TEC2014-53176-R HAVideo (2015-2017)

High Availability Video Analysis for People Behaviour Understanding

D2.1 v3

People Behaviour understanding in single and multiple camera settings

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

AUTHORS LIST

Álvaro García Martín (agm) alvaro.garcia@uam.es

HISTORY

Version	Date	Editor	Description
0.9	7 December 2015	Álvaro García	Final Working Draft
1.0	11 December 2015	José M. Martínez	Editorial checking
1.9	14 December 2016	Álvaro García	Final Working Draft version 2
2.0	19 December 2016	José M. Martínez	Editorial checking
2.1	8 February 2017	Álvaro García	Third and Fourth semester updates
2.2	10 December 2017	Álvaro García	Fifth and Sixth semester updates
2.9	12 December 2017	Álvaro García	Final Working Draft version 2
3.0	15 December 2017	José M. Martínez	Editorial checking

CONTENTS:

1. Introduction

1.1. Motivation

This work package 2 (WP2) aims at developing the required tools, models and control signals in order to enable the development of adaptive and collaborative approaches for video-based understanding of people behaviour. In particular, the goal of this task is to provide the required video analysis tools to fulfil the objectives of the project. It considers both comprehensive related work studies and the development of novel approaches for long-term analysis.

This deliverable describes the work related with the task T.2.1 Analysis Tools for human behaviour understanding. The people behaviour understanding in this project has been already designed as a sequential combination of object segmentation, people detection, object tracking and behaviour recognition. In particular, during this three years of the project there have been a focus on developing different approaches for segmentation, people detection, tracking and behaviour recognition.

1.2. Document structure

This document contains the following chapters:

- Chapter 1: Introduction to this document
- Chapter 2: Object segmentation analysis tools
- Chapter 3: People detection analysis tools
- Chapter 3: Conclusions

2. Object segmentation analysis tools

2.1. Long-Term Stationary Object Detection Based on Spatio-Temporal Change Detection

Stationary Object Detection (SOD) has recently experienced extensive research [1] due to its contribution to prevent terrorist attacks by detecting abandoned objects [1] and illegal parked vehicles [2]. SOD aims to detect the objects in the scene that remain stationary after previous motion. Typically, a Background Subtraction (BS) algorithm extracts the objects and SOD decides whether they are stationary or not [3]. However, current BS algorithms present many shortcomings to label foreground and background regions in real situations [4], thus highly determining the SOD accuracy. In this paper [5] we propose a block-wise approach to detect stationary objects based on spatio-temporal change detection without using BS (see Figure 1).

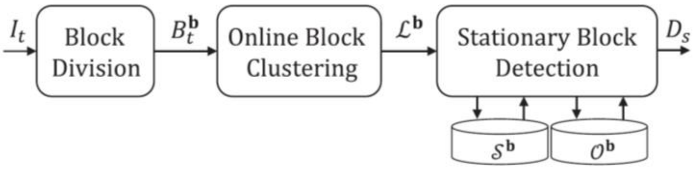

I_t
Block Division
$B_t^{\mathbf{b}}$
Online Block Clustering
$\mathcal{L}^{\mathbf{b}}$
Stationary Block Detection
D_s
$\mathcal{S}^{\mathbf{b}}$
$\mathcal{O}^{\mathbf{b}}$

Figure 1. Block diagram of the proposed approach.

Firstly, a Block Division stage decomposes each frame into non-overlapping $N \times N$ blocks $B_t^{\mathbf{b}}$ at each instant t, where $\mathbf{b}$ denotes the block location. Secondly, an Online Block Clustering stage models each location $\mathbf{b}$ over time, updating a cluster partition $\mathcal{L}^{\mathbf{b}}$. This stage handles the temporal adaptation to scene changes, by assigning each incoming block to one cluster of the partition or creating a new one. Only stationary blocks (i.e., without motion with respect to $B_{t-1}^{\mathbf{b}}$) are analyzed at this stage. This clustering provides robustness against illumination changes by considering pixel ratios at block level which groups blocks even if their illumination has changed. Finally, a Stationary Block Detection stage outputs a result image D_s with stationary objects (see Figure 2), where s defines the sampling instant each k frames. Data associated to the last stable cluster $\mathcal{S}^{\mathbf{b}}$, old stable clusters $\mathcal{O}^{\mathbf{b}}$ and the alarm time T is used to respectively detect the spatio-temporal stability changes, discard those changes caused by previously visualized clusters (i.e. empty scene or previous detections) and detect stationarity for changes longer than the alarm time.

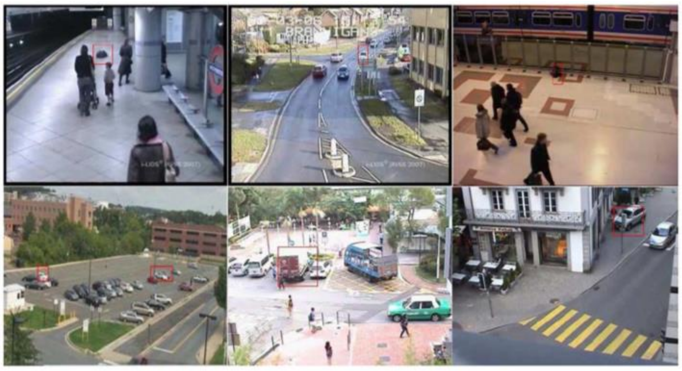


Figure 2. Examples of D_s image in different datasets. Detections are marked in red.

The last stage improves the state-of-the-art by reducing false alarms due to intermittent object motion and allowing to detect stationarity for objects not fully visible during T. Figure 3 presents an example of the scene analysis.

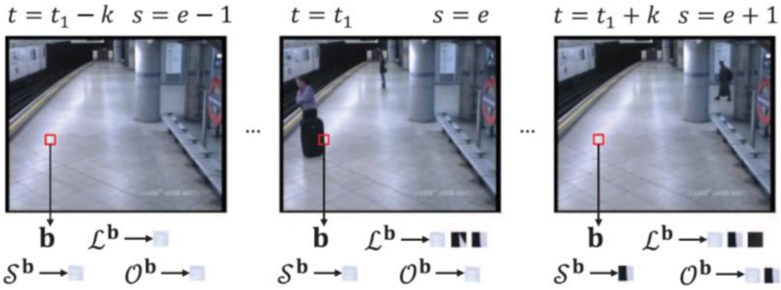

$t = t_1 - k \quad s = e - 1$
$t = t_1 \quad s = e$
$t = t_1 + k \quad s = e + 1$
...
...
$\mathbf{b}$
$\mathcal{L}^{\mathbf{b}} \rightarrow$
$\mathcal{S}^{\mathbf{b}} \rightarrow$
$\mathcal{O}^{\mathbf{b}} \rightarrow$
$\mathbf{b}$
$\mathcal{L}^{\mathbf{b}} \rightarrow$
$\mathcal{S}^{\mathbf{b}} \rightarrow$
$\mathcal{O}^{\mathbf{b}} \rightarrow$
$\mathbf{b}$
$\mathcal{L}^{\mathbf{b}} \rightarrow$
$\mathcal{S}^{\mathbf{b}} \rightarrow$
$\mathcal{O}^{\mathbf{b}} \rightarrow$

Figure 3. Example of the temporal analysis for a block location where the stability is modified changing from the empty scene to a suitcase.

2.2. Comparative evaluation of Points of Interest techniques for detection and description in images

We have decided to include the keypoint detection and description algorithms in this section. Although they are not exactly object segmentation approaches, their purpose is similar. Local descriptors [32] isolate and describe points of interest and feeds the other system stages: segmentation, object detection, tracking, etc.

One of the main objectives of this project [33] has been to develop a framework for the evaluation of keypoint detection and description algorithms, in order to update the references of the state of the art in this topic. Based on this framework, the other main objective has been to perform a comparative evaluation of the state of the art algorithms in local features field.

To this aim, three main stages have been faced. The first one has been to propose a new dataset, together with an evaluation methodology, based on an analysis of the strengths and weaknesses of previous frameworks in the state of the art. This proposal will set a new evaluation framework for the following stages.

The second one, an exhaustive study of the state of the art allowed to select the main techniques and categorize them according to its properties.

Finally, those selected techniques were tested on the proposed evaluation framework.

