TEC2014-53176-R HAVideo (2015-2017)

High Availability Video Analysis for People Behaviour Understanding

D2.3v1

Online quality analysis of people behaviour understanding

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by

AUTHORS LIST

Juan Carlos San Miguel Avedillo	juancarlos.sanmiguel @uam.es
Diego Ortego Hernández	diego.ortego@uam.es

HISTORY

Version	Date	Editor	Description
0.1	23/06/2016	Diego Ortego	First working draft
0.2	28/06/2016	Juan Carlos San Miguel	Final working draft
1.0	30/06/2016	José M. Martínez	Editorial checking
1.1	01/06/2016	Juan Carlos San Miguel	First working draft v2
1.2	14/06/2016	Diego Ortego	Contributions
1.3	18/06/2016	Juan Carlos San Miguel	Contributions
2.0	24/06/2017	José M. Martínez	Editorial checking

CONTENTS:

1. Introduction

This document summarizes the work during the first and half year(s) of the project for the task T2.3 "Online quality analysis of people behavior understanding." (WP2 Video analysis tools, models and performance indicators) whose goal is to propose new methods or strategies for online evaluation of analysis tools results or task relevance for each camera in multi-camera settings in order to provide additional or quality control information to the self-configurable approaches.

This task T2.3 depends upon developments within WP1 (Video Analysis Framework) and WP2 (T2.1 Analysis tools for human behavior understanding). The results of this task T2.3 will be employed for WP3 Self-configurable approaches for long-term analysis and WP4 Evaluation framework, demonstrators and dissemination.

1.1. Document structure

The document is structured in the following chapters:

- Chapter 2: description of the contributions
- Chapter 3: Conclusions and future work

2. Contributions

This chapter compiles the contributions developed in the scope of the task T2.3.

2.1. Multiple trackers quality: study of metrics

The main objective of this work consists on studying metrics to compare algorithms for tracking objects in video sequences using a set of search or tracking algorithms. The results of this work is a degree thesis in the University Autonoma of Madrid [1]. A summary is provided here.

First, an initial study of existing works is carried out with emphasis on the basic aspects of detection, search, tracking and the combination of various search algorithms. Then, a set of scientific papers that are focused on the same topic have been studied. These papers use different techniques to achieve the desired objective. After this theoretical learning stage, the Matlab tool is used to test a set of selected algorithms and sequences, comparing the effectiveness of the algorithms by the error based on ground-truth also to compare them with a series of measures, or comparing an algorithm himself in different time instants to estimate their reliability.

These measures employ the distance between centers and the overlap between areas detected in the corresponding frame. These measures give us an ability to visualize which algorithms are optimal in a possible combination of these, and which are not. Finally, confidence maps are used to estimate results' stability in the possible selection of the optimal algorithms for their combination. These maps are extracted from each algorithm and represent areas of a size similar to the frames which show the probability that the object is located at each point of the map. Experimental results show the strengths and weaknesses of each algorithm on the set of selected sequences using the proposed set of reliability measures. Once the results obtained from the measures described above, we proceed to combine the best algorithms as proposed.

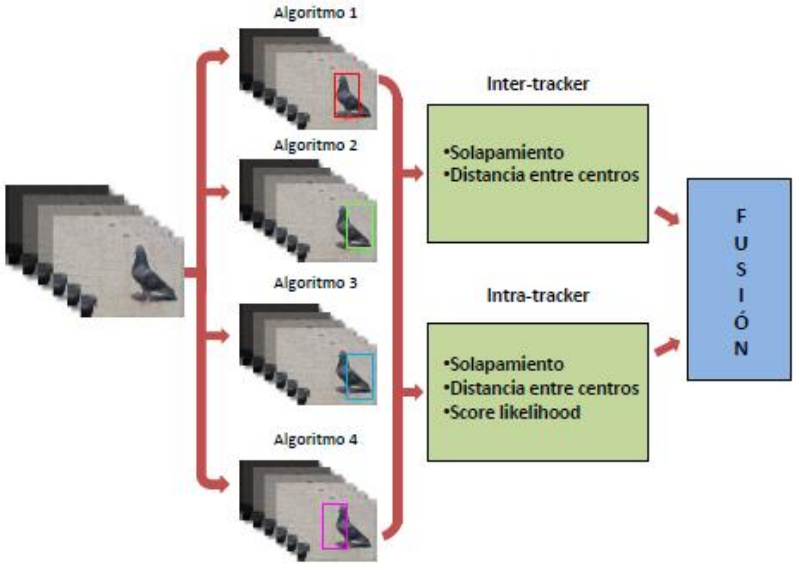
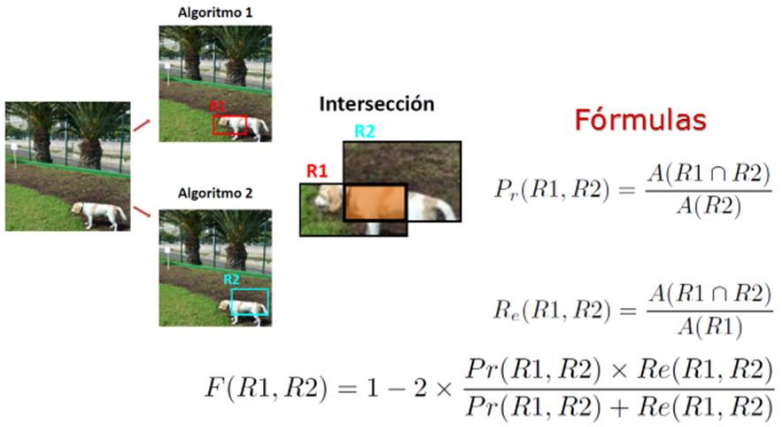
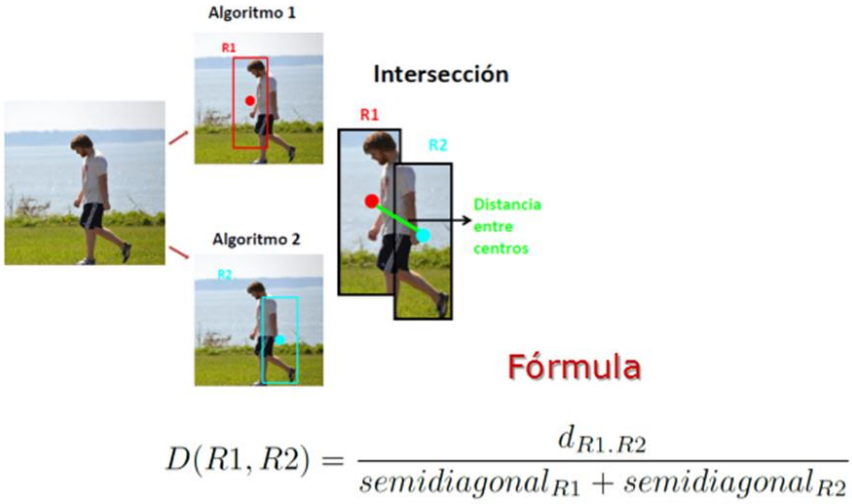


Figure 1. Block diagram of the proposed approach to combine four trackers.

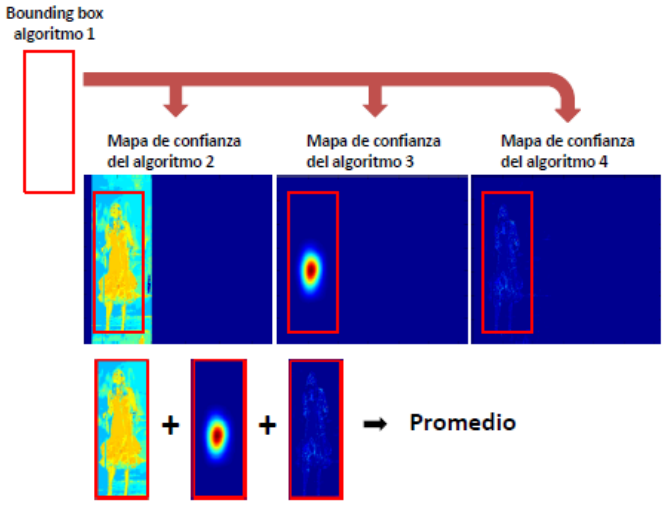
Examples of the studied metrics are presented in the following figure



(a)



(b)



(c)

Figure 2. Metrics studies for the similarity between tracker results: (a) spatial overlap,(b) centroid distance and (c) score likelihood.

To show the applicability, a simple fusion algorithm has been implemented taking those trackers with high similarity (e.g. values over 0.7). Two fusion versions are considered: equal and similarity-weighted combinations.

For the experiments, a set of 10 sequences has been extracted from the Tracker Benchmark 1.0 (https://sites.google.com/site/trackerbenchmark/benchmarks/v10). As trackers, this study used the following:

- **MS** (PSU R.Collins, CSE. Mean-shift Tracking. 2006)
- **CBWH** (David Zhang y Chengke Wu Jifeng Ning, Lei Zhang. Robust mean shift tracking with corrected background-weighted histogram. IET CVI 2010)
- **PFC** (Fabian Kaelin, An Adaptive Color-Based Particle Filter. ECCV 2010)
- **ACA** (Michael Felsberg y Joost van de Weijer Martin Danelljan, Fahad Shahbaz Khan. Adaptive color attributes for real-time visual tracking. CVPR 2014)

Sample results are show in the following figures/tables.

