Exploiting Spatio-Temporal Coherence for Video Object Detection in Robotics

D. Fernandez-Chaves[1,2][0000−0002−5596−1103],
J.L. Matez-Bandera[1][0000−0003−4123−7330],
J.R. Ruiz-Sarmiento[1][0000−0002−9929−5309], J. Monroy[1][0000−0001−7869−7811],
N. Petkov[2][0000−0003−2163−8647], and J. Gonzalez-Jimenez[1][0000−0003−3845−3497]

[1] Machine Perception and Intelligent Robotics group (MAPIR). Dept. of System Engineering and Automation. Biomedical Research Institute of Malaga (IBIMA). University of Malaga. Spain.
{davfercha,josematez,jotaraul,jgmonroy,javiergonzalez}@uma.es
[2] Johann Bernoulli Institute of Mathematics and Computing Science, University of Groningen, The Netherlands n.petkov@rug.nl

Abstract. This paper proposes a method to enhance video object detection for indoor environments in robotics. Concretely, it exploits knowledge about the camera motion between frames to propagate previously detected objects to successive frames. The proposal is rooted in the concepts of planar homography to propose regions of interest where to find objects, and recursive Bayesian filtering to integrate observations over time. The proposal is evaluated on six virtual, indoor environments, accounting for the detection of nine object classes over a total of ∼7k frames. Results show that our proposal improves the recall and the F1-score by a factor of 1.41 and 1.27, respectively, as well as it achieves a significant reduction of the object categorization entropy (58.8%) when compared to a two-stage video object detection method used as baseline, at the cost of small time overheads (120ms) and precision loss (0.92).

1 Introduction

The detection of the objects appearing in a sequence of images (i.e. a video) is of paramount importance for many applications, such as those involving mobile robots [4, 8, 17]. For this particular problem, the exploitation of the spatio-temporal information inherent in the sequence of images, is considered an important factor to boost the object detection performance [2, 3, 15].

Previous works have proposed the use of Spatio-Temporal Networks (STNs) such as tubelets-based [19, 10, 11] or memory-based approaches [20, 1]. Yet, these techniques share a common drawback: they either use a fixed-length temporal window or apply a post-processing phase to the whole video sequence to integrate the observations over time. The latter prevents their use in real-time applications, like the ones relying on a mobile robot, as they require to take decisions upon the detected objects.

Multiple contributions have addressed these handicaps by including motion-guided propagation (MGP) algorithms such as object tracking networks [15, 5] or

optical flow [12, 10, 21, 22]. However, relying on visual information alone is prone to failures under challenging conditions like frames with motion blur, occlusions or appearance changes [10].

In this paper we propose an alternative method that, assuming knowledge about the camera motion between successive frames, leverages this information to enhance the detection of objects in a sequence of images. Concretely, we consider a typical two-stage object detection method consisting of a Region Proposal Network (RPN) that yields regions of interest where an object can be found, followed by an Object Classifier Network (OCN) that processes each region and returns a probability distribution over a given set of object classes for each one. Thus, the method outcome after processing each frame is a set of observations, each one corresponding to a region in the image, and their associated probability distributions. To provide temporal and spatial coherence to these results, our method introduces a motion-guided propagation model based on planar homography, obtained from the camera motion, to propagate previous observations to the frame being processed. Next, we perform a correspondence step that try to match the regions provided by the RPN in the frame t with the previous propagated observations from $t-1$. In the case that an object observation becomes orphaned in the matching process (i.e. a detected object in frame $t-1$ is not proposed by the RPN in frame t), we propose it as a new region of interest. This results in an enhanced set of regions to be classified by the OCN. Finally, to provide further temporal coherence, the probability distributions of matched observations are integrated through a recursive Bayesian filter.

In order to evaluate the benefits of our proposals we have conducted multiple experiments over the robotic dataset Robot@VirtualHome [7]. We show that our combined method boosts video object detection by significantly increasing the recall (i.e. the number of unnoticed objects) while presenting a minor reduction of the precision and a very small time overhead.