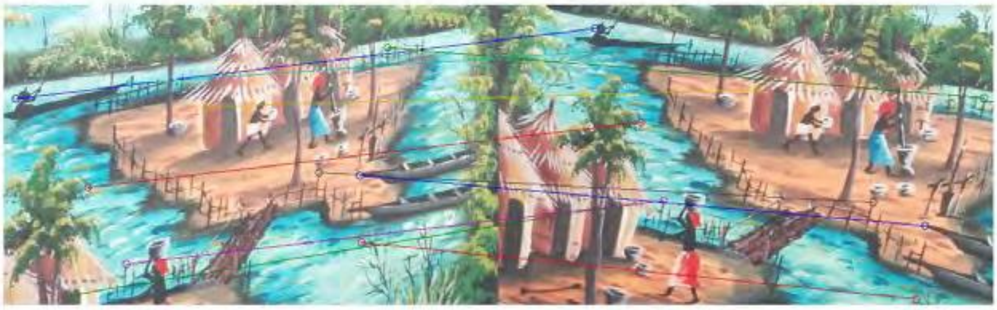


Figure 4. Visual example of keypoint detection, description and matching between two images with different point of view.

3. People detection analysis tools

3.1. People density estimation in crowded environments

Currently, the use of artificial vision systems has acquired great relevance due to the advancement of digital image and video processing technologies and the cheapening of capture tools. The crowd density estimation as part of artificial vision systems has an important niche market in video-surveillance. The crowd density estimation is an important tool to detect abnormal situations in public places such as fights, disturbances, violent protests, panic or congestion. Density information could be also helpful for creating a business strategy according to the distribution of people in public places or shopping centers and the distribution of people over time.

So far, a large number of crowd density estimators has been implemented. A significant part of these estimators uses background subtraction and extract features from foreground pixels [11] [12]. All of them use foreground-background segmentation getting good results for the studied scenarios. This work [8][22] introduces the use of people-background [13] segmentation for crowd density estimation. With the goal of comparing both types of segmentation, one algorithm of the state of the art for density estimation has been implemented and then, results for both types of segmentation, foreground-background and people-background segmentation, has been compared in several scenarios (see Figure 5 and Figure 6).

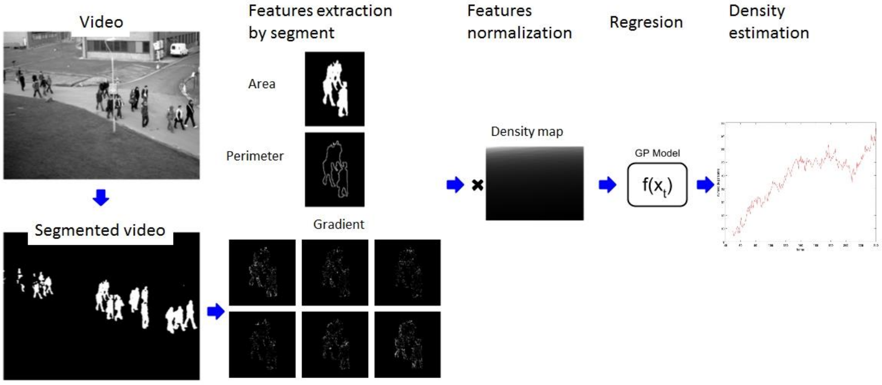
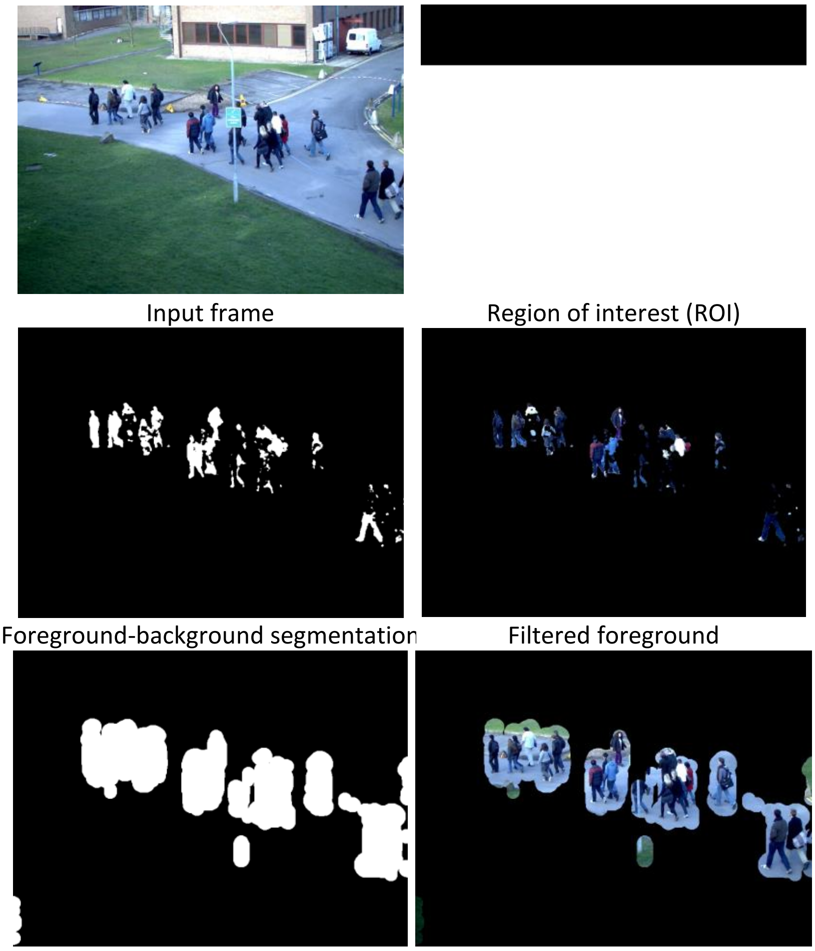


Figure 5. People density estimation system based on foreground-background segmentation.



Input frame

Region of interest (ROI)

Foreground-background segmentation

Filtered foreground

People-background segmentation

Filtered foreground

Figure 6. Examples of foreground-background segmentation and people-background segmentation.

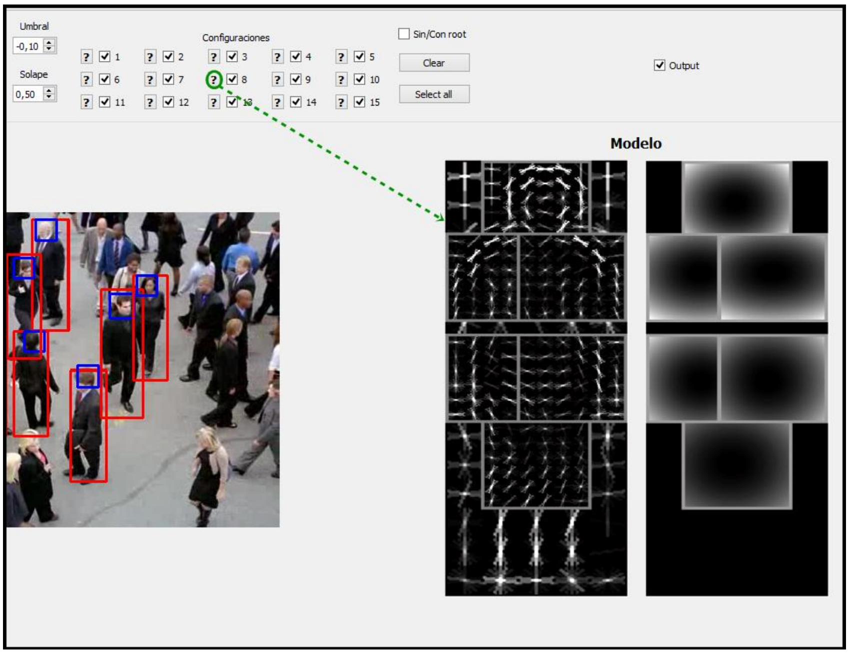
3.2. People detection in groups

In video signal processing, people detection is one of the most difficult tasks, even though there are some algorithms that give good results. However, when the detection takes part in complex environments the quality of the performance decreases, that is why the aim of this work [9] is the implementation of a people detection in groups algorithm in C++, as well as the integration on the proprietary video analysis platform called Distributed Video Analysis (DiVA). This platform allows us the execution with online and off-line videos, with the objective of verifying the functionality and efficiency in crowded environments.

