Deep Fusion of Appearance and Frame Differencing for Motion Segmentation [Supplementary Material]

Marc Ellenfeld[1,3], Sebastian Moosbauer[1], Ruben Cardenes[2], Ulrich Klauck[3,4], Michael Teutsch[1]

[1] Hensoldt Optronics GmbH, Germany [2] Hensoldt Analytics GmbH, Germany
{sebastian.moosbauer, ruben.cardenes, michael.teutsch}@hensoldt.net
[3] Aalen University of Applied Sciences, Germany [4] University of the Western Cape, South Africa
ulrich.klauck@hs-aalen.de

Abstract

In this supplementary material, we present more details on the adaptation of the CDNet-2014 dataset for motion segmentation. Furthermore, we present further experiments as well as further discuss experiments dscribed in the paper. Therefore, we show additional plots and tables.

1. Adaptation of the CDNet-2014 Dataset

CDNet-2014 is a dataset to benchmark the task *change detection*. This means that moving objects that stop moving during the sequence remain labeled as *motion* until the end of the sequence. This is not suitable to evaluate motion segmentation as we consider it in our paper. Hence, we describe an algorithm to re-label that dataset automatically to label objects that stop as *background* after some frames. In addition to the utilization of tracking, Intersection over Union (IoU), and Center of Mass (CoM), we needed some further adjustments that are not described in the paper.

When a GT object previously identified as static gets into contact with another GT object, they get merged and the individual objects can no longer be distinguished. This causes the CoM to change significantly, which in turn causes the object to be detected as being in motion. This causes static objects to sometimes reappear even though they do not move. As soon as the objects are separated again, the static object is identified as being static and is removed from the GT. This problem only occurs in very few frames of scenes like *sofa* of the category *intermittentObjectMotion*. Objects that are removed in one frame due to being static, and then immediately reappear in the next frame because of motion cause the GT to flicker. Two additional thresholds T_s and T_r are introduced to adjust the sensitivity of the algorithm and prevent this flickering effect. T_s is used to adjust how fast an object is removed from the GT if it stopped. If the object stops for more than T_s frames, it is removed. In a similar way T_r is used to decide how fast a static object should reappear in the GT. If the object is moving in the next T_r frames, it reappears in the GT.

Table 1. Configuration parameters for our adaption of CDNet-2014. T_s: Threshold in frames to declare region as static; T_{CoM}: Threshold how far Center of Mass (CoM) must move to not be static; T_{IoU}: Threshold for IoU to consider region being static (proven by T_{CoM}); T_r: Threshold for how many frames the region must move to be moving (again).

Scene	T_{IoU}	T_{CoM}	T_s	T_r
office	1	0.2	20	2
PETS2006	0.95	0.3	15	5
streetLight	0.99	0.1	20	5
tramstop	0.9	0.1	15	4
parking	1	0.001	22	0
abandonedBox	0.8	0.1	5	2
sofa	0.8	0.2	15	2
tunnelExit_0_35fps	1	0.1	1	3
copyMachine	0.95	0.175	15	3
diningRoom	0.99	0.03	20	5
corridor	0.9	0.004	20	1
lakeSide	1	0.0005	25	1
library	0.99	0.016	15	7
turbulence2	0.875	1	42	5

The final adaptation algorithm requires four parameters in total that need to be set for each scene individually. These parameter are:

- The threshold T_{IoU}, which is used to decide whether an object became static based on the IoU.

- The threshold T_{CoM}, which is used to decide whether an object became static based on the shift in position of the center of mass.

