

event video

TEC2011-25995 EventVideo (2012-2014)

*Strategies for Object Segmentation, Detection and Tracking in Complex
Environments for Event Detection in Video Surveillance and Monitoring*

TR.01

EVALUATION RESULTS AND FUTURE RESEARCH LINES

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

AUTHORS LIST

Jesús Bescós Cano	J.Bescos@uam.es
Luis Caro Campos	Luis.Caro@uam.es
Marcos Escudero Viñolo	Marcos.Escudero@uam.es
Miguel Ángel García García	MiguelAngel.GarciaGarcia@uam.es
Álvaro García Martín	Alvaro.Garcia@uam.es
Rafael Martín Nieto	Rafael.Martinn@uam.es
José M. Martínez Sánchez	JoseM.Martinez@uam.es
Juan Carlos San Miguel Avedillo	JuanCarlos.SanMiguel@uam.es
Fabrizio Tiburzi Paramio	Fabricio.Tiburzi@uam.es

CHANGE LOG

Version	Date	Editor	Description
0.0	3/12/2012	José M. Martínez	Template; Introduction
0.1	12/12/2012	Juan Carlos San Miguel	WP5: event detection
0.2	17/12/2012	Álvaro García-Martín	WP3: people detection
0.3	18/12/2012	Miguel Ángel García	WP4: tracking algorithms description
0.4	27/12/2012	Rafael Martín Nieto	WP4: tracking results
0.5	29/12/2012	Marcos Escudero	WP2: segmentation
0.6	27/01/2013	José M. Martínez	Consolidation; style edition; comments; proposals for enhancements; draft conclusions
0.7	27/01/2013	José M. Martínez	V0.6 without tracking changes
0.8	17/02/2013	Fabrizio Tiburzi	WP2: segmentation T2.2
0.9	03/03/2013	Marcos Escudero	WP2: v0.7 update
0.10	15/03/2013	Álvaro García-Martín	WP3: v0.7 update
0.11	20/03/2013	Rafael Martín	WP4: v0.7 update
0.12	25/03/2013	Juan Carlos San Miguel	WP5: v0.7 update
0.13	28/04/2013	José M. Martínez	Final Working Draft editing
1.0	30/04/2013	José M. Martínez	TR release

CONTENTS:

1	INTRODUCTION	1
1.1	MOTIVATION	1
1.2	DOCUMENT STRUCTURE	1
2	SEGMENTATION IN FIXED CAMERA SCENARIOS	3
2.1	INTRODUCTION	3
2.1.1	<i>Background model nature</i>	<i>3</i>
2.1.2	<i>Background updating nature</i>	<i>3</i>
2.1.3	<i>Proposed VOS algorithms organization.....</i>	<i>4</i>
2.2	SELECTED EVALUATION SCENARIO	4
2.3	ALGORITHMS	9
2.3.1	<i>Splitting Gaussians in Mixture Models (SGMM) [6]</i>	<i>9</i>
2.3.2	<i>Splitting Gaussians in Mixture Models & Static Object Detection (SGMM-SOD) [7] 10</i>	<i>10</i>
2.3.3	<i>Chebyshev inequality based modelling (CHEBYSHEV) [8]</i>	<i>10</i>
2.3.4	<i>Gamma inequality based modelling (GAMMA) [9].....</i>	<i>11</i>
2.3.5	<i>Multilayer modelling updated by Bayes' rule (BAYES MULTI-LAYER) [10]</i>	<i>11</i>
2.3.6	<i>Pixel-Based Adaptive Segmenter (PBAS) [11].....</i>	<i>12</i>
2.3.7	<i>Probabilistic Superpixel Markov Random Fields (PSP-MRF) [12]</i>	<i>12</i>
2.3.8	<i>Self-Organized Background Subtraction with Spatial Coherence (SOBS-SC)[13] 13</i>	<i>13</i>
2.3.9	<i>Enhanced Visual Background Extractor (ViBe+)[14].....</i>	<i>14</i>
2.3.10	<i>Region-based video object segmentation RBVOS [15].....</i>	<i>14</i>
2.4	COMPARATIVE RESULTS	15
2.5	CONCLUSIONS	18
2.6	FUTURE RESEARCH LINES	18
2.6.1	<i>Refinement by post-processing techniques.....</i>	<i>18</i>
2.6.2	<i>Use of alternative features.....</i>	<i>18</i>
2.6.3	<i>Include semantics in the descriptions.....</i>	<i>18</i>
3	SEGMENTATION IN MOVING CAMERA SCENARIOS	21
3.1	INTRODUCTION	21
3.2	SELECTED EVALUATION SCENARIO	21
3.3	ALGORITHMS	23
3.3.1	<i>Robust Global Motion Estimation Oriented to Video Object Segmentation (RGME-VOS) [17].....</i>	<i>23</i>
3.3.2	<i>Robust Global Motion Estimation in presence of Large Objects (GME-LO)</i>	<i>24</i>
3.4	COMPARATIVE RESULTS	24
3.5	CONCLUSIONS	26
4	PEOPLE MODELLING AND DETECTION	27
4.1	INTRODUCTION	27
4.2	SELECTED EVALUATION SCENARIO	27
4.3	ALGORITHMS	28
4.3.1	<i>Edge[23].....</i>	<i>28</i>
4.3.2	<i>Fusion[24].....</i>	<i>28</i>
4.3.3	<i>HOG[25]</i>	<i>28</i>
4.3.4	<i>ISM[26]</i>	<i>29</i>
4.3.5	<i>TUD[27].....</i>	<i>29</i>
4.3.6	<i>DTDP[28]</i>	<i>30</i>
4.3.7	<i>IMM [29].....</i>	<i>30</i>
4.4	COMPARATIVE RESULTS	30

4.4.1	<i>Evaluation dataset A</i>	31
4.4.2	<i>Evaluation dataset B</i>	33
4.4.3	<i>Evaluation dataset B with motion</i>	33
4.5	CONCLUSIONS	34
4.6	FUTURE RESEARCH LINES	34
4.6.1	<i>Expand the evaluation dataset PDs</i>	34
4.6.2	<i>Improve or refine segmentation</i>	34
4.6.3	<i>Appearance and motion fusion</i>	34
5	TRACKING	35
5.1	INTRODUCTION	35
5.2	SELECTED EVALUATION SCENARIO	35
5.3	ALGORITHMS	37
5.3.1	<i>Colour-based mean-shift (MS) [34]</i>	37
5.3.2	<i>Template matching (TM) [35]</i>	39
5.3.3	<i>Lucas-Kanade tracking (LK) [36]</i>	40
5.3.4	<i>Particle filter-based colour tracking (PFC) [37]</i>	42
5.3.5	<i>Corrected background colour-based mean-shift tracker (CBWH) [38]</i>	43
5.3.6	<i>Scale and orientation adaptive mean-shift tracking (SOAMST) [39]</i>	45
5.4	COMPARATIVE RESULTS	46
5.5	CONCLUSIONS	48
5.6	FUTURE RESEARCH LINES	49
5.6.1	<i>Evaluation of more complex algorithms</i>	49
5.6.2	<i>Modify and complete the content set</i>	49
5.6.3	<i>Evaluation of the algorithms with new metrics</i>	49
5.6.4	<i>Fusion</i>	49
6	EVENT DETECTION	51
6.1	INTRODUCTION	51
6.2	ABANDONED AND STOLEN OBJECT DISCRIMINATION	51
6.2.1	<i>Evaluation scenario</i>	52
6.2.2	<i>Approaches</i>	52
6.2.3	<i>Comparative results</i>	53
6.3	HUMAN INTERACTIONS (WITH OBJECTS AND HUMANS).....	56
6.3.1	<i>Evaluation scenario</i>	56
6.3.2	<i>Approaches</i>	59
6.3.3	<i>Comparative results</i>	60
6.4	CONCLUSIONS	63
6.5	FUTURE RESEARCH LINES	63
6.5.1	<i>Stationary object detection in high-density scenarios</i>	63
6.5.2	<i>Human-related interactions enhancements</i>	64
7	CONCLUSIONS AND FUTURE WORK	65
8	REFERENCES	67

1 Introduction

1.1 Motivation

In this Technical Report we evaluate our current algorithms and State-of-Art (SoA) ones using the evaluation framework (including datasets, associated ground-truth and metrics) described in Deliverable 5.3v1 “EventVideo test sequences, ground-truth and evaluation methodology”[1]. The analysis of the obtained results is used to define the research lines for the rest of the project.

1.2 Document structure

This document contains the following chapters:

- Chapter 1: Introduction to this document
- Chapter 2: Segmentation in fixed camera scenarios
- Chapter 3: Segmentation in moving camera scenarios
- Chapter 4: People Modelling and Detection
- Chapter 5: Tracking
- Chapter 6: Event Detection
- Chapter 7: Conclusions and future work.

2 Segmentation in fixed camera scenarios

2.1 Introduction

Commonly named as video object segmentation (VOS), this task is sometimes defined as foreground-background segregation or change detection. Although there are small differences between a change detection algorithm and a VOS algorithm (mainly the inclusion in the background model of foreground objects that remain static for a long time), both terms are commonly synonyms in the state of the art and also along this document. Regardless of the name, the main objective of a VOS system is to detect a set of pixels (foreground) as changes respect to a set of reference pixels (background).

The motion of the camera used to capture the scene limits the applicability of VOS techniques. In the simplest case, the camera is fixed in a spatial position and captures the time evolution of a spatially static shot (a single frame of view). Under these conditions, temporal evolution of the set of background pixels can be modelled or estimated, pixel-based or region-based, and foreground pixels can be detected as deviations from the background set. On the contrary, when the video is being (or has been) recorded allowing camera motion, both time and spatial evolution of the set of background pixels need to be considered. This task is usually performed by camera motion compensation. The foreground is then detected as pixels or regions whose estimated motion differs from the modelled one.

In this technical report we just focus on algorithms devoted to perform VOS in fixed camera scenarios; these are usually tagged as *background subtraction algorithms*. VOS by background subtraction, while being the simplest and the most studied situation, is still an unresolved task

VOS algorithms are classically characterized as parametric or non-parametric. Although this organization is still valid, it should be extended in order to account for recent trends in this task. In essence, VOS algorithms can be classified by two main aspects: how they model the background and how they update the background model.

2.1.1 Background model nature

A background model is tagged as parametric if it tries to adjust the input data to a predefined probability distribution function at each unit of analysis (a single Gaussian, a mixture of Gaussians, etc.). Then, foreground is detected by evaluating new instances against their corresponding model and thresholding low probabilities. On the contrary, either if the model is not predefined and is estimated by a set of samples, or if the set of samples is sprightly used as the background model, the VOS algorithm can be considered non-parametric. However, this classification does not imply that non-parametric VOS algorithms are completely parameter free. Non-parametric models are supposed to be more flexible (but also more sensitive) to the nature of the input data.

2.1.2 Background updating nature

To account for temporal variations (sudden or gradual changes in illumination, aggregation of new objects, uncovering of unobserved background, etc.) of the modelled background, the model should be updated. Classically, the foreground discrimination and the model updating processes are performed at pixel level, that is, the classification of a pixel either as background or foreground just relies on its temporal evolution and only affects the updating of the model at such pixel. However, solutions that use neighbouring pixel information to discriminate the

foreground and/or update the neighbouring pixels independently of their classification have been proven to obtain promising results.

Additionally, these decision and updating mechanisms can be performed either entirely at pixel level or can rely on higher level analysis modules (e.g. at blob, region or object level) that refine or correct pixel-based decisions.

2.1.3 Proposed VOS algorithms organization

According to the aforementioned aspects, and irrespective of the features and the distances used to analyze the scene and to discriminate the foreground, we can organize most of the existing VOS algorithms in the categories described in Table 1. In this organization, classical parametric algorithms as the Mixture of Gaussians MoG [2] would be classified as a VOSC1 algorithm while first non-parametric algorithms as the Kernel Density Estimation (KDE) [3] would be included in the VOSC5 category.

Background Model Nature	Parametric				Non-parametric			
	Single Pixel		Group of pixels		Single Pixel		Group of pixels	
Updating Mechanisms	Pixel Level	Higher Levels	Pixel Level	Higher Levels	Pixel Level	Higher Levels	Pixel Level	Higher Levels
Category	VOSC1	VOSC2	VOSC3	VOSC4	VOSC5	VOSC6	VOSC7	VOSC8
Examples (in this document)	[6]	[7], [8]	[9]	[10]	[11]	[12]	[13]	[14],[15]

Table 1 – Proposed organization of Video Object Segmentation algorithms

With this organization in mind, the rest of this chapter is organized as follows. Section **Error! Reference source not found.** describes the evaluation scenario and its associated complexity factors. Section **Error! Reference source not found.** categorizes and briefly describes the algorithms chosen to perform the comparative study presented at section 2.4. Finally, section **Error! Reference source not found.** derives some conclusions from the analysis of the study while section **Error! Reference source not found.** includes a set of future research lines.

2.2 Selected evaluation scenario

To date, many change detection algorithms have been developed that perform well in some types of videos but not in many others. No single algorithm seems to be able to simultaneously address all the key challenges that accompany real-world (non-synthetic) videos. This is due, in part, to the absence of a realistic large-scale dataset with accurate ground truth, which would help designing such general purpose algorithm.

The Change Detection Dataset (referenced in **Error! Reference source not found.** as <http://www.changedetection.net/>, a more detailed description can be found at [4]) was proposed as part of the CVPR 2012 Change Detection Workshop as a rigorous and comprehensive academic benchmarking effort for testing and ranking existing and new algorithms. This dataset aims to provide a balanced coverage of the range of challenges present in the real world. The main advantages of this dataset respect to the others detailed in **Error! Reference source not found.** can be summarized as following:

- It presents a compound of real-world videos (including thermal) and it is representative of indoor and outdoor visual data captured today in surveillance and smart environment scenarios.

- It includes a comprehensive set of carefully human-annotated ground truth change/motion areas to enable a precise quantitative comparison and ranking of various algorithms.
- It divides the content in 6 categories selected to include diverse motion and change detection challenges.

In **Error! Reference source not found.**, videos in this dataset were assigned an estimated complexity: S1 or S2, S1 corresponding to low complex background videos and low foreground density, and S2 including similar foreground but background scenarios of high complexity¹. However, the complexity of scenarios labelled as S2 should be further analysed. Which are the challenges (CH) that face VOS algorithms in the analysis of real-world videos, even if the foreground is of low density? The exhaustively surveys presented at [5] and [6] summarize them:

CH1: Light changes. A VOS system should adapt to illumination changes whether gradual changes (e.g., time of day in outdoor scenarios) or sudden changes (e.g., light switch in indoors)

CH2: Moving background. A good VOS algorithm should handle the relocation of background objects, non-stationary background objects (e.g. waving trees), and image changes due to small camera motion which is common in outdoor applications (e.g. camera jitter by wind load).

CH3: Cast Shadows. Shadows share the same motion patterns and have a similar magnitude of intensity change as that of the foreground objects. Since cast shadows can be as big as the actual objects, their incorrect classification as foreground results in inaccurate detection and severely harm the overall results of a particular algorithm. Self-shadows are less problematic as they affect an internal part of the foreground, then they are equivalent to foreground in the quantification of the algorithm performance.


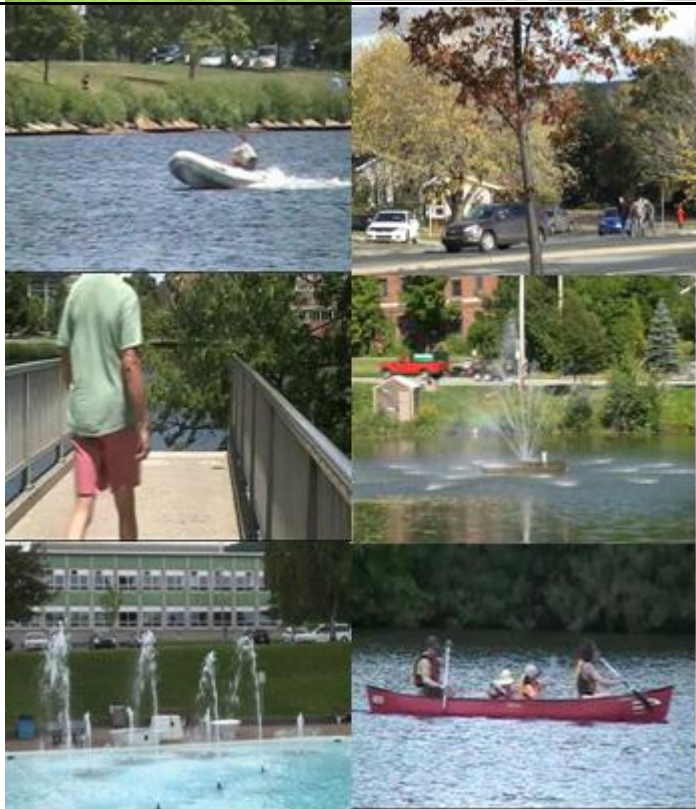
CH4: Model initialization (bootstrapping). Most of the background models are built on a set of initial parameters that come out from a short sequence (or from the beginning of a sequence), in which no foregrounds objects are present. This is a too strong assumption, because in some situations it is difficult or impossible to control the area being monitored (e.g., public zones), which might be characterized by a continuous presence of moving objects which can also be of considerable size respect to the field of view.

CH5: Camouflage. As a foreground object might have similar characteristics as the background, it becomes difficult to distinguish between them while being robust to small variations of the background. Camouflage is a particular case of the classical sensitivity discriminability trade-off.

Numerically, the Change Detection Dataset [4] contains 31 real-world videos adding up to over 80,000 frames organized in the categories depicted in Figure 1.

Dataset Category (number of videos)	Sample Frames	Predominant Complexity factors
---	---------------	--------------------------------------

¹ We are not aware of any VOS dataset including videos of categories S3 and S4. As defined in **Error! Reference source not found.**, scenarios tagged with those complexity factors combine crowded foregrounds and simple and complex backgrounds respectively).

<p>1. Baseline (4)</p>		<p>CH 3 CH4 CH5</p>
<p>2. Dynamic Background (6)</p>		<p>CH 1 CH 2 CH5</p>

<p>3. Intermittent Object Motion (6)</p>		<p>CH 2 CH 3 CH 5</p>
<p>4. Shadow (6)</p>		<p>CH 1 CH 3 CH 4 CH 5</p>

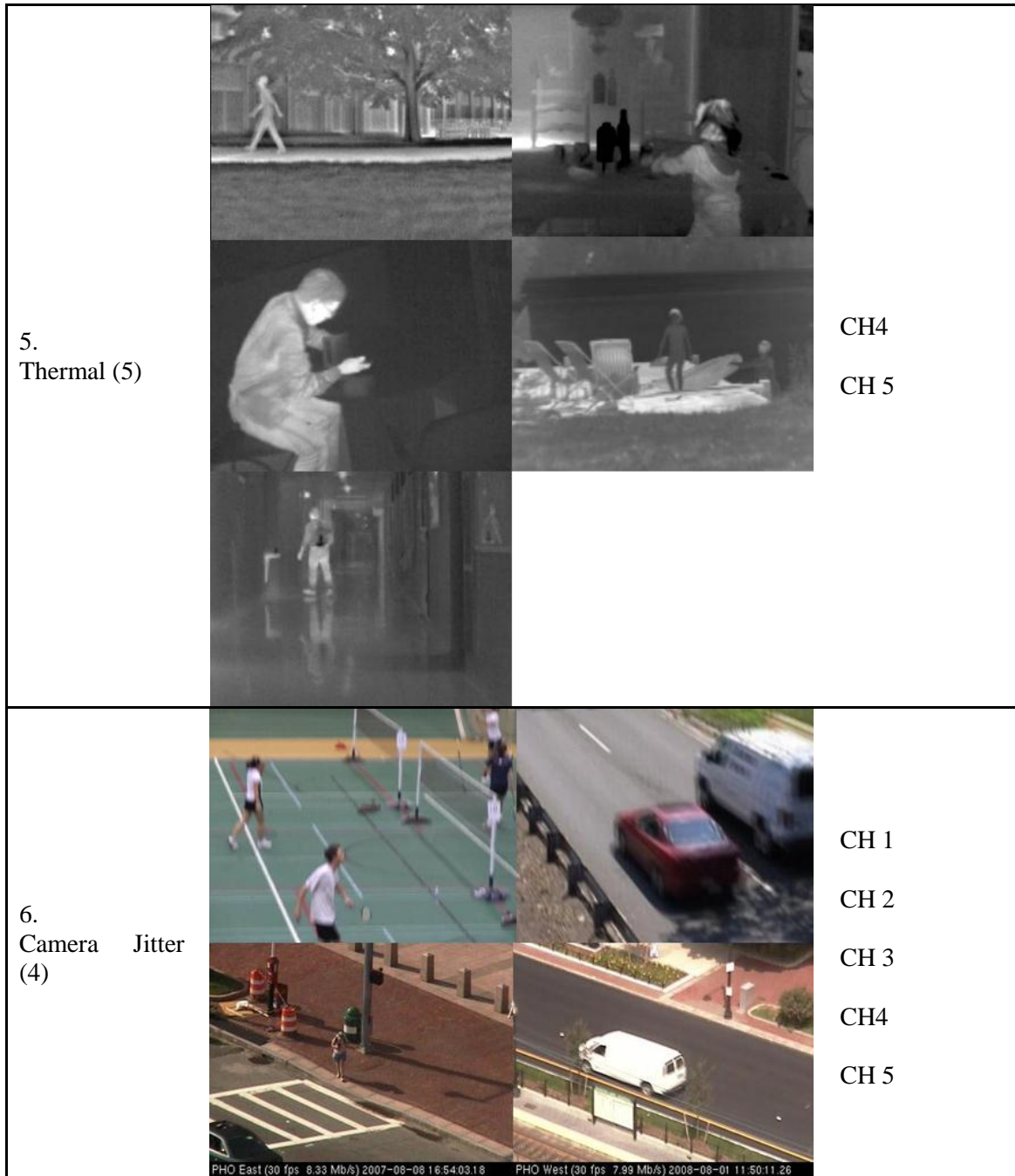


Figure 1 – VOS dataset description (challenges of main relevance are highlighted in bold)

Several metrics were computed in [4] aiming to perform a faithful comparison among the algorithms. The ones used in this technical report are listed at Table 2.

Metrics used for results comparison		
Name	Description	Equivalent metric in Error! Reference source not found.
TP	True Positives	TP

FP	False Positives	FP
FN	False Negatives	FN
TN	True Negatives	TN
RE	Recall: $TP / (TP + FN)$	R1
SP	Specificity: $TN / (TN + FP)$	R0
FPR	False Positive Rate: $FP / (FP + TN)$	1-R0
FNR	False Negative Rate : $FN / (TP + FN)$	1-R1
PWC	Percentage of Wrong Classifications: $100 * (FN + FP) / (TP + FN + FP + TN)$	-
F-M	F-Measure : $(2 * Precision * Recall) / (Precision + Recall)$	FS1
PR	Precision : $TP / (TP + FP)$	P1

Table 2 – Metrics used for result comparison

2.3 Algorithms

Top-cited VOS algorithms have been tested over the Change Detection Dataset. We include here a list of the results obtained by the top-ranked in each category, preceding each by a brief overview of their operation. Results were straightly extracted from [4], where the parameters of the algorithms were adequately tuned to the dataset sequences. Additionally, we have included two more VOS algorithms developed (one also designed) at the VPULab.

2.3.1 Splitting Gaussians in Mixture Models (SGMM) [6]

2.3.1.1 Algorithm overview

The authors claim that Gaussian mixture models often suffer from the problem of converging to poor solutions if the main mode stretches and thus over-dominates weaker distributions. Based on the results of the Split and Merge EM algorithm, they propose a solution to this problem. They define a new splitting operation and the corresponding criterion for the selection of candidate modes in order to avoid over-dominating ones. They also propose a heuristic to adaptively compute a value for the correct initialization of the variance parameters of new created modes. This heuristic is based on the estimation of the variance of new modes from the median of all the observed variances until its creation.

