian lane) events.

Yu *et al*. [15] proposed ReMotENet, a light-weight event detector that leverages spatial-temporal relationships between objects in adjacent frames. It uses 3D CNNs ("spatial-temporal attention modules") to jointly model these video characteristics using the same trainable network. A frame differencing process allows for the network to focus exclusively on relevant, motion-triggered regions of the input frames. This simple yet effective architecture results in fast processing speeds and reduced model sizes [15].

3. Proposed approach

We propose a hybrid method for the automatic detection of context-aware, ecologically relevant behavior in sablefish. We first describe our method for tracking sablefish in video, then propose the use of 4 track-wise features to constrain sablefish startle behaviors. Finally, we describe a Long Short Term Memory (LSTM) architecture that performs classification using the aforementioned features.

Object detection and tracking. We use the YOLOv3 [12] end-to-end object detector as the first step of this hybrid method. The detector was completely re-trained to perform a simplified detection task were only the class *fish* is targeted. We use a novel 600-image dataset (detailed in 4.1) of sablefish instances composed of data from Ocean Networks Canada (ONC) to train the object detector.

The detection of each frame offer a set of bounding boxes and spatial centers. In order to track organisms we associate these detection between frames. Our association loss value consists simply of the distance between detection centers in two subsequent frames. We employ the Hungarian Algorithm [13] to generate a loss minimizing associations between detection in two consecutive frames. We then remove any associations where the distance between various detection is greater than 15% of the frame resolution. Tracks are terminated if no new detection is associated with them for 5 frames (*i.e*. 0.5 seconds—see 4.1).

Behavior Specific Features. We propose the use of a series

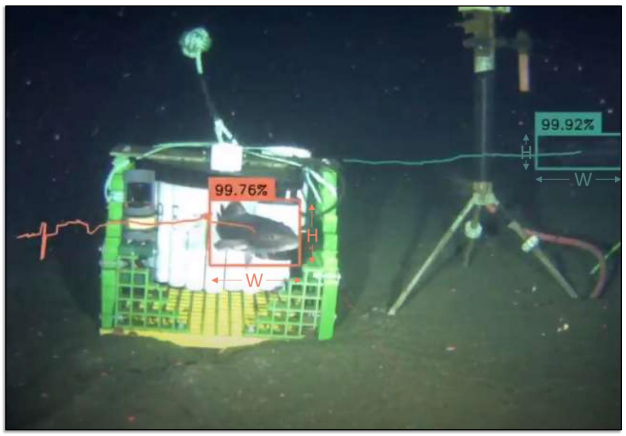

Figure 2. Sample movement tracks and object detection confidence scores. The bounding boxes highlight the *fish* detection in the current frame. Each color represents an individual track.

of four domain-specific features that describe the startle behavior of sablefish. Each feature conveys independent and complementary information, and the limited number of features (4) prevents over-constraining the behavior detection problem.

The first two features quantify the *direction of travel* and *speed* from a track. These track characteristics were selected because often the direction of travel changes and the fish accelerates at the moment of a startle motion.

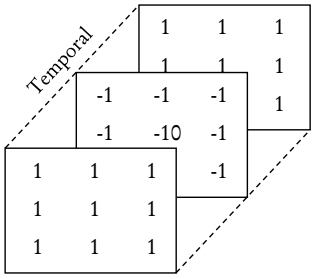

Figure 3. LMCM kernel designed to extract fast changes in sequential images.

A third metric considers the *aspect ratio* of the detection bounding box of a *fish* instance over time. The reasoning behind this feature is the empirical observation that sablefish generally take on a "c" shape when startling, in preparation for moving away from their current location. The final *Local Momentary Change Metric (LMCM)* feature seeks to find fast and unstained motion, or temporal visual impulses, associated with startle events. This feature is obtained by convolving the novel 3-dimensional LMCM kernel, depicted in Figure 3, over three temporally adjacent frames. This spatially symmetric kernel was designed to produce high output values where impulse changes occur between frames. Given its zero-sum and isotropic properties, the kernel outputs zero when none or only constant temporal changes are occurring. We observe that the LMCM kernel efficiently detects leading and lagging edges of motion. In order to associate this feature with a track we average the LMCM output magnitude inside a region encompassed by a given fish detection bounding box.