Comparación algoritmos	FaceOcc1	Basketball	Bolt	I1_basic	I2_basic	I3_cars	Rolling	Singer1	Skiing	Walking
MS-CBWH	.01±.002	.20±.005	.15±.004	.99±.001	.09±.007	.47±.109	.50±.041	.44±.062	.97±.011	.28±.039
MS-PFC	.34±.003	.34±.015	.45±.012	.85±.064	.41±.006	.60±.083	.36±.024	.67±.069	.92±.047	.39±.018
MS-ACA	.11±.001	.12±.003	.90±.043	.95±.035	.11±.001	.20±.024	.82±.075	.31±.019	.90±.058	.26±.012
CBWH-PFC	.35±.006	.23±.009	.53±.029	.00±.000	.36±.009	.66±.061	.61±.038	.82±.079	.95±.019	.36±.030
CBWH-ACA	.12±.002	.18±.007	.92±.025	.88±.035	.12±.005	.38±.060	.94±.015	.65±.128	.90±.036	.42±.039
PFC-ACA	.19±.004	.40±.001	.86±.073	.97±.012	.33±.002	.47±.039	.88±.059	.32±.001	.89±.050	.64±.034

(a)

Comparación algoritmos	FaceOcc1	Basketball	Bolt	I1_basic	I2_basic	I3_cars	Rolling	Singer1	Skiing	Walking
MS-CBWH	.03±.008	.30±.009	.33±.016	.04±.004	.17±.026	.10±.033	.58±.079	.50±.059	.04±.038	.39±.052
MS-PFC	.90±.006	.79±.003	.69±.005	.20±.128	.79±.0060	.54±.150	.68±.014	.75±.088	.09±.062	.69±.009
MS-ACA	.22±.001	.21±.012	.12±.070	.01±.006	.21±.0050	.37±.044	.13±.069	.46±.040	.07±.040	.53±.034
CBWH-PFC	.86±.038	.78±.005	.83±.026	.00±.000	.76±.010	.27±.127	.65±.118	.23±.128	.08±.063	.58±.057
CBWH-ACA	.24±.006	.28±.013	.15±.104	.03±.025	.23±.020	.10±.031	.14±.108	.44±.163	.12±.079	.61±.060
PFC-ACA	.89±.009	.79±.004	.15±.088	.02±.018	.69±.002	.75±.021	.15±.094	.69±.003	.13±.084	.51±.209

(b)

Table 1. Average results for the inter-tracker and intra-tracker distance using the spatial overlap. Higher values indicate that the trackers have similar results

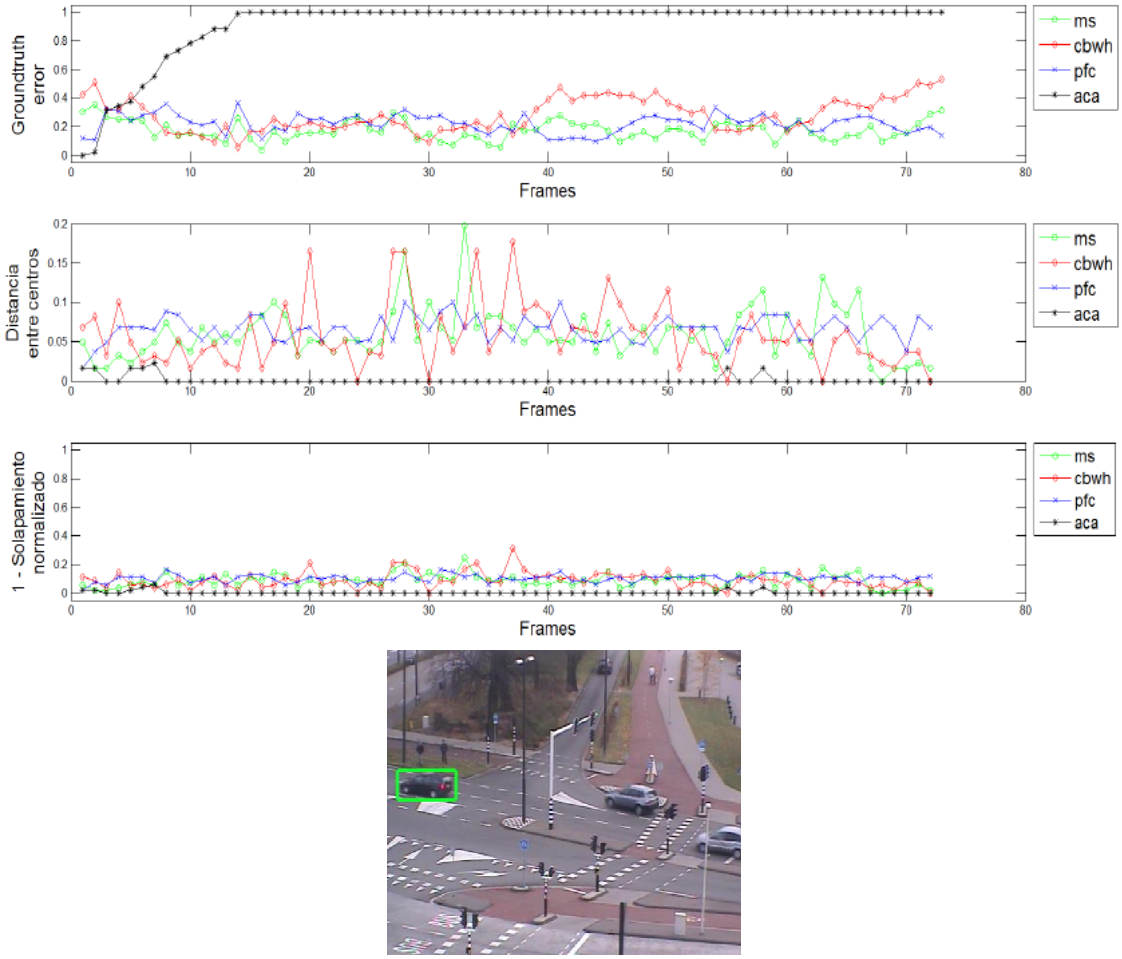


Figure 3. Similarity results for the sequence **I3_car_basic_2**. Top: ground-truth error, spatial overlap and normalized overlap score. Bottom: sample frame.

The final results of the fusion approaches are show in the following Table

Algoritmo	FaceOcc1	Basketball	Bolt	I1_basic	I2_basic	I3_cars	Rolling	Singer1	Skiing	Skiing
MS	.23±.004	.49±.013	.84±.076	.22±.001	.26±.014	.17±.005	.78±.048	.62±.086	.96±.020	.43±.030
CBWH	.20± .012	. 61±.035	.99±.032	.24±.006	.35±.020	.28±.013	.82±.042	.75±.105	.95±.013	.73±.095
PFC	.19± .008	. 50±.019	.28±.007	.66±.059	.35±.020	.22±.004	.77±.106	.54±.110	.91±.048	.16±.003
ACA	.08± .002	. 19±.031	.97±.017	.00±.000	.16±.005	.91±.048	.86±.081	.47 ±.024	.90±.060	.16±.004
Fusión Equitativa	.11± .001	. 35±.013	.91±.058	.16±.002	.21±.009	.13±.005	.79±.053	.49±.069	.90±.064	.23±.013
Fusión ponderada	.09±.001	. 47±.021	.99±.004	.08±.001	.27±.0152	.87±.083	.98±.003	.41±.057	.99±.004	.56±.175

Table 2. Average accuracy results for proposed fusion approaches based on the inter-tracker and intra-tracker distance computed previously.

2.2. Multiple trackers quality: spatio-temporal correlation

Visual tracking is widely used in applications such as video surveillance, human-computer interaction, activity recognition and video indexing. A tracker faces several challenges such as occlusions, clutter, changes in target scale or appearance and variations in scene illumination. Because no individual tracker can still provide accurate results for all challenges[2], fusing complementary trackers whose expected failures are uncorrelated can increase robustness.

Decision-level fusion combines the output of multiple trackers in cascade or in parallel. A cascade for fusión defines an execution order where each tracker output is used by the next tracker. Examples include the combination of two trackers (region and shape) and two detectors (head and motion for people tracking) [3]; the sequential execution of the template-based Mean Shift (MS) and appearancebased trackers [4]; and the integration of three PFs and one Kalman filter (KF) [5]. Moreover, trackers can be integrated within the framework of another tracker. For instance, a head tracker uses MS to improve the PF tracker predictions [6]. In parallel tracker fusion, two trackers may be combined using target motion [7] or probability density functions [8]. Moreover, tracker performance within a parallel framework can be measured as disagreement with other trackers [9] or can be used to select the best tracker [10]. Other approaches may use tracker correlation to improve the overall tracking performance by correcting PFs and KFs[10][11]. These approaches determine the accuracy as the spatial uncertainty of hypotheses whose value may vary across trackers, thus making tracker fusion difficult.