VOS category: VOSC1
 Explicitly designed to tackle challenges: CH1 CH2

Observations: The system seems to work properly at shadowed scenarios. However, its authors do not include any special module to tackle the shadows challenge.

2.3.1.2 Results

Category	RE	SP	FPR	FNR	PWC	F-M	PR
Baseline	0.8680	0.9949	0.0051	0.1320	1.2436	0.8594	0.8584
Dynamic Background	0.7715	0.9933	0.0067	0.2285	0.9132	0.6380	0.6665
Intermittent object motion	0.5013	0.9853	0.0147	0.4987	4.9180	0.5397	0.6993
Shadow	0.8580	0.9889	0.0111	0.1420	1.7965	0.7944	0.7617
Thermal	0.5363	0.9970	0.0030	0.4637	3.9394	0.6481	0.9263
Camera Jitter	0.7088	0.9869	0.0131	0.2912	2.3761	0.7251	0.7752
Overall	0.7073	0.9910	0.0090	0.2927	2.5311	0.7008	0.7812

Table 3 – SGMM [6] results as stated in [4]

2.3.2 Splitting Gaussians in Mixture Models & Static Object Detection (SGMM-SOD) [7]

2.3.2.1 Algorithm overview

Authors combine two models; one is devoted to detect motion while the other aims to achieve a representation of the empty scene. The differences in foreground detection of the complementary models are used to identify new static regions. A higher-level module is used to detect if a static object was placed or removed from the scene. Static objects are prevented from being incorporated into the empty scene model while removed objects are dropped from both models.

VOS category:

VOSC2

Explicitly designed to tackle challenges:

CH1 CH2 CH4

Observations: This algorithm is explicitly devoted to run over very crowded scenarios. However, with the available datasets we cannot evaluate its operation in scenarios of estimated complexity factors S3 and S4. Again, no special treatment of the shadows is performed, even while being this algorithm one of the top performing at the Shadow category.

2.3.2.2 Results

Category	RE	SP	FPR	FNR	PWC	F-M	PR
Baseline	0.9361	0.9971	0.0029	0.0639	0.5578	0.9169	0.9018
Dynamic Background	0.7538	0.9964	0.0036	0.2462	0.6121	0.6861	0.7391
Intermittent object motion	0.7198	0.9833	0.0167	0.2802	3.0501	0.6873	0.7737
Shadow	0.9184	0.9903	0.0097	0.0816	1.2583	0.8613	0.8187
Thermal	0.5941	0.9966	0.0034	0.4059	1.8926	0.6949	0.9521
Camera Jitter	0.6310	0.9920	0.0080	0.3690	2.1632	0.6988	0.8273
Overall	0.7589	0.9926	0.0074	0.2411	1.5890	0.7576	0.8354

Table 4 – SGMM-SOD [7] results as stated in [4]

2.3.3 Chebyshev inequality based modelling (CHEBYSHEV) [8]

2.3.3.1 Algorithm overview

The background model is based on a Chebyshev probability inequality. The model is supported with peripheral and recurrent motion detectors. The system additionally uses a shadow detection module as well as relevance feedback from higher-level object tracking and object classification to further refine the segmentation accuracy.

VOS category:

VOSC2

Explicitly designed to tackle challenges:

CH1 CH2 CH3 CH4

Observations: As stated by the authors, some of the higher-level modules conflict with the techniques designed at pixel level processing.

2.3.3.2 Results

Category	RE	SP	FPR	FNR	PWC	F-M	PR
Baseline	0.8266	0.9970	0.0030	0.1734	0.8304	0.8646	0.9143
Dynamic Background	0.8182	0.9976	0.0024	0.1818	0.4086	0.7520	0.7339
Intermittent object motion	0.3570	0.9807	0.0193	0.6430	6.4700	0.3863	0.7688
Shadow	0.8670	0.9887	0.0113	0.1330	1.5561	0.8333	0.8103

Thermal	0.6887	0.9963	0.0037	0.3113	1.4283	0.7230	0.8906
Camera Jitter	0.7223	0.9725	0.0275	0.2777	3.6203	0.6416	0.5960
Overall	0.7133	0.9888	0.0112	0.2867	2.3856	0.7001	0.7856

Table 5 – Chebyshev [8] results as stated in [4]

2.3.4 Gamma inequality based modelling (GAMMA) [9]

2.3.4.1 Algorithm overview

The background model is dynamically generated based on temporal information. Foreground is detected by performing a significance test over the difference between a particular frame and the background model. The difference is supposed to be caused just by camera noise in absence of foreground and therefore it is modelled by a Gaussian distribution. The distribution of the difference at a spatial neighbourhood of a pixel is compared to a dynamic threshold modelled by a ratio of Gamma functions.

VOS category: VOSC3
Explicitly designed to tackle challenges: CH1

Observations: We use our own implementation of this algorithm to obtain the results included at Table 6. Eight different configurations of the algorithm’s learning rate and sensitivity to foreground have been tested. Specifically, we have swept the learning rate value from 5% to 20% with increments of 5% at each step and we have evaluated the algorithm with foreground’s sensitivities of 7, 13 and 26. Included results correspond to the best configuration at each category in terms of average F Score Measure (*F-M*).

2.3.4.2 Results

Category	RE	SP	FPR	FNR	PWC	F-M	PR
Baseline	0.8241	0.9980	0.0020	0.1759	0.6707	0.8773	0.9406
Dynamic Background	0.7849	0.9587	0.0413	0.2151	4.2557	0.3301	0.2348
Intermittent object motion	0.8972	0.7975	0.2025	0.1028	18.163	0.5060	0.4275
Shadow	0.7986	0.9831	0.0169	0.2014	2.4146	0.7069	0.6549
Thermal	0.7381	0.9901	0.0099	0.2619	1.8812	0.7509	0.7724
Camera Jitter	0.6839	0.9408	0.0592	0.3161	6.9153	0.4639	0.3609
Overall	0.7878	0.9447	0.0553	0.2122	5.7169	0.6059	0.5652

Table 6 – Gamma [9] results

2.3.5 Multilayer modelling updated by Bayes’ rule (BAYES MULTILAYER) [10]

2.3.5.1 Algorithm overview

This algorithm works by modelling the different appearances of a pixel in a set of independent layers. Its main contribution with respect to the existing approaches is the use of an a priori classification scheme that classifies the pixel before updating the background model. This scheme isolates the pixel instances that belong to the foreground, hence avoiding their influence in the model updating and discrimination processes of the subsequent frames. Additionally, authors propose the inclusion of a foreground model driven by a tracking module. The model updating is performed over a group of pixels by feeding back the results obtained at higher levels.

VOS category: VOSC3

Explicitly designed to tackle challenges:

CH1 CH2 CH4

Observations: We use our own implementation of this algorithm to obtain the results included at Table 7. Configuration parameters have not been specifically tuned.

2.3.5.2 Results

Category	RE	SP	FPR	FNR	PWC	F-M	PR
Baseline	0.6547	0.9914	0.0086	0.3453	2.1523	0.6658	0.6967
Dynamic Background	0.8294	0.9601	0.0399	0.1706	4.1094	0.5160	0.4850
Intermittent object motion	0.4867	0.9677	0.0323	0.5133	6.7053	0.4887	0.5579
Shadow	0.6633	0.9683	0.0317	0.3366	4.6287	0.5841	0.5292
Thermal	0.6566	0.9931	0.0069	0.3433	2.6112	0.7162	0.8607
Camera Jitter	0.8177	0.9910	0.0090	0.1823	1.0681	0.5912	0.5668
Overall	0.6847	0.9786	0.0214	0.3152	3.5458	0.5937	0.6161

Table 7 – BAYES MULTI-LAYER [10] results

2.3.6 Pixel-Based Adaptive Segmenter (PBAS) [11]

2.3.6.1 Algorithm overview

It follows a non-parametric background modelling approach. Background is modelled by a history of recently observed pixel values. The foreground detection depends on a decision threshold. The background update is based on a learning parameter. Both parameters are extended to dynamic per-pixel state variables and dynamic controllers were introduced to control them by estimation of the background dynamics.

VOS category:

VOSC3

Explicitly designed to tackle challenges:

CH1 CH2 CH3

Observations: Up to nine parameters should be tuned. The complexity of its configuration may complicate its use over untrained videos as well as its use as an on-the-fly VOS algorithm.

2.3.6.2 Results

Category	RE	SP	FPR	FNR	PWC	F-M	PR
Baseline	0.9594	0.9970	0.0030	0.0406	0.4858	0.9242	0.8941
Dynamic Background	0.6955	0.9989	0.0011	0.3045	0.5394	0.6829	0.8326
Intermittent object motion	0.6700	0.9751	0.0249	0.3300	4.2871	0.5745	0.7045
Shadow	0.9133	0.9904	0.0096	0.0867	1.2753	0.8597	0.8143
Thermal	0.7283	0.9934	0.0066	0.2717	1.5398	0.7556	0.8922
Camera Jitter	0.7373	0.9838	0.0162	0.2627	2.4882	0.7220	0.7586
Overall	0.7840	0.9898	0.0102	0.2160	1.7693	0.7532	0.8160

Table 8 – PABS [11] results as stated in [4]

2.3.7 Probabilistic Superpixel Markov Random Fields (PSP-MRF) [12]

2.3.7.1 Algorithm overview

In this work, the authors proposed a post-processing framework to improve a given VOS mask with the use of Probabilistic Superpixel Markov Random Fields. First, they convert the input pixel-based VOS into a probabilistic superpixel (similar-in-shape regions) representation. Based

on these probabilistic superpixels, a Markov random field exploits structural information and similarities to improve the VOS mask.

VOS category: VOSC6

Explicitly designed to tackle challenges: -

Observations: The algorithm starts from pre-computed results obtained by different state-of-the-art VOS systems. No explicit mention to the one used at each category of the Change Detection Dataset [3] is made along the paper.

2.3.7.2 Results

Category	RE	SP	FPR	FNR	PWC	F-M	PR
Baseline	0.9319	0.9978	0.0022	0.0681	0.4127	0.9289	0.9261
Dynamic Background	0.8955	0.9859	0.0141	0.1045	1.4514	0.6960	0.6576
Intermittent object motion	0.7010	0.9530	0.0470	0.2990	6.0594	0.5645	0.5727
Shadow	0.8736	0.9829	0.0171	0.1264	2.2414	0.7907	0.7281
Thermal	0.5991	0.9962	0.0038	0.4009	1.9189	0.6932	0.9218
Camera Jitter	0.8211	0.9825	0.0175	0.1789	2.2781	0.7502	0.7009
Overall	0.8037	0.9830	0.0170	0.1963	2.3937	0.7372	0.7512

Table 9 – PSP-MRF [12] results as stated in [4]

2.3.8 Self-Organized Background Subtraction with Spatial Coherence (SOBS-SC)[13]

2.3.8.1 Algorithm overview

It is based on the neural background model automatically generated by a self-organizing method, without prior knowledge about the involved patterns. Such adaptive model is supposed to handle scenes containing moving backgrounds, gradual illumination variations and camouflage, and to add into the background model cast-shadows generated by moving objects. Moreover, the introduction of spatial coherence into the background update procedure provides further robustness against false detections.

VOS category: VOSC7

Explicitly designed to tackle challenges:

CH1 CH2 CH3 CH5

Observations: Nature of the feature vector has been selected to provide robustness against cast shadows. On the contrary, there are not devoted parts of the work designed to avoid the influence of camouflage, neither at feature selection stage nor by means of a specific module.

2.3.8.2 Results

Category	RE	SP	FPR	FNR	PWC	F-M	PR
Baseline	0.9327	0.9980	0.0020	0.0673	0.3747	0.9333	0.9341
Dynamic Background	0.8918	0.9836	0.0164	0.1082	1.6899	0.6686	0.6283
Intermittent object motion	0.7237	0.9613	0.0387	0.2763	5.2207	0.5918	0.5896
Shadow	0.8502	0.9834	0.0166	0.1498	2.3000	0.7786	0.7230
Thermal	0.6003	0.9957	0.0043	0.3997	1.9841	0.6923	0.8857
Camera Jitter	0.8113	0.9768	0.0232	0.1887	2.8794	0.7051	0.6286
Overall	0.8017	0.9831	0.0169	0.1983	2.4081	0.7283	0.7315

Table 10 – SOBS-SC [13] results as stated in [4]

2.3.9 Enhanced Visual Background Extractor (ViBe+)[14]

2.3.9.1 Algorithm overview

This technique models the background with a set of samples for each pixel and compares new frames, pixel by pixel, to determine if a pixel belongs to the background or to the foreground. In its original version, the scope of ViBe was limited to background modelling. In this extension, a set of modifications that alter the working of ViBe were introduced. They include the inhibition of propagation to the background model of samples placed around internal borders and the distinction between the updating and segmentation masks. Finally they include a model to post-process the output by some operations on the connected components.

VOS category:

VOSC8

Explicitly designed to tackle challenges:

CH2 CH3

Observations: Some of the new modules seem to rely excessively on configuration parameters. The selected higher level modules may harm the flexibility provided at the previous version of the algorithm. However, in overall it performs better than its predecessor.

2.3.9.2 Results

Category	RE	SP	FPR	FNR	PWC	F-M	PR
Baseline	0.8283	0.9974	0.0026	0.1717	0.9631	0.8715	0.9262
Dynamic Background	0.7616	0.9980	0.0020	0.2384	0.3838	0.7197	0.7291
Intermittent object motion	0.4729	0.9820	0.0180	0.5271	5.4282	0.5093	0.7513
Shadow	0.8108	0.9910	0.0090	0.1892	1.6516	0.8153	0.8302
Thermal	0.5411	0.9974	0.0026	0.4589	2.8201	0.6646	0.9477
Camera Jitter	0.7293	0.9908	0.0092	0.2707	1.8473	0.7538	0.8064
Overall	0.6907	0.9928	0.0072	0.3093	2.1824	0.7224	0.8318

2.3.10 Region-based video object segmentation RBVOS [15]

2.3.10.1 Algorithm overview

This technique is based on region-level analysis. A robust-to-illumination region segmentation is used as the analysis entity of a post-processing framework for region matching. A multi-layer region-based background model is used to account for multimodality (region variability). Differently than in [12], the temporal evolution of the segments (here regions, there superpixels) is used during the analysis, e.g. a new background model is built during the post-processing.

VOS category:

VOSC8

Explicitly designed to tackle challenges:

CH2 CH3 CH4

Observations: We use our own implementation of this algorithm to obtain the results included at Table 11. Results have been obtained by refining those resulting from the application of BAYES [10]. Configuration parameters have not been specifically tuned.

2.3.10.2 Results

Category	RE	SP	FPR	FNR	PWC	F-M	PR
Baseline	0.8032	0.9887	0.0113	0.1968	1.9858	0.7566	0.7192
Dynamic Background	0.9097	0.9889	0.0111	0.0903	1.2024	0.6128	0.5685
Intermittent object motion	0.5594	0.9614	0.0386	0.4406	6.7657	0.5136	0.5408
Shadow	0.8554	0.9767	0.0233	0.1446	2.8145	0.7079	0.6123

Thermal	0.7236	0.9937	0.0063	0.2764	2.2947	0.7515	0.8500
Camera Jitter	0.6319	0.9838	0.0162	0.3681	2.5390	0.5651	0.5110
Overall	0.7472	0.9822	0.0178	0.2528	2.9337	0.6512	0.6337

Table 11 – RBVOS [15] results

2.4 Comparative results

In the light of the results included in **Error! Reference source not found.**, **Error! Reference source not found.** and **Error! Reference source not found.** we observe that the overall operation of the evaluated algorithms takes place in an F-score range between 0.6 and 0.8. However, their performances among the video categories present remarkable differences.

The *baseline* category is supposed to be the simplest one. There is no dynamism in the background and the camera is fully static during the whole recording. However, there are two factors that slightly worsen the operation of two algorithms: Gamma [9] and BAYES [10]. Foreground objects in two of the sequences remain static for a long time, and these two algorithms do not handle correctly this situation, including them into the background model at some frames (see the False Negative Rate graph of **Error! Reference source not found.**). This complexity factor (CH4) harms the rest of their statistics. SGMM-SOD [7] performs the best in this category in terms of F-Score measure. PSP-MRF [12] and SOBS-SC [13] also obtain excellent results (**Error! Reference source not found.**). RBVOS [15] clearly improve the results of [10]. However, its overall performance in this category is harmed by the pre-computed results. This behaviour is repeated in all categories except for the Jitter, this exception is discussed below.

The *dynamic background* category is one of the most challenging one. However, analysed algorithms present adequate results, according to the complexity of the scenarios. The Gamma [9] was not explicitly designed to be robust to background dynamics (CH2), but still performs better than PABS [11], that was supposed to adequately handle this challenge. They both fail in the adequate classification of the dynamic parts of the background (check the Percentage of Wrong Classifications bar graph in **Error! Reference source not found.**). Furthermore, by comparing recall and precision results in **Error! Reference source not found.**, we see that most of the algorithms that achieve the best results in recall, significantly operate worse (5-10%) in precision terms. In our opinion, this is one of the most relevant unresolved problems in VOS. A system that provides a flexible solution to model the background including in the model its non-stationary parts would be less accurate in the foreground detection mainly due to this flexibility. As a consequence, the frame-to-model comparison is less demanding and some foreground samples might be incorrectly used to feed and update the background model. The Chebyshev [8] is the algorithm that best faces this problem, and consequently the one that obtains the best results in this category (**Error! Reference source not found.**).

The *intermittent object motion* category includes videos where objects remain static for a long time and scenarios where background is uncovered after a high amount of frames. These complexity factors (CH2 and CH4) severely harm the operation of those algorithms that do not explicitly include techniques to overcome them: Chebyshev [8] and Gamma [9] (see **Error! Reference source not found.**). The same problem in the trade-off between background modelling flexibility and foreground discriminability arises in this category, especially in the operation of the BAYES [10] and the PABS [11] algorithms. The SGMM-SOD [7] is the most equilibrate, in overall, in this category. Observe that the operation of RBVOS [15] does not allow the recovery of areas which were completely wrong labelled by the feeding algorithm (BAYES [10]).

According to the *shadow* category (shadows are also slightly present at the Baseline category), algorithms not considering them (mainly Gamma [9], but also BAYES [10] and PABS [11]), obtain the worst scores in this category; see how the shadows penalize these algorithms in the

Percentage of Wrong Classifications graph of **Error! Reference source not found.** Furthermore, it is noticeable the improvement in the results produced by post-processing [10] with the robust-to-illumination RBVOS [15].

The *thermal* category is the most suitable to contain camouflage situations (CH5). The Gamma [9] algorithm is the only one that obtains recall rates over 0.7 in these scenarios. This can be mainly explained by its neighbouring analysis strategy and by the nature of the features used during its analysis. Camouflage is by far, as declared by the dataset designers, the less studied complexity factor. Observe, for instance, that according to the feature of analysis, there is low diversity among the evaluated systems. Almost all use the RGB colour vector (or even just the luminance) of the pixel; just one of them (SOBS-SC [13]) moves to a robust to illumination colour space; only one includes the gradient as an additional feature (PABS [11]) and just one related the feature value with its surroundings (RBVOS [15]).

The *camera jitter* category seemed to be one of the most difficult, as camera is moving by an unpredictable wind. This complexity entails disparity of operation among the algorithms (observe **Error! Reference source not found.**). RBVOS [15] decreases the performance of BAYES [10] in this scenario. This is mainly due to its regional strategy. Where BAYES misclassifies pixels, RBVOS misclassified associated regions, then increasing the quantity of wrong classified pixels. On the contrary, the performance of the other regional based approach (PSP-MRF [12]) is the best for this scenario. Such difference in operation can be related to the pixel’s aggrupation nature. [12] segments the image in superpixels, while [15] uses regions.

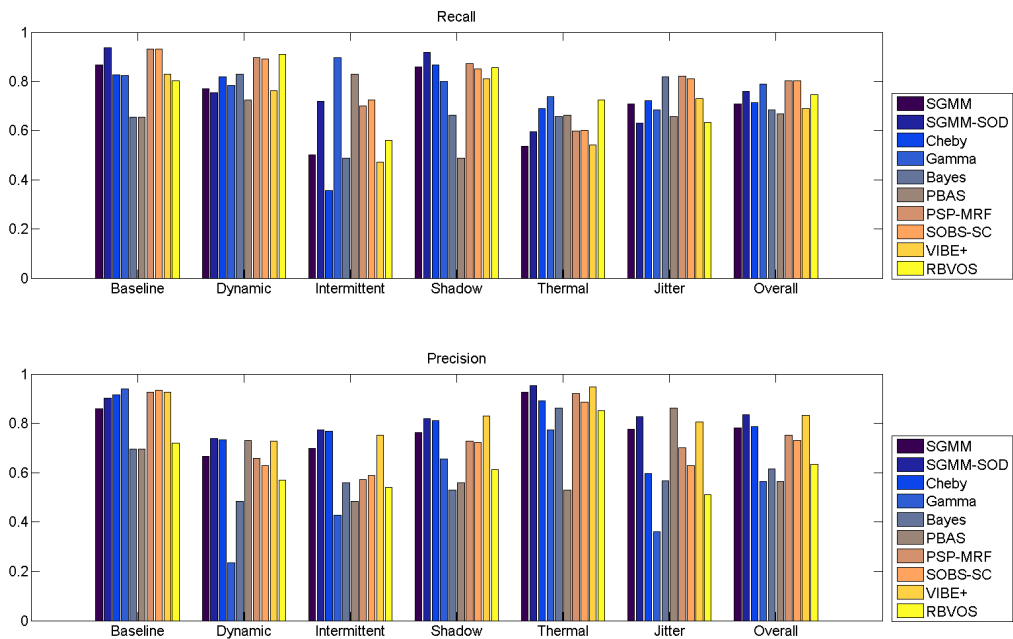


Figure 2 – Comparative results Recall and Precision

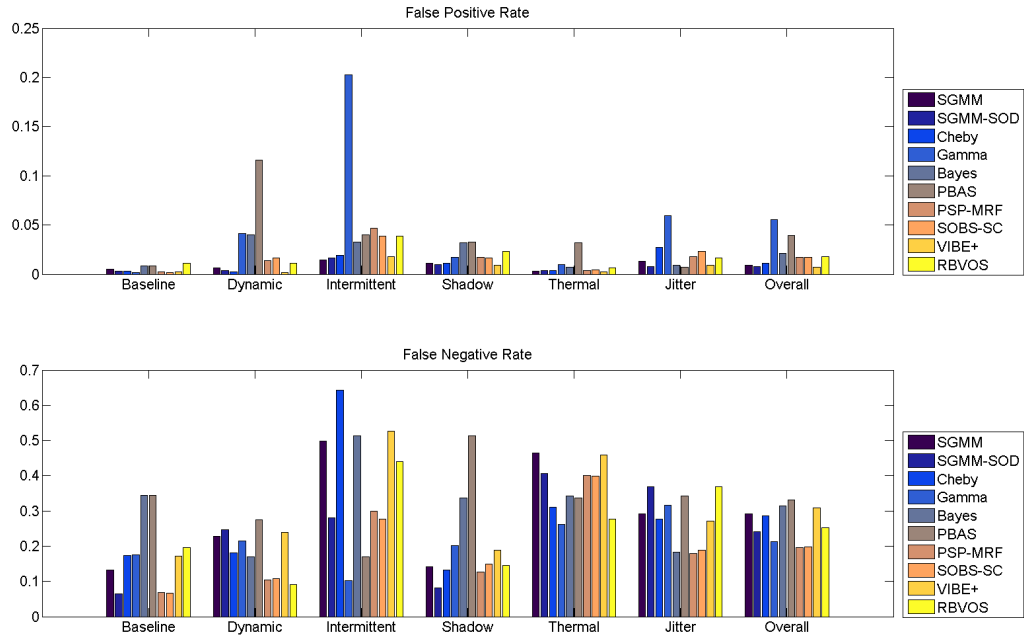


Figure 3 – Comparative results False Positive and False Negative Rate

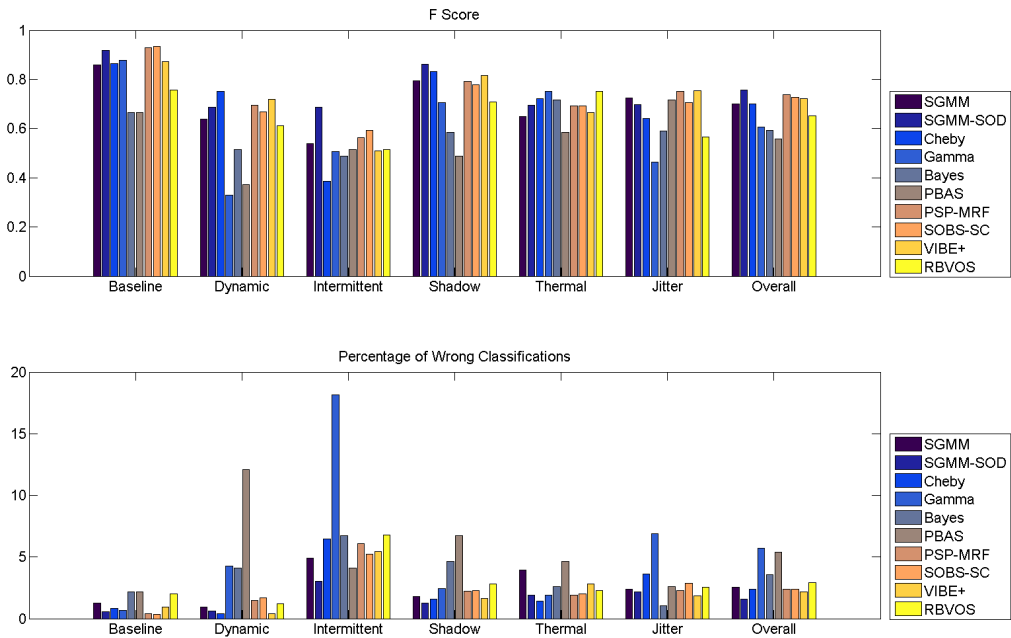


Figure 4 – Comparative results F-Score Measure and Percentage of Wrong Classifications

2.5 Conclusions

In recent years researchers have improved the classical VOS algorithms in terms of quality, exportability to different scenarios, efficiency and applicability. Existing challenges in VOS have been discovered, analysed and categorized, and both parametric and non-parametric algorithms perform well enough in their presence if they include specific techniques to face them. Nevertheless, there exist severe problems that, considering the influence of the VOS results in high level tasks as tracking or object and event recognition, need to be solved.