The algorithm implemented, known as Multi-configuration Part-based Person Detector [14], is an adaptation of the algorithm known as DTDP (Discriminatively Trained Part Based

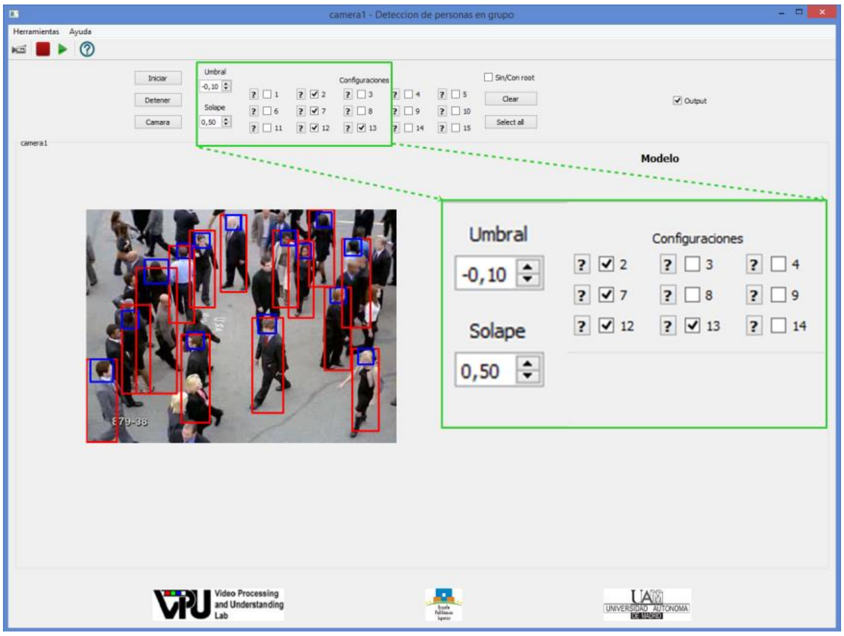
Models) [15] to improve the detection in these types of environments. It is founded in an exhaustive search of person models, which are formed by the mixture of different body parts. Thanks to this search it obtains good results despite of the processing time, therefore the analysis of this algorithm is not possible in real time.

In order to complete our goal, we have developed a user interface so that any user can interact with the algorithm and prove easily the functionality of the algorithm in different environments as well as test the efficiency of the different models defined (see Figure 7 and Figure 8).



Umbral -0,10
Solape 0,50
Configuraciones ☑ 1 ☑ 2 ☑ 3 ☑ 4 ☑ 5 ☑ 6 ☑ 7 ☑ 8 ☑ 9 ☑ 10 ☑ 11 ☑ 12 ☑ 13 ☑ 14 ☑ 15
□ Sin/Con root
☑ Output

Figure 7. Visual example of the body parts configuration including six body parts, named configuration 8, following [16].


camera1 - Deteccion de personas en grupo
Herramientas
Ayuda
Iniciar
Detener
Camara
Clear
Select all
camera1
Modelo
E79-38
Video Processing and Understanding Lab
UAM UNIVERSIDAD AUTONOMA DE MADRID

Umbral -0,10

Solape 0,50

Configuraciones

? ☐ 1 ? ☑ 2 ? ☐ 3 ? ☐ 4 ? ☐ 5
? ☐ 6 ? ☑ 7 ? ☐ 8 ? ☐ 9 ? ☐ 10
? ☐ 11 ? ☑ 12 ? ☑ 13 ? ☐ 14 ? ☐ 15

☐ Sin/Con root

☑ Output

Figure 8. User interface example with selected configurations 2,7,12 and 13, following [16].

3.3. People detection in presence of groups

In this work [9] we address one of the most typical problems of people detection in presence of groups of people: in this kind of scenarios, traditional people detectors have difficulties dealing with several occlusions. In order to deal with this problem, we propose the use of two different hierarchies. The first one consists of a hierarchy of people, i.e., the use of the detections of different people belonging to a group in order to refine the individual's detections. The second one consists of a hierarchy of parts [14], i.e., the use of different combinations of body parts in order to refine the final detection.

Our main aim is to be robust to different groups configurations, camera point of view, scene constraints, etc. Therefore, we propose to update this hierarchies structures frame by frame, so we can adapt the detection system to specific scene variations.

In order to update both the hierarchy of people and the hierarchy of parts, we study the detection results and determine which have been the most typical or representative configurations over time (scales and body parts configurations). Using only the last most representative configuration, we are able to reduce the computational cost and false positive detection and therefore the global detection results.

3.4. People detection in residential and hospitalary environments

There is a large demand in the area of video-surveillance, especially in detecting people, which has caused a large increase in the number of researches in this field. For elderly people detection, the detector must have into account different positions such as sitting or in a wheelchair. Also is important the cost involved in making these detectors.

Therefore, this work [16] has two main objectives. The first has been to develop a sitting person model with the aim of completing a detector for a nursing home scenario. The second one was based on reducing the amount of resources needed and save the cost of having to record sequences for a detector in this scenario. To achieve this, three synthetic images dataset were created in order to perform three different models, evaluating which model is optimal and finally analyzing its feasibility by comparing it with the people detector in wheelchairs. Figure 9 shows an example of each image dataset.

The result, as expected, is worse than the performance of the trained detector with real images, but a functional detector has been obtained without having to record the real object. This method can be useful in situations where there it is not possible to record a dataset of the desired object type, or in cases in which obtaining it has a high cost. Other examples in which this technique is applicable, considering people detection, could be for people riding horses or people in the supermarket with shopping carts.

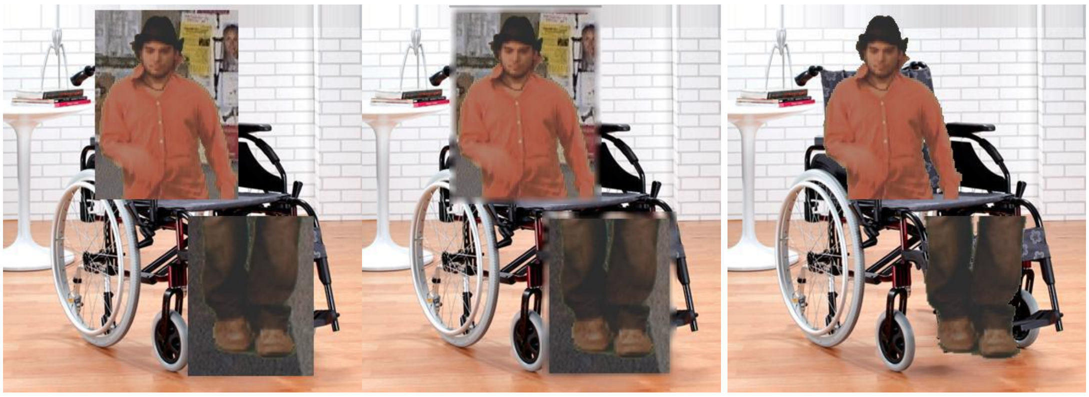


Figure 9: Image example of each created dataset.

3.5. Traffic Signs Detection and Recognition

Lately, a huge evolution has taken place within the automotive world. We can find many improvements in this area, being the search of the autonomous driving the most important one. Despite the fact that there still remain a lot of years of progress to encounter autonomous driving in the street, more and more vehicles are now equipped with systems providing some useful information to the driver. These pieces of information are taken from the road itself.

Some new companies and specific departments within car brands have arisen in search of this very general target, creating systems that, in the ordinary usage, are able to perform autonomous driving. Besides, these systems are starting to be sold as ADAS (Advanced Driver Assistance Systems), allowing their testing and upgrading in a real driving environment. One of the above said systems, called TSR (Traffic Sign Recognition), consists in the recognition of the detected traffic sign.

This work [35] aims to obtain a first implementation of an own algorithm of vertical traffic signs detection and recognition, (see Figure 10). This detection is not easy due to the changing conditions of the road and the circulation, such as weather or light changes or the difficulty to receive traffic signs from a moving car.

Although other detection and recognition systems use more than one camera or sensor, we will be making use of just one camera, trying to create a system whose best quality be its adaptive capacity to the environment. This way, it could be included in a modern car software or in a Smartphone and then used from a car dashboard or a motorbike wind deflector. The system aspires, not to stand out among the ones described in the State of the art section, but to propose a simple solution to this existent question.