We propose a decision-level approach to group trackers into clusters based on the spatiotemporal pair-wise correlation of their short-term trajectories. Then, we evaluate performance based on reverse-time analysis with an adaptive reference frame and define the cluster with trackers that appear to be successfully following the target as the on-target cluster. The proposed approach uses standard tracker outputs and can therefore combine various types of trackers.
The proposed approach is inspired by the test and select framework [12] for ensemble combination where accurate classifiers are fused assuming that their errors are diverse. Considering trackers as classifiers, we extend this framework to video tracking by introducing spatio-temporal correlation and adaptive online performance evaluation (Figure 4).

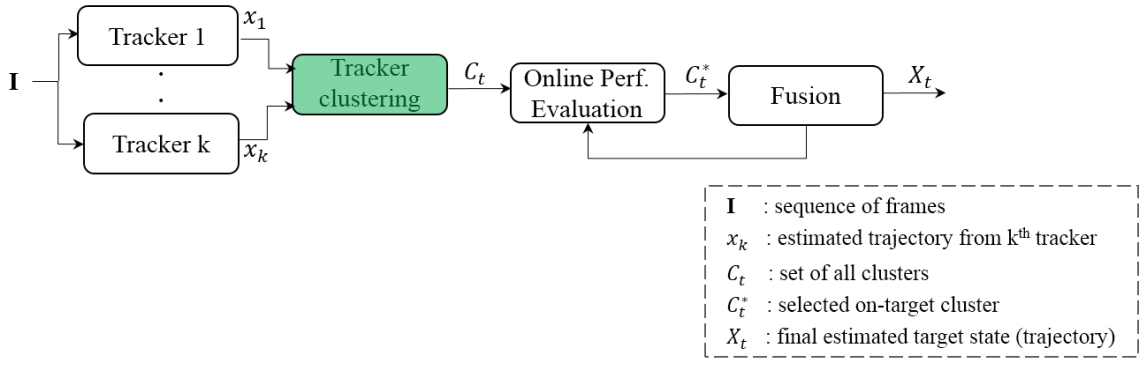


Figure 4. Block diagram of the proposed approach to fuse the output of K trackers.

Let $\boldsymbol{I} = \{I_t\}_{t=1}^{T}$ be an image sequence of *T* frames and $\boldsymbol{F} = \{F^k\}_{k=1}^{K}$ be a set of *K* trackers. Let the target state x_t^k be a bounding box, with the position of the target and the width and height of the target. Using a set of observations $\boldsymbol{Z} = \{Z_t\}_{t=1}^{T}$ for each F^k and the target appearance model at frame I_t, ϕ_t^k, each tracker F^k estimates the target state.

Let *ON* and *OFF* be the labels when the tracker is successfully following the target (*on-target*) or not (*off-target*), respectively. The goal is to identify *on-target* trackers given the tracker output x_t^k and assign the binary label l_t^k to each tracker. We determine l_t^k by recognising groups of trackers (clusters) following the same region and identifying the cluster C_t^* with the *on-target* trackers. The framework starts with with a single cluster as all trackers are initialized at the same position. The visual challenges cause the trackers to fail and therefore divide them into different clusters over time, where only one (C_t^*) or none of them correctly tracks the target.
For each frame I_t, we *generate hypothesis for partitioning* the *K* trackers into clusters based on their pair-wise spatio-temporal relationships measured as similarity of spatial location and direction of movement over a time window Δt_1. After *validating the best cluster partition* P_t^*, the on-target cluster C_t^* is determined by online performance evaluation of the trackers using reverse tracking [13]. In comparison to other methods of online track quality evaluation [5], reverse tracking provides a generic method for evaluation across different trackers. Reverse analysis is carried out in the past using a sliding time window Δt_2. Fig. \ref{fig:TimeWindows} presents both time windows Δt_1 and $\Delta t_2$where information from future and past are used, respectively. Finally, we propagate the selected C_t^* to the next frame until it splits or merges into other clusters indicating that the *on-target* trackers may have failed due to visual challenges.

The following figures shows examples of the clustering for the trackers where it can be observed that four groups of trackers exist.

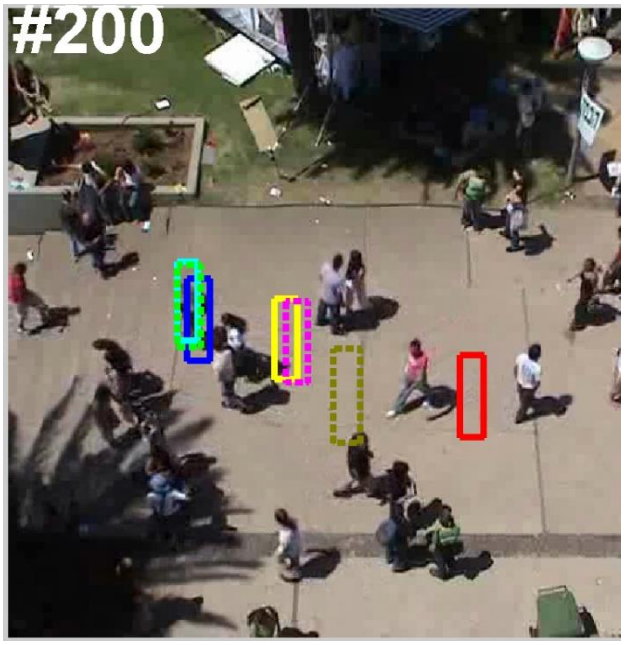


Figure 5 Visual example for clusters of trackers

The following figure shows an example of the results achieved to improve reverse tracking evaluation of video tracking.

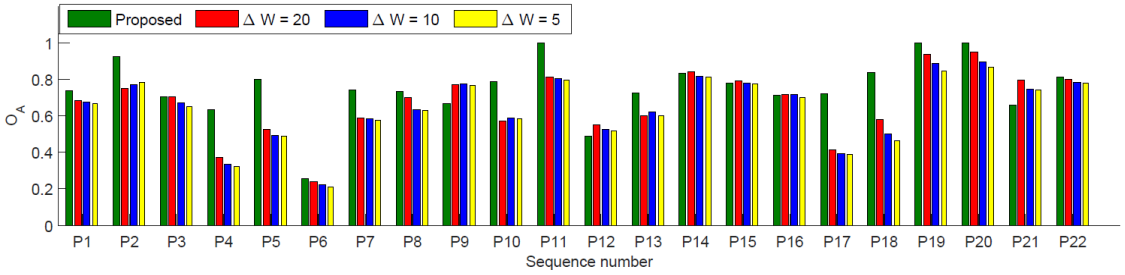


Figure 6 Comparison of $\boldsymbol{I_{ref}}$ selected by the proposed approach and the original approach based on fixed time windows $\boldsymbol{\Delta W}$ = 5; 10; 20.

The results of this work have been accepted for publication in the prestigious IEEE Transactions on Circuits and Systems for Video Technology [14].

2.3. Stand-alone quality estimation of Background Subtraction algorithms

Background Subtraction (BS) is a key stage in multiple computer vision applications, where existing algorithms are commonly evaluated making use of ground-truth data. Reference-free or stand-alone evaluations that estimate segmented objects quality are an alternative methodology that overcomes the limitations inherent to ground-truth based approaches. In this work, we explore existing measures proposed from the literature to determine good object properties for segmentation quality estimation.

First, we have analysed the literature and proposed a new taxonomy to organize stand-alone methods. Second, we discuss the available measures in the object segmentation literature in order to identify the properties of good object segmentation masks. Finally, we evaluate a set of 21 measures using four state-of-the-art BS algorithms in CDNET2014 dataset [15].

We have extended the classical empirical-analytical taxonomy for performance evaluation [16][17] for stand-alone evaluations including new categories for stand-alone evaluations: assisted, specific and generic.

Subjective-vs-Objective evaluation: existing approaches are frequently classified into subjective and objective, denoting whether human perception is or not used to quantify performance [18]. Furthermore, the objective evaluation is divided into analytical and empirical, where the former evaluates an algorithm considering its theoretical description and the latter uses algorithm results. Although there are some analytical methods, the evaluation in BS has been mainly studied empirically, either by using ground-truth (discrepancy evaluation) or not (stand-alone evaluation).

Discrepancy evaluation. Ground-truth based evaluations assess algorithm performance through comparisons between expected and segmented object masks. These evaluations include traditional measures such as Precision, Recall, F-score [15] or segmentation accuracy [19].

Stand-alone assisted evaluation. Assisted measures improve segmented object masks by employing the results of external algorithms, such as visual tracking algorithms. Specifically, tracking feeds segmentation providing effective sequential motion and structure constraints, while segmentation improves tracking thanks to accurate local appearance and information [20][21].

Stand-alone specific evaluation. These measures detect challenging situations with an expected decrease in performance, such as illumination changes [22][23], shadows [24][25] or dynamic background [26][27].

Reference	Measure	Symbol	Category	Fixed range
[32]	Shape Regularity	SH	Shape	NO
	Spatial Uniformity	SU	Uniformity	NO
	Motion Uniformity	MU	Uniformity	NO
	Local Contrast to Neighbors	LN	Contrast	YES
[30]	Spatial Color Contrast	SC	Contrast	YES
	Motion Difference	MD	Contrast	YES
[29]	Boundary Turning Points	BT	Shape	YES
	Boundary Curvature	BC	Shape	NO
	Local Contrast	LC	Contrast	YES
	Separability	SE	Density	NO
	Edge fitness	E1	Fitness	YES
[33]	Spatial Clique Potential	SP	Contrast	NO
	Temporal Clique Potential	TC	Contrast	NO
	Edge fitness	E2	Fitness	YES
[34]	Boundary Complexity	BX	Shape	NO
	Color Contrast	CC	Contrast	YES
	Superpixel Straddling	SS	Fitness	YES
	Motion Contrast	MC	Contrast	YES
	Edge Density	ED	Density	YES
	Color Homogeneity	CH	Uniformity	YES
	Motion Homogeneity	MH	Uniformity	NO

Table 3. Selected quality measures.