We have previously mentioned the need to confront the sensitivity-discriminability problem. The question is: can we adequately account for highly dynamic backgrounds without degrading the system performance in foreground discrimination and, inversely, are we capable of designing a system accurate enough to discriminate camouflaged foreground while maintaining its capability to adapt to changing backgrounds?

In our opinion, it is important to evaluate the limits of applicability of pixel level segregation. We think that the inclusion of accurate and robust high level modules may improve the results and would take complexity out of the models (as it is suggested by the results of regional approximations [12] and [15]). However, the use of this kind of post-processing modules increases the computational cost of the system while its operation is severely conditioned to their preliminary VOS stage.

Finally, it is necessary to remark that presented techniques might not work correctly in crowded scenarios (categories S3 and S4 of **Error! Reference source not found.**) as they were designed under the premise that background samples of each pixel are majority along the video. We need to rethink its applicability, take advantage of their scopes and overcome their limitations.

2.6 Future research lines

Considering the analysed results, the aforementioned conclusions and the existing problems, we propose three main lines of future research.

2.6.1 Refinement by post-processing techniques

Usually, some kind of post processing technique is used to improve VOS's results. Among them, one of the most common is refinement by morphological operations. These are used either to discard the classification as foreground of noisy pixels (erosion) or to fill small misclassifications due to camouflage (dilatation). The problem arises when these operations severely degrade the tightness of the segregation at the object's boundaries. In our opinion, techniques similar to the ones proposed at [12] and [15], which refine by using tight to foreground superpixels or regions, would better improve the quality of the obtained segregation masks without degrading them.

2.6.2 Use of alternative features.

Aiming to improve the operation of VOS systems in camouflage situations while maintaining their robustness in background modelling, we propose to follow the paved path of [12] and [15] and enhance the regional features with the pixel surrounding texture. Moreover, we aim to research alternative modelling strategies including deformable background models.

2.6.3 Include semantics in the descriptions

Nowadays, high level tasks should not rely in the output from VOS systems when analysing crowded scenarios. Obtained masks and per-pixel classification are inaccurate at the object's boundaries; results usually include several objects of interest joined at the same connected component. Although the use of feed-back strategies might improve the whole process path in

scenarios of categories S3 and S4, the definition of these strategies is complex and its use entails the inclusion of heavy computational looping modules. Alternatively, the use of higher-semantic features directly in the segregation process (such as person-not person classification, floor or ceiling location, dynamic backgrounds isolation, etc.,) may provide a simpler and more efficient scheme to operate in these scenarios. The segments (either regions or superpixels) may provide a robust entity to include this new knowledge in the analysis modules.

3 Segmentation in moving camera scenarios

3.1 Introduction

In sequences with camera motion, the static background can no longer be inferred from the unchanging frame regions, which renders fixed-camera VOS algorithms "per se" useless even in the simplest scenarios. Nevertheless, changes in the frames induced by camera motion can usually be described by compact parametric models. This aspect is exploited by most moving camera VOS algorithms, which start estimating the camera motion by solving a regression problem over the motion data (spatio-temporal derivatives, frame correspondences, motion vectors...). Then, the estimated parameters can be used to find the moving objects by: (a) compensating the camera motion and applying some adapted fixed camera segmentation technique, (b) identifying frame regions with significant deviation from the camera motion in the frame optical flow, or (c) some combination of both approaches.

Improvements on segmentation algorithms mostly focus on the proper segmentation stage, solving camera motion estimation (CME) with just standard well known techniques and rarely questioning the results. But the simplicity of the usual procedures to compute this camera motion and the compactness of its representation can be misleading. Actually, the proper camera parameters are not always easy to obtain and the derived values will probably misrepresent the real camera motion unless the assumptions implicitly made by the employed techniques are conveniently met in the scene. Worst of all, wrongly estimated camera motion parameters can have a devastating effect in segmentation: unlike any of the challenges discussed in section 2.2 (i.e. shadows, dynamic background), which mostly affect the involved frame areas, problems in CME can ruin the whole segmentation result.

This strongly suggests focusing our evaluation in challenging situations for the CME stage that, if improperly handled, will prevent the recovery of the camera motion parameters, eventually impairing the performance of any segmentation algorithm. The influence of additional complexity factors (independent from camera motion) on segmentation is already being studied in the fixed camera section, from where general results can be extended in most cases.

3.2 Selected evaluation scenario

In the design of this evaluation scenario we have analyzed which elements might be particularly challenging for the CME techniques typically used in segmentation algorithms. Most common techniques are based on the optimization of robust estimators (most times, M-Estimators) starting from (non-robust) Least-Squares solutions. This is known to properly handle outliers from small objects, but large ones can seriously degrade the result. In particular, motion from large dominating objects can be mistaken for the real camera motion (which may be an understandable result, but not the one most segmentation algorithms expect). On the other hand, large yet non-dominant objects -which are even more common- can also cause problems, making the estimators yield bridging fits: trade-off estimates averaging the motion from several structures and/or gross outliers).

Large objects are not rare in every-day video sequences. However, CME algorithms are not typically evaluated in this scenario, which makes it difficult to find public data-sets with sequences specifically involving camera motion and large objects. Let alone complementary ground-truth data in the form of parameters (which is unusual, except in cases of synthetically generated sequences) or objects masks (necessary to evaluate the parameters via the motion compensation error in the background areas).

Therefore, we have compiled a small dataset of 5 sequences with its object masks specifically for this evaluation. These sequences were cut from different videos from the MPEG-7 data set [16] and the object masks obtained at every frame by manual segmentation. One of the advantages of this dataset is that it includes sequences from several genres, which favors the representativity of the results.

Despite the small number of sequences considered, they still cover a number of different situations which may affect the performance of CME algorithms. This includes cases of isolated and simultaneous objects, similar and different concurrent motions, and various levels of homogeneity in the background. Intuitively, situations where the camera/background motion becomes virtually smaller than other motions in the frame will make CME harder. This may occur when several simultaneous objects have similar motions (which are likely to be absorbed into a single larger motion) or when the background has large homogeneous areas. Table 12 summarizes the characteristics of the 5 sequences in our test set and Figure 5 shows some example frames from each of them.

Seq. ID	Number frames	Simultaneous objects	Similar motions	Background homogeneity
1	188	No	-	Medium
2	57	No	-	Low
3	114	No	-	High
4	75	Yes	Yes	Medium
5	143	Yes	No	Medium

Table 12 – Characteristics of the sequences in our test-set



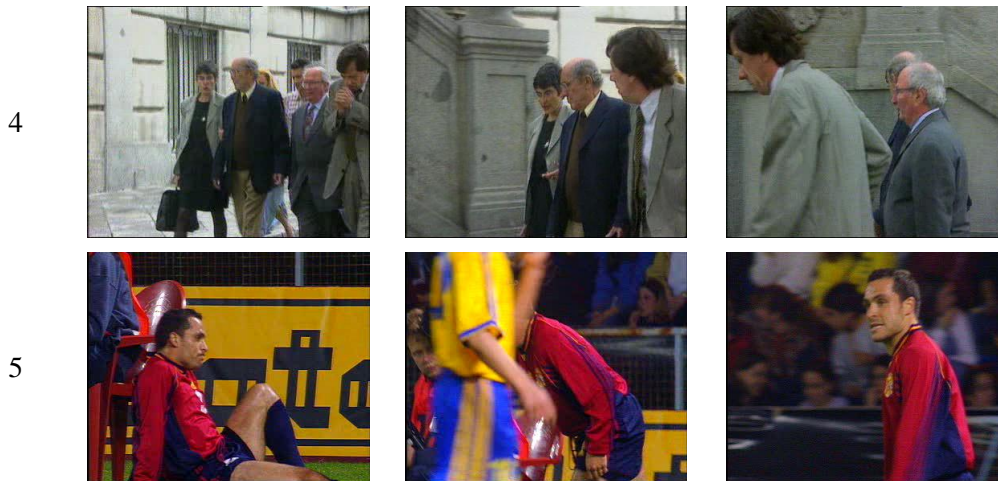


Figure 5 –Sample frames of the sequences in our test-set

The quality of the camera motion parameters will be assessed via the mean absolute error (MAE) between the current frame and the motion compensated previous frame evaluated just in the background regions of the current frame. Note that the segmentation-based metrics presented in section 2.2 will not be used. This is because CME -which does not derive the segmentation masks required by these metrics- is being studied separately from segmentation, so also avoiding the influence of the many additional factors influencing segmentation quality.

The last consideration in our evaluation scenario is the temporally independent operation. This imposes a restriction to the evaluated CME algorithms: camera motion between two consecutive frames must not rely on any information derived in the past (as object masks or motion parameters). Whilst we agree that in real applications this information can benefit robustness and efficiency, it also makes more difficult to draw conclusions on the strengths of a particular algorithm: good performance in a particular frame may only be due to previous information derived in less challenging situations.

3.3 Algorithms

In this section we briefly describe the two algorithms that we have considered for evaluation. The first one is a GME technique which follows a standard optimization and interestingly, is oriented to segmentation. The second one is an innovative technique currently being developed in the VPULab intended to robustly extract camera motion even in presence of very large objects (not yet available in the literature; the results in this document come from our last version)

3.3.1 Robust Global Motion Estimation Oriented to Video Object Segmentation (RGME-VOS) [17]

RGME-VOS uses a classic hierarchical differential scheme to estimate the camera motion. Overall, a robust function of the motion compensation residuals is optimized via a Newton–Raphson technique at the different levels of a multiresolution pyramid. The robust function definition depends on whether object information from previous frames is available or not. However, because of the temporal constraint of our scenario, we just use the version proposed for the absence of temporal information (which is employed in the first frame in the original algorithm). This function exploits the fact that frames may contain objects via an outlier rejection mechanism based on the analysis of local neighborhoods in the error between the current frame and the motion-compensated previous frame.

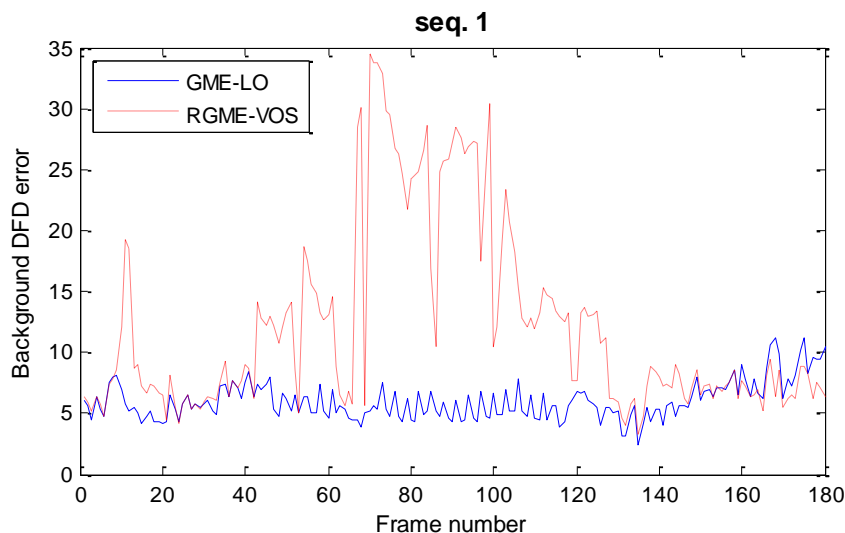
3.3.2 Robust Global Motion Estimation in presence of Large Objects (GME-LO)

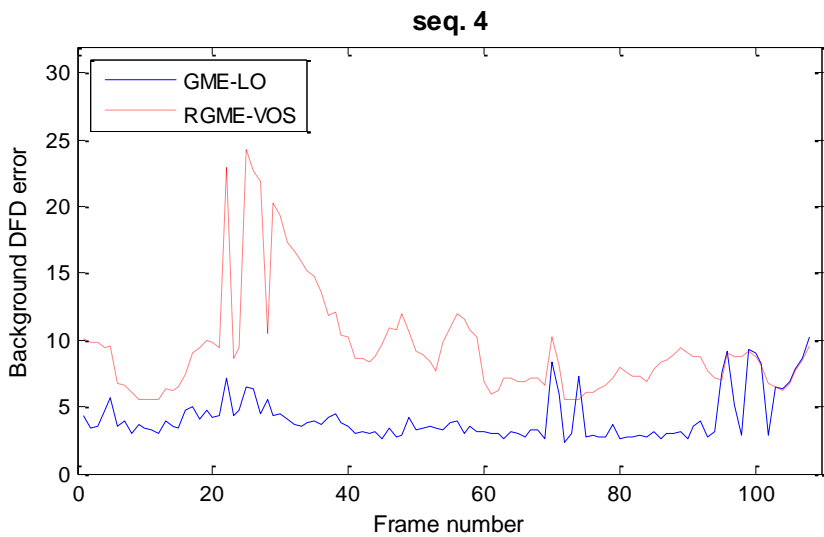
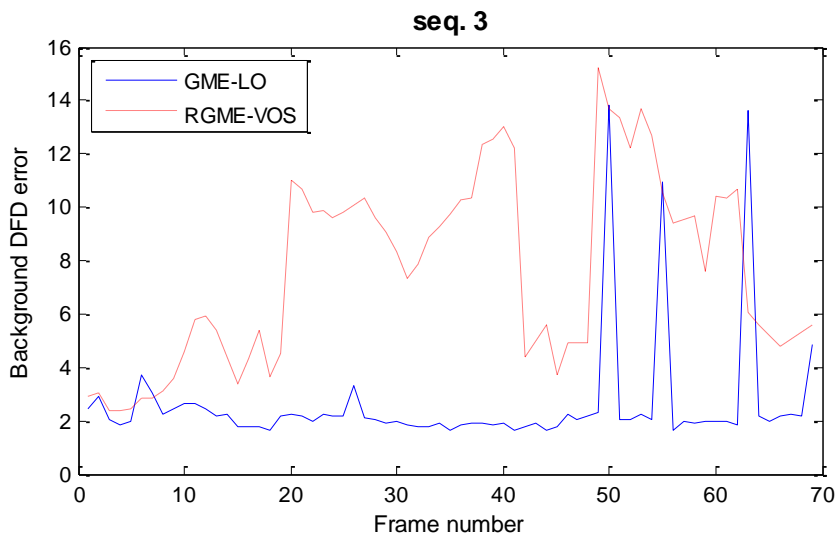
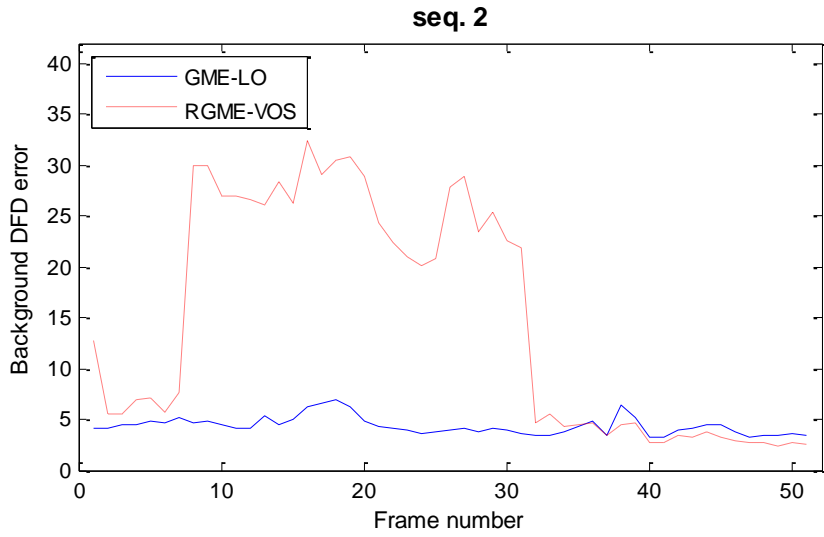
This algorithm is intended to robustly extract the camera motion even if large objects are present. For doing so, it uses specific tools and procedures to avoid: (a) mistaking the motion of large objects for the camera motion and (b) the convergence to bridging fits. The main elements of the algorithm are:

- A random sampling optimization scheme to bypass the initialization problem (which is convenient because initialization in presence of large objects is difficult and improper initialization is one the main causes of bridging fits)
- An effective objective function that:
 - o Can have minima for populations with only the relative majority of the data (i.e. it can find the camera motion even if the background size is smaller than 50 %).
 - o Does not need an a-priori scale as other estimators like RANSAC (which is convenient because the scale value is difficult to estimate in presence of large objects).
 - o Is inspired in non-parametric density power techniques that can tolerate very large amounts of outliers (in the order of 85 %) and avoid bridging fits
- A heuristic to tell the background from the objects that uses alternative criteria to the number of inliers. With this heuristic it can identify the camera motion even if some object is larger than the background.

3.4 Comparative results

In the different plots of Figure 6 we show the MAE error at every frame for all the sequences of the data set considered for evaluation. Frames where GME-LO more clearly outperforms RGME-VOS (yielding much lower errors) are precisely those where moving objects occupy a large percentage of the frame space. On the other hand, frames with similar performance either do not include large objects or these objects remain mostly static (which includes the case where motion is limited to some small parts).





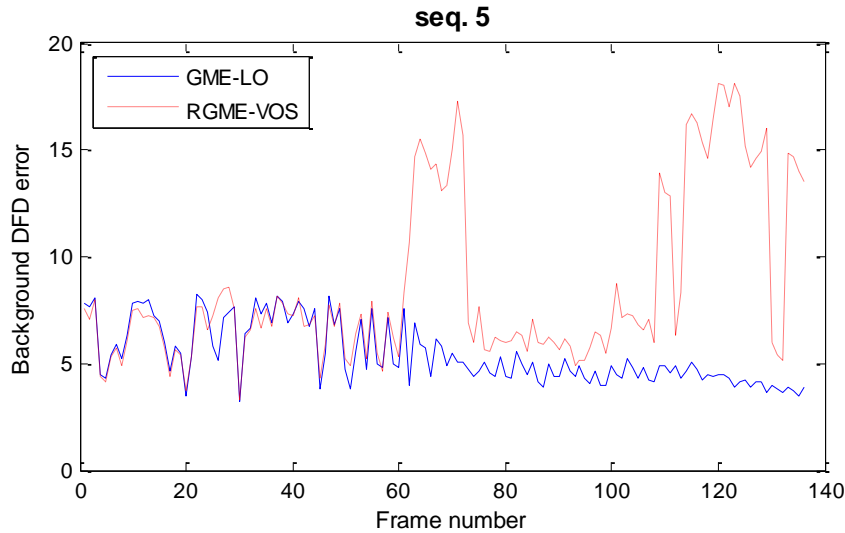


Figure 6 – CME accuracy (via the background MAE error) at frame resolution for the 5 sequences of our data set

3.5 Conclusions

Whilst commonly neglected, the accuracy of the CME stage can have tremendous influence in the whole segmentation result. Given this importance, as well as the many additional factors also influencing segmentation, it seems reasonable to isolate the evaluation of CME stage from the segmentation itself. In this evaluation we have found that situations involving large objects – which are completely normal in every-day videos– can make standard techniques used for CME fail. However, we should also note that, if the restriction of not using previously derived information were removed, these standard techniques would yield much better results than reported in this document. This could be seen, for instance, in the –perfectly natural– circumstance of large objects entering the scene slowly. Without the temporal restriction, the number of outliers that the involved robust estimator would have to recognize would be small, as outliers also existing in previous frames could be discarded beforehand. Nevertheless, it is extremely valuable to have CME techniques that can provide robustness to large objects even in absence of temporal information. These techniques, which will often be more computationally demanding, can always be used when temporal information is unavailable (e.g. initial frame) or becomes unreliable (e.g. after shot changes or when a previously static object starts to move).

4 People Modelling and Detection

4.1 Introduction

Automatic people detection in video sequences [18][19][20] is one of the most challenging problems in computer vision. The complexity of the people detection problem is mainly based on the difficulty of modelling persons because of their huge variability in physical appearances, articulated body parts, poses, movements, points of view and interactions among different people and objects. This complexity is even higher in typical real world surveillance scenarios such as airports, malls, etc., which often include multiple persons, multiple occlusions and background variability.

4.2 Selected evaluation scenario

The chosen experimental corpus, Person Detection dataset (PDds) [21], mainly excels other datasets in the amount of sequences (90 videos) and variability of sequences. It has been divided in two evaluation datasets. The first dataset, named A, has been selected to evaluate the different approaches at every complexity level, it includes the first 29 sequences from the experimental corpus. These sequences include the five different complexity categories depending on the described/defined people detection critical factors (from C1 to C5) or according to the previously described scenario classification (from S1 to S3), i.e, S1 includes the complexity categories C1 and C2, S2 includes the C3 and C4 and S3 includes the C5. The experimental dataset includes both non-rigid and rigid people/objects differing in size, motion and textural appearance. These people/objects are involved in a number of interactions and in different contexts, like typical every-day situations or surveillance video scenarios. Regarding the backgrounds, it includes in-door and out-door scenarios with different background complexities. The second dataset, named B, has been selected to evaluate more thoroughly the highest complexity category, i.e., only category C5 or scenario S3 (It is necessary to remark that after some preliminary experiments, the described techniques cannot work correctly in crowded and complex scenarios ,category S4 of **Error! Reference source not found.**, as they are designed. We need to rethink its applicability, take advantage of their scopes and overcome their limitations). It includes the following 61 sequences from the experimental corpus. The sequences have been extracted from the TRECVID 2008 dataset [22], namely, the ones for the surveillance TRECVID event detection task recorded at London Gatwick International Airport. This dataset contains highly crowded scenes, severely cluttered background, people at different scales and people completely static along the whole sequences. Due to the small size of the objects at the top of the image, during the annotation of sequences, the top 15% of the images has been discarded.