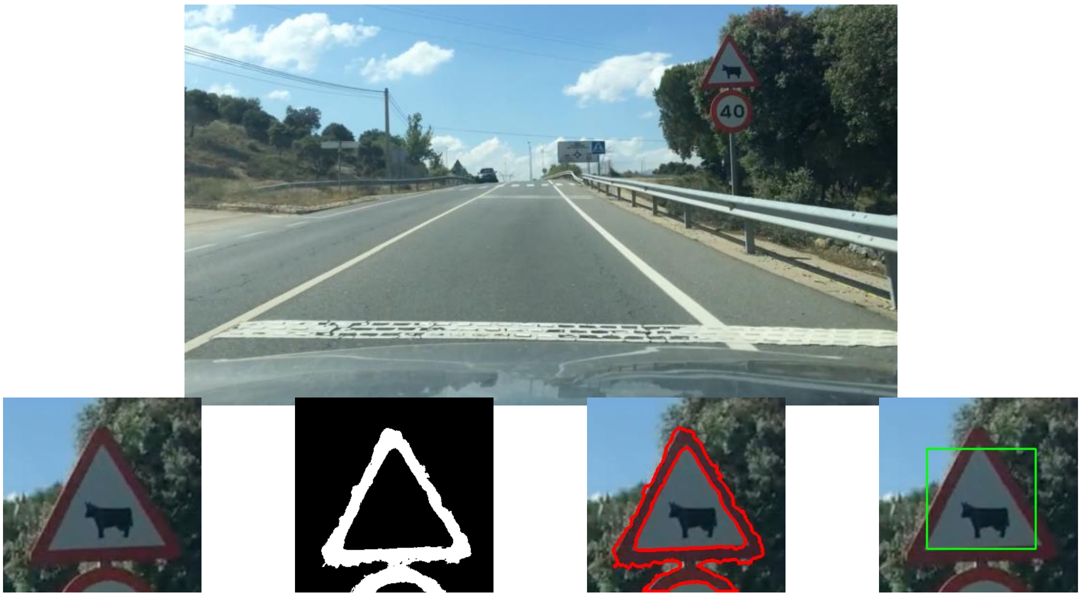


Figure 10: Traffic signs detection and recognition example. Input image, signal detection, color segmentation, edge extraction and sign recognition.

3.6. People Detection using Convolutional Neural Networks

Nowadays we live in a period in which the people detection is on the rise, being very important in some aspects of society like video security. As a matter of fact, many researchers create their own algorithms for object detection obtaining excellent results over databases chosen by themselves. But the problem arises, mainly, at the time of generating a model that includes as much characteristics as possible, to make it capable of overcoming changes in posture, movement, illumination, the interaction among people or changes in the point of view. To train the models, is becoming more popular the usage of a new kind of learning, know as Deep Learning, this handles huge amounts of data that solves those problems trying to simulate the behaviour of neurons of the human brain. New algorithms are based in Deep Learning for people detection, which are named as detectors based on convolutional networks.

With all this being said, the aim of this work [36] is to offer a comparative between the traditional algorithms of people detection (Deformable Parts Model, DPM, and Aggregate Channel Features, ACF) and the modern ones based on convolutional networks (Faster Region-based Convolutional Network, FASTER R-CNN), studied in the state of the art. For the achievements of this, different models of people will be designed and implemented and the obtained results will be evaluated by the chosen detectors over a database and metrics in common, in equal conditions.

3.7. Object-Background Segmentation

This work [37] aims to develop an object-background segmentation capable of generating segmentation masks where the images can be divided into the element to be segregated and the remaining ones. This has been achieved through the generalization of the previous work from which we proceed, the person-background segmentation, whose main feature is the use of detection in order to generate such masks. The difference to be highlighted between the two segmentation is the model they use for detection. In the person case, the model is always the same with a single pose, while the object-background segmentation can use any model with an indeterminate number of poses (see Figure 11). To accomplish our system, two application procedures will be analyzed in this work: one method in which the independent parts of the object model can be detected, and another one in which these parts could be combined. Finally, the results of both methods will be put to the test, and the system will be assessed.

To obtain the tests and validation of our algorithm it is necessary to conduct a selection of sequences of images or datasets and of a set of object models to detect whether they are useful for their application to our system or not.

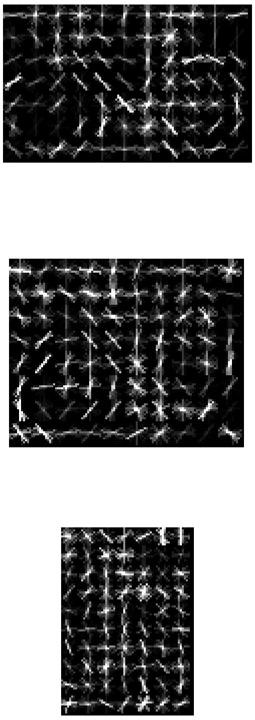


Figure 11: Example of bike model including three different poses: rear, semi-rear and front view.

3.8. Platform for People Detection Algorithms Evaluation

The objective of this work [39] has been to develop an automatic web platform for the evaluation of algorithms for detecting people. The detection algorithms are currently booming, in this case the topic of people detection has been chosen. First, it has been studied of the platforms ChangeDetection.net (CDNET) and People Detection Benchmark repository (PDbm) of the Video Processing Understanding Lab (VPULab). Both platforms are oriented to the evaluation of algorithms. After studying both platforms, it was decided to automate the PDbm based on CDNET. The design of the system has been made in four functional blocks: Web server, database, evaluation application and mail manager. In addition, we adopt the layered programming model on the web server (see Figure 12).

During the development of the project, several specific technologies have been used for each block: Windows 7, Apache, MySQL, PHP, Matlab, Outlook. Subsequently, a control module is implemented for the platform administrator. Then, platform is tested in a local pre-production space and verified to be working properly. Finally, we analyze the possibility of continuing to develop the platform in a modular way due to the design of the system.

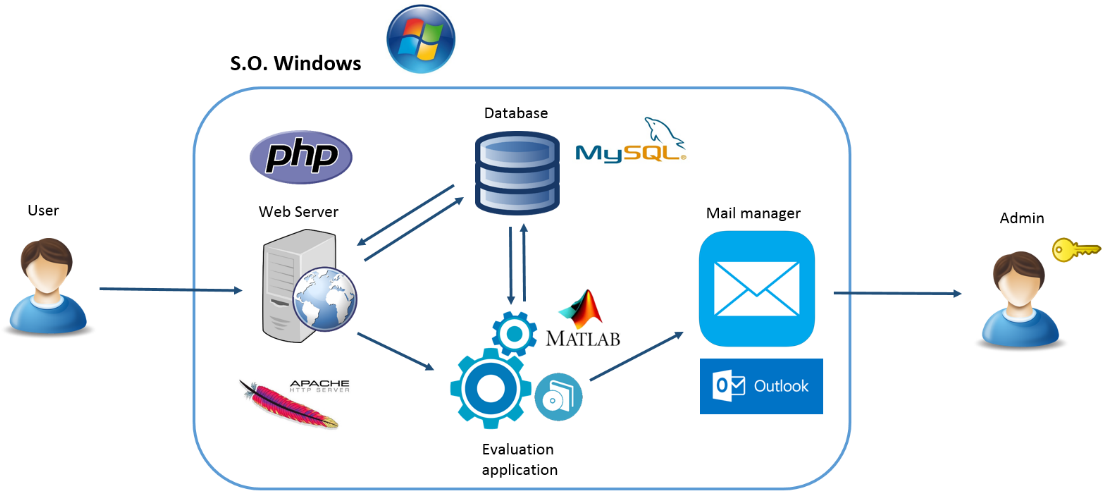


Figure 12: Platform for People Detection Algorithms Evaluation: System overview.

4. Object tracking analysis tools

4.1. Object tracking based on TILDE points

The aim of this work [17] is the design and development of a video object tracker based on point-of-interest (PoI). Specifically, we focus on the potential benefits of including a recently published PoI detector in the core of the tracking process. The design of the algorithm starts by studying existing tracking algorithm emphasizing in methods using PoI in any stage of their tracking algorithm. In order to provide a flexible framework on which to develop potential improvements, the algorithm's design builds on the basic PoI-based tracking scheme. From the state-of-the-art, we propose to use TILDE (Temporally Invariant Learned Detector [18]) as the PoI detection method. TILDE is a train-based PoI detection method which is claimed to provide stable results to change in illumination and appearance. TILDE-driven correspondences are used to spatially constrain the target position between consecutive frames. Final result is refined by means of a classic cross-correlation method. The tracker is experimentally evaluated in a generic evaluation corpus. From this evaluation we aim to discuss on their benefits and drawbacks. Furthermore, we also compare the tracker against state-of-the-art trackers, in order to quantitatively contextualize its operation. Experimental results (see Figure 13) indicate that the designed tracker performs significantly better than a common references (Normalised-Cross-Correlation ncc [20]) and slightly worse than a deep-learning-based approach (MDnet [19]); hence, results partially validate the design and development of the algorithm and suggest that the use of TILDE may help to generate robust constraining schemes for trackers based on cross-correlation.