Stand-alone generic evaluation. Generic measures estimate quality by inspecting certain properties of the object masks. These stand-alone generic measures have been weakly explored for the evaluation of BS. However, closely related areas, such as image segmentation [28], image co-segmentation [29], video object segmentation [30] or object recognition [31] have studied stand-alone generic measures. Furthermore, existing measures can be classified into five groups, namely contrast, uniformity, shape, fitness and density. We have studied and evaluated representative measures from each sub-category (see Table 3). Contrast measures compute spatial or temporal contrast between internal and external regions of object masks, establishing that higher contrast indicates higher performance; uniformity measures analyse the internal homogeneity of the object mask region in terms of colour or motion, being such homogeneity considered as a high quality indicator; shape measures estimate an object quality through its shape complexity, as complex shapes are associated with poor segmentation; fitness measures identify high quality with the adjustment of the object mask to image regions and contours; and density measures associate external and internal density properties of an object to, respectively, low and high quality.

To perform experiments, we use CDNET2014 dataset [1] which provides an extensive set of common BS challenges with ground-truth data. We select eight of the eleven categories (PTZ, Thermal and Turbulence are excluded) as our current target are colour images from static cameras. These eight categories include 40 video sequences (113848 frames in total). To extract blobs from the video sequences, we employ four relevant BS algorithms according to their CDNET2014 results (ordered in increasing ground-truth performance): GMM [35], MBS [36], FTSG [37] and SuBSENSE [27]. We use the results provided by the authors in CDNET2014 and apply the quality measures on every 30th frame, obtaining approximately 87000 blobs.

To assess the stand-alone generic measures, we define new blob-level ground-truth based metrics. This is necessary as objects may be fragmented into several blobs and no prior knowledge is assumed for the correspondence between blobs and objects. We define such ground-truth measures for true and false positive objects, computing Precision and Recall at blob-level and combining them to obtain a unique blob-level ground-truth measure F.
To analyse stand-alone generic measures from Table 3, we have identified useful complementary ones, studied their potential to distinguish among different performance levels and analysed their algorithm ranking capabilities.
In order to identify useful and complementary measures, we have applied agglomerative hierarchical clustering guided by the cross-correlation among measures obtaining that fitness measures perform better than the rest.
From a practical viewpoint, a quality measure must be able to replicate the ranking of algorithms given by the ground-truth evaluation. Additional experiments have been done to explore these capabilities. In particular, we have analysed the four algorithms to see if stand-alone measures replicate ground-truth performance, leading to a better behaviour of fitness measures.
In conclusion, in this work we provide a comprehensive study on stand-alone measures to estimate the quality of segmented objects in Background Subtraction. We have selected from related literature a diverse set of measures that are thoroughly analysed in terms of correlation with ground-truth and algorithm ranking capabilities. Experiments in a large Background Subtraction dataset shows superior potential of fitness measures to approximate ground-truth performance.

2.4. Background initialization in video sequences

Nowadays, the automatic analysis of video-surveillance sequences is a relevant research field due to the need of increasing security in public and private facilities. Many applications start their operation by detecting moving and stationary objects in the scene, task that is commonly performed by Background Subtraction (BS) algorithms [38]. The first stage of BS is Background Initialization (BI), that has been weakly studied as it is commonly assumed as an easy task. BI consists in estimating a background image given a set of frames where objects may be occluding the background, thus invalidating the assumption of BI being an easy task. In this Graduate thesis [39], we implement two algorithms [40][41] from the state-of-the-art and we develop a new dataset considering different challenges for evaluation purposes.

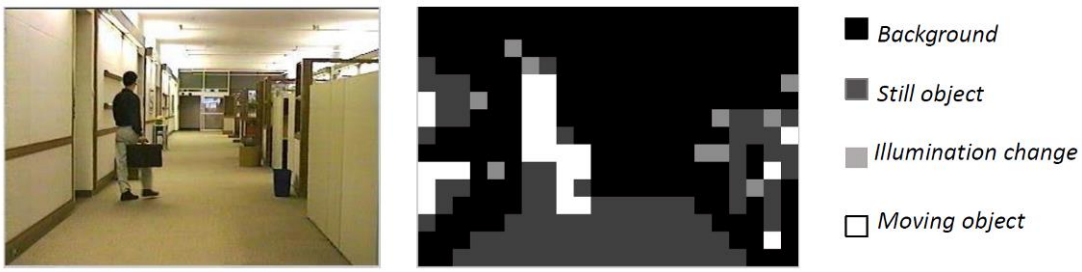


Figure 7. Example of Block classification output stage. Left: Image under analysis. Right: Image of each block labels: Background, Still object, Illumination change and Moving object.

The first algorithm, proposed in [40], performs a spatio-temporal block-wise and online analysis of the scene. To analyse each incoming frame, the algorithm is divided in four stages: Block division, Motion estimation, Block classification and Background updating. First, an image is divided into non-overlapping square blocks in the Block division stage. Then, the Motion estimation stage performs a block-wise motion estimation between temporal adjacent frames to distinguish among moving and static blocks. Subsequently, Pearson's correlation coefficient is used in the Block classification stage to classify each block into four categories depending on the

result of the motion estimation stage (see Figure 7). Static blocks could be labelled as background or still object, whereas moving blocks are labelled as moving object or illumination change. Finally, the updating stage performs a different background updating strategy depending on the block label obtained, taking into account spatial and temporal constraints. Figure 8 shows in the first row an example of images from the video to initialize, while in the second row it presents the reconstructed background for each frame of the first row. Note that initially the background is empty and it is updated frame by frame by incorporating static or background blocks information.

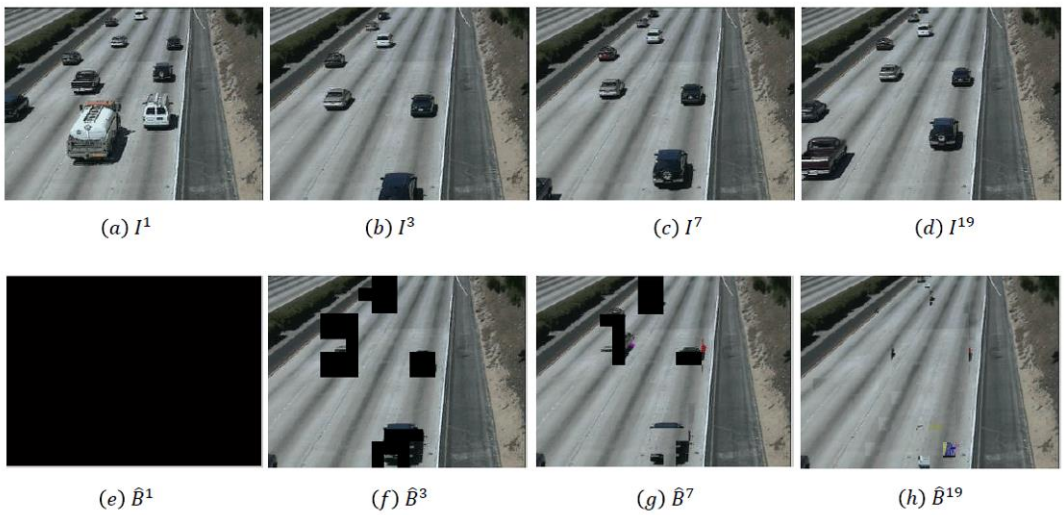


Figure 8. Example of initialization procedure from [40]. First row: Images under analysis along time. Second row: Generated background along time.

The second algorithm, proposed in [41], performs a pixel-wise spatio-temporal online analysis of the scene to estimate a background image. This algorithm models the statistical variation of pixel luminance in order to determine those variations induced by background pixels. The operation is divided in three stages: Salient pixels filtering, Maximum Likelihood Set (MLS) and Background Updating. First, the Salient pixels filtering stage, discards pixels with high inter-frame variations as they are considered not useful to characterize the statistical background variations.

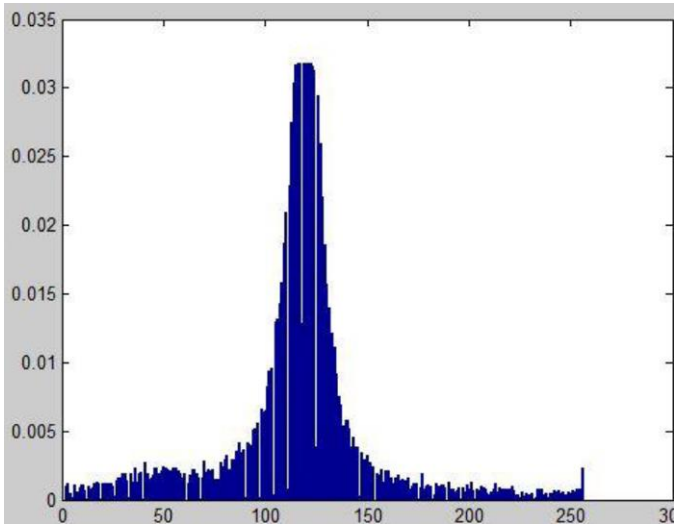


Figure 9. Example of estimated distribution of pixel luminance variations for level 120.