A summary of the complexity levels and scenario classification of both evaluation datasets is shown in Table 13.

Category	Scenario	#Sequences		Complexity	
		Dataset A	Dataset B	Classification	Background
C1	S1	6	0	Low	Low
C2	S1	6	0	Medium	Low
C3	S2	4	0	Medium	Medium
C4	S2	5	0	High	Low
C5	S2	4	0	High	Low
C5	S3	4	61	High	High

Table 13 – Sequences categorization people detection evaluation datasets.

In order to evaluate different people detection approaches, the evaluation methodology described in [1] has been followed.

4.3 Algorithms

In this section, we describe different approaches from the state of the art. We have selected seven diverse people detection approaches: Edge [23], Fusion [24], HOG [25], ISM [26], TUD [27], DTDP [28] and IMM [29].

4.3.1 Edge[23]

4.3.1.1 Algorithm overview

The Edge people detector is based on the body part-based algorithm proposed in [30], but proposing modifications in order to achieve real time performance in video surveillance scenarios. An individual human is modelled as an assembly of natural body parts. The main idea consists of identifying characteristic edges of each body part and generating four edge models of body parts (body, head, torso and legs). The object detection approach is a combination of segmentation and exhaustive search: the initial objects candidates to be person are extracted using background subtraction and then those selected candidates are scanned with four independent edge feature detectors previously trained.

4.3.1.2 Results

	Dataset A					Dataset B
AUC-PR (% Δ)	S1.C1	S1.C2	S2.C3	S2.C4	S3.C5	S3.C5
Edge	0.98	0.93	0.85	0.89	0.70	0.59

Table 14 – Edge results.

4.3.2 Fusion[24]

4.3.2.1 Algorithm overview

The Fusion people detector is a real time detection approach based on segmentation and a holistic person model. The initial objects candidates to be person are extracted using background subtraction and the holistic person model is the combination or fusion at decision level of three simple person models: ellipse fitting [31], ghost [32] and aspect ratio.

4.3.2.2 Results

	Dataset A					Dataset B
AUC-PR (% Δ)	S1.C1	S1.C2	S2.C3	S2.C4	S3.C5	S3.C5
Fusion	0.78	0.81	0.60	0.69	0.48	0.44

Table 15 – Fusion results.

4.3.3 HOG[25]

4.3.3.1 Algorithm overview

The HOG people detector is based on exhaustive search and a holistic person model. It consists in scanning the full image looking for similarities with the chosen person model, evaluating

different detection windows with a classifier at multiple scales and locations. The chosen person model is based on appearance information using the Histogram of Oriented Gradients. And the final decision is based on a previously trained SVM classifier.

4.3.3.2 Results

	Dataset A					Dataset B
AUC-PR (% Δ)	S1.C1	S1.C2	S2.C3	S2.C4	S3.C5	S3.C5
HOG	0.92	0.86	0.74	0.82	0.71	0.66

Table 16 –HOG results.

4.3.4 ISM[26]

4.3.4.1 Algorithm overview

The ISM is a generative model for object detection and has been applied to a variety of object categories including cars, motorbikes, animals and pedestrians. The ISM people detector is based on exhaustive search and a holistic person model. It consists in scanning the full image looking for similarities with the chosen person model at multiple scales and locations by local features matching. The chosen person model is based on appearance information using the SIFT features.

4.3.4.2 Results

	Dataset A					Dataset B
AUC-PR (% Δ)	S1.C1	S1.C2	S2.C3	S2.C4	S3.C5	S3.C5
ISM	0.95	0.91	0.80	0.84	0.71	0.69

Table 17 – ISM results.

4.3.5 TUD[27]

4.3.5.1 Algorithm overview

The TUD people detector is based on feature-based exhaustive search and a part-based person model. It is a part-based adaptation of the original ISM using pictorial structures. The appearance of body parts is modelled using densely sampled shape context descriptors and discriminatively trained AdaBoost classifiers. As a result, it presents a strong discriminatively trained appearance model and a flexible kinematic tree prior on the configurations of body parts.

4.3.5.2 Results

	Dataset A					Dataset B
AUC-PR (% Δ)	S1.C1	S1.C2	S2.C3	S2.C4	S3.C5	S3.C5
TUD	0.93	0.88	0.75	0.84	0.67	0.56

Table 18 – TUD results.

4.3.6 DTDP[28]

4.3.6.1 Algorithm overview

The DTDP people detector is based on scanning-based exhaustive search and a part-based person model. It is a part-based adaptation of the original HOG detector. It proposes a object detection system based on mixtures of multiscale deformable part models where each deformable body part is modelled as the original HOG detector.

4.3.6.2 Results

	Dataset A					Dataset B
AUC-PR (% Δ)	S1.C1	S1.C2	S2.C3	S2.C4	S3.C5	S3.C5
DTDP	0.96	0.92	0.81	0.86	0.74	0.68

Table 19 – DTDP results.

4.3.7 IMM [29]

4.3.7.1 Algorithm overview

The IMM people detector is based on feature-based exhaustive. The chosen person model is based in the characteristic movements of people using the Implicit Shape Model (ISM) Framework and the MoSIFT interest points detector and descriptor. It consists in scanning the full image looking for similarities with the chosen person model at multiple scales and locations by local features matching. The chosen person model is based on motion information using the MoSIFT features.

4.3.7.2 Results

	Dataset B with motion
AUC-PR (% Δ)	S3.C5
IMM	0.60
Edge+IMM	0.62
Fusion+IMM	0.49
HOG+IMM	0.68
ISM+IMM	0.67
TUD+IMM	0.62
DTDP+IMM	0.70

Table 20 –IMM results and appearance and motion combinations.

4.4 Comparative results

In this section, we describe the experiments performed over the experimental dataset and including different approaches that cover all the people detection issues identified from the state of the art. As we have already commented, we have selected seven diverse people detection approaches: Edge, Fusion, HOG, ISM, TUD, DTDP and IMM. According to the chosen object detection approach, Edge combines segmentation and exhaustive search, Fusion is based only on segmentation and the rest of them are based on exhaustive search. According to the chosen person model, the IMM includes the use of motion, appearance and their combination, the rest

of them are based only on appearance: holistic (Fusion, HOG, ISM) or part-based (Edge, TUD, DTDP).

	Object Detection		Person Model		
	Segmentation	Exhaustive Search	Motion	Appearance	
				Holistic	Part-based
Edge	✓	✓			✓
Fusion	✓			✓	
HOG		✓		✓	
ISM		✓		✓	
TUD		✓			✓
DTDP		✓			✓
IMM		✓	✓	✓	

Table 21 – Proposed People Detectors classification.

The Edge, Fusion and IMM results have been obtained with the original code, the HOG results have been obtained using the available binaries (<http://pascal.inrialpes.fr/soft/olt/>), the ISM results have been obtained using the available code and binaries (<http://www.vision.ee.ethz.ch/~bleibe/index.html>), the TUD results have been obtained using the available code (http://www.d2.mpi-inf.mpg.de/andriluka_cvpr09) and the DTDP results have been obtained using the available code (<http://www.cs.brown.edu/~pff/latent/>).

Despite the fact that all algorithms performance depends on the hit rate, or confidence level of the decision, we only classify objects detected in previous stages as person or non-person. Consequently, the maximum/minimum Recall and Precision will be limited by previous stages. Edge and Fusion are mainly limited by the segmentation step. Moreover, HOG, ISM, TUD, DTDP and IMM, are limited by the image scanning.

4.4.1 Evaluation dataset A

Firstly, we evaluate and compare the appearance based approaches at every complexity level using the evaluation dataset A. Figure 7 shows the averaged detection performance in terms of Recall vs. (1-Precision) curves and Table 22 shows the results in terms of AUC-PR, in both cases the results are for each complexity category included within the used video dataset A.

The results show clearly that all algorithms perform worse at higher complexity categories (from C1 to C5). However, it is observed that all approaches obtain generally worse results at category C3 than at category C4, due to the great influence of the background complexity in category C3 and thus the generation or extraction of the initial object hypotheses or candidates to be a person in the scene is more difficult. On the other side, the complexity of the category C4 lies on the classification of those initial candidates.

The Fusion approach gets the worst results. The use of segmentation makes easier the classification stage, allowing the approach to reach high recall results, but the use of such a simplified person model and all the segmentation problems (under/over segmentation) reduce the global precision rate. The Edge approach gets good results in all complexity categories and similar to the other approaches not based on segmentation. It is due to the use of a more complex person model and the combination of segmentation and exhaustive search. Despite the fact that the combination of segmentation and exhaustive search reduces the segmentation problems, these problems are magnified in complex background scenarios (C3-C5) where it is quite difficult to obtain a reliable segmentation.

The exhaustive search approaches are more robust to scale and pose variations and therefore more reliable in complex environments than those based on segmentation. Even so, the background complexity still has a negative impact in the results (C3). Moreover, unlike the

previous case, the classification stage is not simplified, it is even more complex because the approach must deal with a great number of negative examples (potential false positive detections), reducing the recall rate in order to maintain the precision rate. The HOG and TUD approaches show similar results in all complexity categories but the ISM and DTDP get better results. The ISM is a holistic approach but with a great flexible person model based on spatial feature probability distribution, and the DTDP is a body part-based variation of the HOG approach.

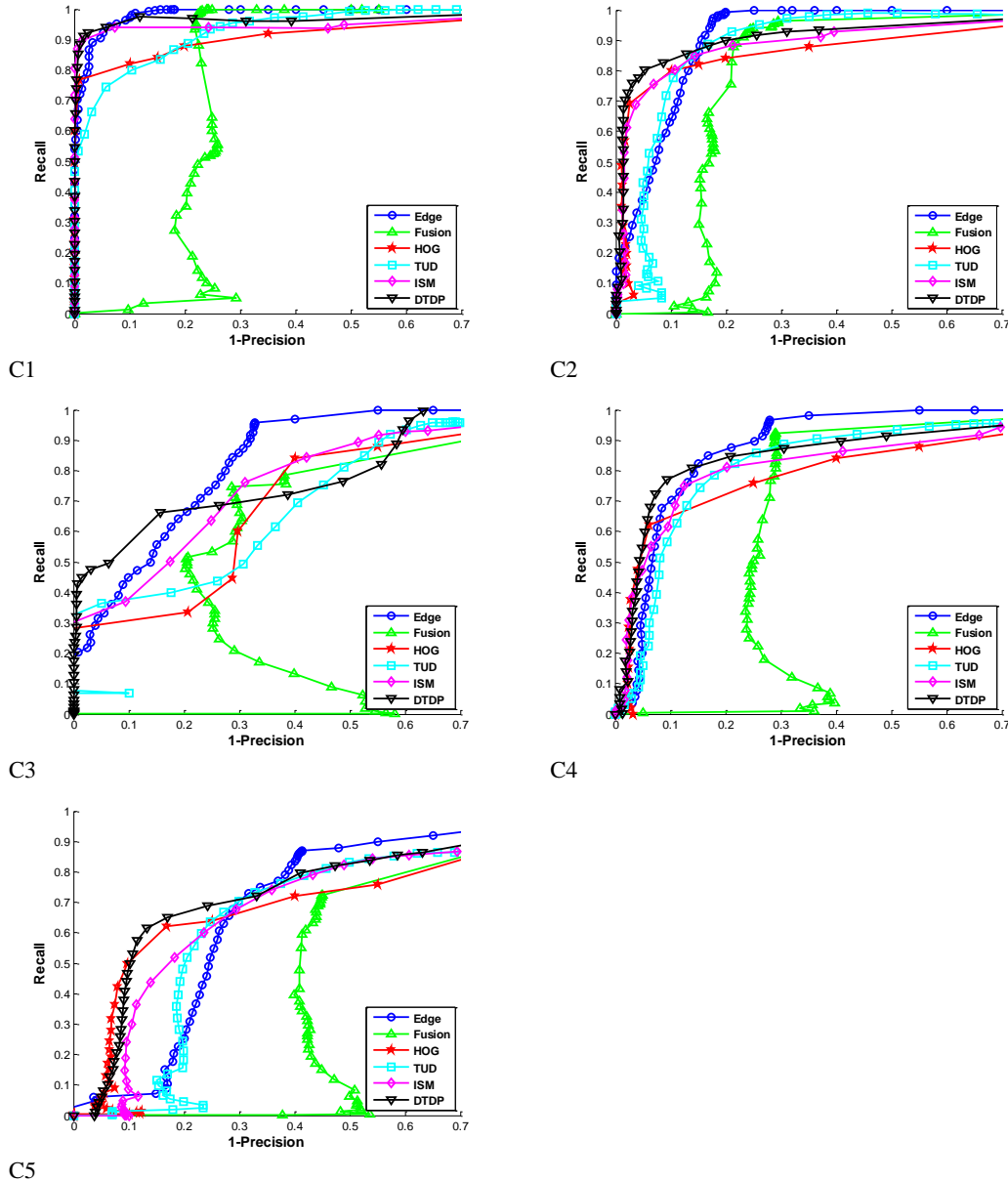


Figure 7 – Precision-Recall results per complexity category of people detection dataset A.

AUC-PR (%Δ)	Edge	Fusion	HOG	ISM	TUD	DTDP	IMM
S1.C1	0.98	0.78(-20)	0.92(-6)	0.95(-3)	0.93(-5)	0.96(-2)	-
S1.C2	0.93	0.81(-13)	0.86(-8)	0.91(-2)	0.88(-5)	0.92(-1)	-
S2.C3	0.85	0.60(-29)	0.74(-13)	0.80(-6)	0.75(-12)	0.81(-5)	-
S2.C4	0.89	0.69(-22)	0.82(-8)	0.84(-6)	0.84(-6)	0.86(-3)	-
S3.C5	0.70(-5)	0.48(-35)	0.71(-4)	0.71(-4)	0.67(-9)	0.74	-

Table 22 – Area under the Precision-Recall curve (AUC-PR) average for each complexity category of evaluation dataset A. Percentage increase (%Δ) calculated with respect to the best result for each complexity category. (IMM results do not apply).

4.4.2 Evaluation dataset B

In this section, we evaluate more thoroughly the highest complexity category (C5) using the dataset B. Table 23 shows the results in terms of AUC-PR of dataset B. Due to the greater complexity of the sequences extracted from TRECVID (the content set contains challenging scenarios, crowds and a wide range of scales), the results are worse than those obtained in the dataset A.

In this case, due to the higher sequences complexity, all the approaches get worse results than with the dataset A. Both approaches based on segmentation, the Edge and Fusion, obtain worse results than the other approaches from the state of the art. As already commented, the main problem of these approaches is the difficulty of making a reliable segmentation (foreground/background) in complex scenarios. However, the sequences extracted from TRECVID present an additional difficulty to both approaches: the sequences include people completely static along the whole sequences. Both approaches extract the objects candidates to be a person using motion information (background subtraction), therefore it is not able to extract those static objects/people, reducing the Recall rate and therefore the overall performance.

The results also show that the approaches based on exhaustive search also get worse results than with dataset A. However, except the TUD approach, they are more stable in more complex scenarios because they are more robust to scale and pose variations and more robust to the background complexity.

AUC-PR (% Δ)	Edge	Fusion	HOG	ISM	TUD	DTDP	IMM
S3.C5	0.59(-13)	0.44(-35)	0.66(-3)	0.69(-1)	0.56(-18)	0.68	-

Table 23 – Area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B. Percentage increase (% Δ) calculated with respect to the best result. (IMM results do not apply).

4.4.3 Evaluation dataset B with motion

In this section, we evaluate the dataset B including the people detector based on motion IMM and all the appearance and motion combinations (Edge+IMM, Fusion+IMM, HOG+IMM, ISM+IMM, TUD+IMM and DTDP+IMM). In order to train the people motion model, the evaluation dataset B has been divided in training and test. To be homogeneous, the appearance based detectors approaches also have been evaluated on the same video sequences, the test dataset. As in the experiments in [29], the training dataset is composed of 25 sequences and the test dataset is composed of the other 36 sequences. Table 24 shows the results in terms of AUC-PR of test dataset.

The results show that the IMM approach gets good results in complex or realistic scenarios and comparable to the other approaches from state of the art. The IMM is based only on motion, so it is only able to detect moving people. For this reason, the IMM approach in general is able to get high precision rates but low recall rates. Even so, in environments as complex as these ones, the use of motion information obtains results close to the use of appearance information. The combination of appearance and motion information (Edge+IMM, Fusion+IMM, HOG+IMM, ISM+IMM, TUD+IMM and DTDP+IMM) improves the global results in all the cases. Thus, it is clear that human motion provides useful information for people detection and independent from appearance information.

AUC-PR (% Δ^1)	Edge	Fusion	HOG	ISM	TUD	DTDP	IMM
S3.C5	0.58(-13)	0.46(-31)	0.66(-1)	0.64(-4)	0.56(-16)	0.67	0.60(-10)

AUC-PR (% Δ^2)	Edge+IMM	Fusion+IMM	HOG+IMM	ISM+IMM	TUD+IMM	DTDP+IMM
S3.C5	0.62(+7)	0.49(+7)	0.68(+3)	0.67(+5)	0.62(+11)	0.70(+4)

Table 24 – Area under the Precision-Recall curve (AUC-PR) average of evaluation dataset B without and with motion information. Percentage increase (% Δ^1) calculated with respect to the best result or percentage increase (% Δ^2) calculated with respect to single appearance versions.

4.5 Conclusions

The experimental results over the evaluation dataset show the people detection problems in video sequences. According to the chosen object detection approach, the use of segmentation makes easier the classification stage. However, they must deal with all the segmentation problems (under/over segmentation). The combination of segmentation and exhaustive search reduces these problems but they are still a drawback especially in complex scenarios where these problems are magnified. The exhaustive search approaches are more reliable in complex environments than those based on segmentation. However, unlike the previous case, the classification task is not simplified, it is even more complex because the approach must deal with a great number of negative examples (potential false positive detections), reducing the recall rate in order to maintain the precision rate. According to the chosen person model, in general, the use of simplified person models gets worse results mainly in terms of Precision than those more complex person models. And finally, the motion information is less characteristic than the appearance of the people, but the combination of motion and appearance shows to be useful even in complex scenarios.

4.6 Future research lines

Based on the results and discussions of this document, we plan the following future research lines:

4.6.1 Expand the evaluation dataset PDds

Expand the evaluation dataset PDds. The proposed experimental dataset PDds includes a great variability of scenarios with different background complexities and it also includes a great variability of people appearance and multiple interactions with objects and/or persons. However, we propose to extend the contents of the dataset and make use of every sequence from the CVSG dataset[33] (recorded in a chroma studio and having the possibility of combining the foreground with different backgrounds), in order to be able to analyse independently the background and foreground factors.

4.6.2 Improve or refine segmentation

As noted in the experimental results, our combination of segmentation and exhaustive search reduces the segmentation problems (under/over segmentation), but these problems are magnified in complex scenarios where it is quite difficult to obtain a reliable segmentation. So, we propose the study of techniques for multimodal background modelling, noise removal, shadows detection, etc, in order to refine the background subtraction in complex scenarios.

4.6.3 Appearance and motion fusion

We propose the study of different fusion/combination techniques between the appearance and motion detectors to improve the Recall without compromising the Precision, or even the creation of a single integrated Implicit Shape-Motion Model (ISMM), using the full MoSIFT description.

5 Tracking

5.1 Introduction

Video tracking algorithms basically aim at identifying the candidate position (pixel coordinates) within a video frame where a target model is most likely to be present. The target model is usually a predefined region of interest (e.g., rectangle, circle, ellipse) either automatically or semi-automatically delimited within a given image. The different algorithms basically differ in the representation of the target model and the method applied in order to compare the representations of both that target model and the models corresponding to the possible candidate positions. The aforementioned methods typically take into account visual cues such as colour, as well as the history of candidate positions that have been found throughout the video sequence.

Multi-tracking algorithms have the same underlying principles as described above, although they support several target models for a same video sequence. A multi-tracking algorithm can trivially be obtained by running multiple independent instances of a single-target tracking algorithm, one per target model. However, algorithms specifically tailored to multiple tracking can be more effective by taking into account the possible interactions among target models, such as occlusions and groupings, which are the main challenges in video tracking.

5.2 Selected evaluation scenario

For video object tracking, the Single Object Video Tracking dataset - SOVTds (see [1] for further details), was selected, as it was created focusing on the main problems that affect video object tracking in surveillance videos. As a tracker can operate in different conditions in which the same problem appears, we propose to organize them into four situations ranging from completely controlled (e.g., synthetic sequences) to uncontrolled (e.g., real-world sequences). Moreover, the complexity of the tracking problems is estimated for each set of sequences:

1. Synthetic sequences (L1): It is composed of synthetic sequences that provide controlled testing conditions allowing to isolate each problem. They consist on a moving ellipse in a black background that can contain squares of the same or different colour (acting as, respectively, similar or occlude objects).
2. Laboratory sequences (L2): It provides a natural extension of the L1 situation by representing real test data in a laboratory setup under controlled conditions. An object with a simple colour pattern was used for generating such data.
3. Simple real sequences (L3): It includes data from previously existing datasets that have been captured in uncontrolled conditions. We have extracted clips from the original sequences that contain isolated tracking problems.
4. Complex real sequences (L4): The last situation contains the most complex sequences, which are clips from other datasets that include several problems. Once the algorithms are tested for each problem individually, it is a good idea to check the performance in more realistic (and complex) situations.

Each of the first three sets of sequences (L1, L2 and L3) are divided into seven subcategories. Each of these subcategories corresponds to the major problems of tracking: Complex movement, illumination local, illumination global, noise, occlusion, scale changes and similar objects. Also in L2 category there are three videos for each of the seven subcategories, corresponding to low, medium and high difficulty. The last set of sequences (L4) is divided into three subcategories: cars, faces and people.

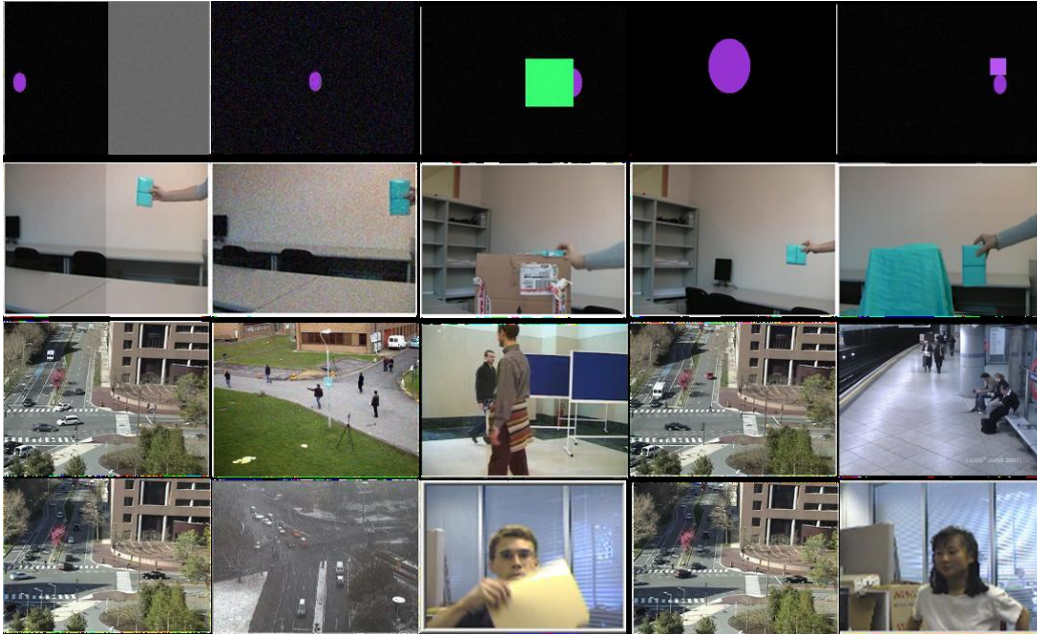


Figure 8 – Sample frames for the situations of the proposed dataset (from top row to bottom row): synthetic (L1), laboratory (L2), Simple real (L3) and Complex real (L4). In addition, samples of some tracking-related problems are also presented for each column (from left to right): abrupt illumination change, noise, occlusion, scale change and (colour-based) similar objects.