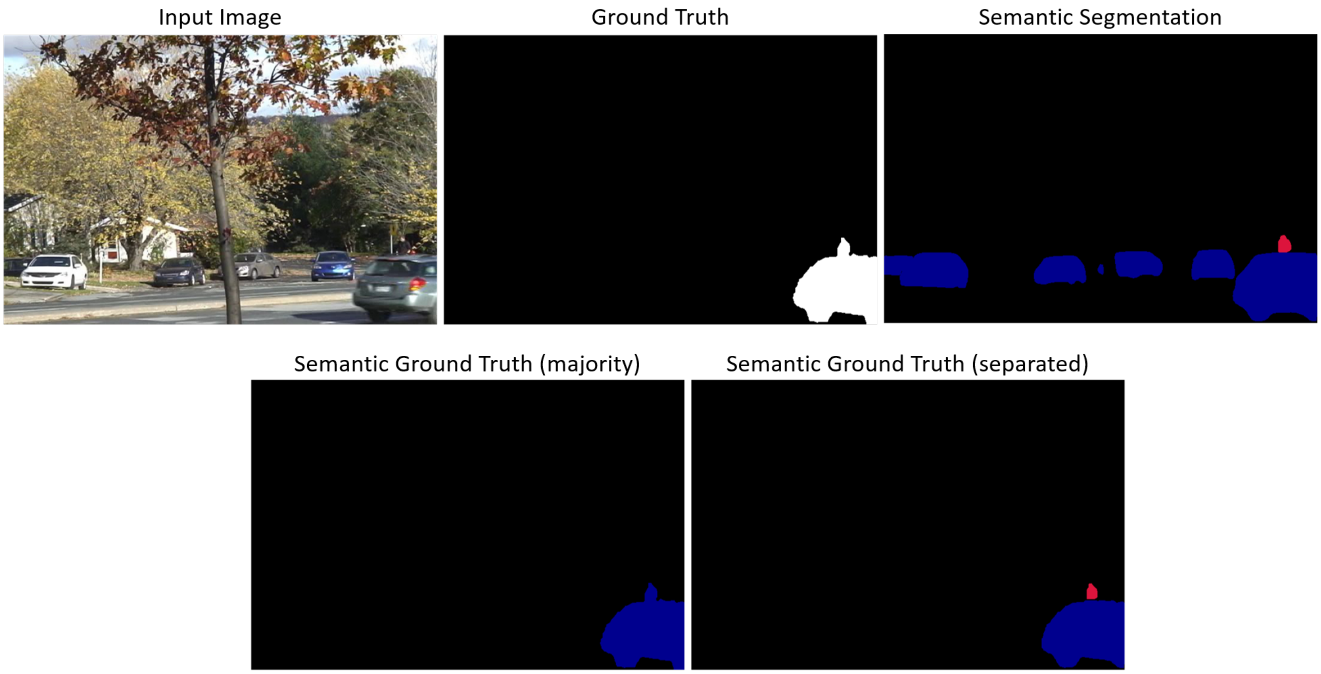
Input Image
Ground Truth
Semantic Segmentation
Semantic Ground Truth (majority)
Semantic Ground Truth (separated)

Figure 1. Creation of the new semantic GT labels based on majority voting and semantic class separation. Blue pixels indicate cars while red pixels indicate persons.

- The threshold T_s, to remove objects from the GT if they are static for more than the specified number of frames.
- The threshold T_r, to let static objects reappear in the GT if they move for more than the specified number of frames.

Especially in scenes without many overlapping objects, the parameters of the algorithm can be fine-tuned to match the behavior of the occurring objects. In scenes, where moving objects overlap with static objects, it happens that annotations of static objects reappear for a short period of frames. The big change in bounding box IoU and the position of the center of mass, causes the annotations of the static object to reappear for a small number of frames. An overview of the scene-specific paramterization of the four parameters is provided in Table 1.