Then, in the Maximum Likelihood set stage, the distribution of the luminance levels variations for non-salient pixels is estimated (see Figure 9), leading to the MLS of expected (non-significant) and non-expected (significant) variations of each luminance level ([0…255]). Finally, the background updating procedure is carried out depending of the nature of each pixel, significant or non-significant, by integrating spatial and temporal constraints. An example of generated Background is presented in Figure 10.

Figure 10. Example of reconstructed background images from [41].

Additionally, a dataset containing four challenges or categories (Baseline, Clutter, Low framerate and Static objects) has been developed. Each category contains 10 sequences, leading to 40 sequences for the whole dataset.

2.5. Background initialization in video sequences with stationary objects

Several state-of-the-art BI approaches easily capture the background by assuming the availability of a set of frames without foreground objects (training frames) [38]. This assumption may not be correct in many video-surveillance scenarios (e.g. shopping malls, airports or train stations) where many foreground objects may exist due to crowds and stationary objects, making very challenging the capture of the background. In general, BI faces two problems related with spatio-temporal scene variations: Background visibility and photometric factors. To overcome these limitations, we propose a block-level BI approach based on a temporal-spatial strategy that reconstructs an object-free background in presence of moving and stationary objects.

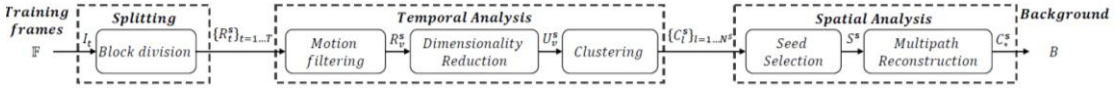


Figure 11. Overview of the proposed approach.

The proposed approach [42] performs a temporal-spatial analysis at block level (see Figure 11 and Figure 12) over a set of T training frames I_t, $\mathbb{F} = \{I_1 \dots I_T\}$, to extract the reconstructed background image B free of moving and stationary objects.

First, the *Splitting* module divides each I_t It into non-overlapping blocks $R_t^{\mathbf{s}}$ of size $W \times W$, where $\mathbf{s}$ is the bi-dimensional index for the spatial location of each block.

Second, the *Temporal Analysis* module creates a number of background candidates $C_l^{\mathbf{s}}$ for each spatial location $\mathbf{s}$, where $l \in \{1 \dots N^{\mathbf{s}}\}$ and $N^{\mathbf{s}} \leq T$ is the number of candidates. To that end, the Motion filtering stage discards the $R_t^{\mathbf{s}}$ blocks where moving objects exist using frame difference.

Then, the Dimensionality Reduction stage applies PCA to reduce the amount of the data to analyse as the useful information to generate background candidates is represented by the block variance. Subsequently, the Clustering stage obtains a set of background candidates $C_l^{\mathbf{s}}$, via threshold-free agglomerative hierarchical clustering, to be the background B for each location $\mathbf{s}$. This clustering represents one of the main contributions of this work, as the background candidates are clustered without using any threshold. This is done by grouping the PCA-reduced data into clusters $K_l^{\mathbf{s}}$ which are structured as partitions $\mathbb{P}_{N^{\mathbf{s}}}^{\mathbf{s}} = \{K_1^{\mathbf{s}} \dots K_{N^{\mathbf{s}}}^{\mathbf{s}}\}$ where $N^{\mathbf{s}}$ is the total number of clusters. The candidates $C_l^{\mathbf{s}}$ represent each cluster $K_l^{\mathbf{s}}$ where the best candidate $C_*^{\mathbf{s}}$ is selected in the Spatial Analysis. As the optimum $N^{\mathbf{s}}$ is not known for each $\mathbf{s}$, hypotheses for the partitions are created for different values of $N^{\mathbf{s}}$. The optimal partition, i.e. the one containing the background candidates $C_l^{\mathbf{s}}$, is found by validation indexes that maximize inter-cluster differences and intra-cluster similarities. Note that we compute each background candidate $C_l^{\mathbf{s}}$ as the average of members in cluster $K_l^{\mathbf{s}}$ similarly to the widely used K-means clustering.

Finally, the Spatial Analysis module reconstructs the background of each spatial location $\mathbf{s}$ by the Seed Selection stage that partially initializes B starting from a set of seeds $S^{\mathbf{s}}$(selected background candidates) and by the Multipath Reconstruction stage that iteratively fills each spatial location with the optimal candidate $C_*^{\mathbf{s}}$ via inter and intra-block smoothness constraints.

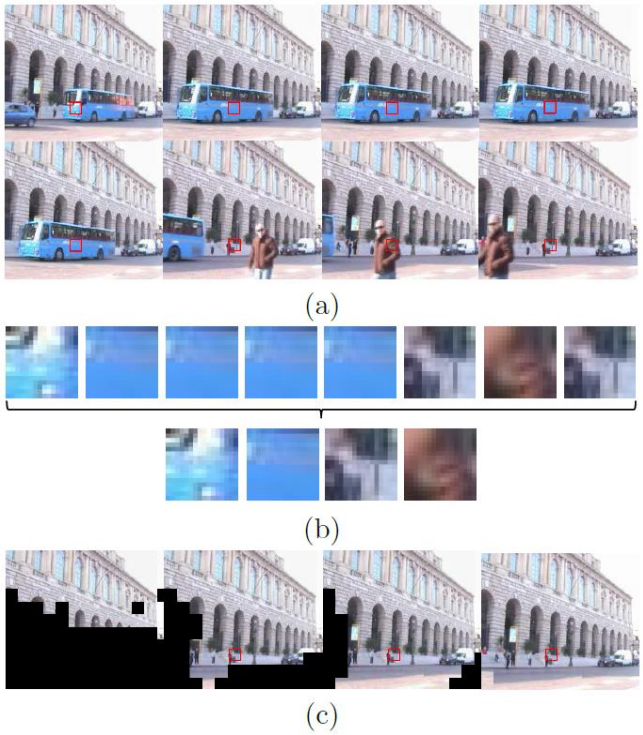


Figure 12. Example of the proposed approach for background initialization of the spatial location s (marked in red). a) Several frames from a video sequence. b) Temporal Analysis example. First row: blocks $\boldsymbol{R}_t^{\mathbf{s}}$ extracted from the frames in a). Second row: background candidates $\boldsymbol{C}_l^{\mathbf{s}}$ obtained by clustering. c) Spatial Analysis example. From left to right: selected seeds $\boldsymbol{S}^{\mathbf{s}}$ to partially approximate the background, two iterations of the multipath reconstruction and the final reconstructed background. In these images, the red rectangle corresponds to the selected candidate $\boldsymbol{C}_*^{\mathbf{s}}$.

The Seed Selection stage performs a unified analysis of stationarity and motion activity along training frames. We compute a seed selection map (see Figure 13) to detect locations $\mathbf{s}$ with low motion or without stationary objects over time as suitable locations to initialize with seeds. Therefore, the initial background approximation is obtained only in locations with minimum score in the seed selection map. Note that the selected seeds $S^{\mathbf{s}}$ conforms the initial background $\tilde{B}$.

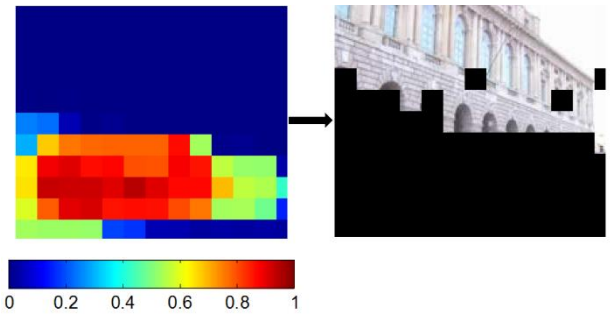


Figure 13. Seed selection example. From left to right: seed selection map, where minimum score represents low stationarity or motion activity and image with selected seeds $\boldsymbol{S}^{\mathbf{s}}$.

The Multipath Reconstruction represents another main contribution of this work as it iteratively estimates the background image B taking into account different hypotheses or paths in such reconstruction. In each iteration, the 4-connected neighbour $\mathbb{V}_4^{\mathbf{s}}$ of an already initialized location $\mathbf{s}$ is filled exploring different paths (see Figure 14) and exploiting intra-block heterogeneity and inter-block colour discontinuity and inter-block dissimilarity to initialise the neighbourhood with

$C_l^{\mathbf{s}}$. These three measures are employed to maximize the smoothness in the background as lower intra-block heterogeneity, inter-block colour discontinuity and inter-block dissimilarity are preferred when selecting background candidates.