2 Method Overview

Given a sequence of frames $\mathcal{F}_0, ..., \mathcal{F}_{t-1}, \mathcal{F}_t$, we propose a spatio-temporal object detection method that incorporates knowledge from frame $\mathcal{F}_{t-1}$ to frame $\mathcal{F}_t$. Specifically, for each new frame $\mathcal{F}_t$, we employ a two-stage detection pipeline (see Figure 1). In a first phase, we rely on a region proposal network (RPN) to obtain regions of interest B_t^i, typically known as bounding boxes. These regions are rectangular boxes in the image whose enclosed pixels form an image patch π_t^i where objects are expected to be found.

In a second phase, all image patches are evaluated by an object classifier, which yields a discrete probabilistic distribution $P(\mathbf{C}_t|\pi_t^i)$ over the considered set of objects class labels $\mathbf{C} = \{C^1, ..., C^N\}$.

It must be noticed that RPNs tend to predict multiple bounding boxes with different shapes and sizes for the same object. Thus, in this work we filter out redundant candidates by selecting the most appropriate one for each object. For this purpose, we employ a Non-Maximum Suppression (NMS) algorithm for bounding boxes. Similarly, related to the classification probabilities of each

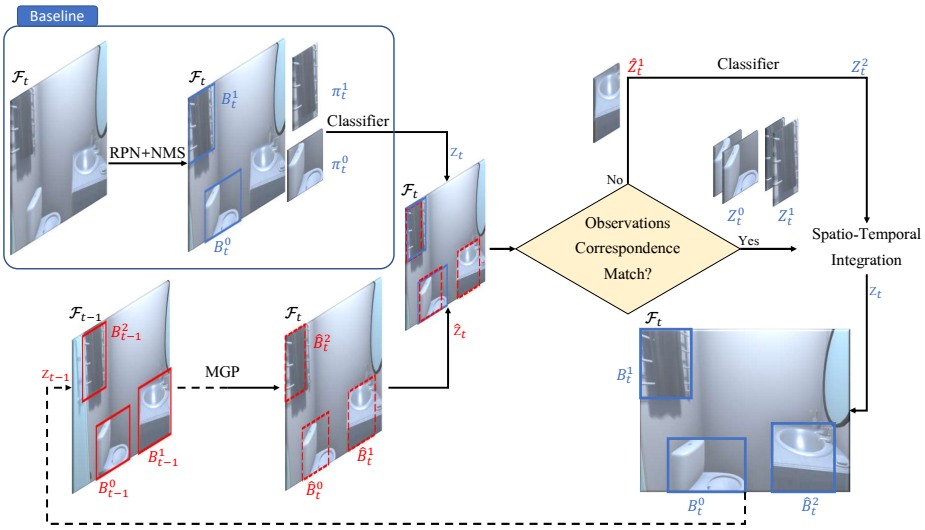

Fig. 1. Diagram showing the proposed method pipeline. Bounding boxes of the detected objects in frame $\mathcal{F}_{t-1}$ (in red) are projected to the next frame $\mathcal{F}_t$ (in blue) through a planar homography. Next, projected bounding boxes are matched with the new proposed bounding boxes given by the RPN. Black dashed line represents the feedback for the next iteration. The knowledge from both is temporally integrated to improve object detections in videos.

image patch, we filter out candidates whose highest probability of belonging to an object class is lower than a given threshold, reducing false detections.

At this point, we define an observation Z_t^i by the pair formed by a bounding box and the probability distribution resulting from the classification of its respective image patch: $Z_t^i = \{B_t^i, P(\mathbf{C}_t|\pi_t^i)\}$. Then, after obtaining the observations corresponding to the current frame $\mathcal{F}_t$, we compute the projections $\hat{\mathbf{Z}}_t$ of previous observations $\mathbf{Z}_{t-1}$ to propagate previous knowledge into the current frame (see Section 2.1). Subsequently, we carry out a correspondence step between projected observations $\hat{\mathbf{Z}}_t$ and new observations $\mathbf{Z}_t$ (see Section 2.2), and finally, we integrate both by applying a recursive Bayesian filter (see Section 2.3).