After the preparation of the data set for motion segmentation, an additional version is created that also includes semantic classes of moving objects. Since the GT contains only *things*, the object classes are determined by means of instance segmentation. This has the advantage that *stuff* objects are not even considered in the segmentation and, thus, greatly simplifies the identification of the semantic class for each object. The terminology *things* and *stuff* is taken from the task of panoptic segmentation [2]. To create the instance segmentation for the adapted CDNet-2014-MotSeg, a pre-trained Mask R-CNN [3] is used. This model is chosen because it is pre-trained on the MS COCO dataset, which includes all necessary object classes that appear in CDNet-2014. The instance segmentation of the model are converted to semantic segmentation since only this information is needed. The model produces accurate segmentation results of all relevant objects. However, the segmentation differs slightly from frame to frame. For higher consistency, the result of the instance segmentation is combined with the GT of the adapted CDNet2014-MotSeg. This is done by using the semantic segmentation of the input frame and merging it with its motion GT of the CDNet2014-MotSeg dataset. For each pixel that is annotated as *motion* in the GT, the predicted semantic class is transferred into the motion GT. Usually, one GT object represents a single object. In this case a majority voting based on the pixels that are part of each GT object is used to determine the semantic class of each object. By reusing the exact shape of the original GT pixels that were not classified by the semantic segmentation are filled in and the new semantic GT stays consistent.

This approach works for most of the scenes, in which multiple different types of objects occur. However, in some scenes, several different types of objects may partially occlude each other. In these scenes, majority-based relabeling does not provide satisfactory results. For example, it sometimes happens that a person is partially occluded by a passing car. Due to the much larger area of the car, majority voting decides that the GT object represents a vehicle. To prevent this, an object separation based on the results of the semantic segmentation is applied to the GT. By adding a border around each individual object class, GT objects that belong to different semantic classes are separated from each other. After this separation the new GT objects are annotated by applying the initial majority based strategy to determine the semantic class of each object. Figure 1 shows the different results of both methods.

2. Further Experimental Results

In this section, we present further experimental results. This can be more evaluation measures for the same experiment already presented in the paper or experiments not mentioned on the paper.

2.1. Frame Differencing Ablation Study

The ablation study on frame differencing in the paper only considers the F_1-score as evaluation measure. In Table 2, we see the same evaluation for all CDNet-2014 measures. It can be seen that none of the two-frame differencing approaches performs best on any measure. Interestingly, the large structuring element (SE) for the morphological operations seems to be beneficial for reduction of false positives, which is indicated by top performances in case of Specificity, False Positive Rate, Percentage of Wrong Classification and Precision. The reason could be that the large structuring element significantly reduces the number of false negatives in the difference image before training the DCNN.

2.2. Optical Flow Experiments

Table 3 shows the results of our optical flow experiments for all measures. As can be seen the approach using two consecutive frames for optical flow calculation and the difference image used for our baseline approach does not only perform best on F_1-score, but on all measures. Interestingly, its FPR is only $0.5\,\%$, which underlines the very high Specificity of $99.4\,\%$.

2.3. Hybrid vs. Late Fusion

The hybrid fusion approach in the paper was not properly discussed. Instead of the hybrid fusion at multiple stages of the modality-specific ResNet-50 backbone, we could simple collect and concatenate the feature maps at the end of each encoder. Such an architecture for late fusion is shown in Fig. 2. Table 4 shows the comparison of our hybrid fusion DCNN with the late fusion DCNN. As expected, the model trained with hybrid fusion strategy performs better on all evaluation measures. Especially on the more challenging categories, the hybrid fusion model benefits from the richer fusion features and outperforms the late fusion approach. Both models use the same difference image input calculated using three-frame differencing with 5 frames gap, summation of difference images, and small structuring element for the morphological operations.

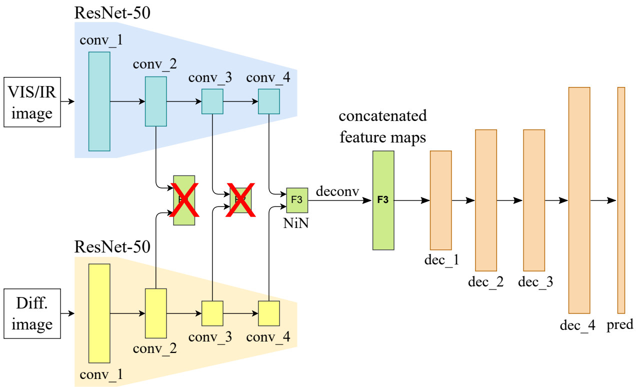

Figure 2. Overview of the modified architecture to perform a late fusion with the concatenated feature maps from both encoder streams taken only from the conv_4 stage.