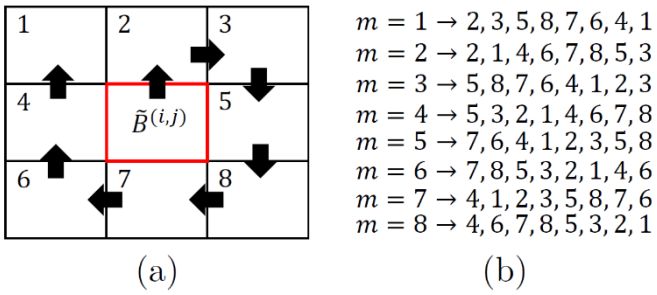


Figure 14. Multipath reconstruction scheme for each iteration of $\widetilde{\boldsymbol{B}}^{\mathbf{s}} = \widetilde{\boldsymbol{B}}^{(\boldsymbol{i},\boldsymbol{j})}$. (a) First path ($\boldsymbol{m} = \mathbf{1}$) to reconstruct $\mathbb{V}_{\mathbf{4}}^{\mathbf{s}}$. Black arrows describe the path direction. (b) Locations explored for all paths, i.e. $\boldsymbol{m} = \mathbf{1} \ldots \mathbf{8}$ paths.

The Multipath Reconstruction represents another main contribution of this work as it iteratively estimates the background image B taking into account different hypotheses or paths in such reconstruction. In each iteration, the 4-connected neighbour $\mathbb{V}_{\mathbf{4}}^{\mathbf{s}}$ of an already initialized location $\mathbf{s}$ is filled exploring different paths (see Figure 14) and exploiting intra-block heterogeneity and inter-block colour discontinuity and inter-block dissimilarity to initialise the neighbourhood with $C_l^{\mathbf{s}}$. These three measures are employed to maximize the smoothness in the background as lower intra-block heterogeneity, inter-block colour discontinuity and inter-block dissimilarity are preferred when selecting background candidates.
We evaluate our approach, RMR, against 13 state-of-the-art algorithms [42] including recent and top background subtraction algorithms in a proposed dataset of 29 short video sequences. We outperform all previous approaches due to our multipath reconstruction scheme. Figure 15 shows such evaluation using the measure Average Error pixels (AE).

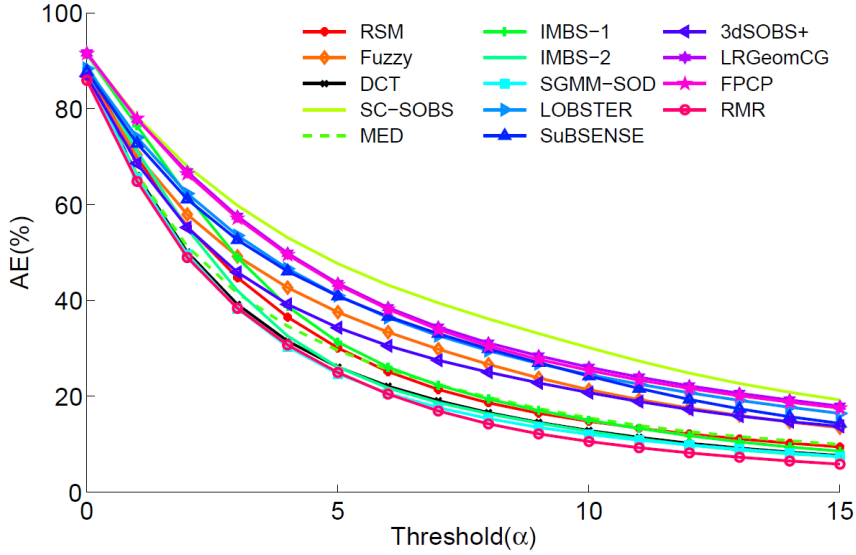


Figure 15. Comparison of the proposed approach RMR against related work. Lower AE means lower error.

2.6. Performance improvement of background subtraction algorithms based on quality

An algorithm for the improvement of Background Subtraction algorithms has been developed following the findings of the study presented in Subsection **¡Error! No se encuentra el origen de la referencia.**. In particular, the foreground segmentation improvement task has been tackled as a post-processing procedure using foreground quality. The use of generic foreground mask quality has the remarkable property of being independent of specific phenomena (e.g. illumination or shadows) or algorithm, thus being suitable to be applied in any condition.
Post-processing techniques in the literature are either model-dependent or model-independent. The former implements techniques that make use specific background model properties [43], whereas the latter only uses image and foreground properties [34], thus being independent of a particular algorithm. Therefore, we have estimated foreground quality based on foreground segmentation mask properties. In particular, fitness of foreground segmentation masks to segmented image regions has been used to expand foreground masks to undetected areas while removing poor fitted isolated foreground (see Figure 16).

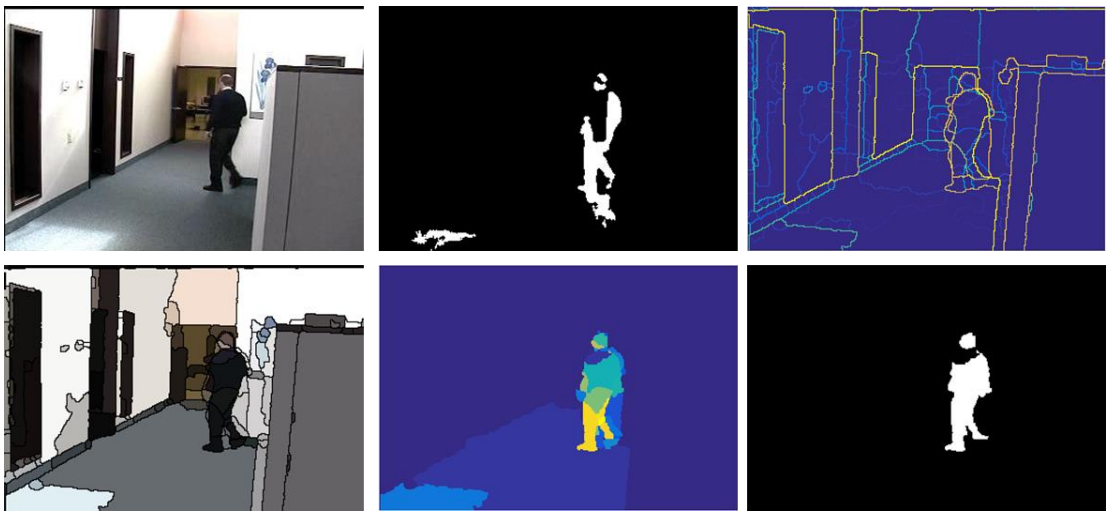


Figure 16. Foreground segmentation improvement. From left to right, first row: Image, its foreground segmentation mask and the UCM computed; second row: segmented image regions obtained from the UCM, foreground quality computed from fitness between foreground and segmented image regions and improved foreground mask by thresholding the foreground quality.

The algorithm has three stages. First, an ultrametric contour map (UCM) [44] is computed to perform image segmentation by thresholding the UCM. Second, fitness between the segmented foreground mask and the segmented image regions is computed to estimate a foreground segmentation quality. Finally, the segmentation quality is thresholded to compute an improved foreground mask. Figure 16 presents a complete example of the algorithm overview.
To validate the improvement we have used 6 sequences from the foreground segmentation dataset CDNET2014 [15]. In particular, we have used the algorithms GMM [35] and SuBSENSE [27] to sweep the threshold applied to obtain the improved foreground mask from the foreground quality. Figure 17 presents the evaluation, where the performance obtained sweeping the threshold demonstrates that an appropriate value improves the original performance (red).

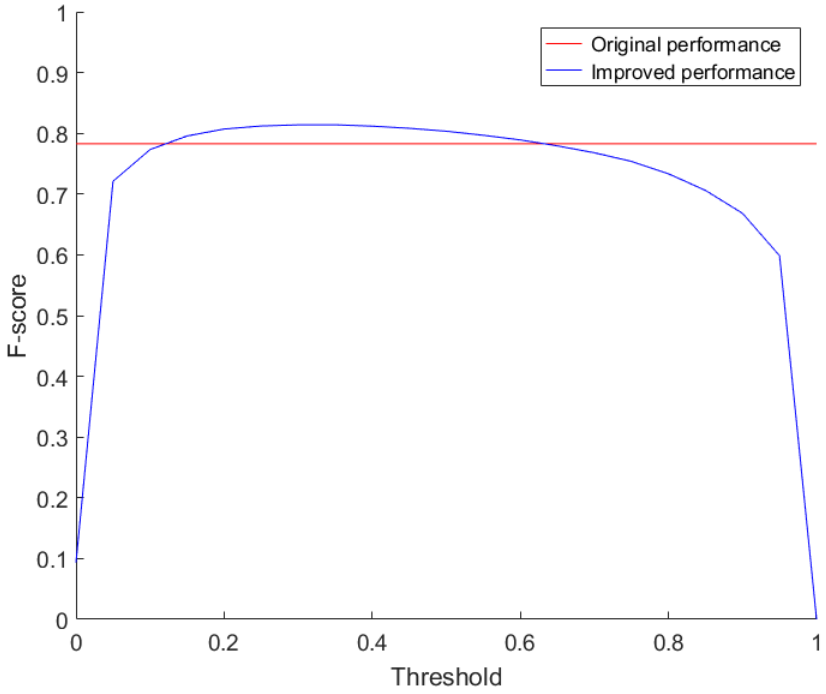


Figure 17. Performance of the post-processing algorithm sweeping the threshold used (blue), compared to the original performance (red). Thresholds approximately between 0.15 and 0.6 improve the original performance.