3.1. LSTM classifier

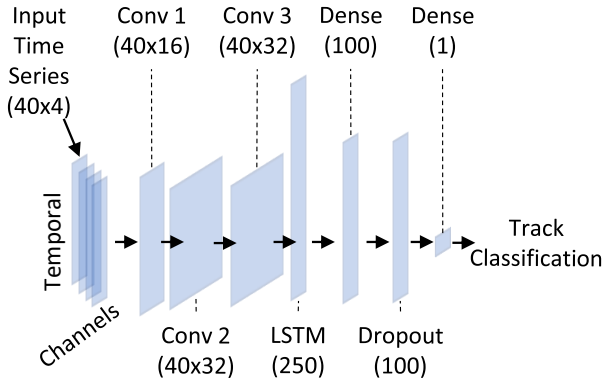

Figure 4. LSTM-based network for the classification of movement tracks. The network receives four features conveying speed, direction and rate of change from each track as input.

In order to classify an individual track, we first combine its four features as a *tensor* data structure of dimensions $(40, 4)$; tensors associated with tracks of less than 40 frames are end-padded with zeros. A set of normalization coefficients calculated empirically using the training set (see 4.1) is then used to normalize each value in the input time-series to the range $[-1, 1]$. A custom-trained long short-term memory (LSTM) classifier receives the normalized tensors as input, and outputs a track classification of *non-startle background* or *startle*, as well as a confidence score. This is done by considering underlying temporal relationships between their values. We chose to use LSTM networks because the temporal progression of values from the features extracted (along 40 frames or 4 seconds) conveys important information for the classification of individual clips/tracks. Figure 4 details the architecture of the LSTM network employed. All convolutional layers employ three-layered 1D kernels.

4. Results and Discussion

We compare our method with a state-of-the-art event detection algorithm [15]. Section Section 4.1 describes our dataset. Our comparison considers standard performance metrics in Section 4.2 and discusses the potential advantages of semantically richer data outputs in ecological research.

4.1. Sablefish Startle Dataset

The data used in this work was acquired with a stationary video camera permanently deployed at 620m of depth at the Barkley Node of Ocean Networks Canada's [1] NEPTUNE marine cabled observatory infrastructure. All videos samples were collected between September 17th and October 24th 2019 because this temporal window contains high sablefish activity. The original monitoring videos are first divided into units of 4-second clips (a sablefish startle is expected to last roughly one second) and down-sampled to 10 frames per second for processing reasons. An initial filtering is carried out using Gaussian Mixture Models [31], resulting in a set composed only by clips which contain motion. For training purposes, these motion clips are then manually classified as possessing startle or not. The Sablefish Startle dataset consists of 446 positive (*i.e*. presenting startle events) clips, as well as 446 randomly selected negative samples (*i.e*. without startle events). Table 1 details the dataset usage for training, validation and testing.

Data Split	Clips	Startle Clips	Tracks	Startle Tracks
Train	642	321	1533	323
Validation	150	75	421	80
Test	100	50	286	50

Table 1. Division of the 4-second clips of the Sablefish Startle Dataset for training, validation and testing purposes.

A second dataset composed of 600 images of sabblefish was created to train the YOLOV3 [12] object detector (*i.e*. first step of the proposed approach). In order to assess the track-creation performance of the proposed system, we use this custom-trained object detector to derive movement tracks from each of the 892 clips composing the Sablefish Startle Dataset. Tracks with length shorter than 2 seconds are discarded. The remaining tracks are manually annotated as startle or non-startle (see Table 1).