2.4. Comparison with the State-of-the-Art

To further compare our approach against state-of-the-art approaches, we provide a category-wise comparison in Fig. 3 to show identify potential weaknesses of our approach. As can be seen our baseline performs best on all categories, which underlines its generalization ability. Interestingly, for the two challenging categories *camera jitter* and *low frame rate* Short Term Background Subtraction (STBGS) is our strongest competitor, while for most other categories it's the 3D convolution based DCNN based approach [1].

2.5. Detailed Comparison with [1]

Figure 4 shows a comparison of our approach against Bosch's [1] DCNN based approach, for different threshold values from 0.1 to 0.9. The radar chart shows a large margin in F_1-score for our approach. Furthermore, it can be seen that our approach is more robust regarding the choice of the confidence threshold as most threshold values show similar or even equal performance, while the compared approach seems to be more dependent on the chosen threshold.

2.6. Additional Qualitative Evaluation

In this section, we show some more example images to show strong results but also weaknesses of our approach. Figure 5 shows good examples for a variety of scenes and challenges. Among them is thermal imagery (upper left), panning motion (lower left), and camera jitter (lower right). All moving objects present in the scene are segmented well.

The scene *port_0_17_fps* is one of the most challenging scenes of the dataset. The extremely low frame rate in combination with low contrast between the very small-scale persons walking across the jetty and the constant swaying of

Table 2. Ablation study on frame differencing variants as input for the multi-modal DCNN. The best performing approach for each measure is indicated in bold font.

Diff. frames	ΔF	Fusion	Morph. SE	Conf. thrs.	Re	Sp	FPR	FNR	PWC	Pr	F_1
3	5	sum	small	0.4	0.751	0.993	0.006	0.242	1.146	0.774	**0.745**
3	2	sum	small	0.3	0.750	0.993	0.006	0.249	1.190	0.755	0.717
3	1	sum	small	0.3	0.722	0.993	0.007	0.277	1.291	0.766	0.710
3	5	sum	large	0.4	0.73	**0.994**	**0.005**	0.264	**1.102**	**0.781**	0.733
3	2	sum	large	0.4	0.756	0.993	0.006	0.243	1.178	0.768	0.739
3	1	sum	large	0.5	0.710	0.992	0.007	0.289	1.386	0.749	0.687
3	5	min	small	0.3	0.755	0.994	0.005	0.244	1.160	0.776	0.735
3	2	min	small	0.3	0.732	0.992	0.007	0.267	1.274	0.756	0.708
3	1	min	small	0.3	0.706	0.990	0.009	0.293	1.542	0.720	0.658
3	5	min	large	0.3	**0.760**	0.993	0.006	**0.239**	1.200	0.776	0.738
3	2	min	large	0.4	0.726	0.991	0.008	0.273	1.411	0.770	0.709
3	1	min	large	0.4	0.707	0.991	0.008	0.292	1.376	0.735	0.668
2	10		small	0.4	0.712	0.994	0.005	0.287	1.231	0.771	0.700
2	5		small	0.3	0.742	0.992	0.007	0.257	1.261	0.761	0.720
2	1		small	0.4	0.695	0.993	0.006	0.304	1.262	0.734	0.681
2	10		large	0.2	0.701	0.989	0.011	0.298	1.658	0.734	0.662
2	5		large	0.5	0.707	0.992	0.007	0.292	1.441	0.715	0.677
2	1		large	0.5	0.674	0.992	0.008	0.325	1.496	0.737	0.641

Table 3. All evaluation measures for our experiment using optical flow as input for the DCNN.