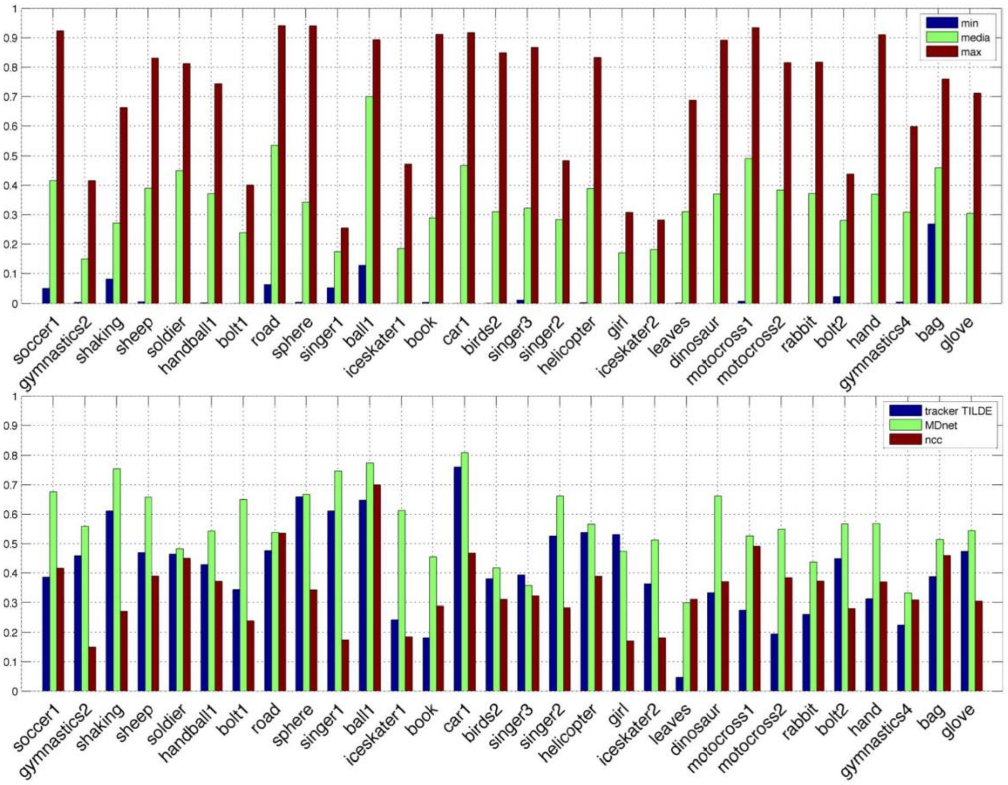


Figure 13: Comparative evaluations in terms of tracking accuracy of the designed tracking approach (tracker TILDE [18]). Top. Operation extrema of the proposed method when analyzing the VOT2016 dataset. Bottom, comparison with a reference method (ncc [20]) and the winner of the competitions, a deep-learning method (MDnet [19]).

4.2. Visual Attention Based on a Joint Perceptual Space of Color and Brightness for Improved Video Tracking

This work [23] proposes a new visual attention model based on a joint perceptual space of both color and brightness, and shows that this model is able to extract more discriminant visual features, especially when dealing with objects that are very similar visually. That joint color and brightness space is based on a biologically inspired theoretical perceptual model originally proposed by Izmailov and Sokolov [24] in the scope of psychophysics. The present paper proposes a computational model that allows the application of Izmailov and Sokolov's theoretical model to digital images, since the original model can only be applied to perceptual data directly drawn from psychophysical experiments. Experimental results with real video sequences show that the proposed visual attention model yields significantly more accurate results in the particular application scope of video tracking than well-known visual attention models that process color and brightness separately.

4.3. Automatization of functions for teacher tracking in lectures broadcasting

The main motivation behind this Project [27] has been to automate the process of tracking a teacher in a classroom so as to be able to broadcast classes via the internet to students who, for various reasons, can't physically attend the class that is being taught.

This project is a continuation of several previous works, as a result of which there is an algorithm that uses images from a video sequence taken by a fixed camera to track, in real time, the position of a teacher in a classroom, and guides in that direction a mobile camera whose video signal is the image the student will see. Furthermore, there is a web application that allows users to view these classes (see **Figure 14**).

The integration of an HOG people detector in the original algorithm is detailed in this project. This detector is useful so as to be able to automate the initialization phase, and it allows a faster recovery of the target once the algorithm determines that it has lost it.

Once the changes that have been made to the tracking algorithm have been explained, two new proposals that improve the mobile camera movements are introduced. With these proposed solutions better movement control and zoom levels are obtained, making seem as if a person is controlling the mobile camera. Of these approximations, one is based on the original rule scheme, while the other attempts to predict the target's future position using a Kalman filter [28].

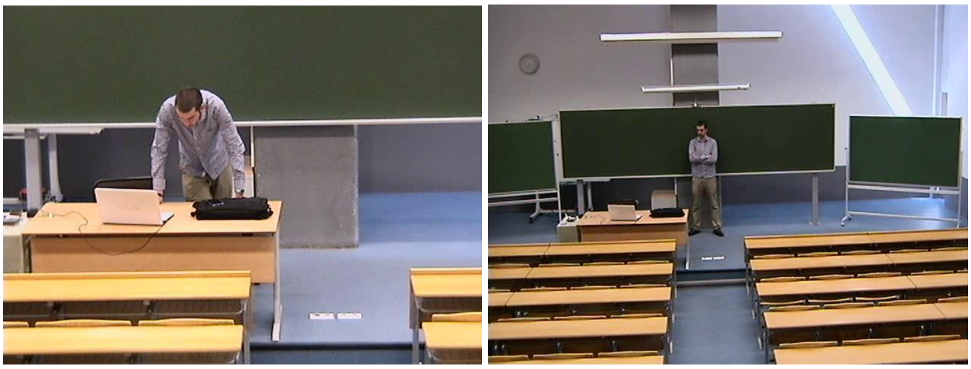


Figure 14: Visual examples of the teacher tracking system.

4.4. Long-term objects tracking in video sequences

In this master thesis [32] an analysis of tracking objects in long-term sequences is proposed. Recently, the development of video tracking algorithms has been focused on short videos. However, the need to operate for long times (e.g. 24/7 video-surveillance) have increased need to study mechanisms to improve and update existing tracking algorithms for their use in long-term sequences.

The main aim of the work is the study, design and evaluation of an algorithm that combines other trackers sequences previously developed both short and long term. For this objective, first it has conducted a study of the state of art related to object tracking, focused on the case of long-term videos. After, this project focuses on the selection and description of the chosen tracking algorithms to evaluate and compare the set of videos of this project. Once these trackers have been studied, a fusion algorithm is implemented which examines the behavior of the combination of algorithm under the long-term framework.

Figure 15 shows one example of different trackers and the proposed combination.