Moreover, filtering false positives induced by camera jitter have been addressed using information of image variations over time together with stationary object detection. Thus, we build a superpixel variation map that we use to filter the foreground segmentation map in order to compute a filtered foreground mask (see Figure 18).

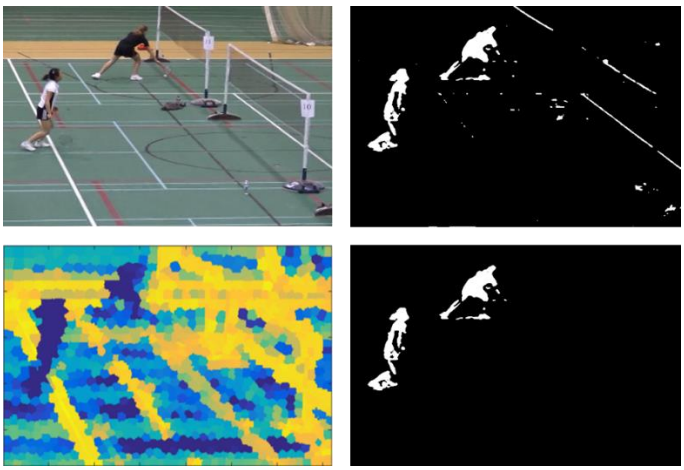


Figure 18. False positive filtering in camera jitter scenarios. From left to right, top row: image and segmented foreground; bottom row: variation map and filtered foreground mask.

Again, evaluating the results in the camera jitter sequences of CDNET2014, we achieve a performance improvement. In Figure 19 we show the improvement of the algorithm SuBSENSE, which is improved for the four CDNET2014 jitter sequences although it already faces camera jitter issues.

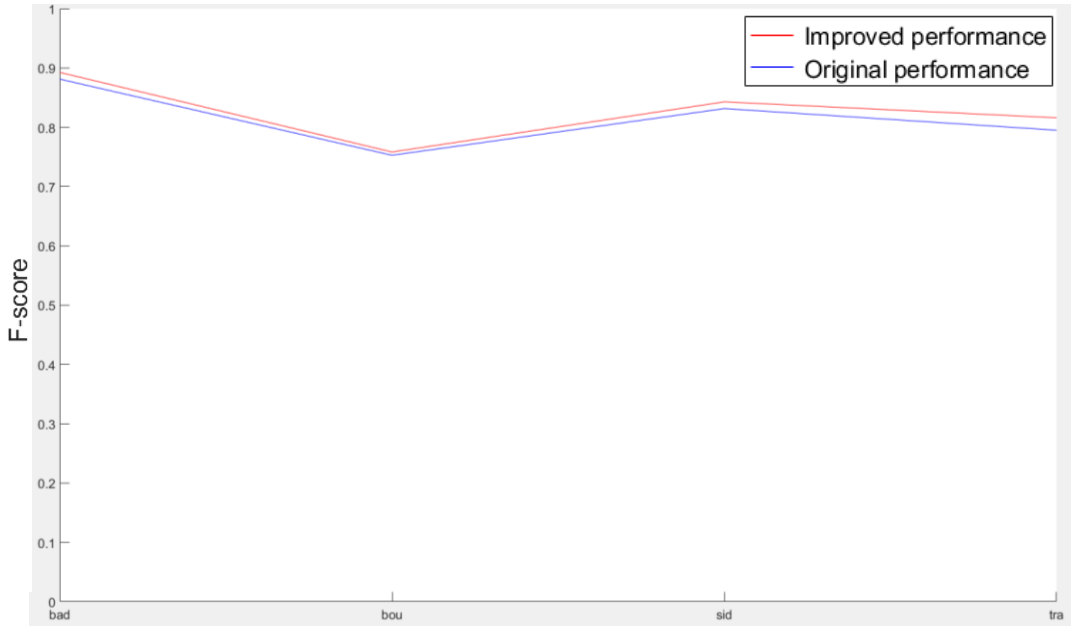


Figure 19. F-score performance of SuBSENSE (blue) against the filtered foreground version of SuBSENSE (red).

2.7. Background initialization based on the median

Continuing with the background initialization works from Subsections **¡Error! No se encuentra el origen de la referencia.** and **¡Error! No se encuentra el origen de la referencia.**, in this Graduate thesis [45] we have tackled the background initialization task from a new perspective introduced in [46]. We have used this algorithm, named LabGen-P, due to its recent success in a scene background modelling contest (http://scenebackgroundmodeling.net/). This algorithm estimates the background image using the temporal median of the pixels with less motion during the training frames. Therefore, it assumes that foreground objects always move and that the background is static. This premise is not always satisfied, as stationary objects are foreground objects with no motion. In consequence, this algorithm easily incorporates stationary objects into the estimated background. Moreover, the moving/non-moving pixel decision is based on frame-difference, which is a non-precise motion information based on spatially expanding a frame difference information. (see Figure 20).

Figure 20. Example of non-precise motion information from LabGen-P. From left to right: Image, frame difference and motion information (spatially-expanded frame-difference). Note that this motion information does not accurately reflect the true image motion.

The LabGen-P algorithm has three stages (see Figure 21): Motion estimation, to determine which pixels are less reliable for the background estimation; candidates selection, to update a buffer of the most reliable pixels over time to estimate a background image; and Background estimation, to compute a background image by applying a per-pixel median over all pixel representations stored in the buffer.

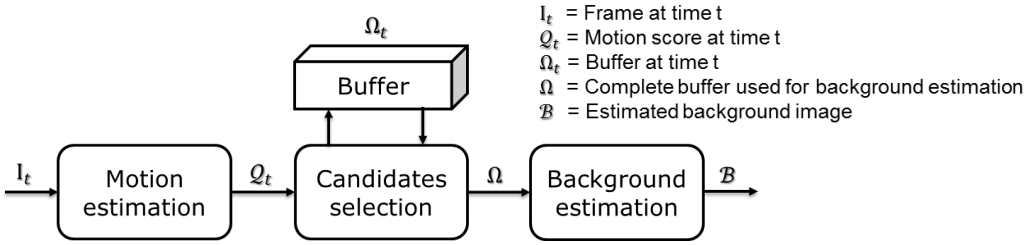


Figure 21. LabGen-P scheme.

We have modified the first and the second stages, aiming to improve the estimated background with respect the aforementioned issues of the original algorithm. On the one hand, we have replaced the frame-difference based motion information by an optical flow, which is capable of estimating more accurate motion information. On the other hand, we have performed a candidate selection that is not only based on motion information, but in a combination of motion and segmented image region sizes. Note that we have used the optical flow presented in [47] and the image regions presented in [48].

The results achieved were tested in the BEds (http://www-vpu.eps.uam.es/DS/BEds/) and the SBMnet (http://scenebackgroundmodeling.net/) datasets. Table 4 presents the improvement achieved for the BEds dataset in each category (Baseline, Clutter, LowFrame and StaticObject), whereas Table 5 presents the results in each category of the SBMnet dataset. Note that in the latter dataset there is an improvement in some categories, while others suffer a performance decrease that deals to no global improvement.

	FDiff	Reg+OF
Baseline	0.97445	0.97503
Clutter	0.97096	0.97284
LowFrame	0.97495	0.97763
StaticObject	0.94991	0.96364
MEAN	**0.9675675**	**0.972285**

Table 4. Results in BEds dataset. FDiff denotes the original algorithm and Reg+OF the proposed algorithm. The metric used for the evaluation is MS-SSIM (the higher the better).

	FDiff	Reg+OF
Basic	0.9739	0.9746
IntermittentMotion	0.9709	0.9675
Clutter	0.8978	0.9345
Jitter	0.8502	0.8266
IlluminationChange	0.9625	0.8120
BackgroundMotion	0.8430	0.8607
Very Long	0.9748	0.9459
Very Short	0.9457	0.9466
MEAN	**0.9273**	**0.9086**

Table 5. Results in SBMnet dataset. FDiff denotes the original algorithm and Reg+OF the proposed algorithm. The metric used for the evaluation is MS-SSIM (the higher the better).

3. Conclusions and future work

3.1. Achievements

As summary the achievements of task 2.3 are:

- Study of good object properties for stand-alone evaluation of Background Subtraction algorithms.
- Study of online people detection algorithms quality analysis using correlation metrics.
- Development of a Background Estimation algorithm for video sequences robust to stationary objects.
- Development of a Background Estimation algorithm for video sequences in a Master Thesis.
- Development of a long-term abandoned object detector robust against sudden illumination changes and stationary pedestrians.
- Implementation of two algorithms for Background Estimation from the literature.

3.2. Future work

As future work, we will focus on the last research milestone for this task T2.3, the "Stand-alone improvement of Background Subtraction algorithms.".

4. References

[1] Moreno De Pablos, E. “Seguimiento de objetos basado en múltiples algoritms”, Trabajo Fin de Grado, Degree ITST, Universidad Autonoma de Madrid, July 2016

[2] A. Smeulders, D. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, “Visual tracking: An experimental survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 7, pp. 1442 – 1468, July 2014.