2.1 Motion-Guided Propagation Model

A MGP model allows us to propagate observations between frames by projecting the bounding box of each observation from one frame B_{t-1}^i to the next $\hat{B}_t^i$ (see Figure 1). In this work, we propose to transform the corner points which define each bounding box by means of a planar homography:

$$p_t = H p_{t-1} \tag{1}$$

where p_t and p_{t-1} are the homogeneous coordinates of the bounding box corners in two consecutive frames. The planar homography matrix H is computed from

the camera motion [9] as:

$$H = K\left(R - \frac{tn^T}{d}\right)K^{-1} \tag{2}$$

where R and t are the rotation matrix and the translation vector of the camera between the two frames, and K is the intrinsic camera matrix. n and d are the normal vector and the distance, respectively, to the 3D plane where the bounding box lies in the scene.

It must be emphasized that applying this planar homography transformation to the bounding boxes of the observations does not transform the objects themselves, because objects are not in a plane. However this transformation can be considered an approximate projection of the object observations. Moreover, for the specific case of mobile robotics, given that most robots are non-holonomic (i.e. their translation is only along the z-axis $t = [0, 0, t_z]$), we can consider that the camera translation between frames w.r.t. the distance to objects d is sufficient small, thus $t_z/d \simeq 0$. Thus, H is approximated as follows:

$$H = KRK^{-1} \tag{3}$$

Upon projecting the bounding boxes of all previous observations, we disregard those that fall outside the image plane.

2.2 Correspondence Step and Region Proposal

For each new frame $\mathcal{F}_t$, we have a set of new observations $\mathbf{Z}_t = \{Z_t^1, ..., Z_t^J\}$ which we desire to integrate with previous observations $\mathbf{Z}_{t-1} = \{Z_{t-1}^1, ..., Z_{t-1}^I\}$. To this end, the MGP model (see Section 2.1) projects previous observations $\mathbf{Z}_{t-1}$ into the current frame $\hat{\mathbf{Z}}_t = \{\hat{Z}_t^1, ..., \hat{Z}_t^I\}$, where $\hat{Z}_t^i = \{\hat{B}_t^i, P(\mathbf{C}_t|\pi_{t-1}^i)\}$.

In this step, we perform a correspondence between $\mathbf{Z}_t$ and $\hat{\mathbf{Z}}_t$ to determine three possible outcomes: (i) an observation in $\mathbf{Z}_t$ refers to a possible new object, (ii) an observation in $\mathbf{Z}_t$ refers to a previously detected one; or (iii) a projected observation in $\hat{\mathbf{Z}}_t$ has not been detected in the current frame. To do so, we measure the similarity s_{ij} between each pair of observations $(\hat{Z}_t^i, Z_t^j)$ as follows:

$$s_{ij}(Z_t^j, \hat{Z}_t^i) = \text{IoU}(B_t^j, \hat{B}_t^i) \tag{4}$$

where $\text{IoU}(\cdot,\cdot)$ is the intersection over union function [16].

For each projected observation $\hat{Z}_t^i$, we select the pair $(Z_t^j, \hat{Z}_t^i)$ that maximizes the similarity function. Then, if the similarity is greater than a threshold $\mathcal{T}$, we integrate both observations by choosing the most recent bounding box B_t^j and updating the probability distributions through a recursive Bayesian filter (see Section 2.3).

However, since observations in $t-1$ may not be proposed again by the RPN in t (e.g. due to motion blur), projected observations $\hat{\mathbf{Z}}_t$ may be left alone, not matching with any new observation Z_t^j. It must be noticed that the latter does not implies that the object is not in the current frame, just that it is not proposed. To address this fact, we classify the image patch $\hat{\pi}_t^i$ associated to the

projection of the previous bounding box $\hat{B}_t^i$ into the current frame $\mathcal{F}_t$, obtaining a new probability distribution $P(\mathbf{C}_t|\hat{\pi}_t^i)$. Next, we update the detection $\hat{Z}_t^i$ to Z_t^j, defining it with $\hat{B}_t^i$ and updating the probability distribution $P(\mathbf{C}_{t-1}|\pi_{t-1}^i)$ to $P(\mathbf{C}_t|\hat{\pi}_t^i)$ through a recursive Bayesian filter.