The following table shows the relation between the SOVTds categories and the proposed scenario classification in the EventVideo project:

Scenario	Complexity	Density	SOVTds sequences
S1	Low	Low	L1, L2, L3, L4(cars), L4(people)
S2	High	Low	L2(high), L3, L4(faces), L4(cars)
S3	Low	High	L3, L4(people)
S4	High	High	-

Table 25 – Relation between the SOVTds categories and the EventVideo categories.

In order to evaluate the accuracy selected tracking algorithms, the chosen metric was SFDA (Sequence Frame Detection Accuracy) which calculates for each frame the spatial overlap between the estimated target location and the ground-truth annotation (see [1] for further details). This metric has been selected because it evaluates the cumulative spatial accuracy (mean) of the algorithm for the complete sequences.

$$SFDA = \frac{\sum_{t=1}^{N_{frames}} FDA(t)}{\sum_{t=1}^{N_{frames}} \exists(N_{GT}^t \text{ OR } N_D^t)}$$

$$FDA(t) = \frac{OverlapRatio}{\frac{N_{GT}^t + N_D^t}{2}}$$

where N_{GT}^t and N_D^t represent the number of ground-truth and target annotations respectively in the t th frame.

5.3 Algorithms

5.3.1 Colour-based mean-shift (MS) [34]

5.3.1.1 Algorithm overview

This single-target tracking algorithm represents the target model by the colour histogram of all pixels belonging to the given elliptical region of interest to be tracked. That histogram is computed in such a way that pixels close to the target's centre have a larger weight than those away from it according to the Epanechnikov kernel function. This weighting is done in order to lower the influence of pixels close to the boundaries of the region of interest, which are assumed to be less confident than those close to the centre.

The candidate position within the current video frame is the one that maximizes the Bhattacharyya distance between its associated colour histogram, which is computed in the same manner as the histogram of the target model, and the latter. That candidate position is found by iterating from the previously known target position until convergence by applying the mean-shift procedure to an image of weights. The larger the weight corresponding to a certain image position the larger the similarity between the colour histograms associated with both that position and the previously known target position.

The algorithm can adapt to scale changes by slightly modifying the width of the Epanechnikov kernel function, thus slightly changing the area of the effective image region over which all histograms are computed. Three widths are considered: the previous width without changes and after both increasing it and decreasing it by 10%. The width that yields the maximum Bhattacharyya distance for the final candidate position is the one that denotes the change of scale.

A simple variation of the algorithm described above, referred to as background-weighted histogram (BWH), aims at reducing the interference of background pixels in the tracking process by taking into account the colour histogram of the background surrounding the target model in order to modulate the colour histograms associated with the target model and the candidate positions. In particular, when a bin from the background histogram has a significant value, the corresponding bins for the target model and the candidate positions are given a low weight. The background histogram is computed in a region three times bigger than the area of the target model.

5.3.1.2 Results

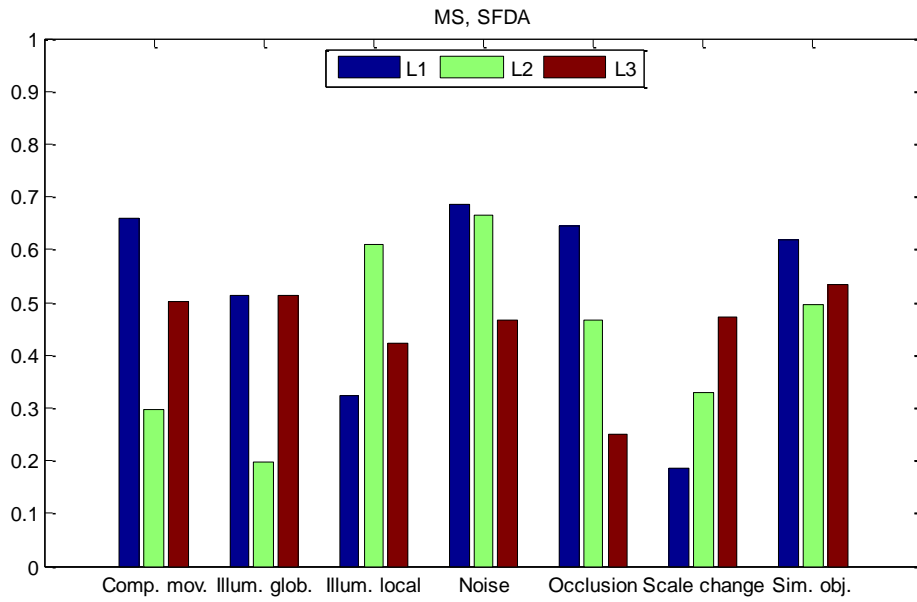


Figure 9 – MS SFDA for L1, L2 and L3 videos.

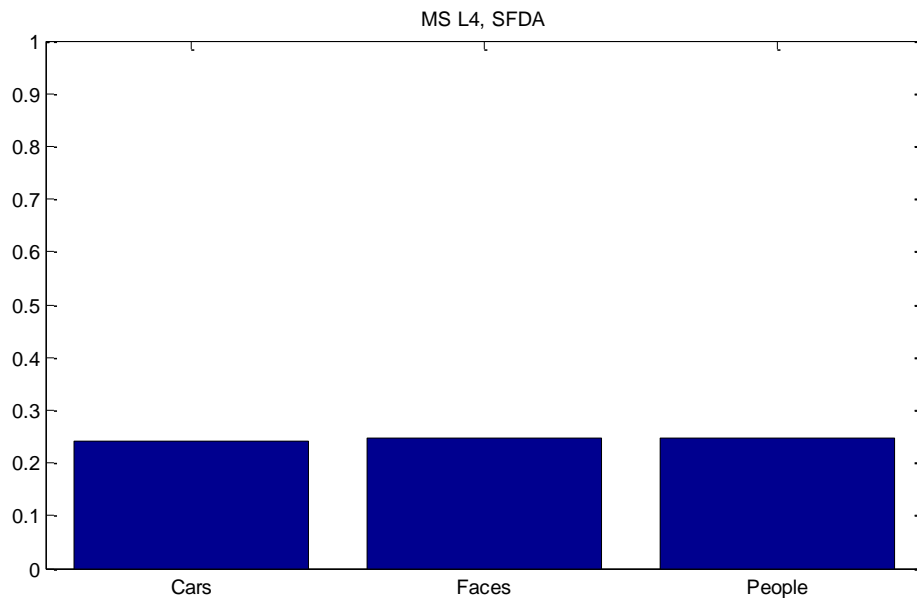


Figure 10 – MS SFDA for L4 videos.

MS algorithm does not stand out in any of the categories of the sequences L1 and L2. Its results are in the top 3 in most of the comparisons of the two cited sequences sets. For the L3 sequences, the results are slightly better than the results of the other algorithms for the categories complex movement and similar objects. In addition, MS gives better results for the scale change category.

5.3.2 Template matching (TM) [35]

5.3.2.1 Algorithm overview

This single-target tracking algorithm represents the target model by the subimage corresponding to the given rectangular region of interest to be tracked. The target model (template) is then searched over the current video frame by applying a convolution process in which the target model is the convolution mask. The candidate position is the location within the current frame with the largest convolution value. The convolution process can be replaced by other types of sum-comparing metrics, such as the sum of absolute differences (SAD), sum of square differences (SSD) and cross-correlation.

Due to its inherent simplicity, this algorithm can be directly implemented in hardware or by taking advantage of vector machine code instruction sets (MMX, SSE, ...), hence making it suitable for real time processing. Its main drawback is its only invariance to translation changes of the target model, which can be the case for targets moving relatively slowly between consecutive frames. Notwithstanding, due to its extremely high computational efficiency, several templates can be generated by applying small rotations and scale changes to the original target model. The algorithm can then be applied to the different templates, finding out the one that yields the best matching with the current frame. Similarly, in order to cope with occlusions, the original target model can be partitioned into several templates that can then be independently matched with the current frame.

5.3.2.2 Results

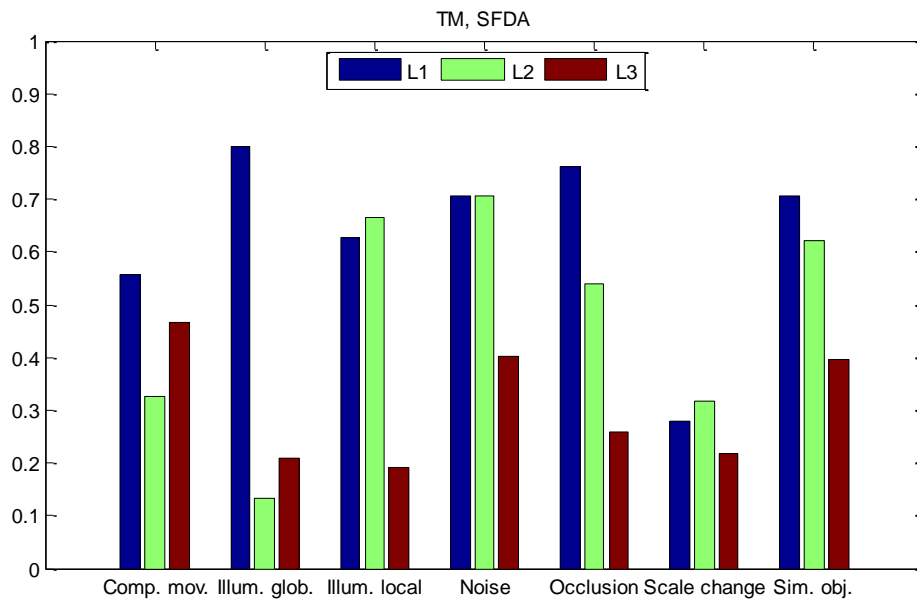


Figure 11 – TM SFDA for L1, L2 and L3 videos.

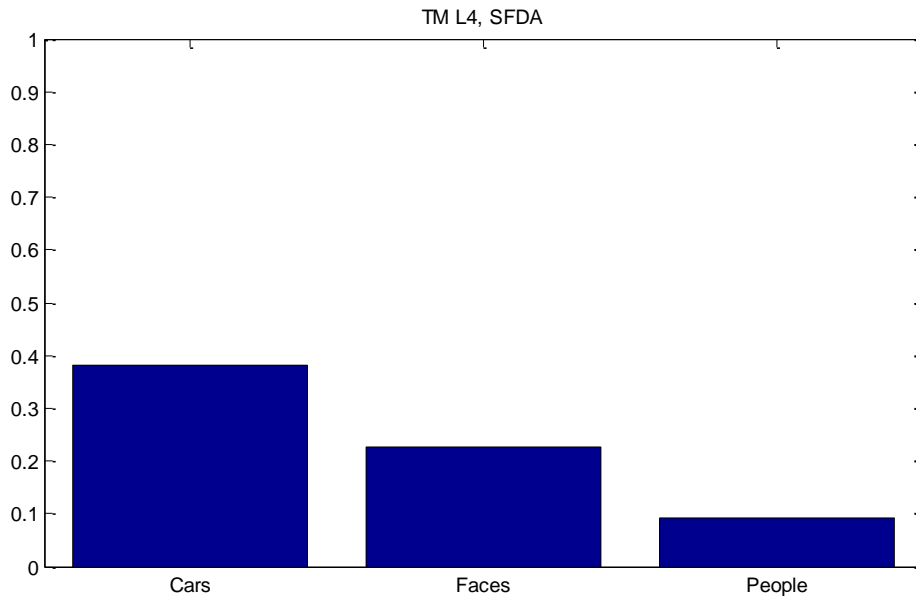


Figure 12 – TM SFDA for L4 videos.

Due to the relative simplicity of the TM algorithm, its best scores are obtained in the sequences belonging to L1 and L2, which correspond to the simplest tracking sequences. This algorithm does not consider the changes of scale, which is reflected in the decrease in performance for sequences of this category in all sets of sequences. For the L1 sequences, TM algorithm gets the highest score in illumination global, illumination local and occlusions categories, and the second best score in noise and similar objects categories.

5.3.3 Lucas-Kanade tracking (LK) [36]

5.3.3.1 Algorithm overview

This single-target tracking algorithm can be considered to be a generalization of the above template matching algorithm that allows for small affine transformations (translation, rotation, scaling, shear mapping, etc.) of the target model. In particular, the target model is represented by the subimage corresponding to the given rectangular region of interest to be tracked. The target model (template) is then searched over the current video frame by finding the parameters of the affine transformation that best aligns the transformed image with the target model. That search is cast as a minimization problem that is iteratively solved by applying gradient descent, starting with an initial estimation of the sought parameters. Since the variation between consecutive video frames is usually small, this initial estimation can simply be the values of the parameters corresponding to the target model in the previous frame or zeroes for the first frame.

5.3.3.2 Results

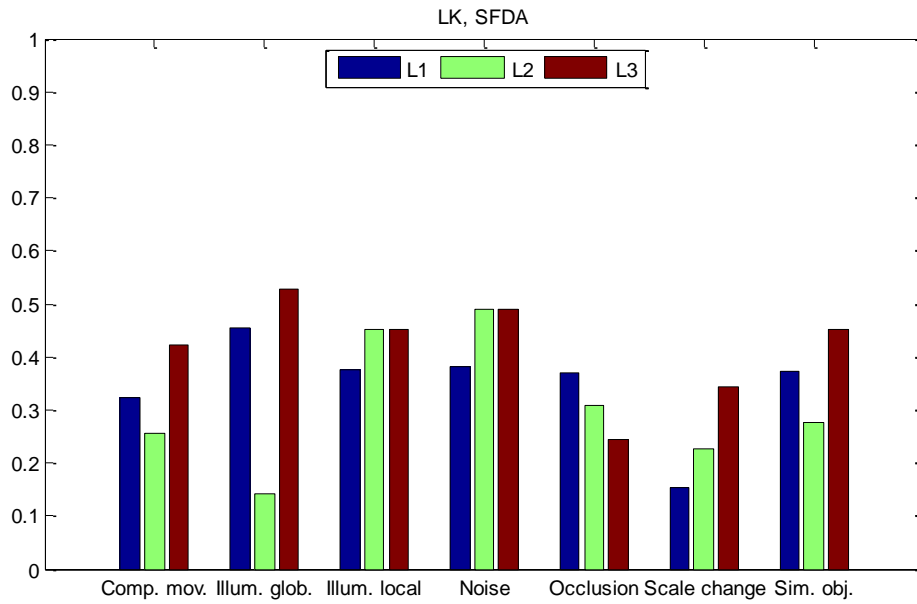


Figure 13 – LK SFDA for L1, L2 and L3 videos.

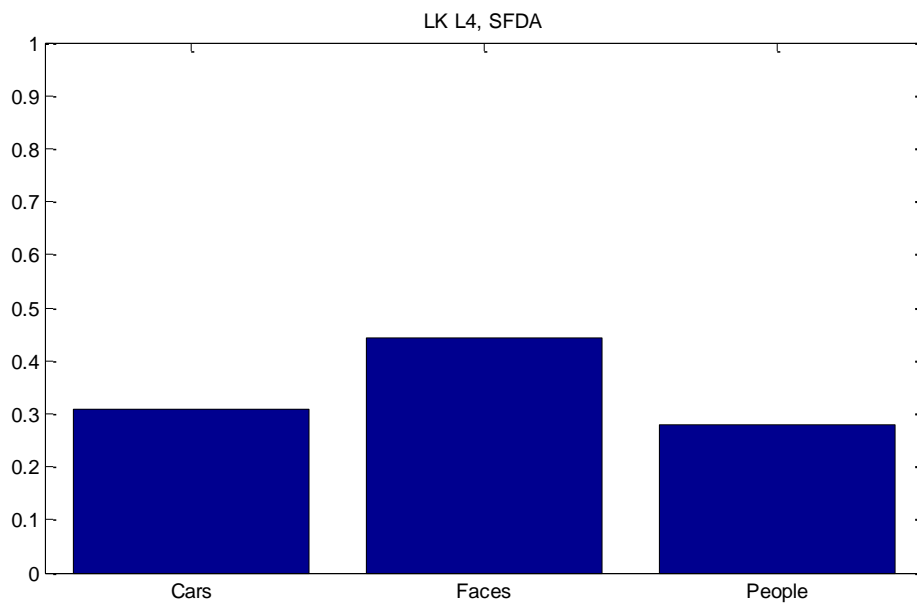


Figure 14 – LK SFDA for L4 videos.

The results obtained by the LK algorithm for the L1 sequences are far from the best results obtained in each of the seven categories. In the case of the L2 sequences, results in each category are the worst of all, except in the global illumination category. In contrast, for L3 and L4 sequences, the LK algorithm presents better results than the average.

5.3.4 Particle filter-based colour tracking (PFC) [37]

5.3.4.1 Algorithm overview

Similarly to the colour-based mean shift tracker summarized above, this single-target tracking algorithm represents the target model by the colour histogram of all pixels belonging to the given elliptical region of interest to be tracked. That histogram is also computed in such a way that pixels close to the target's centre have a larger weight than those away from it according to the Epanechnikov kernel function.

However, differently to the mean shift tracker, the candidate position of the target model in the current video frame is found as a weighted average of alternative candidate positions, each referred to as a particle. Every particle is represented by the position, size and the corresponding first derivatives of a 2D ellipse. The weight associated with each particle is computed according to the Bhattacharyya distance between the colour histograms of both the target model and the ellipse corresponding to that particle, such that the larger the distance, the larger the weight.

Every particle iteratively evolves at every time step by changing its position and size according to its corresponding first derivatives plus a random offset following a zero-mean Gaussian distribution. The derivatives are also changed by applying a random offset. Initially, all particles can be randomly distributed over the video frame in order to cover regions where the target is expected to appear or where an object detection algorithm determines. The iterative algorithm stops when the candidate position converges.

5.3.4.2 Results

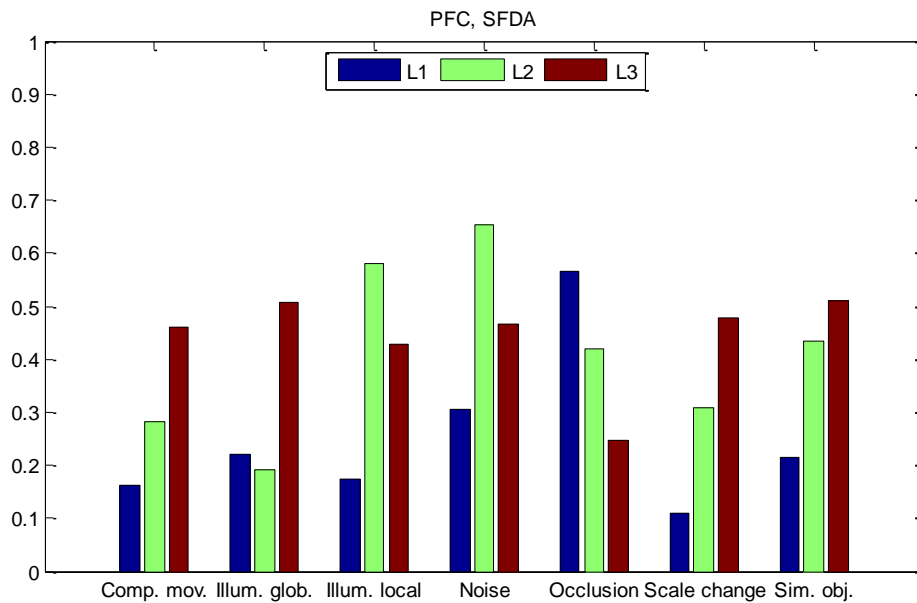


Figure 15 – PFC SFDA for L1, L2 and L3 videos.

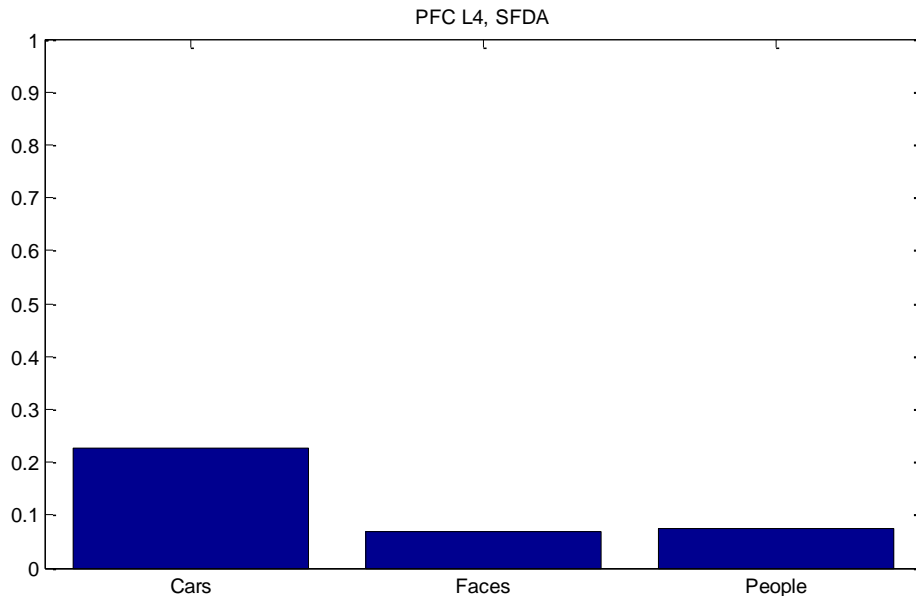


Figure 16 – PFC SFDA for L4 videos.

In the case of PFC algorithm, for the first sequences set (L1) the presented results are the worst of all the algorithms (excluding occlusions category). PFC algorithm does not work properly in synthetic sequences, due to uniform regions cause malfunctions in the particle filter. For the L2 and L3 sequences, medium-high results are obtained.

5.3.5 Corrected background colour-based mean-shift tracker (CBWH) [38]

5.3.5.1 Algorithm overview

This single-target tracking algorithm is a variation of the original colour-based mean-shift technique [34] that modifies the stage that reduces the interference of background pixels, originally referred to as background-weighted histogram (BWH). In particular, the proposed algorithm, referred to as corrected background-weighted histogram (CBWH) only transforms the histogram of the target model, but not the histograms of the candidate positions, thus decreasing the probability of target model features that are prominent in the background. Experimental results show that CBWH can reduce the number of mean-shift iterations, as well as improve the tracking accuracy. One of its main advantages is that it reduces the sensitivity of mean-shift tracking to the target initialization. Therefore, CBWH can robustly track the target even if it is not initialized precisely.