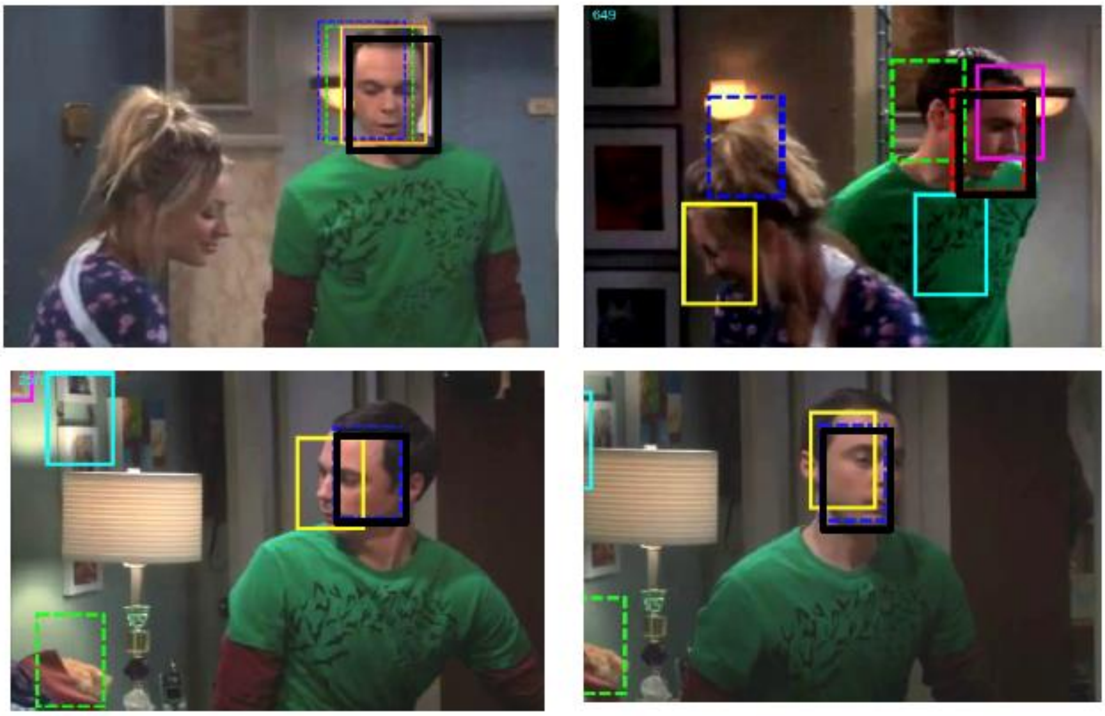


Figure 15: Visual examples of different SoA trackers and the proposed combination (in black color).

5. Behaviour recognition analysis tools

5.1. Integration and evaluation of abandoned-stolen object detection systems in security-video

This final degree project [21] (proyecto final de carrera) proposes a configurable abandoned-stolen object detection system in security-video that integrates the most relevant techniques in each one of its stages. A formalization of the problem is presented, followed by a description of the different analysis stages required for the detection. Firstly, this work analyses the state of the art to know the present date problems about the matter. Secondly, the work focuses on the integration of the most recent and relevant algorithms of the literature in every single phase of the system. It also designs the necessary interfaces for its execution in a sequential order (see Figure 16).

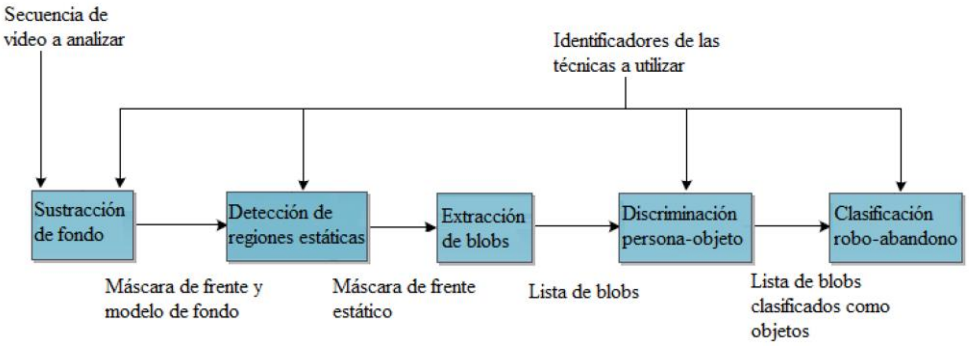

Secuencia de video a analizar
Identificadores de las técnicas a utilizar
Sustracción de fondo
Detección de regiones estáticas
Extracción de blobs
Discriminación persona-objeto
Clasificación robo-abandono
Máscara de frente y modelo de fondo
Máscara de frente estático
Lista de blobs
Lista de blobs clasificados como objetos

Figure 16: Block diagram of the proposed system for evaluation

To conclude, the different configurations of the system regarding the detection of static regions as well are evaluated and compared, while it is classified as abandoned-stolen about a compound of heterogeneous videos sequences. Abandoned/stolen object event detection depends on the parameters that modulate the absorption of the blobs from the static foreground by the background model. The results of abandoned/stolen detection stage are affected by the propagation of errors in the earlier stages as it is the last stage of video analysis system (see Figure 17).

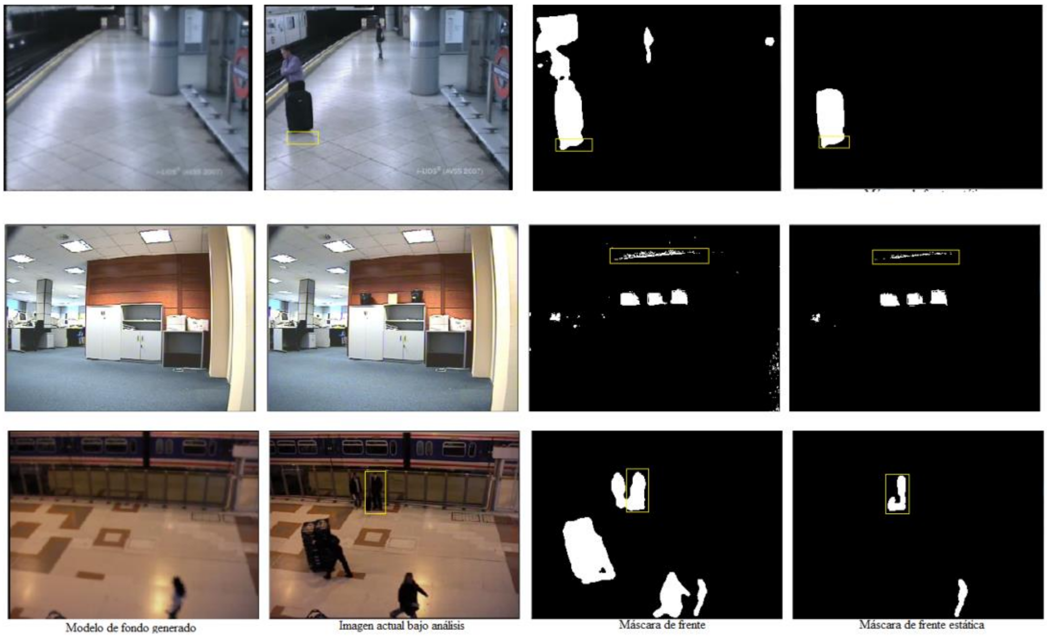


Figure 17: Examples of the output generated by the different configurations of the system.

5.2. Fall detection using video

This Master Thesis Project [25] consists on studying the development of a fall detection video-based system, primarily intended for implementation in home environments to promote independent living for the elderly. Due that falls are one of the main problems of the elderly population, we are in a field with great potential still developing. In the first place, we conducted a comprehensive study of the art of existing methods in the detection of falls in general, as well as exclusive video analysis algorithms. Once detailed the techniques, we chose the one that best fits the needs of our project and provides reasonable results in the detection of falls. Then, we have implemented an algorithm that characterizes the chosen technique. The algorithm is evaluated by a different set of videos with different features and various approaches that allow us to draw conclusions. Figure 18 shows visual examples of the fall detection system.

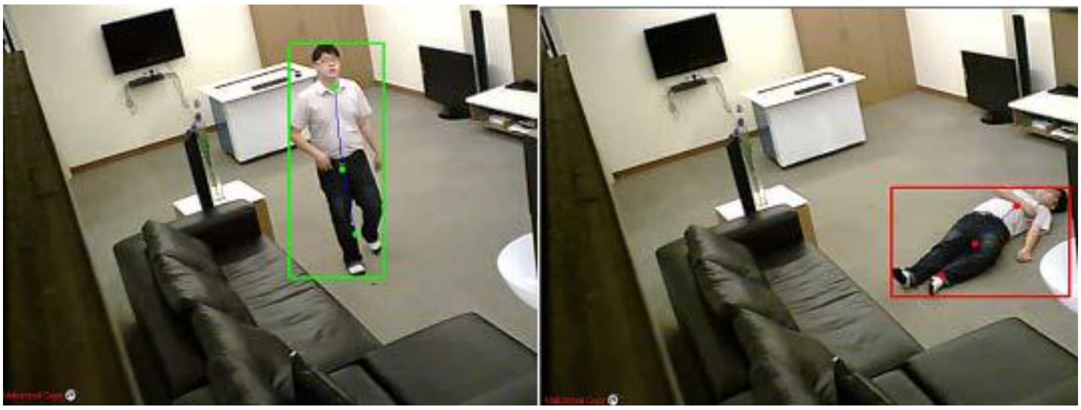


Figure 18: Visual examples of a normal detection (green blob on the left) and a fall detection (red blob on the right).