ΔF	last channel	Conf. thrs.	Re	Sp	FPR	FNR	PWC	Pr	F_1
1	Mag	0.3	0.665	0.979	0.020	0.334	2.642	0.646	0.604
1	Diff	0.3	**0.758**	**0.994**	**0.005**	**0.241**	**1.157**	**0.779**	**0.743**
5	Mag	0.5	0.639	0.985	0.014	0.361	2.299	0.690	0.595
5	Diff	0.3	0.745	0.992	0.007	0.254	1.343	0.736	0.712

the surrounding boats make it exceptionally difficult to distinguish between actual and irrelevant motion. Even for a human, it is difficult to see the moving foreground objects in this scene. Figure 6 shows two examples, where motion is present. In the lower example image our approach misses the moving person.

References

[1] Markus Bosch. Deep learning for robust motion segmentation with non-static cameras. Master's thesis, Ulm University of Applied Sciences, Ulm, Germany, 2021. 3, 5

[2] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic Segmentation. In *IEEE CVPR*, 2019. 2

[3] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 2

Table 4. Overall results of the baseline hybrid fusion and late fusion model. The best results are denoted in bold font.

Model	Conf. Thrs.	Re	Sp	FPR	FNR	PWC	Pr	F_1
Baseline (Hybrid Fusion)	0.4	**0.757**	**0.994**	**0.006**	**0.243**	**1.147**	**0.775**	**0.745**
Late Fusion	0.3	0.709	0.993	0.007	0.291	1.267	0.755	0.693

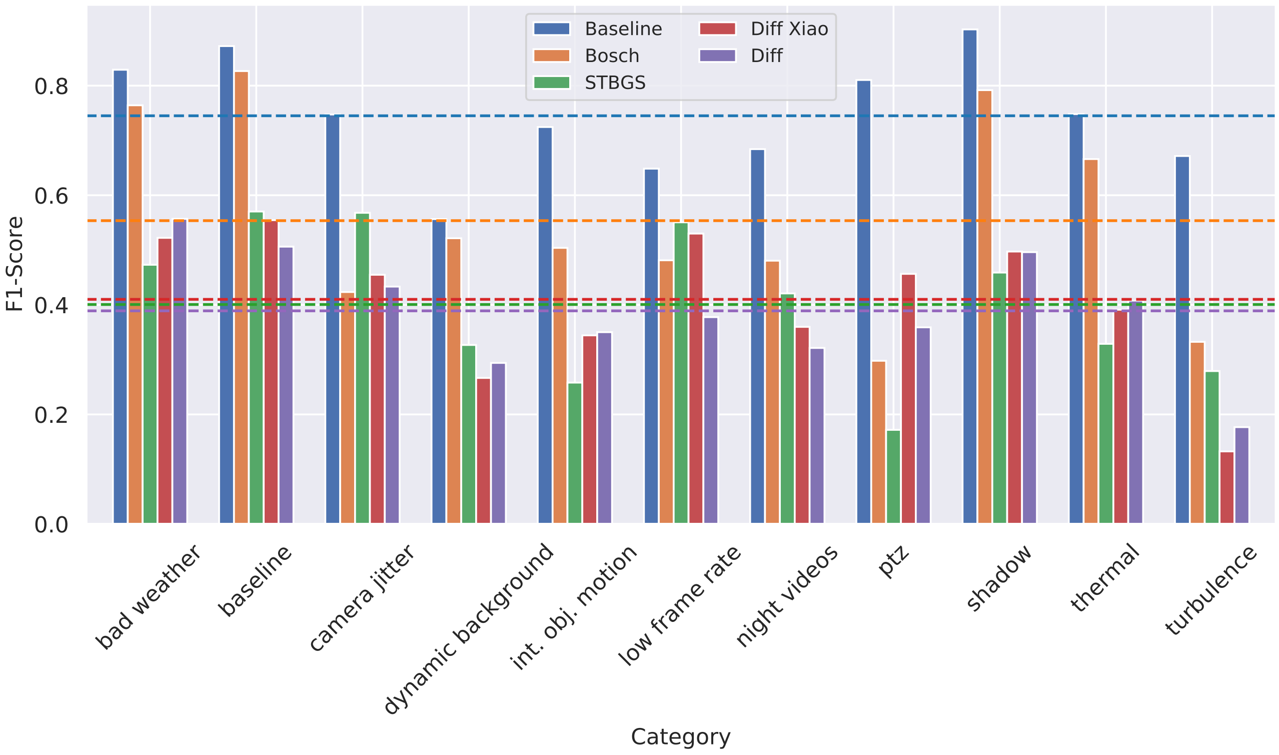

Figure 3. Comparison of the category-wise F_1-scores of each method. Dotted lines represent the overall average F_1-score of each method.