5.3.5.2 Results

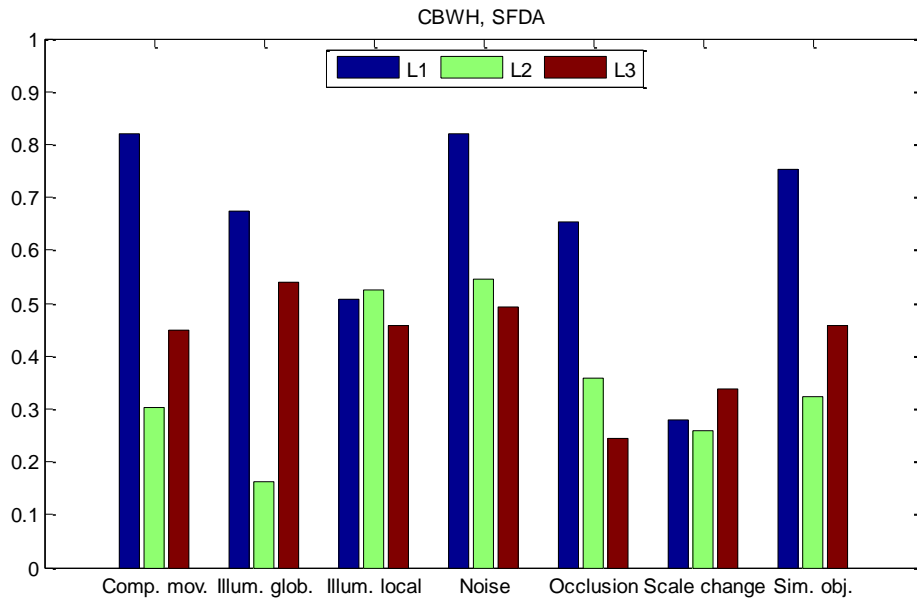


Figure 17 – CBWH SFDA for L1, L2 and L3 videos.

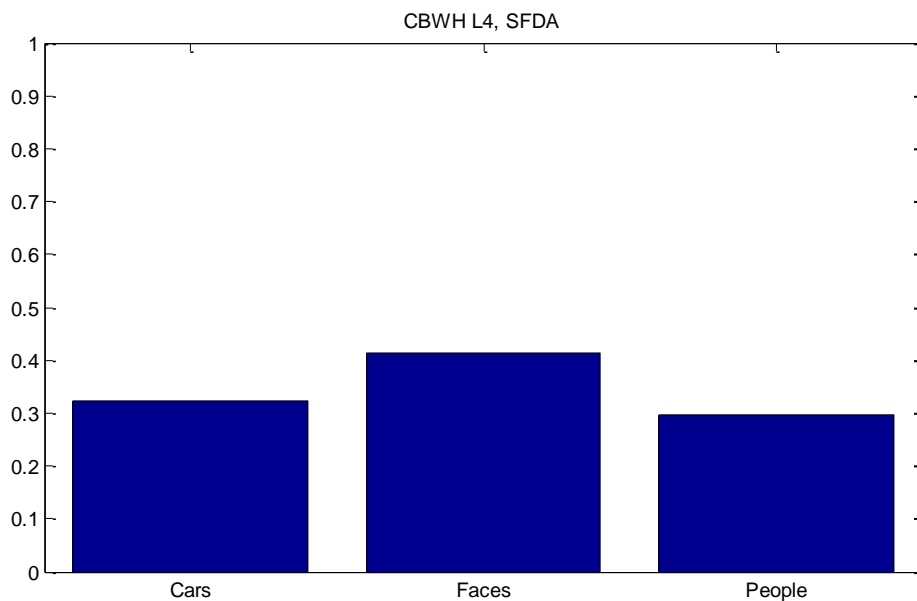


Figure 18 – CBWH SFDA for L4 videos.

The algorithm CBWH presents its best performance for L1 sequences, which gets the highest score in the complex movement, noise and similar objects categories, and the second best score in the global illumination, local illumination and occlusions categories in comparison with the other algorithms. This algorithm studies the background and reduces its influence on the tracked object. For these synthetic sequences, background subtraction is performed with greater precision due to foreground and background are more differentiated, which facilitates its discrimination. However, for L2, CBWH does not present good results in comparison with other algorithms as in several categories has the second lowest score.

5.3.6 Scale and orientation adaptive mean-shift tracking (SOAMST) [39]

5.3.6.1 Algorithm overview

This single-target tracking algorithm is a variation of the original colour-based mean-shift technique [34] that is able to update the scale and orientation of the target model during the tracking process. The original mean-shift tracker only supports discrete changes in the scale of the target model. In the proposed variation, the image of weights generated by the original mean-shift tracker, in which a pixel has a large weight if the colour histogram associated with that candidate position is similar to the histogram of the target model, is utilized to estimate the area and orientation of the target. In particular, the zero-th-order moment of the image of weights is utilized to estimate the area of the target model, and hence its scale, whereas the width, height and orientation changes of the target are estimated using the area estimated before, as well as the second-order centre moment of the image of weights.

5.3.6.2 Results

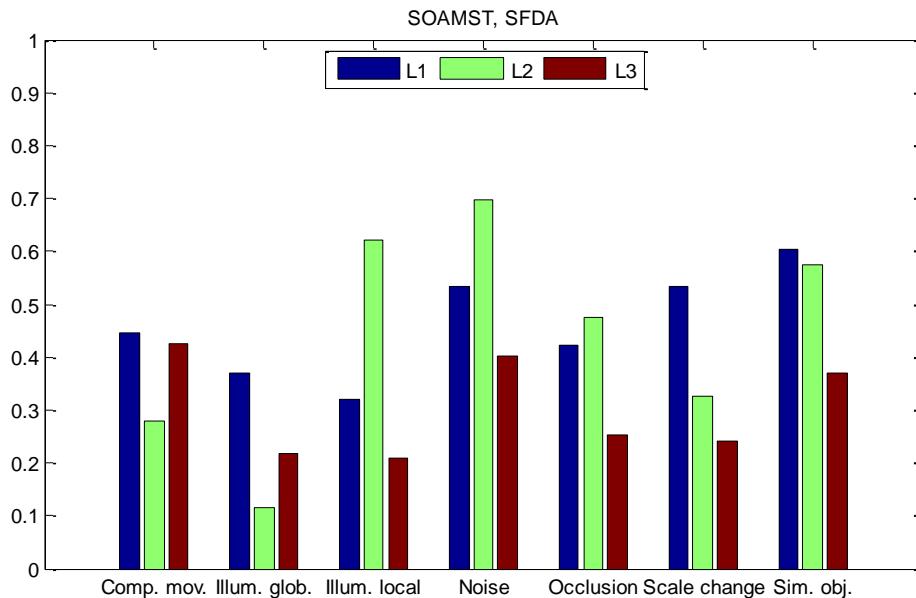


Figure 19 – SOAMST SFDA for L1, L2 and L3 videos.

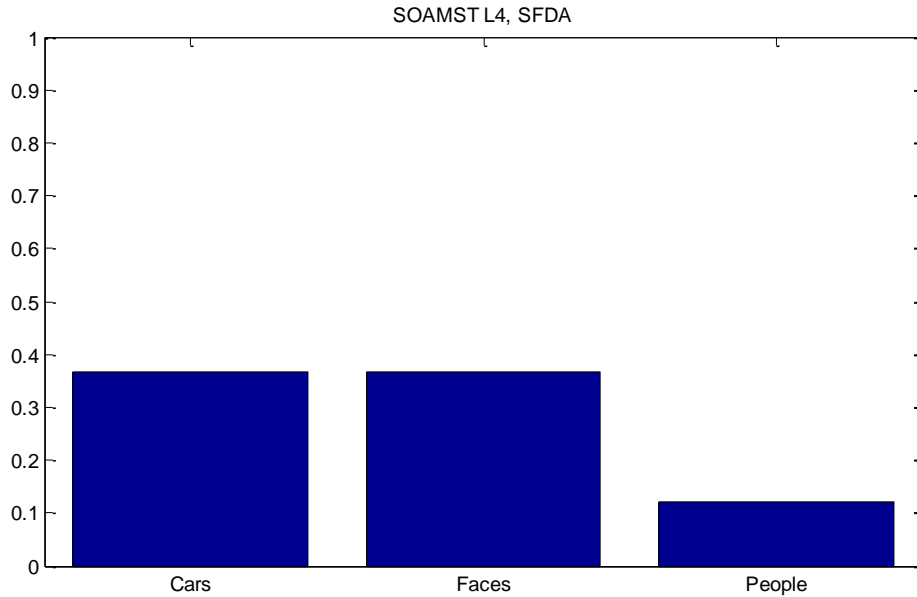


Figure 20 – SOAMST SFDA for L4 videos.

For L1 sequences SOAMST algorithm has better performance than the others algorithms in the L1 sequence in scale change category. This algorithm has been specially designed to withstand scale and orientation changes. Despite this, for L2 sequences this algorithm shows similar results than the others, and for L3 sequences it shows the second worst scores from all the algorithms. For all other categories of L1, SOAMST shows worse results than CBWH, MS and TM.

5.4 Comparative results

Figure 21 to Figure 24 show the comparative results of the evaluated algorithms.

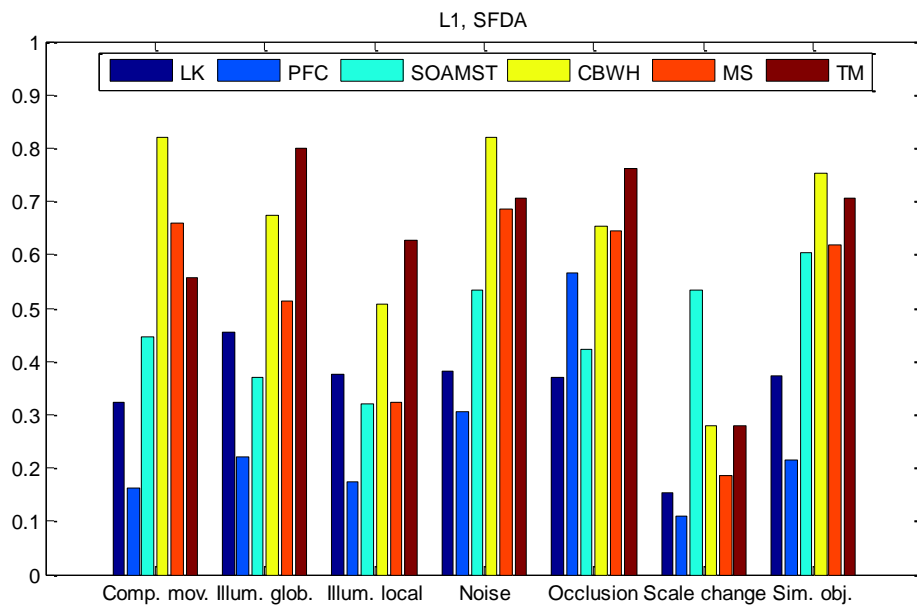


Figure 21 –SFDA for L1 videos.

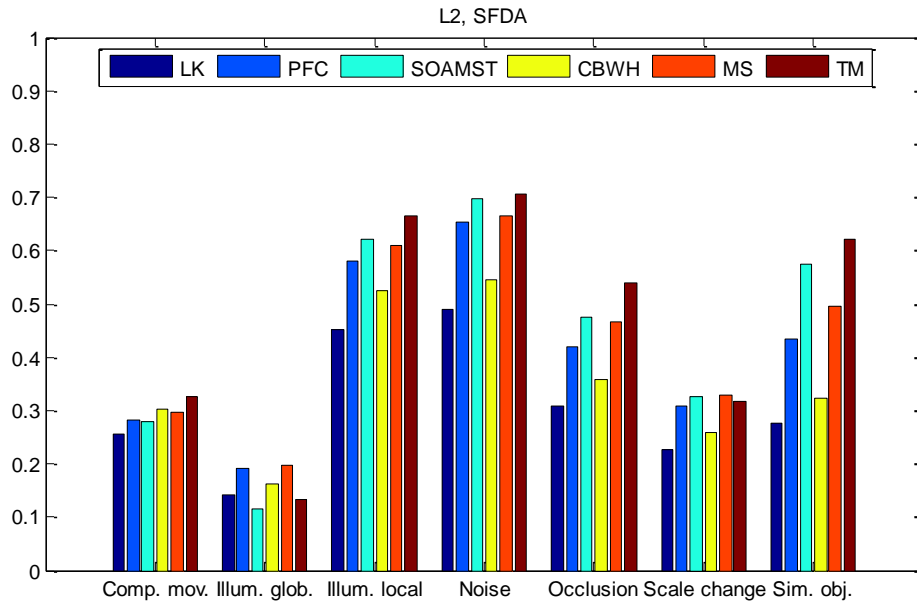


Figure 22 –SFDA for L2 videos.

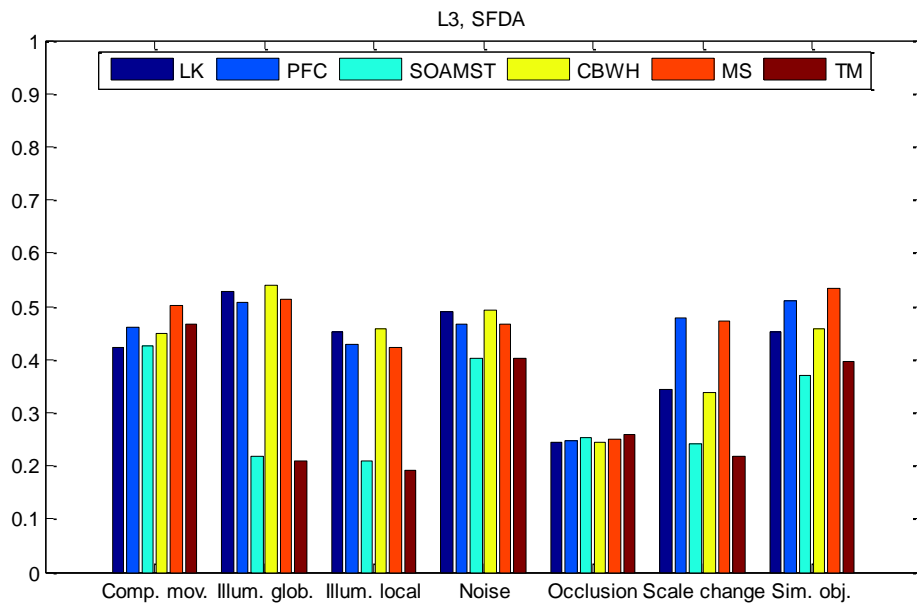


Figure 23 –SFDA for L3 videos.

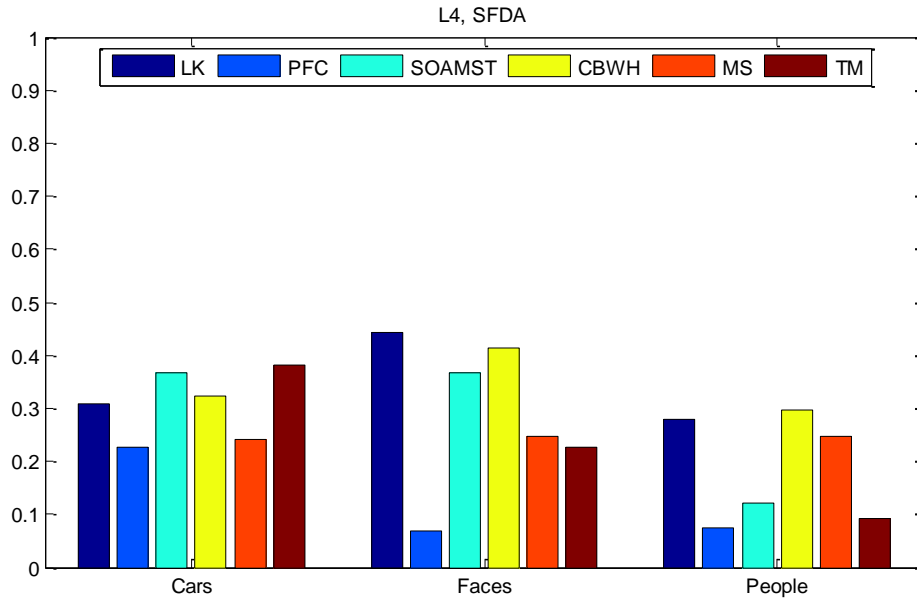


Figure 24 –SFDA for L4 videos.

The best results from the L1 sequence corresponds with the algorithms CBWH, MS and TM. In the scale change category, algorithm SOAMST presents results significantly better than the others. PFC algorithm presents the worse results in each one of the categories of this set of sequences.

For the L2 sequences, the order of the scores for the local illumination, noise, occlusions and similar objects categories is always as follows: TM, SOAMST, MS, PFC, CBWH and LK. For the scale changes and complex movement categories, the obtained scores are similar, except for LK and CBWH, whose scores in the scale changes are slightly worse. Note that the results of all the algorithms in the category of global illumination are quite low.

In the L3 sequences, LK, PFC, DBWH and MS present similar and better results than the other algorithms in most categories. In the scale change category, MS and PFC have the best results, over the rest of the algorithms. For occlusions category, all the algorithms have nearly identical results. SOAMST and TM algorithms have significantly worse results in the categories global illumination, local illumination and scale change. In the case of SOAMST scale change result, it is particularly interesting since the algorithm design attempts to solve the scale change problem and its score is lower than for other algorithms which do not consider this problem.

Finally, for the L4 sequences, the obtained scores are generally low and worse than those obtained in the other 3 sets of sequences. PFC presents the worse results, especially in faces and people categories.

5.5 Conclusions

As expected, none of the algorithms performs well in all categories and subcategories. Furthermore, none of the algorithms work well in the same subcategory of the different categories.

CBWH, MS and TM present the best results for the L1 sequences. In the case of the L2 sequences, the results of the six algorithms are similar results in all of the subcategories except for similar objects subcategory, highlighting the poor performance of the illumination global subcategory. For the L3 sequences, the results of the six algorithms presents similar results except for algorithms SOAMST and TM that present significantly worse results in some subcategories. Finally, for the L4 sequences, there is no algorithm that works well. The worst

overall results have been obtained for this category. This is reasonable since it is the most complex category.

5.6 Future research lines

Based on the results and discussions of this document, we plan the following future research lines:

5.6.1 Evaluation of more complex algorithms

In general, the used algorithms have not been published recently, as seen in the publication dates of its documents. One possible future research line is to test more recent and powerful algorithms to compare their performance with the algorithms that have been already evaluated.

5.6.2 Modify and complete the content set

As seen in Table 25, the video content set used does not cover all the categories of videos in the proposed scenario classification in the EventVideo project. One possible future research line consist of completing the content set extracting some videos from another content sets or even recording new videos that complete the categories where needed.

5.6.3 Evaluation of the algorithms with new metrics

There are many measures that can be added to the evaluation system for extracting more information from the analysis of the algorithms. An analysis of different metrics used in other documents is proposed to select the metrics that best complement to the SFDA.

5.6.4 Fusion

Tracking fusion is a popular topic, studied by many researchers in recent years. The main motivation of this future work line is to study how to improve the object tracking results, focused on the fusion of multiple tracking algorithms.

6 Event Detection

6.1 Introduction

In this section, we describe the evaluation of the event detection task for the EventVideo project. In particular, we present the results, conclusions and future research lines for the two current event-related tasks: discrimination of stationary objects (between abandoned and stolen) and recognition of human interactions (with objects and other humans).

6.2 Abandoned and stolen object discrimination

The detection of stationary objects in video sequences and their discrimination between (abandoned and stolen) is common task in video surveillance. Typically, a system is composed of sequentially-connected stages to detect the object of interest (in this case, a stationary one) such as foreground segmentation, blob tracking & classification and static blob recognition routines. Finally, the discrimination stage is in charge of deciding the type of event that applies to the static object (abandoned or stolen). Figure 25 shows an example of such kind of systems.

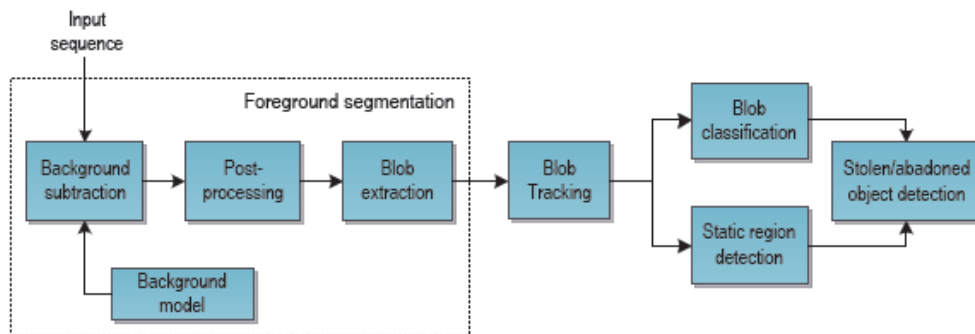


Figure 25 – Diagram of a typical video analysis system for abandoned and stolen object detection.

After extracting each stationary foreground object, we can distinguish two distinct processing stages for performing this discrimination. First, desired features are extracted from the foreground mask, the reference background, the current frame and the location of the static object; as detected by preceding analysis modules. Based on the extracted features, a likelihood measure (score) is then generated for each static object. Then, this score (or set of scores) is used by a classifier that assigns each object to a class (stolen or abandoned). Figure 26 shows an example of the discrimination task and its inputs.

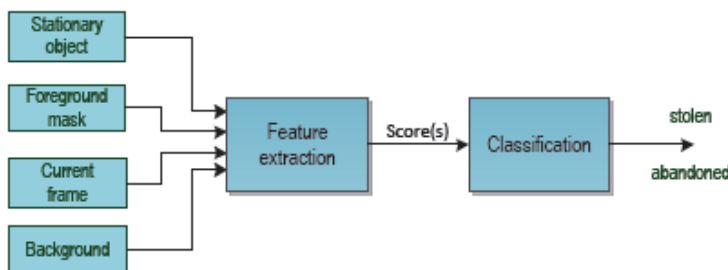


Figure 26 –Stolen/abandoned object discriminator

In this section, we evaluate the accuracy of the discrimination task using state-of-the-art approaches based on colour, edge and contour information. We present the evaluation scenario considered (dataset), the selected approaches and the obtained results.

6.2.1 Evaluation scenario

As evaluation scenario, we have used the dataset ASODDs (Abandoned and Stolen Object Discrimination dataset [40]), which is described in deliverable “D5.3. EventVideo test sequences, ground-truth and evaluation methodology”[1]. This dataset consists of two sets of input data (foreground masks) for the discrimination task for, respectively, real and ideal foreground data with three degrees of complexity (see Table 13).

Category	Number of annotations (blobs)				Complexity
	Annotated sequences		Real sequences		
	Abandoned	Stolen	Abandoned	Stolen	
S1	771	442	756	863	Low
S2	666	316	794	397	Medium
S3	595	174	852	660	High
All	2032	932	2402	1920	

Table 26 – ASODDs dataset description. The categories Sx directly correspond with the Sx scenarios considered in the document D5.3

As metrics, we evaluate the recognition performance of the discrimination task using the area under the ROC (Receiver Operating) curves and the discrimination accuracy as follows:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \approx \frac{\# \text{correctly classified samples}}{\# \text{total samples}}$$

, where *TP* are True Positives, *TN* are True Negatives, *FP* are false positives and *FN* are false negatives. Each term is defined as follows:

Observed classification	Predicted classification	
	abandoned(+)	stolen(-)
abandoned(+)	True positive (TP)	False Negative (FN)
stolen(-)	False positive (FP)	True negative (TN)

Table 27 – Definition of confusion matrix (TP, TN, FP, FN) for the discrimination task

6.2.2 Approaches

In order to cover the related state-of-the-art, we have selected the following approaches: Based on region-level colour information (CHIST), Based on edge information (GH and GL), Based on combining edge and colour information (FUS), Based on contour information [40] (PE, GR and GE), and Based on pixel colour information (PCC).

6.2.2.1 Based on region-level colour information (CHIST) [41]

We have selected an approach [41] based on measuring the colour similarity between the regions delimited by the foreground mask (internal and external regions of the bounding box of the stationary blob) in both the background and the current frame. The assumption is that in the current frame, stolen objects show a higher colour similarity (in the current frame) between these two regions than abandoned objects. Analogous reasoning is applied to the background frame and therefore, abandoned objects present high colour similarity between these regions in the background frame.

6.2.2.2 Based on edge information (GH and GL) [42]

Two approaches are selected based on comparing the values of the image gradient along the contour of the object (as obtained from the foreground mask) in the background and current frame images. It assumes that this ‘edge’ energy is high in the current frame for abandoned objects and low for stolen objects. The implemented approaches are similar to the one described in [42].