This dual annotation approach (*i.e*. clip- and track-wise) employed with the Sablefish Startle Dataset allows for a two-fold performance analysis: 1) *clip-wise classification*, where an entire clip is categorized as possessing startle or not, and 2) *track-wise classification*, which reflects the accuracy in the classification of each candidate track as startle or non-startle.

4.2. Experimental Results

We calculate the Average Precision (AP), Binary Cross Entropy (BCE) loss and Recall for both track- and clip-wise outputs of the proposed system using only the Test portion of the Sablefish Startle Dataset. A threshold of 0.5 (in a $[0, 1]$ range) is set to classify a candidate movement track as positive or negative with respect to all of its constituent points. In order to measure the performance of the clip-wise classification and compare it with the baseline method (ReMotENet [15]), we consider that the "detection score" of a clip is that of its highest-confidence movement track (if any). Thus, any clip where at least one positive startle movement track is identified will be classified as positive.

The conversion from track-wise classification to clip-wise classification is expected to lower the overall accuracy of our proposed approach. A "true" startle event might create only short, invalid tracks, or sometimes no tracks at all. This situation would lower the clip-wise classification performance, but would not interfere with the track-wise one. The track-wise metrics are applicable only to our approach and they mainly reflect the difference between the manual and automatic classification of the tracks created in the dataset by the proposed system, thus evaluating the ability of the LSTM network to classify tracks. Table 2 shows that the LSTM portion of our method performs well for classifying startle tracks (AP of 0.85). Clip-wise, our method outperformed a state-of-the-art DL-based event detector [15] with an AP of 0.67.

Method	Track AP	Track BCE	Clip AP	Clip Recall
Ours	**0.85**	**0.412**	**0.67**	**0.58**
ReMotENet [15]	N/A[1]	N/A[1]	0.61	0.5

[1]: ReMotENetdoes not perform track-wise classification.

Table 2. Performance comparison between the proposed method and a state-of-the-art event detector.

Our proposed system generates more semantically rich data by detecting and describing behavior rather than just marking a clip as containing the behavior; this may be helpful for ecological and biological research. Our approach provides instance-level information such as track-specific average speed, direction and rate of change. These extra data may allow for further analyses considering inter- and intra-species behaviors. Also, there are clips that contain more than one startle, and our approach is able to identify all startle instances in such clips. Simply classifying a clip as startle or non-startle would, of course, not allow for the detection of multiple startle instances within the same clip.

5. Conclusion

We propose an automatic detector of fish behavior in videos that performs semantically richer tasks than typical computer vision-based analyses: instead of specimens counting, it identifies and describes a complex biological event (startle events of sablefish). Our intent is to enable long-term studies on changes in fish behavior that could be caused by climate change (*e.g*. temperature rise and acidificaton).

A dataset composed of 892 4-second positive (startle) and negative (non-startle) clips, and associated tracks were manually annotated. Experiments using this data show that the proposed detector identifies and classifies well individual tracks of motion as startle or not (AP of 0.85). Furthermore,

[1] www.oceannetworks.ca/

the performance of our clip-wise classification is compared to that of a state-of-the-art event detector, ReMotENet [15]. Our system outperforms ReMotENet with an AP of 0.67 (against 0.61).

Future work will address more fish behaviors (*e.g.* predation, spawining) and will adapt DL-based event detectors such as ReMotENet [15] and NoScope [24] to that end.

References

[1] Thomas F Stocker, Dahe Qin, G-K Plattner, Melinda MB Tignor, Simon K Allen, Judith Boschung, Alexander Nauels, Yu Xia, Vincent Bex, and Pauline M Midgley. Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of ipcc the intergovernmental panel on climate change, 2014.

[2] Nathaniel L Bindoff, Peter A Stott, Krishna Mirle AchutaRao, Myles R Allen, Nathan Gillett, David Gutzler, Kabumbwe Hansingo, G Hegerl, Yongyun Hu, Suman Jain, et al. Detection and attribution of climate change: from global to regional. 2013.