[3] J. Yoon, D. Kim, and K.-J. Yoon, “Visual tracking via adaptive tracker selection with multiple features,” in European Conference on Computer Vision, 2012, pp. 28–41.

[4] X. Zhang, W. Hu, H. Bao, and S. Maybank, “Robust head tracking based on multiple cues fusion in the kernel-bayesian framework,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no. 7, pp. 1197–1208, July 2013.

[5] J. Kwon and K. Lee, “Tracking by sampling and integrating multiple trackers,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 7, pp. 1428–1441, July 2014.

[6] Y. Wu and T. Huang, “Robust visual tracking by integrating multiple cues based on co-inference learning,” International Journal of Computer Vision, vol. 58, pp. 55–71, November 2002.

[7] J. SanMiguel, A. Cavallaro, and J. Martinez, “Adaptive online performance evaluation of video trackers,” IEEE Transactions on Image Processing, vol. 21, no. 5, pp. 2812–2823, May 2012

[8] M. Heber, M. Godec, M. Ruther, P. Roth, and H. Bischof, “Segmentation-based tracking by support fusion,” Computer Vision and Image Understanding, vol. 117, no. 6, pp. 573–586, June 2013

[9] Z. Kalal, K. Mikolajczyk, and J. Matas, “Forward-backward error:Automatic detection of tracking failures,” in International Conference on Pattern Recognition, 2010, pp. 2756–2759

[10] K. Shearer, K. Wong, and S. Venkatesh, “Combining multiple trackingalgorithms for improved general performance,” Pattern Recognition, vol. 34, no. 6, pp. 1257–1269, June 2001

[11] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan, “Whose vote should count more: Optimal integration of labels from labelers of unknown expertise,” in Proceedings of Neural Information Processing Systems, 2009, pp. 2035–2043

[12] N. Anjum and A. Cavallaro, “Multifeature object trajectory clustering for video analysis,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, no. 11, pp. 1555–1564, Nov 2008.

[13] H. Wu, A. Sankaranarayanan, and R. Chellappa, "Online empirical evaluation of tracking algorithms," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 8, pp. 1443–1458, Aug 2010.

[14] O. Khalid, Juan c. SanMiguel, A. Cavallaro, "Multi-tracker partition fusión", IEEE Trans. On Circuits and Systems for Video technology, 2016

[15] Goyette, N., Jodoin, P.M., Porikli, F., Konrad, J., Ishwar, P., "A novel video dataset for change detection benchmarking", IEEE Transactions on Image Processing, vol. 23, nº 11, pp. 4663-4679, 2014.

[16] SanMiguel, J.C., Martinez, J.M, "On the evaluation of background subtraction algorithms without ground-truth", in Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 80-187, 2010.

[17] Shi, R., Ngan, K., Li, S., Paramesran, R., Li, H., "Visual quality evaluation of image object segmentation: Subjective assessment and objective measure", IEEE Transactions on Image Processing, vol. 24, nº 12, pp. 5033-5045, 2015.

[18] Villegas, P., Marichal, X., "Perceptually-weighted evaluation criteria for segmentation masks in video sequences", IEEE Transactions on Image Processing, vol. 13, nº 8, pp. 1092-1103, 2004.

[19] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., "The Pascal Visual Object Classes (VOC) challenge", International Journal of Computer Vision, vol. 88, nº 2, pp. 303-338, 2010.

[20] Wen, L., Du, D., Lei, Z., Li, S.Z., Yang, M.H., "JOTS: Joint Online Tracking and Segmentation", in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2226-2234, 2015.

[21] Salti, S. Lanza, A., Stefano, L., "Synergistic change detection and tracking", IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, nº 4, pp. 609-622, 2015.

[22] Cheng, F.C., Huang, S.C., Ruan, S.J., "Illumination-sensitive background modeling approach for accurate moving object detection", IEEE Transactions on Broadcasting, vol. 57, nº 4, pp. 794-801, 2011.

[23] Ramirez-Quintana, J., Chacon-Murguia, M., "Self-adaptive SOM-CNN neural system for dynamic object detection in normal and complex scenarios", Pattern Recognition, vol. 48, nº 4, pp. 1137-1149, 2015.

[24] Al-Najdawi, N., Bez, H., Singhai, J., Edirisinghe, E., "A survey of cast shadow detection algorithms", Pattern Recognition Letters, vol. 33, nº 6, pp. 752-764, 2012.

[25] Huerta, I., Holte, M., Moeslund, T., Gonzalez, J., "Chromatic shadow detection and tracking for moving foreground segmentation", Image and Vision Computing, vol. 41, pp. 42-53, 2015.

[26] Pham, D.S., Arandjelovic, O., Venkatesh, S., "Detection of dynamic background due to swaying movements from motion features", IEEE Transactions on Image Processing, vol. 24, nº 1, pp. 332-344, 2015.

[27] St-Charles, P.L., Bilodeau, G.A., Bergevin, R., "SuBSENSE: A universal change detection method with local adaptive sensitivity", IEEE Transactions on Image Processing, vol. 24, nº 1, pp. 359-373, 2015.

[28] Zhang, H., Fritts, J., Goldman, S., "Image segmentation evaluation: A survey of unsupervised methods", Computer Vision and Image Understanding, vol. 110, nº 2, pp. 260-280, 2008.

[29] Li, H., Meng, F., Luo, B., Zhu, S., "Repairing bad co-segmentation using its quality evaluation and segment propagation", IEEE Transactions on Image Processing, vol. 23, nº 8, pp. 3545-3559, 2014.

[30] Erdem, C., Sankur, B., Tekalp, A., "Performance measures for video object segmentation and tracking", IEEE Transactions on Image Processing, vol. 13, nº 7, 937-951, 2004.

[31] Zitnick, C., Dollar, P., "Edge Boxes: Locating Object Proposals from Edges", in Proceedings of European Conference on Computer Vision (ECCV), pp. 391-405, 2014.

[32] Correia, P., Pereira, F., "Stand-alone objective segmentation quality evaluation", EURASIP Journal on Advances in Signal Processing, vol. 4, pp. 1-12, 2002.

[33] Min, C., Zhang, J., Chang, B., Sun, B., Li, Y., "Spatio-temporal segmentation of moving objects using edge features in infrared videos", Optik-International Journal for Light and Electron Optics, vol. 125, nº 7, pp. 1809-1815, 2014.

[34] Giordano, D., Kavasidis, I., Palazzo, S., Spampinato, C., "Rejecting False Positives in Video Object Segmentation", in Proceedings of International Conference on Computer Analysis of Images and Patterns (CAIP), vol. 9256 (LNCS), pp. 100-112, 2015.

[35] Staufer, C., Grimson, W., "Adaptive background mixture models for real-time tracking", in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 246-252, 1999.

[36] Sajid, H., Samson Cheung, S.C., "Background subtraction for static & moving camera", in Proceedings of IEEE International Conference on Image Processing (ICIP), 4530-4534, 2015.

[37] Wang, R., Bunyak, F., Seetharaman, G., Palaniappan, K., "Static and moving object detection using flux tensor with split gaussian models", in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 420-424, 2014.

[38] Bouwmans, T., "Traditional and recent approaches in background modeling for foreground detection: An overview", Computer Science Review, vol. 1112, pp. 31-66, 2014.

[39] Fernández-Predraza, C., “Reconstrucción de fondo de escena a partir de secuencias de vídeo”, Graduate thesis of Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Universidad Autónoma de Madrid, 2016.

[40] Hsiao, H.-H., Leou, J.-J., “Background initialization and foreground segmentation for bootstrapping video sequences”, EURASIP Journal on Image and Video Processing, vol. 2013, no. 1, pp. 1-19, 2013.

[41] R. Zhang, W. Gong, A. Yaworski, and M. Greenspan, “Nonparametric on-line background generation for surveillance video”, Proceedings of International Conference on Pattern Recognition (ICPR), pp. 1177-1180, 2012.

[42] Ortego, D., SanMiguel, J.C., and Martínez, J.M., “Rejection based Multipath Reconstruction for Background estimation in Video Sequences with Stationary Objects”, Computer Vision and Image Understanding, vol. 147, pp. 23-37, 2016.

[43] Sanin, A., Sanderson, C., and Lovell, B.C., "Shadow detection: A survey and comparative evaluation of recent methods", Pattern Recognition, vol. 45, pp. 1684-1695, 2012.

[44] Arbelaez, P., Maire, M., Fowlkes, C., Malik, J., "From contours to regions: An empirical evaluation", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2294-2301, 2009.

[45] Gómez García, E., “Reconstrucción de fondo de escena basada en la mediana”, Graduate thesis of Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Universidad Autónoma de Madrid, 2017.

[46] Laugraud, B., Piérard, S., and Van Droogenbroeck, M., "LaBGen-P: A pixel-level stationary background generation method based on LaBGen", Proceedings of International Conference on Pattern Recognition (ICPR), pp. 107-113, 2016.

[47] Brox, T. and Bruhn, A. and Papenberg, N. and Weickert, J., "High Accuracy Optical Flow Estimation Based on a Theory for Warping", Proceedings of European Conference on Computer Vision (ECCV), pp. 25-36, 2004.

[48] Dollár, P., Zitnick, C. L., "Structured Forests for Fast Edge Detection" Proceedings of IEEE International Conference on Computer Vision (ICCV), pp. 1841-1848, 2013.