5.3. Activity analysis in multicamera sports videos

Sport video-content analysis systems are on the rise both from the commercial view point and the researching viewpoint. In this scope, the video processing group (VPU-Lab) developed a prototype for sport video-content analysis previously to the beginning of this master thesis. This prototype performs the detection and tracking of players in sport videos, and provides statistical information about their behaviour.

This prototype achieved good results in quantitative and qualitative terms, presenting some deficiencies which motivate this mater thesis. These deficiencies were mainly related to aspects as usability, system interaction, results visualization and fine-tuning.

This project [26] has been focused in providing a solution for those problems by working on three main tasks. Firstly, the work focused in compacting the prototype. In origin it was divided in modules which have to be manually linked and which are programmed in different languages. As a result of the project there is a unified prototype fully programmed in C++, full working and portable.

Secondly, efforts were aimed to improve the usability, interactions and results visualization of the prototype. Two applications were developed and adapted to guarantee specific sports support, football and tennis. They allow a non-expert user to fully control the prototype and visually obtain its results, via a Graphical User Interface (GUI) (see Figure 19).

Finally, keeping in mind that this products use to work under supervision in commercial applications, the prototype and the interface have been equipped with tools to allow the online interaction with its results. This improvement allows a supervisor to control the application and correct its results when necessary, obtaining more reliable results than any other automatic system.

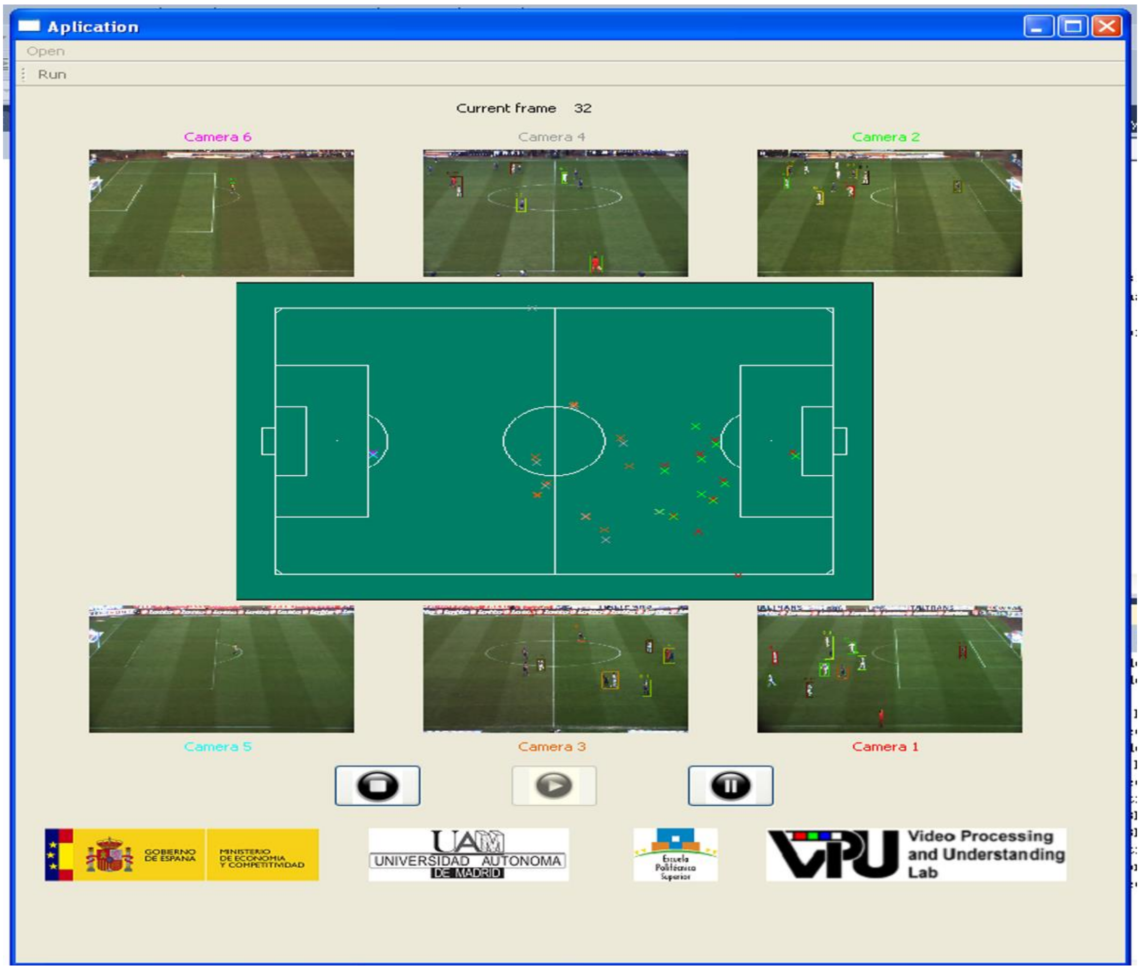

Aplication
Open
Run
Current frame 32
Camera 6
Camera 4
Camera 2
Camera 5
Camera 3
Camera 1
GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
UAM UNIVERSIDAD AUTONOMA DE MADRID
Escuela Politécnica Superior
Video Processing and Understanding Lab

Figure 19: Graphical User Interface (GUI) application.

5.4. Activity analysis in basketball videos

The video processing group (VPU-Lab) developed a prototype for sport video-content analysis previously to the beginning of this master thesis [26]. This prototype performs the detection and tracking of players in sport videos, and provides statistical information about their behaviour. The application had been created both for individual and collective sports, but in this last case, it only worked for soccer videos. Due to this, a new line of investigation arose, that consisted in the adaptation of this prototype to another sport such as basketball [31].

Therefore, this Project [31] has worked among the same lines, solving the existing issues and creating a similar prototype on which to continue investigation and further development. Following this, the Graphical User Interface (GUI) has been modified, (see Figure 20), to be able to optimally represent the newly obtained results (basketball), maintaining the full functionality of the previous prototype.