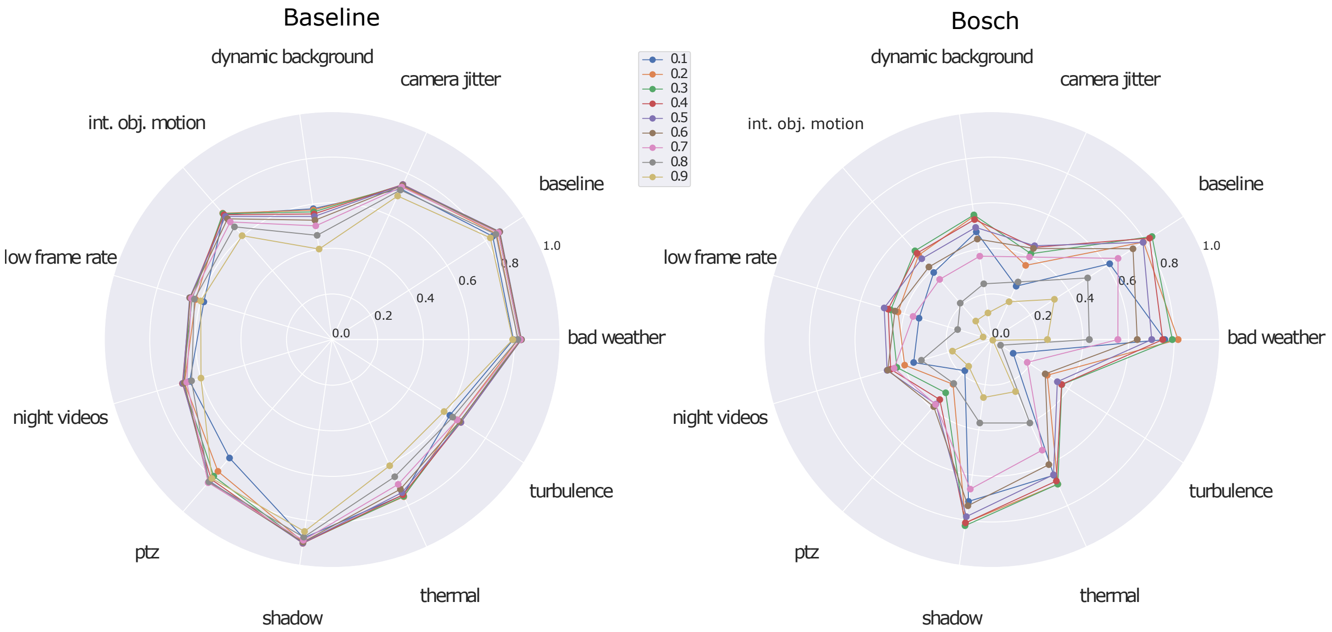

Figure 4. Category-wise F_1-score of the baseline model and the approach by Bosch [1] for different confidence thresholds.