6.2.2.3 Based on combining edge and colour information (FUS) [43]

One approach has been selected that combines colour and edge results [43]. It computes likelihood models for each feature (colour and edge) and case (abandoned or stolen). It assumes the scores follow a Gaussian distribution which can be estimated using training data. Then, final scores are obtained by combining the likelihood models for each case (abandoned and stolen), selecting the final score with maximum value.

6.2.2.4 Based on contour information (PE, GR and GE) [40]

The main limitation of the edge and colour based approaches is that they need homogeneous properties in the regions of the background close to the static object (in terms of colour, motion and edges) and rely on precise foreground segmentation masks. Therefore, their accuracy is reduced in complex situations. Recently, contour-based approaches have been proposed to increase the robustness in complex situations by applying adjustments of the object contour (e.g. active contours) using pixel or region information. If the contour adjustment is reduced to small contour (compared to the initial one) in the current image, the object is stolen whereas if it occurs in the background image, abandonment has happened. In this study we have selected a pixel-based (PE) and two region-based (GR and GE) active contour approaches described in [40].

6.2.2.5 Based on pixel colour information (PCC) [44]

Due to the iterative nature of contour-based approaches, their use in real-time video surveillance is limited. Derived from the object contour, the colour contrast along the object contour at pixel level can be used where the colour difference is measured between both contour sides [44] (Figure 27).

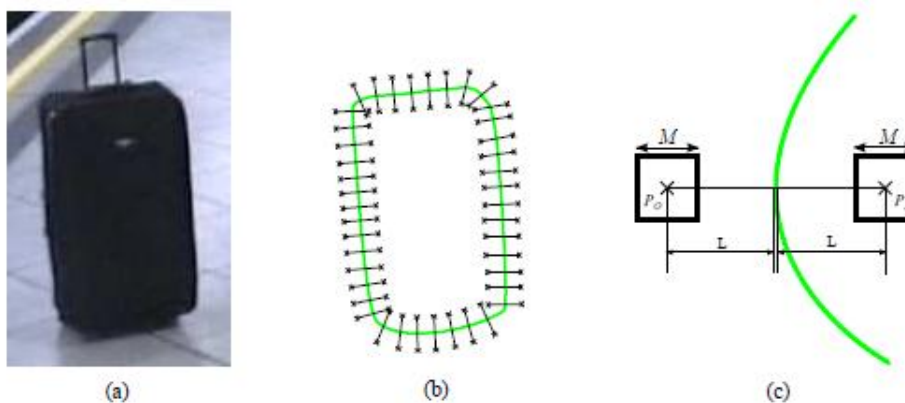


Figure 27 – Pixel colour contrast detector: (a) static foreground object, (b) analysed points along the boundary and (c) analysed contour point.

6.2.3 Comparative results

For the annotated data, a summary of the results is shown in Table 28 (accuracy) and Figure 28 (ROC). CHIST performs generally well on all categories given an accurate (ideal) foreground mask. However, some problems have been observed due to quantization noise introduced by the video compression scheme. In some cases, it causes colour information to leak beyond object

boundaries. We have seen that this problem affects smaller objects, which explains why it performs poorly on category 1 (low complexity); as this category includes mostly small objects. Additional problems have been observed due to the fact that the colour histograms are only computed on the Hue channel of the HSV colour space (assuming that this channel gives enough contrast). GH and GL show very good results for blobs of both classes. By using a window-based approach instead pixel-based one, GL has shown to provide better results than GH. Both discriminators, however, are affected by the presence of strong edges near the object boundaries that can be attributed to the background, as this causes the discriminators to produce a score that would correspond to objects from the opposite class. Active contours discriminators (PE, GR and GE) outperform the previous ones. PE has been effective in all cases, attaining perfect classification for all blobs in the annotated data set. GR and GE produce similar scores. However, region based approaches (GR and GE) are more robust to small changes between frames are attributed to noise (camera noise, compression noise...). This has not been observed for the PE discriminator, which seems to be more vulnerable to noise. PCC discriminator has proven very robust in situations in which the other discriminators have shown weaknesses. Finally, FUS discriminator demonstrates that combining different approaches is not an easy task

Discrim.	C1		C2		C3		ALL	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
CHIST	.864±.031	.933±.020	.819±.031	.959±.016	.947±.035	1.0±.0	.870±.018	.941±.013
GH	.892±.026	.973±.009	.837±.040	.925±.033	.804±.058	.932±.038	.852±.018	.940±.015
GL	.965±.016	.997±.002	.911±.013	.961±.012	.872±.044	.990±.009	.923±.010	.982±.004
PE	.999±.002	1.0±.0	.979±.014	.993±.006	.978±.018	.995±.007	.988±.008	.996±.002
GR	.962±.014	.996±.002	.920±.037	.971±.016	.942±.025	.987±.010	.943±.012	.985±.006
GE	1.0±.0	1.0±.0	.999±.003	1.0±.0	1.0±.0	1.0±.0	1.0±.001	1.0±.0
PCC	1.0±.0	1.0±.0	.991±.008	.999±.001	1.0±.0	1.0±.0	.997±.002	1.0±.0
FUS	.902±.030	.922±.035	.818±.051	.829±.019	.805±.037	.815±.024	.850±.022	.900±.025

Table 28. Discrimination results for annotated data. (Key. ACC:accuracy, AUC:Area Under Curve).

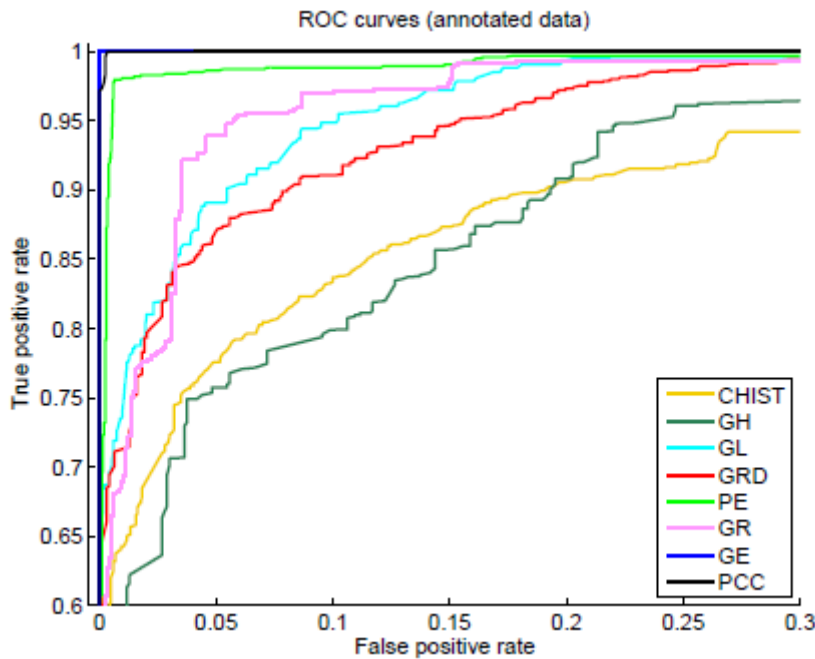


Figure 28 – ROC analysis for single-feature discrimination on annotated data

For the real data, the results are depicted in Table 29 (accuracy) and Figure 29 (ROC). For CHIST, we can see a decrease in accuracy of roughly 10% as compared to results on annotated data. This is explained by the fact that the CHIST completely relies on correct segmentation to obtain the histograms of inner and outer regions of the object to analyse. If there are errors in these regions, the extracted histograms may be too similar and the discriminator is incapable of producing a good score. For GL and GH, their reduction in accuracy for the gradient discriminators is not as significant as with other discriminators. This can be primarily attributed to the contour adjustment operation applied to the initial extracted contour, as it drives the contour to match the actual boundaries, as well as the small neighbouring window in which the analysis is performed. We can conclude that these edge-based algorithms (GL and GH) are affected by imprecise segmentation with a ~5% decrease.

Discrim.	C1		C2		C3		ALL	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
CHIST	.755±.035	.845±.026	.712±.055	.858±.036	.848±.025	.910±.019	.777±.024	.856±.019
GH	.918±.024	.964±.017	.791±.039	.867±.025	.743±.041	.792±.030	.821±.022	.879±.014
GL	.970±.020	.995±.004	.817±.029	.902±.032	.825±.030	.919±.026	.877±.016	.947±.013
PE	.882±.026	.949±.015	.768±.031	.821±.030	.806±.027	.905±.022	.824±.020	.90±.016
GR	.857±.018	.945±.010	.80±.031	.839±.037	.748±.042	.845±.027	.803±.020	.882±.015
GE	.960±.010	.996±.002	.952±.027	.984±.012	.929±.016	.954±.009	.947±.011	.981±.004
PCC	.967±.014	1.0±.0	.943±.013	.987±.006	.951±.013	.996±.002	.954±.009	.994±.001
FUS	.737±.052	.802±.042	.556±.034	.606±.019	.665±.052	.688±.027	.662±.025	.679±.035

Table 29. Discrimination results for real data. (Key. ACC:accuracy, AUC:Area Under Curve).

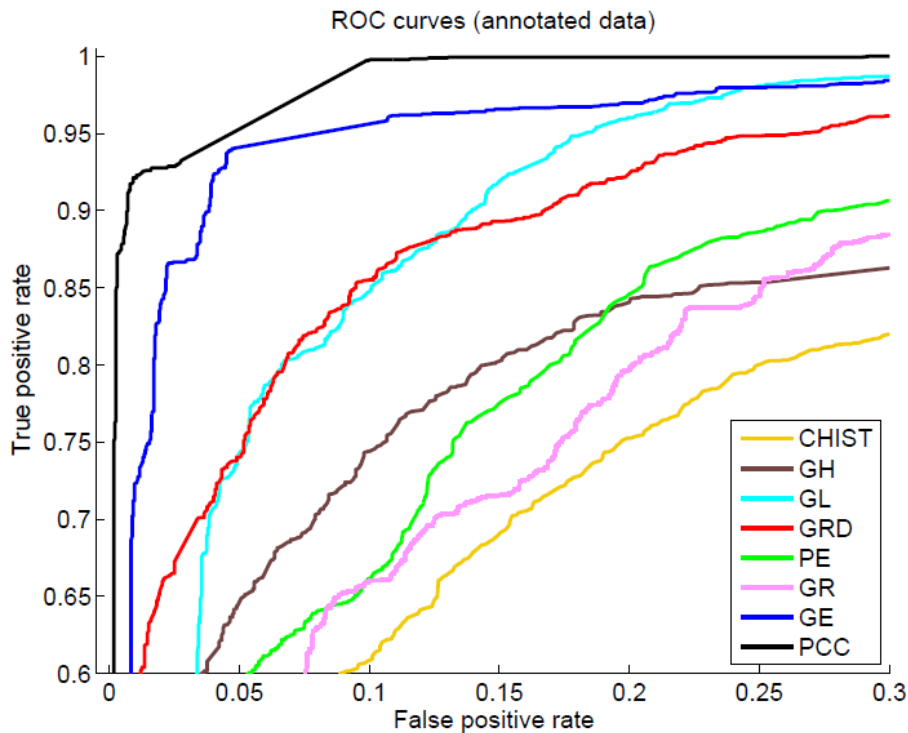


Figure 29 – ROC analysis for single-feature discrimination on real data

For the active contours discriminators, we can observe that the accuracy of the PE and GR discriminators is reduced due to incorrect segmentation demonstrating a dependency on the contour initialization. In contrast, the GE has shown robustness against incorrect foreground segmentation. We can attribute this to the fact that the adjusted contour generally tends to shrink unless the object boundaries are nearby. This is usually the case, even for inaccurate foreground masks. We have observed, however, that if the foreground mask is initialized inside a uniformly

coloured object, the contour tends to shrink or even disappear. When this happens in both the current frame and the background the resulting score is very close to 0.0. This problem is more evident for smaller blobs, as the contours tend to quickly disappear in both images. This explains the presence of scores with value 0.0 (Figure 6.11), that in the vast majority of cases correspond with very small foreground masks due to over-segmentation.

PCC discriminator is the least affected for incorrect segmentation. We could explain this by taking into consideration that the measures are taken at a distance from the corresponding contour pixel (parameter L), and averaged inside a small window (parameter M), which leave the discriminator a margin to overcome segmentation inaccuracies.

Table 30 shows the obtained computational costs of all the evaluated discriminators. Maximum and minimum values correspond to large and small objects, respectively. As it is shown, CHIST, GL and GH discriminators have a lower computational cost than the proposed active contours ones (PE, GE and GR). This is due to the complexity of the employed active contours algorithms (iterative nature). Among all evaluated discriminators, PCC has shown the lowest computational cost, improving existing approaches, due to the simplicity of the performed analysis (average colour contrast).

	Time (ms)							
	CHIST	GH	GL	GRD	PE	GR *	GE	PCC
Min.	5.67	0.15	0.14	0.29	2.30	96.89	20.78	0.19
Max.	44.57	133.80	255.78	354.33	1401.80	8207.50	1187.10	8.66
Avg.	23.23	28.35	57.14	84.61	234.47	901.40	246.39	1.78

Note: times for the GR discriminator are given for the Matlab implementation

Table 30. Computational cost comparative

6.3 Human interactions (with objects and humans)

The recognition of human-related events has emerged as a very promising research area due its multiple applications such as video surveillance. These events can be broadly divided into activities (e.g., jump, run) and interactions (e.g., get an object). Moreover, human interactions can be divided if they consider an object (e.g., leave bag) or humans (e.g., shake hands).

Typically, a complete event detection system is composed of several stages such as foreground detection, blob tracking, feature extraction and event recognition. Each stage provides data that is used for recognizing the event in the last stage. Figure 30 shows an example of such kind of systems.

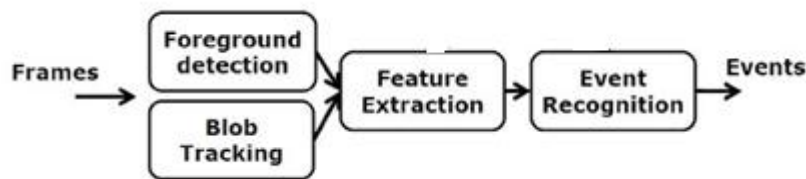


Figure 30 – Common block diagram for a generic event detection system

In this section, we evaluate the accuracy of the event detection task using state-of-the-art approaches for recognizing human-object interactions and human-human interactions. We present the evaluation scenario considered (dataset), the selected approaches and the results.

6.3.1 Evaluation scenario

As evaluation scenario, we have focused on indoor scenarios where spatial information (e.g., location of static parts of the scene such as tables and doors) can be defined and applied to improve accuracy. We used the dataset EDds (Event Detection dataset [45]), which is described in the document of the EventVideo project “D5.3. EventVideo test sequences, ground-truth and evaluation methodology”. It has 17 sequences focused with two types of human-related events: interactions (Leave, Get and Use object) and activities (Hand Up and Walking).

Different degrees of complexity are contained in the dataset (see Table 31). As we are only interested in human interactions, activities of EDs (Hand Up and Walking) are not considered.

Category	Event Occurrences					Complexity			
	Interactions			Real sequences		Segmentation	Tracking	Classification	Event
	LEA	GET	USE	HUP	WLK				
S1	18	13	9	9	54	Medium	Low	Medium	Medium
S2	7	7	10	14	44	Medium	Medium	Medium	High
S3	14	14	22	20	10	High	High	High	High

Table 31 – EDs dataset description. The categories Sx directly correspond with the Sx scenarios considered in the document D5.3

In order to increase the variability and quantity of data for evaluating the performance of this evaluation scenario, we have included two additional datasets: LIRIS [46] and SSG [47]. Moreover, the use of the LIRIS dataset allows participating in the “Human activities recognition and localization competition” (ICPR - HARL 2012, and getting a performance comparison with other existing approaches.

The LIRIS dataset contains several human-object and human-human interactions in a controlled indoor settings captured with a static camera at a 720x576 resolution (25 fps). Each sequence contains 1-5 humans performing actions in short sequences (500-3000 frames). This dataset can be considered as a very realistic scenario as the events are performed in a natural way presenting several occlusions in most of the situations. Moreover, different viewing angles and distances to the camera were considered in the sequences. Besides colour and depth information were provided for the data composing two datasets: D1 (colour+depth) and D2 (colour). Sample frames are shown in Figure 31. More details can be found at <http://liris.cnrs.fr/harl2012/>

The SSG dataset mimics the number of events contained LIRIS dataset whilst reducing their complexity. Hence, simple backgrounds are considered with fewer humans involved (1-3). Actions are performed in a simple way without occlusions considering the camera viewing angle. The distance to the action is short and therefore, facilitates the event recognition.



Figure 31 – Examples of human-related events of the LIRIS dataset (KEY. DI:Discussion, GI:Give Object. BO: Take Object. EN: Enter through a door. ET: Try to unlock a door. LO: Unlock a door. HS: Hand Shake. UB:Unattended Bag. KB: Keyboard typing. TE:Talking with telephone).

Dataset	Equivalency in document D5.3	Event Occurrences							Complexity
		BO	EN	LO	UB	HS	KB	TE	
SSG	S1	32	3	9	13	3	8	10	Low

ED	S2	46	44	-	-	9	20	-	Medium
LIRIS-train	S3	9	20	-	6	11	12	5	High

Table 32 summarizes the material used for evaluating the performance of the event recognition task and an estimation of the overall complexity of each dataset. As it can be observed, the LIRIS dataset present the highest complexity.

Dataset	Equivalency in document D5.3	Event Occurrences							Complexity
		BO	EN	LO	UB	HS	KB	TE	
SSG	S1	32	3	9	13	3	8	10	Low
ED	S2	46	44	-	-	9	20	-	Medium
LIRIS-train	S3	9	20	-	6	11	12	5	High

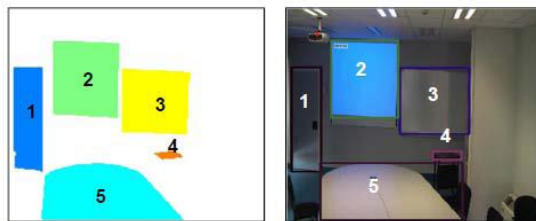
Table 32 – Dataset description for evaluating human interactions

As metrics, we evaluate the event detection performance using the standard spatio-temporal overlap between the detections generated by the system and the event annotations (ground-truth). This measure considers standard Precision (P) and Recall (R) as described in the document of deliverable “D5.3. EventVideo test sequences, ground-truth and evaluation methodology”[1].

6.3.2 Approaches

6.3.2.1 VPULab approach[47]

The approach developed within the EventVideo project is based on the event detection system that uses contextual information [48]. The VPULab approach contains the typical analysis stages (foreground segmentation, blob tracking, feature extraction and event recognition) and an additional one that considers contextual information that allows improving the event recognition rate. It detects 10 human-object and human-human interactions (all the events defined in the ICPR-HARL competition) based on features extracted from foreground blobs: blob velocity, blob trajectory, people likelihood, blob compactness, people skin and relative distances to contextual objects (tables, chairs, walls...). More details can be found at [47]. Figure 32 shows an example of the contextual information and Figure 33 depicts the block diagram of the approach.



Object categories

- 1 Obj\ContextObject\FixedObject\Door
- 2 Obj\ContextObject\FixedObject\ProjectionA
- 3 Obj\ContextObject\FixedObject\Blackboard
- 4 Obj\ContextObject\FixedObject\Table
- 5 Obj\ContextObject\FixedObject\Table

Figure 32 – Example of contextual information used by the system

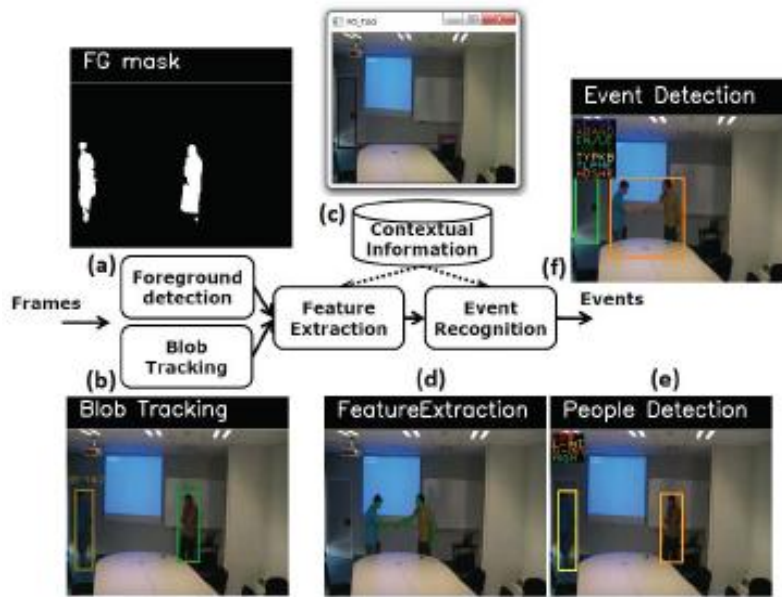


Figure 33 – Diagram of the VPULab event detection system

6.3.2.2 Additional approaches in the ICPR-HARL 2012 competition

Initially 70 teams were registered in the competition and downloaded the dataset. Finally, only 4 participants submitted results that are briefly summarized:

- ADSC-NUS-UIUC participant is a collaboration between Advanced Digital Sciences Center (Singapur), National University of Singapore (Singapur) and University of Illinois at Urbana-Champaign (EEUU). It only used D1 dataset (colour+depth) and the approach is based on combining human detection (using HOG), human pose detection, specific object recognition, interaction attributes and scenario type classification. The combination scheme is a Bayesian network.
- TATA-ISI participant is collaboration between two indian institutions: India: Innovation Lab, Tata Consultancy Services e Indian Statistical Institute. It also used D1 dataset. Its approach is based on two stages. First, moving object segmentation is performed through depth information and key pose recognition is used to recognize the events as described in [49].
- IACAS participant is the Chinese Academy of Sciences (Beijing, China). It focused on the D2 dataset (only colour). Its approach is based on STIPs (Space-Time Interest Points) as defined in [50]. Then, STIPs are used to train SVMs for each event to recognize. Then, event recognition is determined as the output of the SVM with maximum likelihood. The localization is performed similarly to [51].

A detailed participant description is available at <http://liris.cnrs.fr/harl2012/algorithms.html>

6.3.3 Comparative results

Evaluation of the VPULab approach is done in two phases. The first one aims to determine the strengths and weaknesses of the approach in the three considered datasets. The second provides a comparison with state-of-the-art approaches within the ICPR-HARL competition.

6.3.3.1 Phase 1

Results are summarized in Table 33 for the VPULab approach. Although results for SSG and ED present good performance, it can be observed that the performance for LIRIS-train is highly reduced. Figure 34 shows some of the most common failures of the approach. First row shows that wrong foreground segmentation (background is not correctly represented) prevents the correct detection of objects. Moreover, KB (keyboard typing) is wrongly detected as there is an

overlap between the skin of the arm and the keyboard. Second row presents a situation where in the point of interest of the handshake, the hands are not detected and therefore, no skin can be associated to both humans. Hence, the event is not detected. Third row shows that the dependence on the size of interacting objects (mobile phone is not detected whereas the laptop is detected and included in the background).