2.3 Bayesian Filtering over Object Class Labels

Seeking to capitalize on the temporal correlation inherent in the posterior distributions of matched observations along a sequence of images, we resort to a recursive Bayesian filter to estimate the accumulated belief over the object classes $Bel(\mathbf{C}_t) = P(\mathbf{C}_t|\pi_{1:t})$:

$$Bel(\mathbf{C}_t) \propto P(\pi_t|\mathbf{C}_t, \mathbf{C}_{1:t-1}) \sum_{n=1}^{N} P(\mathbf{C}_t|\mathbf{C}_{1:t-1}^n) Bel(\mathbf{C}_{t-1}^n) \tag{5}$$

where $P(\pi_t|\mathbf{C}_t, \mathbf{C}_{1:t-1})$ is the conditional density at time t, N is the number of object classes, and $P(\mathbf{C}_t|\mathbf{C}_{1:t-1})$ is the transition probability. Assuming first order Markov properties, i.e. independence between object classes and between observations, we have $P(\mathbf{C}_t|\mathbf{C}_{1:t-1}) = P(\mathbf{C}_t|\mathbf{C}_{t-1})$ and $P(\pi_t|\mathbf{C}_t, \mathbf{C}_{1:t-1}) = P(\pi_t|\mathbf{C}_t)$. Thus, our accumulated belief is simplified to:

$$Bel(\mathbf{C}_t) \propto P(\pi_t|\mathbf{C}_t) \sum_{n=1}^{N} P(\mathbf{C}_t|\mathbf{C}_{t-1}^n) Bel(\mathbf{C}_{t-1}^n) \tag{6}$$

The transition probability function $P(\mathbf{C}_t|\mathbf{C}_{t-1})$ is the function that controls how the object classes evolve over time. We expressed this function as follows:

$$P(\mathbf{C}_t|\mathbf{C}_{t-1}) = \begin{cases} p_c \, s_{ij} & \text{if } \mathbf{C}_t = \mathbf{C}_{t-1} \\ \dfrac{1 - p_c \, s_{ij}}{N-1} & \text{otherwise} \end{cases} \tag{7}$$

where s_{ij} is the similarity score between the bounding boxes of both observations and p_c is the probability that given two consecutive observations, the object class with maximum probability of both observations is the same. The latter value should be set with a higher probability in order to model the fact that in a video, two observations (with similar position, shape and size) from two consecutive frames have a high likelihood to be from the same object class.

Finally, note that the Bayesian filter requires the conditional density $P(\pi_t|\mathbf{C}_t)$, but the object classifier yields the posterior probability $P(\mathbf{C}_t|\pi_t)$. However, both probabilities are related through Bayes theorem as follows:

$$P(\pi_t|\mathbf{C}_t) \propto \frac{P(\mathbf{C}_t|\pi_t)}{P(\mathbf{C}_t)} \tag{8}$$

where $P(\mathbf{C}_t)$ is the marginal class probability. This probability encodes the probability of finding each object class in an environment, hence it is a prior that can be learned from experimental data. For example, in a household, objects such as *chairs* that are found in most rooms must have a greater probability than less common objects such as *microwaves* that are only typically found in kitchens.

3 Experiments

This section covers a set of comparative experiments aimed to evaluate the performance of the proposed method, and the contribution of each of its stages. Concretely, we present a comparison of the following incremental methods: i) B: baseline, ii) B + BF: including Bayesian filter, iii) B + BF + P: adding propagation without homography and iv) Our method: improving the previous one by using the motion-guided propagation model with homography.

3.1 Experimental Setup

To assess the performance of the proposed method we have conducted experiments with data from the state-of-the-art Robot@VirtualHome dataset [7]. This is a robotic dataset that includes sequences of images taken by a mobile robot while navigating through different virtual environments. In addition, the dataset provides the camera motion between frames and segmentation masks for the objects in the images.