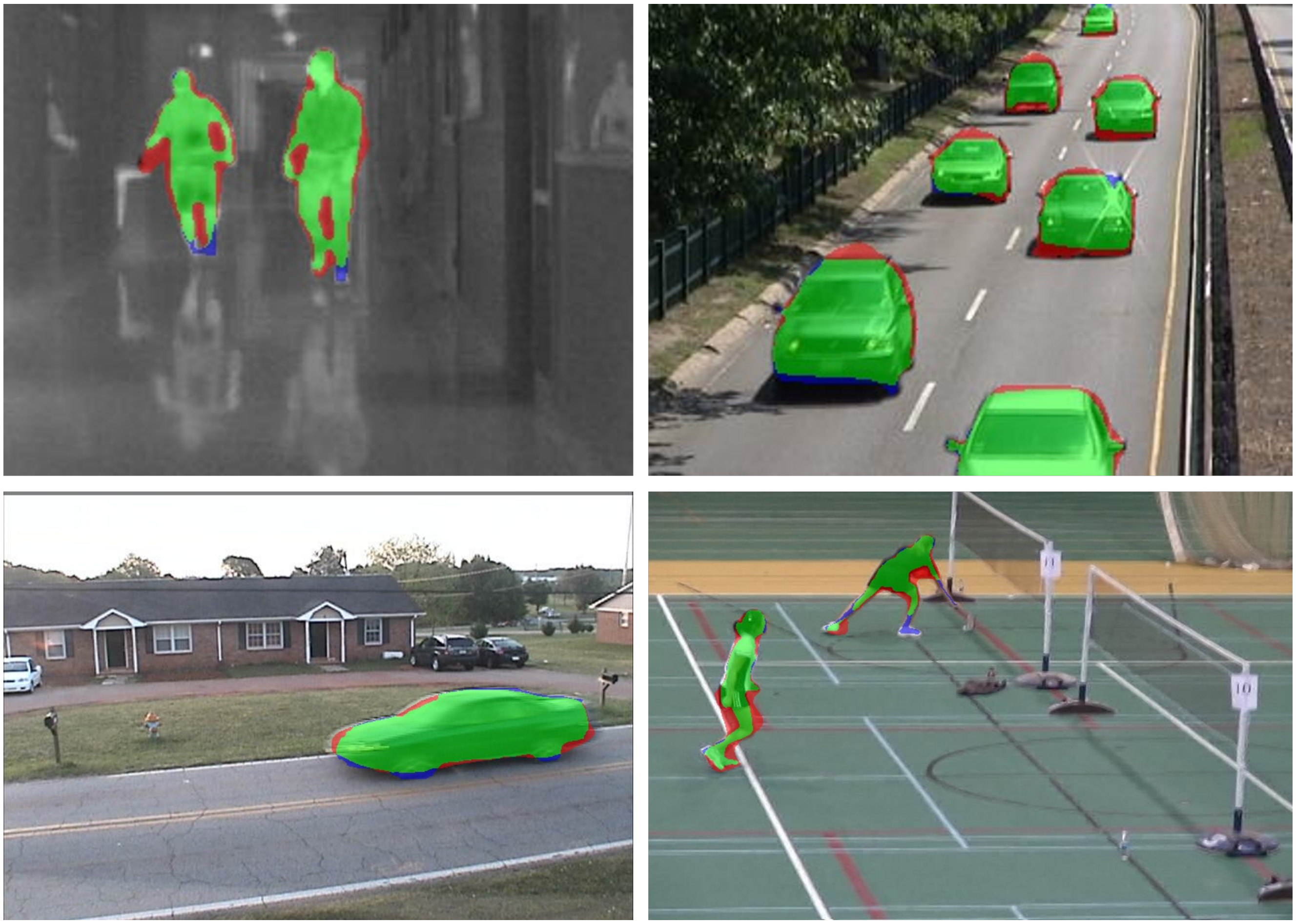

Figure 5. Examples of motion segmentation created by our approach. Images are taken from the scenes *Thermal - corridor*, *Baseline - highway*, *PTZ - continuousPan* and *Camera Jitter - badminton*. Green pixels indicate true positives, red pixels indicate false positives, and blue pixels indicate false negatives.