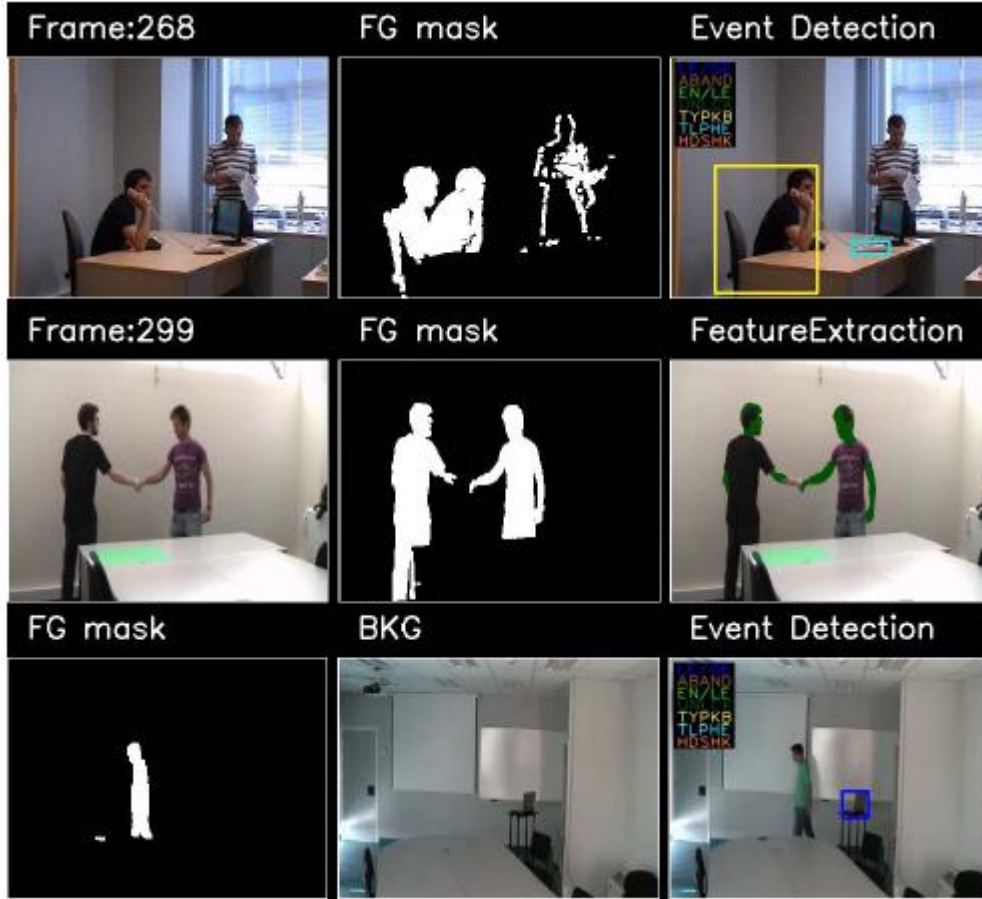


Figure 34 – Failure examples of the VPULab approach

Dataset	BO		EN		LO		UB		HS		KB		TE		Total		
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	F
SSG	0.45	0.94	0.75	1	1	0.44	0.69	0.81	1	0.33	0.88	0.88	0.90	0.90	0.60	0.81	0.69
ED	0.56	0.74	0.89	0.77	-	-	-	-	0.88	0.78	0.95	0.95	-	-	0.74	0.79	0.76
LIRIS-train	0.56	0.56	0.40	0.10	-	-	0.50	0.67	0.67	0.36	1	0.08	1	0.60	0.59	0.30	0.40
Mean	0.52	0.75	0.68	0.62	1	0.44	0.60	0.74	0.85	0.49	0.94	0.64	0.95	0.75	0.64	0.63	0.62

Table 33 – Results of the VPULab for detecting human-object and human-human interactions

6.3.3.2 Phase 2

Here we present the results of the VPULab approach in the ICPR-HARL competition using the D2 dataset (only colour) and a comparison with other participants.

The first comparison is given in Table 34 for the recognition task without spatial and temporal localization (e.g., bounding box and number of frames), that is, indicating if the event has been detected considering the entire sequence.

It can be observed that the VPULab system presents an acceptable performance level compared to other participants. Although it obtained a low recall (36%), it has the second better value and it is the most precise system (66%). Compared to experiments in phase 1, we can observe that the VPULab approach has decreased its precision but increased the recall. Comparing the VPULab with other participants, we can observe that without using depth information, our

system is able to achieve similar performance to the participants using D1 datasets (where foreground objects can be easily extracted based on depth data). Moreover, it can be observed that the two systems using key pose analysis present lower performance compared to the other approaches that do not use such technique. This can indicate that key pose estimation is a not sufficiently discriminative feature to differentiate events from other. Moreover, most of the participants used training data (from LIRIS-train) to detect in the test dataset. The low performance also indicated that pure machine learning methods are not suitable for event recognition as the variability in the executions of the same event is very high

Equipo	Dataset	Recall	Precision	F-Score
ADSC-NUS-UIUC	D1	0.74	0.41	0.53
TATA-ISI	D1	0.08	0.17	0.11
VPULABUAM	D2	0.36	0.66	0.46
IACAS	D2	0.30	0.46	0.36

Table 34 – Results of the ICPR-HARL 2012 competition (without localization). A description of the participant teams is available at <http://liris.cnrs.fr/harl2012/>

The second comparison is given in Table 35 for the recognition task with spatial and temporal localization (e.g., bounding box and number of frames). If we observe the results, it looks that the VPULab system is not accurate. However, it has to be considered the accuracy of the annotations and rules used to annotate the events. Spatio-temporal localization requires a considerable overlap between detections and annotations. As no particular rules were provided to participants for event annotations, we decided to generate an output considering the object of interest. For example, Figure 35 shows an example of UB (unattended bag) event where it can be observed that the competition organizers provided an annotation different from the expected output in such kind of systems. It is expected that an alarm is generated after the event has happened (the abandonment) and not during it. Moreover, the event should consider the object of interest as the abandoned event might involve that the owner of the object exits the scene (as if the owner is close to the object, it is not unattended). Therefore, the obtained results have to be reasonably considered.

Equipo	Dataset	Recall	Precision	F-Score
ADSC-NUS-UIUC	D1	0.63	0.33	0.44
TATA-ISI	D1	N/A	N/A	N/A
VPULABUAM	D2	0.04	0.08	0.05
IACAS	D2	0.03	0.04	0.03

Table 35 – Results of the ICPR-HARL 2012 competition (with localization). A description of the participant teams is available at <http://liris.cnrs.fr/harl2012/>

In summary, this second evaluation does not give sufficient hints about spatio-temporal localization performance of the VPULab approach. In fact, only one the four participants obtained good localization results and it is not directly comparable to the VPULab approach as it uses different data (colour + depth, versus our approach that only uses colour).

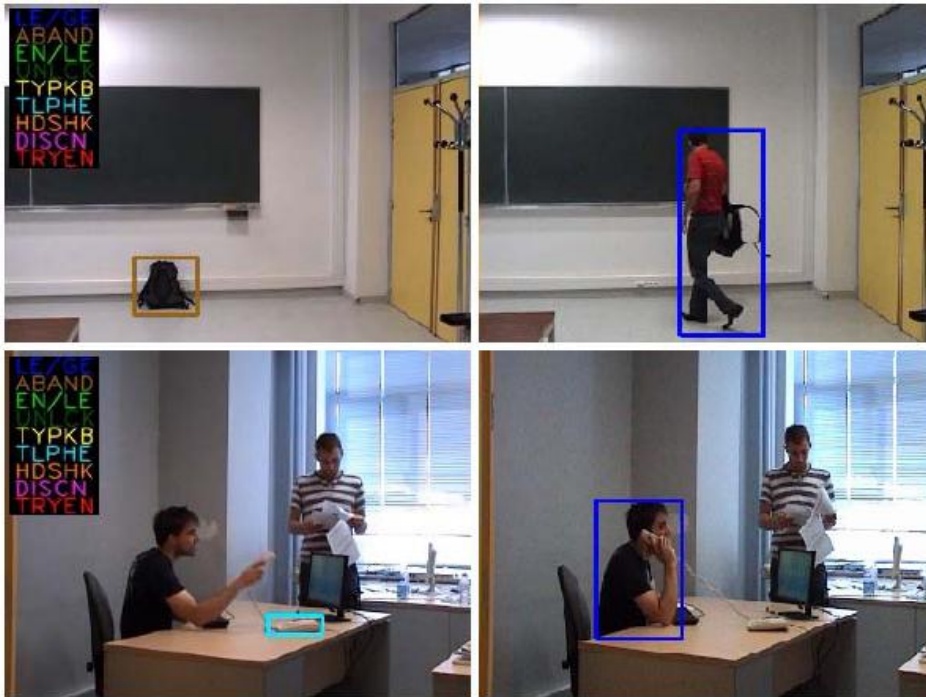


Figure 35 – Visual comparative of VPULab detection versus annotations given by ICPR-HARL

6.4 Conclusions

In this section we have presented an evaluation of two tasks related with the recognition of events in video: discrimination for abandoned/stolen objects and human-based interactions.

The first evaluation has been performed on annotated and real data over a representative set of state-of-the-art measures. Reported results have demonstrated high accuracy with both types of data. Although a slight reduction is observed with real data, a recognition rate higher than 99% was achieved with two algorithms based on active contour adjustment (GE) and pixel colour contrast (PCC). PCC also presented the lowest computational cost and therefore, being the best choice for the discrimination task between abandoned and stolen.

The second evaluation demonstrated that performance of current systems for recognizing human-object and human-human interactions is far from being successful. The VPULab approach was applied to three similar datasets obtaining a mean accuracy of 60% (in both precision and recall). VPULab participation in ICPR-HARL 2012 competition (using a dataset with a significantly higher complexity) showed that although the performance of the system is still low, it is higher than the state-of-the-art systems based on colour and some of them based on colour and depth data. In both experimental phases, the main problems of the system were due to the wrong detection (or misdetection) of objects and humans of interest that performed the actions. Models of the events apparently worked properly in most of the situations.

6.5 Future research lines

As future research lines, we propose the following for the two evaluated tasks.

6.5.1 Stationary object detection in high-density scenarios

For systems devoted to detecting abandoned or stolen objects, the discrimination task seems to be completed (due to the high recognition rates achieved in the evaluation) and the complexity of recognizing the events seems to be in the extraction of the stationary object in crowded scenes (which is an open problem in the current literature). Hence, future work in this task would be directed towards developing efficient algorithms for stationary object detection in videos where there is high density of moving objects. One possible research line in this

direction would be the extraction of stationary regions without performing explicit tracking. Techniques like [52] are able to detect moving objects with unusual motion patterns in the scene, including stationary objects. An extensive evaluation of the algorithm in crowded scenes would be necessary to assess the suitability of such techniques for this particular task. Moreover, comparison with state-of-the-art approaches should be considered.

6.5.2 Human-related interactions enhancements

For recognizing human-related interactions, we have identified two main drawbacks of the VPULab system. The first considers the low-level data extraction (i.e., foreground detection and tracking). It was observed that most of the problems were due to false or missed detections. Development of robust approaches should be thoroughly considered in order to improve current performance. The second one considers the reduction of the parameters for configuring the event recognition models as they currently require manual fine-tuning by an expert.

7 Conclusions and future work

In this Technical Report we have evaluated the algorithms currently considered within the EventVideo project, both State-of-Art (SoA) ones and the ones developed within VPULab. The evaluation has been guided by the evaluation framework (datasets, associated ground-truth and metrics) described in Deliverable 5.3v1 “EventVideo test sequences, ground-truth and evaluation methodology”[1].

With respect to segmentation, the work reported is divided in two depending on the cases of fixed or moving cameras.

In the fixed camera scenario, there are several algorithms that perform well enough if they include specific techniques. Nevertheless, there exist still problems to be solved. There is a need to confront the sensitivity-discriminability problem: adequately identify highly dynamic backgrounds without degrading the system performance in foreground discrimination and, inversely, design systems accurate enough to discriminate camouflaged foreground while maintaining its capability to adapt to changing backgrounds. Additionally, it is important to evaluate the limits of applicability of pixel level segregation: the use of post-processing approaches may be a solution, but there is a high increase in the computational cost of the system and its operation is severely conditioned to their preliminary VOS stage. Finally, it is necessary to remark that presented techniques might not work correctly in crowded scenarios (categories S3 and S4 of **Error! Reference source not found.**) as they were designed under the premise that background samples of each pixel are majority along the video.

In the moving camera scenario, it has been demonstrated that the accuracy of the Camera Motion Estimation (CME) stage can have tremendous influence in the whole segmentation result. Given this importance, as well as the many additional factors also influencing segmentation, it seems reasonable to isolate the evaluation of CME stage from the segmentation itself. In the evaluation we have found that situations involving large objects –which are completely normal in every-day videos– can make standard techniques used for CME fail. Therefore, it is extremely valuable to have CME techniques that can provide robustness to large objects even in absence of temporal information. These techniques, which will often be more computationally demanding, can always be used when temporal information is unavailable (eg. initial frame) or becomes unreliable (eg. after shot changes or when a previously static object starts to move).

The work in people detection provides several conclusions. The use of segmentation stages, even introducing the problems of under/over-segmentation, makes easier the classification stage than in approaches working with exhaustive search. The combination of segmentation and exhaustive search reduces these problems but they are still a drawback especially in complex scenarios where these problems are magnified. Therefore, exhaustive search approaches are more reliable in complex environments, but it must deal with a great number of negative examples (potential false positive detections), reducing the recall rate in order to maintain the precision rate. As expected, the use of simplified person models gets, mainly, worse results mainly in terms of Precision than those using more complex person models. Finally, although the motion information is less characteristic than the appearance for people detection, the combination of motion and appearance shows to be useful also in complex scenarios.

With respect to tracking, as expected, none of the algorithms performs well in all categories and subcategories. Furthermore, none of the algorithms work well in the same subcategory of the different categories. This demonstrates that different algorithms are suited for different scenarios, without having found any working correctly in all of them, although of course there are algorithms that outperform others from a general point of view. For complex scenarios (L4 sequences) scenarios, there is no algorithm that works well.

Work in event detection has been done in two scenarios: abandoned/stolen objects detection and human-based interactions.

With respect to abandoned/stolen object detection, the work has been focused in discrimination of abandoned/stolen object (after a segmentation considering moving and static objects). Results show high accuracy in both cases using two algorithms, one based on active contour adjustment (GE) and the other one on pixel colour contrast (PCC). As the later presents the lowest computational cost, it is the best choice for the discrimination task between abandoned and stolen.

The evaluation of human-based interactions has demonstrated that the performance of current systems for recognizing human-object and human-human interactions is far from being successful. The VPULab approach was applied to three moderate-complexity datasets obtaining a mean accuracy of 60%, whilst its use over a dataset with a significantly higher complexity (ICPR-HARL 2012) showed a very low performance. Nevertheless, its performance was higher than other state-of-the-art systems based on colour, and some of them based on colour and depth data. In both experimental phases, the main problems of the system were due to the wrong detection (or misdetection) of objects and humans of interest that performed the actions. Models of the events (apparently) worked properly in most of the situations. Therefore, it can be concluded that as in the case of abandoned/stolen object detection, the previous analysis stages (mainly segmentation, but also object detection and tracking) are the main drawback for improving event detection, as the models work properly when the previous stages do.

Considering the analysed results, the aforementioned conclusions and the existing problems, we propose several main lines of future research.

- General
 - Expand the evaluation framework (datasets and metrics)
- Segmentation (see section 2.6)
 - Refinement by post-processing techniques
 - Use of alternative features.
 - Include semantics in the descriptions
- People detection (see section 4.6)
 - Improve or refine segmentation
 - Appearance and motion fusion
- Tracking (see section 5.6)
 - Incorporation of more complex algorithms
 - Fusion approaches
- Event detection (see section 6.5)
 - efficient algorithms for stationary object detection in crowded scenes
 - incorporation of robust segmentation, detection and tracking approaches
 - optimization of parameters configuration

8 References

- [1] EventVideo Deliverable 5.3v1 “EventVideo test sequences, ground-truth and evaluation methodology”, July 2012.
- [2] C. Stauffer, W. Grimson, “Adaptive Background Mixture Models for Real-Time Tracking”, Proc. of CVPR 1999.
- [3] A. Elgammal, R. Duraiswami, D. Harwood, L. Davis, “Background and foreground modelling using nonparametric kernel density estimation for visual surveillance”, Proceedings of the IEEE, 90(7):1151-1163, Jul. 2002.
- [4] N. Goyette, P. Jodoin, F. Porikli, J. Konrad, P. Ishwar, “Changetection.net: A new change detection benchmark dataset”, Proc. of CVPR 2012.
- [5] M. Piccardi, “Background subtraction techniques: a review”, Proc. of SMC 2004.
- [6] R. Evangelio, M. Patzold, T. Sikora, “Splitting Gaussians in Mixture Models”, Proc. of AVSS 2012.
- [7] R. Evangelio, T. Sikora, “Complementary background models for the detection of static and moving objects in crowded environments”, Proc. of AVSS 2011.
- [8] A. Morde, X. Ma, S. Guler, “Learning a background model for change detection”, Proc. of CVPR 2012.
- [9] A. Cavallaro, T. Ebrahimi, “Video object extraction based on adaptive background and statistical change detection”, Proc. of SPIE Vol.4310.
- [10] A. Colmenarejo, M. Escudero-Viñolo, J. Bescós, “Class-driven Bayesian background modelling for video object segmentation”, Electronics Letters, 47(18): 1023 -1024, Sep. 2011.
- [11] M. Hofmann, P. Tiefenbacher, G. Rigoll, “Background segmentation with feedback: The Pixel-Based Adaptive Segmenter”, Proc. of CVPR 2012.
- [12] A. Schick, M. Bauml, R. Stiefelhagen, “Improving foreground segmentations with probabilistic superpixel Markov random fields”, Proc. of CVPR 2012.
- [13] L. Maddalena, A. Petrosino, “The SOBS algorithm: What are the limits?”, Proc. of CVPR 2012.
- [14] M. Van Droogenbroeck, O. Paquot, “Background subtraction: Experiments and improvements for ViBe”, Proc. of CVPR 2012.
- [15] M. Escudero-Viñolo, J. Bescós, “A robust framework for region based video object segmentation”, Proc. of ICIP 2010.
- [16] MPEG Requirements Group, "Guide to obtaining the MPEG-7 Content Set," Doc. ISO/MPEG N2570, Dec. 1998.
- [17] B. Qi, M. Ghazal, A. Amer, "Robust Global Motion Estimation Oriented to Video Object Segmentation," IEEE Trans. on Image Processing, 17(6):958-967, Jun. 2008.
- [18] A.M. Enzweiler, D. M. Gavrilu, “Monocular pedestrian detection: Survey and experiments”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(12):2179–2195, Dec. 2009.

- [19] D. Geronimo, A. M. Lopez, A. D. Sappa, T. Graf, “Survey of pedestrian detection for advanced driver assistance systems”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1239–1258, Jul. 2010.
- [20] P. Dollar, C. Wojek, B. Schiele, P. Perona, “Pedestrian detection: An evaluation of the state of the art”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, Apr. 2012.
- [21] A. García-Martín, J.M. Martínez, and J. Bescós, “A corpus for benchmarking of people detection algorithms”, *Pattern Recognition Letters*, 33(2):152–156, Jan. 2012.
- [22] TRECVID, “Trecvid 2008 evaluation for surveillance event detection, <http://www-nlpir.nist.gov/projects/trecvid/>.”
- [23] A. García-Martín, J.M. Martínez, “Robust real time moving people detection in surveillance scenarios”, *Proc. of AVSS 2010*.
- [24] V. Fernández-Carbajales, M.A. García, and J.M. Martínez, “Robust people detection by fusion of evidence from multiple methods”, *Proc. of WIAMIS 2008*.
- [25] N. Dalal, B. Triggs, “Human detection using oriented histograms of flow and appearance”, *Proc. of ECCV 2006*.
- [26] B. Leibe, E. Seemann, B. Schiele, “Pedestrian detection in crowded scenes”, *Proc. of CVPR 2005*.
- [27] M. Andriluka, S. Roth, B. Schiele, “Pictorial structures revisited: People detection and articulated pose estimation”, *Proc. of CVPR 2009*.
- [28] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, “Object detection with discriminatively trained part-based models”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sep. 2010.
- [29] A. García-Martín, A. Hauptmann, J.M. Martinez, “People detection based on appearance and motion models”, *Proc. of AVSS 2011*.
- [30] B. Wu, R. Nevatia, “Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors”, *Proc. of ICCV 2005*.
- [31] F. Xu, K. Fujimura, “Human detection using depth and gray images”, *Proc. of AVSS 2003*.
- [32] I. Haritaoglu, D. Harwood, L.S. Davis. “Ghost: A Human Body Part Labelling System Using Silhouettes”, *Proc. of ICPR 1998*.
- [33] F. Tiburzi, M. Escudero, J. Bescós, J.M. Martinez, “A ground truth for motion based video-object segmentation”, *Proc. of ICIP 2008*.
- [34] D. Comaniciu, V. Ramesh, P. Meer, “Kernel-based object tracking”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(5):564-577, May 2003.
- [35] R. Brunelli, “Template Matching Techniques in Computer Vision: Theory and Practice”, Wiley Publishing, 2009.
- [36] S. Baker, I. Matthews, “Lucas-Kanade 20 Years On: A Unifying Framework”, *International Journal of Computer Vision*, 56(3):221-255, Mar. 2004.
- [37] K. Nummiaro, E. Koller-Meier, L.J. Van Gool, “An adaptive colour-based particle filter”, *Image and Vision Computing*, 21(1):99-110, Jan. 2002.

- [38] J. Ning, L. Zhang, D. Zhang, C. Wu, “Robust mean shift tracking with corrected backgroundweighted histogram”, IET Computer Vision, 6(1):62-69, Jan. 2011.
- [39] J. Ning, L. Zhang, D. Zhang, C. Wu, “Scale and orientation adaptive mean shift tracking”, IET Computer Vision, 6(1):52-61, Jan. 2012.
- [40] L. Caro, J.C. SanMiguel, J.M. Martínez. “Discrimination of abandoned and stolen object based on active contours“, Proc. of AVSS 2011.
- [41] S. Ferrando, G. Gera, C. Regazzoni, “Classification of unattended and stolen objects in videosurveillance system”, Proc. of AVSS 2006.
- [42] P. Spagnolo, A. Caroppo, M. Leo, T. Martiriggiano, T. D'Orazio“, An abandoned - removed objects detection algorithm and its evaluation on pets datasets”, Proc. of AVSS 2006.
- [43] J.C. SanMiguel, J.M. Martinez, “Robust unattended and stolen object detection by fusing simple algorithms”, Proc. of AVSS 2008.
- [44] J.C. SanMiguel, L. Caro, J.M. Martínez, "Pixel-based colour contrast for abandoned and stolen object discrimination in video surveillance", Electronic Letters, 48(2):86-87, Feb. 2012
- [45] J.C. SanMiguel, M. Escudero-Viñolo, J.M. Martínez, J. Bescós, “Real-time single-view video event recognition in controlled environments”, Proc. CBMI 2011.
- [46] C. Wolf, J. Mille, L.E. Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E. Dellandrea, C.-E. Bichot, C. Garcia, B. Sankur, “The liris human activities dataset and the icpr 2012 human activities recognition and localization competition”, Technical Report Technical Report RR-LIRIS-2012-004, LIRIS Laboratory, March 2012.
- [47] S. Suja, “Análisis de interacciones y actividades en entornos controlados”, Proyecto Fin de Carrera, Ingeniería de Telecomunicación, Universidad Autónoma de Madrid, Dec. 2012.
- [48] J.C. SanMiguel, J.M. Martínez, “A semantic-based probabilistic approach for real-time video event recognition”, Computer Vision and Image Understanding, 116(9):937–952, Sep. 2012.
- [49] S. Mukherjee, S.K. Biswas, D.P. Mukherjee, “Recognizing human action at a distance in video by key poses“, IEEE Trans. On Circuits and Systems for Video Technology, 21(9):1228 –1241, Sep. 2011.
- [50] I. Laptev, “On space-time interest points“, International Journal of Computer Vision, 64(2-3):107–123, Sep. 2005.
- [51] J. Yuan, Z. Liu, Y. Wu, “Discriminative subvolume search for efficient action detection”, Proc. of CVPR 2009.
- [52] P.M. Jodoin, V. Saligrama, J. Konrad, “Behavior Subtraction”, IEEE Transactions on Image Processing, 21(9):4244-4255, Sep. 2012