We conducted experiments on six indoor environments from the dataset, which are composed by a total of $6,929$ frames with a resolution of 640×480 px. All images were captured by a frontal camera placed on the robot at a height of 1.59m and 10° rotation in the pitch-axis.

3.2 Evaluation Metrics

To measure the performance of the competing methods we resort to three commonly used metrics: average precision (AP), recall (R) and F1-score [14]. Moreover, we consider an observation as right (i.e. a true positive) when its top-1 classification probability is greater than 0.5 and its associated object class matches the ground-truth label provided by the dataset. In addition, to evaluate temporal coherence, we compute the entropy of the probability distribution associated with each observation as a measure of uncertainty.

3.3 Implementation Details

The implementation of the proposed method has been carried out according to the following aspects:

- For the region proposal network, we rely on the DeepMask architecture with the weights from [13].
- To filter out multiple bounding boxes candidates for each object, we apply the NMS algorithm from [6].
- For the object classification stage, we used the state-of-the-art EfficientNet CNN with the pretrained model *EfficientNet-Lite4* [18]. This classifier yields a discrete probability distribution over the object classes from the ImageNet dataset, from which we considered 9 relevant indoor object types: *toilet*, *chair*, *bed*, *table*, *microwave*, *washbasin*, *closet*, *washer* and *burner*.
- Regarding the parameters of the proposed method, we set empirically the threshold $\mathcal{T}$ as 0.3 and p_c as 0.6, which control when there is a match between two bounding boxes and the transition probability of the Bayesian filter, respectively.

Table 1. Averaged metric results for each evaluated method over the 6 indoor environments. B: Baseline, BF: Bayesian Filter and P: Propagation (without homography). Our method is composed by the baseline, the bayesian filter and the motion-guided propagation model based on planar homography.

	Precision	**Recall**	**F1-score**	**Entropy**	**Time (s)**
B	70.13%	19.66%	30.13%	0.51	**0.55**
B + BF	**70.60%**	19.72%	30.26%	0.31	0.56
B + BF + P	63.90%	25.66%	36.19%	0.31	0.61
Our method	64.73%	**27.63%**	**38.16%**	**0.30**	0.67

Table 2. Comparative results of propagating with/without planar homography facing different camera rotations.

	Rotation 10°			**Rotation 15°**		
	AP	R	F1	AP	R	F1
B + BF + P	62.87%	23.91%	34.27%	58.53%	25.60%	34.97%
Our method	**67.60%**	**27.37%**	**38.44%**	**72.96%**	**32.31%**	**43.83%**

- All experiments have been carried out using a computer with an Intel Core i7-8750H processor at 2.20 *GHz*, a 16 *GB* DDR4 RAM memory at 1333 *MHz*, and a graphic card NVIDIA GeForce GTX 1070 with 8 *GB* of memory.

3.4 Experimental Results

Table 1 presents the average performance of the evaluated methods over the six tested environments. As can be seen, the baseline (first row) shows the second best precision with a 70.13% and the lowest processing time per frame (0.55s). However, this method achieves the lowest recall (19.66%) and F1-score (30.13%), together with the maximum averaged entropy of the probability distributions (0.51). The inclusion of the Bayesian filter (second row) considerably decreases the averaged entropy a 58.8% w.r.t. the baseline, obtaining a value of 0.31, which implies the reduction of the uncertainty associated to the predicted classes.

A considerable performance improvement is appreciated when including the propagation (without planar homography) of previous observations (i.e. the position of bounding boxes in previous frames is preserved for next frames). In this case, the method boosts both recall and F1-score by a factor of 1.31 and 1.20 respectively, while increasing the processing time by 60ms and reducing precision by 6.23% w.r.t. baseline. The recall-precision trade-off is represented by the F1-score, which in this case shows an improvement of the performance.

Finally, the full pipeline where previous observations are propagated to next frames through planar homography yield the best results in terms of recall (27.63%), F1-score (38.16%) and average entropy (0.3). The recall enhancement reveals that the full pipeline improves object detection by proposing a considerable number of observations from previous frames that where unnoticed by the baseline, as shown in Figure 2.