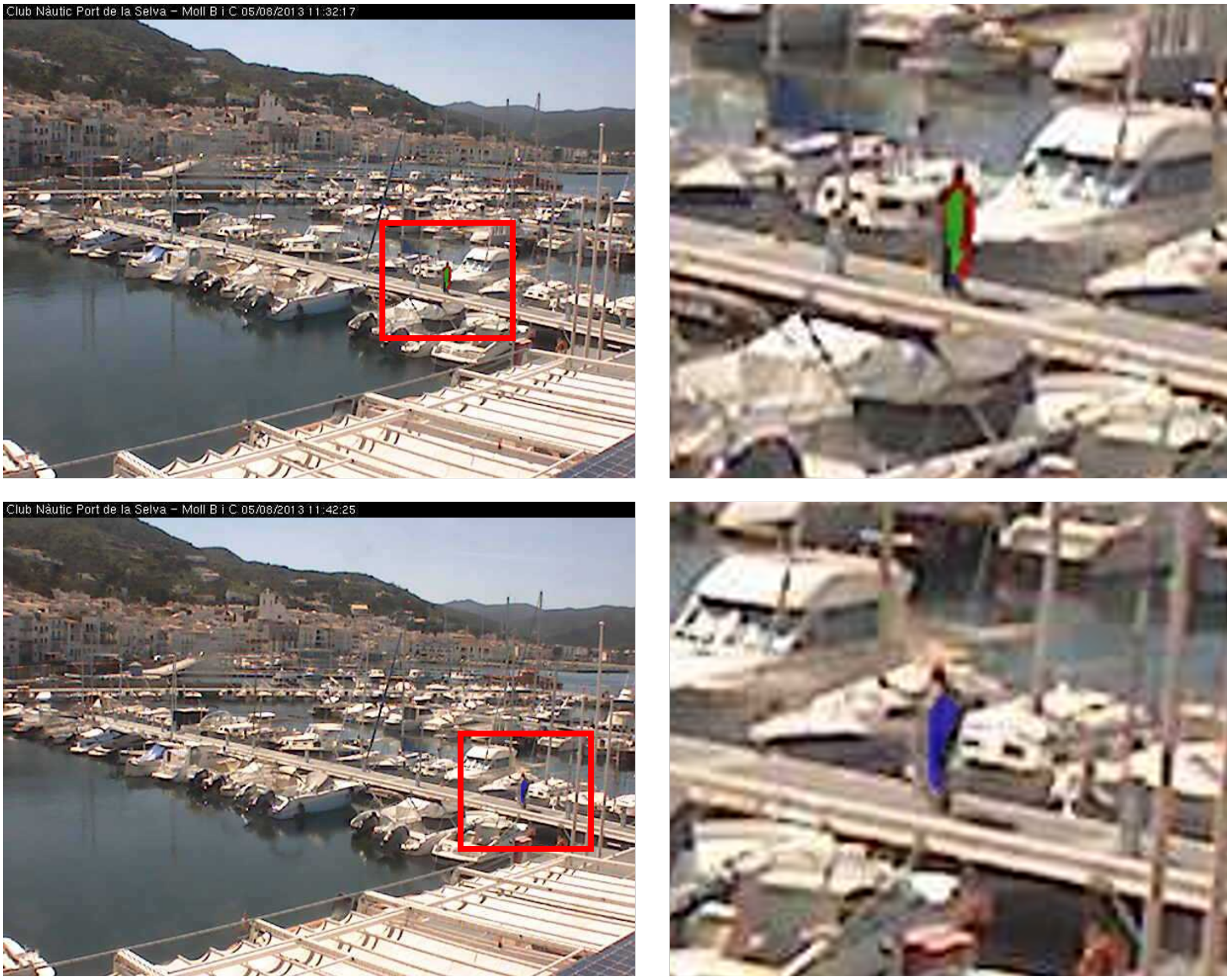
Club Nàutic Port de la Selva – Moll B i C 05/08/2013 11:32:17
Club Nàutic Port de la Selva – Moll B i C 05/08/2013 11:42:25

Figure 6. Examples of the scene *port_0_17_fps*. The walking person in the lower row is missed by our approach. Green pixels indicate true positives, red pixels indicate false positives, and blue pixels indicate false negatives.