[3] Delphine Mallet and Dominique Pelletier. Underwater video techniques for observing coastal marine biodiversity: a review of sixty years of publications (1952–2012). *Fisheries Research*, 154:44–62, 2014.

[4] Tunai Porto Marques, Alexandra Branzan Albu, and Maia Hoeberechts. A contrast-guided approach for the enhancement of low-lighting underwater images. *Journal of Imaging*, 5(10):79, 2019.

[5] Jacopo Aguzzi, Carolina Doya, Samuele Tecchio, Fabio De Leo, Ernesto Azzurro, Cynthia Costa, Valerio Sbragaglia, Joaquin del Rio, Joan Navarro, Henry Ruhl, Paolo Favali, Autun Purser, Laurenz Thomsen, and Ignacio Catalán. Coastal observatories for monitoring of fish behaviour and their responses to environmental changes. *Reviews in Fish Biology and Fisheries*, 25:463–483, 2015.

[6] Tunai Porto Marques and Alexandra Branzan Albu. L2uwe: A framework for the efficient enhancement of low-light underwater images using local contrast and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 538–539, 2020.

[7] Cosmin Ancuti, Codruta Orniana Ancuti, Tom Haber, and Philippe Bekaert. Enhancing underwater images and videos by fusion. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 81–88. IEEE, 2012.

[8] J. Aguzzi1, D. Chatzievangelou, J.B. Company, L. Thomsen, S. Marini, F. Bonofiglio, F. Juanes, R. Rountree, A. Berry, R. Chumbinho, C. Lordan, J. Doyle, J. del Rio, J. Navarro, F.C. De Leo, N. Bahamon, J.A. García, R. Danovaro, M. Francescangeli, V. Lopez-Vazquez1, and Ps Gaughan. The potential of video imagery from worldwide cabled observatory networks to provide information supporting fish-stock and biodiversity assessment. *ICES Journal of Marine Science*, In press.

[9] YH Toh, TM Ng, and BK Liew. Automated fish counting using image processing. In *2009 International Conference on Computational Intelligence and Software Engineering*, pages 1–5. IEEE, 2009.

[10] Concetto Spampinato, Yun-Heh Chen-Burger, Gayathri Nadarajan, and Robert B Fisher. Detecting, tracking and counting fish in low quality unconstrained underwater videos. *VISAPP (2)*, 2008(514-519):1, 2008.

[11] Song Zhang, Xinting Yang, Yizhong Wang, Zhenxi Zhao, Jintao Liu, Yang Liu, Chuanheng Sun, and Chao Zhou. Automatic fish population counting by machine vision and a hybrid deep neural network model. *Animals*, 10(2):364, 2020.

[12] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[13] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[15] Ruichi Yu, Hongcheng Wang, and Larry S Davis. Remotenet: Efficient relevant motion event detection for large-scale home surveillance videos. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1642–1651. IEEE, 2018.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[18] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[19] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[22] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[24] Daniel Kang, John Emmons, Firas Abuzaid, Peter Bailis, and Matei Zaharia. Noscope: optimizing neural network queries over video at scale. *arXiv preprint arXiv:1703.02529*, 2017.

[25] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. *arXiv preprint arXiv:1608.01529*, 2016.

[26] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015.

[27] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2019.

[28] Serhan Coşar, Giuseppe Donatiello, Vania Bogorny, Carolina Garate, Luis Otavio Alvares, and François Brémond. Toward abnormal trajectory and event detection in video surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):683–695, 2016.

[29] Huijuan Xu, Boyang Li, Vasili Ramanishka, Leonid Sigal, and Kate Saenko. Joint event detection and description in continuous video streams. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 396–405. IEEE, 2019.

[30] Duc Phu Chau, François Brémond, Monique Thonnat, and Etienne Corvée. Robust mobile object tracking based on multiple feature similarity and trajectory filtering. *arXiv preprint arXiv:1106.2695*, 2011.

[31] Chris Stauffer and W Eric L Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 2, pages 246–252. IEEE, 1999.