However, as can be seen in Table 1, the results between propagating with planar homography (Our method) and without (B + BF + P) is similar. This

Fig. 2. Example frames where our method boost the video object detection performance during a robot rotation movement. Solid-line bounding boxes are proposed by the RPN. Dashed-line bounding boxes are projections of previous bounding boxes. The effect of Bayesian filter can be highlighted from $\mathcal{F}_{t+1}$ in advance. For the method B + BF + P can be seen how the bounding box is propagated inaccurately, so the *toilet* class probability decreases. In contrast, our method propagates more accurately the bounding box, so the observation of the *toilet* is kept and integrated over time. Note that for simplicity, we show only the object class with maximum probability.

fact is due to the assumption made in the propagation model that only considers the rotation of the camera, so that when the robot only translates, both methods are equivalent. Since in our experiments only 12% of frames show important rotations, the planar homography effect is not highlighted. To analyze the performance impact of the planar homography, we evaluated both methods considering only consecutive frames exhibiting camera rotations larger than a certain angle. The obtained results are shown in Table 2, where we can observe that the higher rotations, the higher the benefit. For example, for 15° rotations, our method outperforms the propagation without planar homography by increasing a 14.43% precision, 6.81% recall and 8.86% F1-score.

Finally, Figure 3 illustrates the average F1-score and entropy results for each considered object class. Note that there is an inverse correlation between entropy and F1-score, as classes with higher entropy have lower F1-score, so their precision and recall are lower. This is mainly due to misclassification errors, such as classifying *chairs* as *tables*. The interested reader can see the proposed method in action in the following video: `https://youtu.be/oNmGG3d0BM4`.

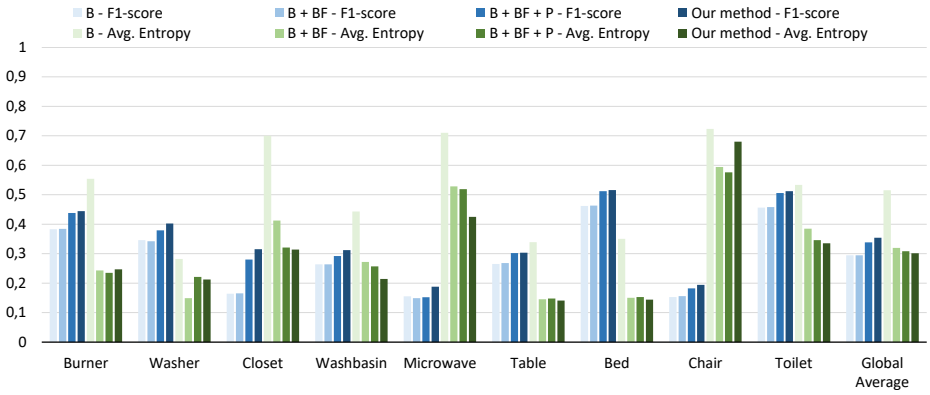

Fig. 3. F1-score (in blue) and entropy (in green). Results are obtained averaging over the 6 indoor environments, per object class and method. At the right, the global average for all object classes. Note that, for all cases, we desire a high F1-score but a low entropy.

4 Conclusions and Future Work

In this work, we have introduced a novel method to boost the detection of objects in a sequence of images given knowledge of the camera motion. Particularly, we have focused on the case of mobile robots operating in indoor environments. Our method uses a MGP model based on planar homography to spatially propagate observations from one frame to the next, allowing an efficient matching with new observations. Finally, a Bayesian filter is introduced to temporally integrate matched observations, yielding a posterior probability distribution or belief over the object classes.

Experimental validation has demonstrated how our proposal improves video object detection w.r.t. the baseline by increasing 8.03% F1-score and 7.97% recall, which implies that our method detect more objects than the baseline. Besides, our method reduces entropy by 58.8% on average, which proves the effect of the Bayesian filter by reducing the uncertainty about object classes over time. However, as drawbacks, the proposed method reduces the average precision by a factor of 0.92 w.r.t. the baseline due to the fact that also wrong detections are propagated over time.

In future work, we plan to extend this method to use a dynamic frame rate object detection based on the robot motion. Thus, each new frame will be taken after a certain robot movement, hence releasing resources such as the CPU and GPU while the view has little changes. In this way, we will reduce the computational cost, which is highly limited in robotics.

Acknowledgements. This work was supported by the research projects WISER (DPI2017-84827-R) and ARPEGGIO (PID2020-117057), the Spanish grant program FPU19/00704 and the UG PHD scholarship program from the University of Groningen.

References

1. Bertasius, G., Torresani, L., Shi, J.: Object detection in video with spatiotemporal sampling networks. In: ECCV. pp. 331–346 (2018)
2. Bosquet, B., Mucientes, M., Brea, V.M.: Stdnet-st: Spatio-temporal convnet for small object detection. Pattern Recognition **116**, 107929 (2021)
3. Chen, Y., Cao, Y., Hu, H., Wang, L.: Memory enhanced global-local aggregation for video object detection. In: IEEE/CVF CVPR. pp. 10337–10346 (2020)
4. Erol, B.A., Majumdar, A., Lwowski, J., Benavidez, P., Rad, P., Jamshidi, M.: Improved deep neural network object tracking system for applications in home robotics. In: CIPR, pp. 369–395. Springer (2018)
5. Feichtenhofer, C., Pinz, A., Zisserman, A.: Detect to track and track to detect. In: IEEE ICCV. pp. 3038–3046 (2017)
6. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence **32**(9), 1627–1645 (2009)
7. Fernandez-Chaves, D., Ruiz-Sarmiento, J., Petkov, N., Gonzalez-Jimenez, J.: Robot@virtualhome, an ecosystem of virtual environment tools for realistic indoor robotic simulation (2021), submitted
8. Fernandez-Chaves, D., Ruiz-Sarmiento, J.R., Petkov, N., Gonzalez-Jimenez, J.: From object detection to room categorization in robotics (jan 2020)
9. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision (2000)
10. Kang, K., Li, H., Yan, J., Zeng, X., Yang, B., Xiao, T., Zhang, C., Wang, Z., Wang, R., Wang, X., et al.: T-cnn: Tubelets with convolutional neural networks for object detection from videos. IEEE TCSVT **28**(10), 2896–2907 (2017)
11. Kang, K., Ouyang, W., Li, H., Wang, X.: Object detection from video tubelets with convolutional neural networks. In: IEEE CVPR. pp. 817–825 (2016)
12. Li, H., Chen, G., Li, G., Yu, Y.: Motion guided attention for video salient object detection. In: IEEE/CVF ICCV. pp. 7274–7283 (2019)
13. Pinheiro, P.O., Collobert, R., Dollár, P.: Learning to segment object candidates. In: NIPS (2015)
14. Powers, D.M.: Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061 (2020)
15. Ray, K.S., Chakraborty, S.: Object detection by spatio-temporal analysis and tracking of the detected objects in a video with variable background. Journal of Visual Communication and Image Representation **58**, 662–674 (2019)
16. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: IEEE/CVF CVPR. pp. 658–666 (2019)
17. Ruiz-Sarmiento, J.R., Guenther, M., Galindo, C., Gonzalez-Jimenez, J., Hertzberg, J.: Online context-based object recognition for mobile robots. In: ICARSC (2017)
18. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML. pp. 6105–6114. PMLR (2019)
19. Tang, P., Wang, C., Wang, X., Liu, W., Zeng, W., Wang, J.: Object detection in videos by high quality object linking. IEEE TPAMI **42**(5), 1272–1278 (2019)
20. Xiao, F., Lee, Y.J.: Video object detection with an aligned spatial-temporal memory. In: ECCV. pp. 485–501 (2018)
21. Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y.: Flow-guided feature aggregation for video object detection. In: IEEE ICCV. pp. 408–417 (2017)
22. Zhu, X., Xiong, Y., Dai, J., Yuan, L., Wei, Y.: Deep feature flow for video recognition. In: IEEE CVPR. pp. 2349–2358 (2017)