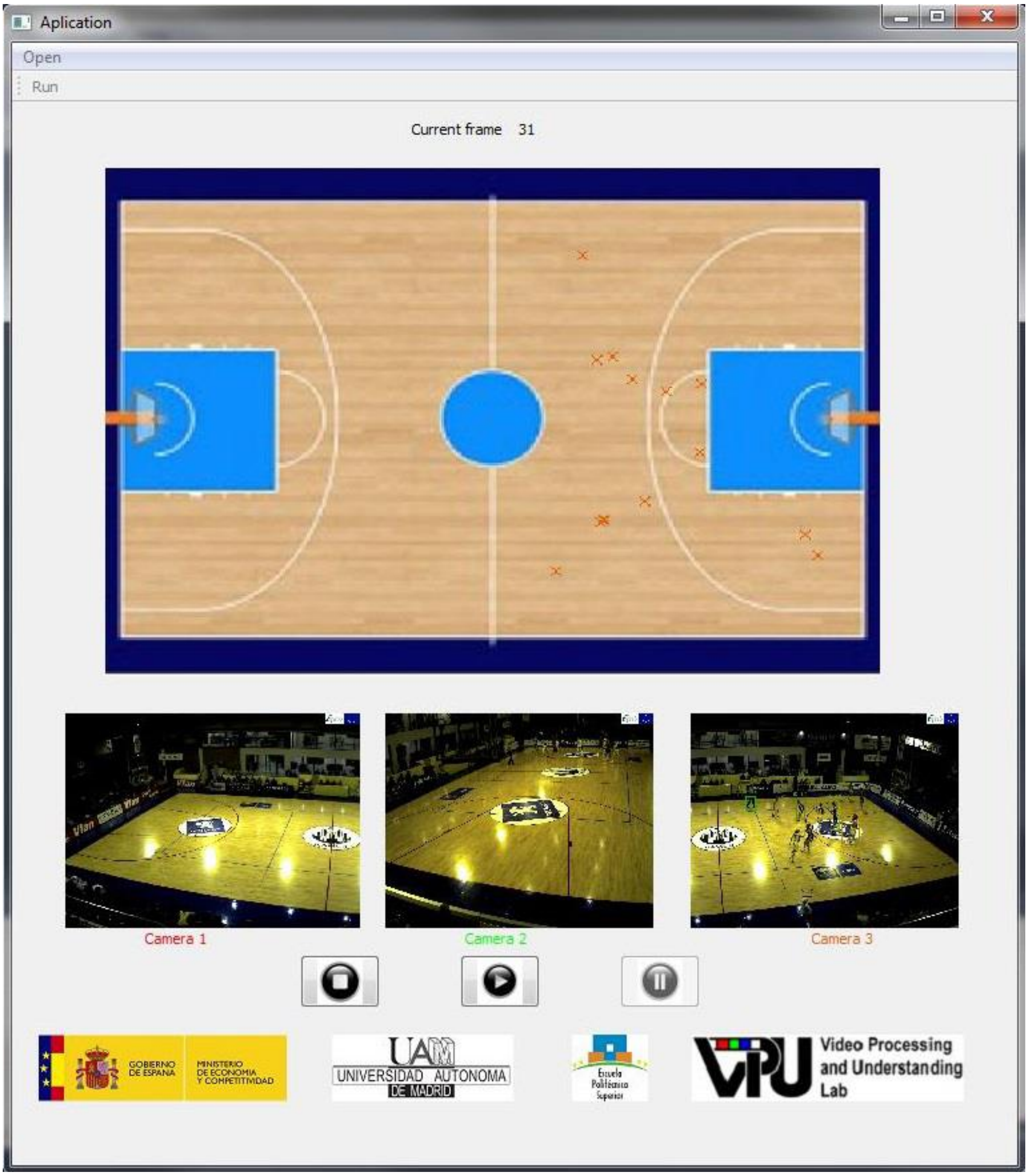
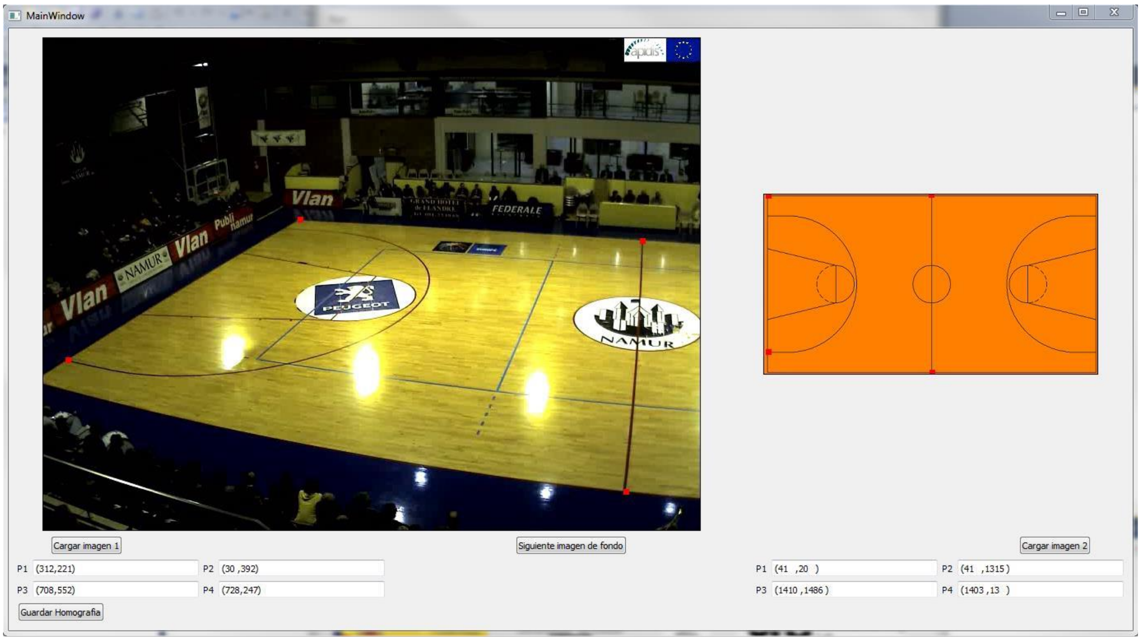

Aplication
Open
Run
Current frame 31
Camera 1
Camera 2
Camera 3
GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
UNIVERSIDAD AUTONOMA DE MADRID
Escuela Politécnica Superior
Video Processing and Understanding Lab

Figure 20: Basketball Graphical User Interface (GUI) application.

Furthermore, an improvement of the obtained results in these new videos was researched, with the system that was being implanted. The goal was to obtain the best possible results in both detection and tracking by adjusting different parameters.

Lastly, due to the interface being the visible part of this project, there have been updates made to improve it, by including a new homography creation module that simplifies the use by the end user even more, (see Figure 21).



P1 (312,221) P2 (30 ,392)
P3 (708,552) P4 (728,247)

P1 (41 ,20) P2 (41 ,1315)
P3 (1410 ,1486) P4 (1403 ,13)

Figure 21: Homography creation module Graphical User Interface (GUI).

5.5. Automatic classification of videos using features descriptors

This work aimed to obtain a classifier capable of discriminating between violent and non-violent videos using only the information of the frames of the video sequences, without additional information such as audio, subtitles or any other context information. In order to achieve this, different tools and algorithms of computer vision have been used, whose performance has increased in recent years thanks to the automation of processes like the one that is treated in this project.

To achieve this goal, a detailed study of the state of the art has been carried out. Within this study we have reviewed methods that pursue the same objective as ours, such as ViF, MoSIFT or LaSIFT. There has also been a previous study of the methods of detection and description of interest points as well as the basic concepts of motion extraction between frames of a video and the algorithms existing for that extraction.

After this study, and due to the good results obtained in its publication, it was decided to use the algorithm SP-SIFT developed in the VPULab of the Universidad Autónoma de Madrid. In this way we could measure the performance against its predecessor algorithm, SIFT, in a different task than the recognition of objects in static images. In the same way, thanks to the collaboration of the Technische Universität Berlin, it was decided to use the Lagrangian theory in estimating the motion between the frames of each video.

In summary, the algorithm developed [34] can be divided into three clearly differentiated phases, (see Figure 22**Figure 22**: Blocks diagram.). In the first one, the interest points of each frame will be extracted and described, and the movement of the scene will be estimated and described by the use of the direction Lagrangian measures. In order to measure the performance in a comparative way SIFT and SP-SIFT will be used to describe both the appearance and the motion. In the second phase of our algorithm will be used the Bag of Words (BoW) model to extract a representative set of appearance and movement descriptors. From them will be calculated a histogram of appearance frequencies of these descriptors by each one of the videos. Thus we will be able to reduce the dimensionality of the problem by facilitating the classification task. In the third and final phase of the algorithm the previous histograms will be used to train a classifier capable of labelling a new video entry as violent or non-violent.

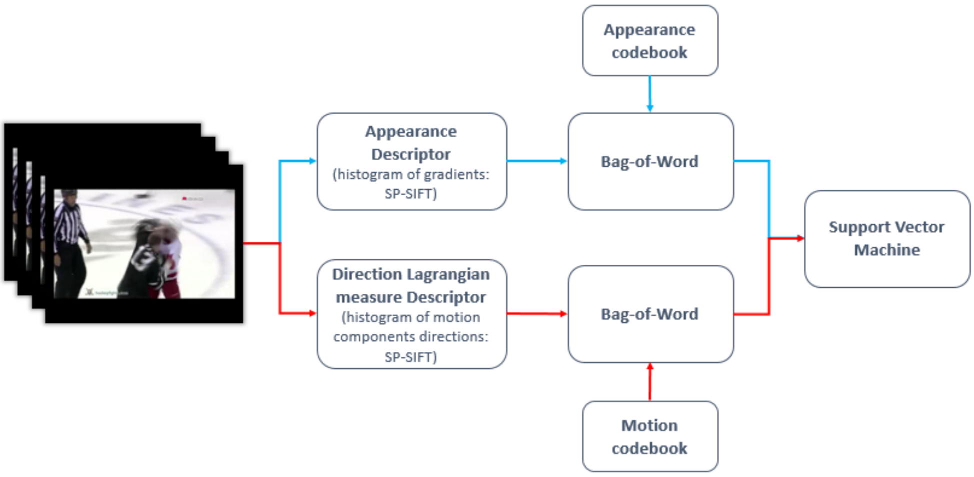

Appearance codebook
Appearance Descriptor (histogram of gradients: SP-SIFT)
Bag-of-Word
Support Vector Machine
Direction Lagrangian measure Descriptor (histogram of motion components directions: SP-SIFT)
Bag-of-Word
Motion codebook

Figure 22: Blocks diagram.

To evaluate the performance of the algorithm, the cross-validation method has been used in order to obtain reliable results, (see Figure 23). The main measure of error used is the percentage of success in the test phase. The matrices of confusion have also been extracted to know more accurately the procedence of the mistakes made.

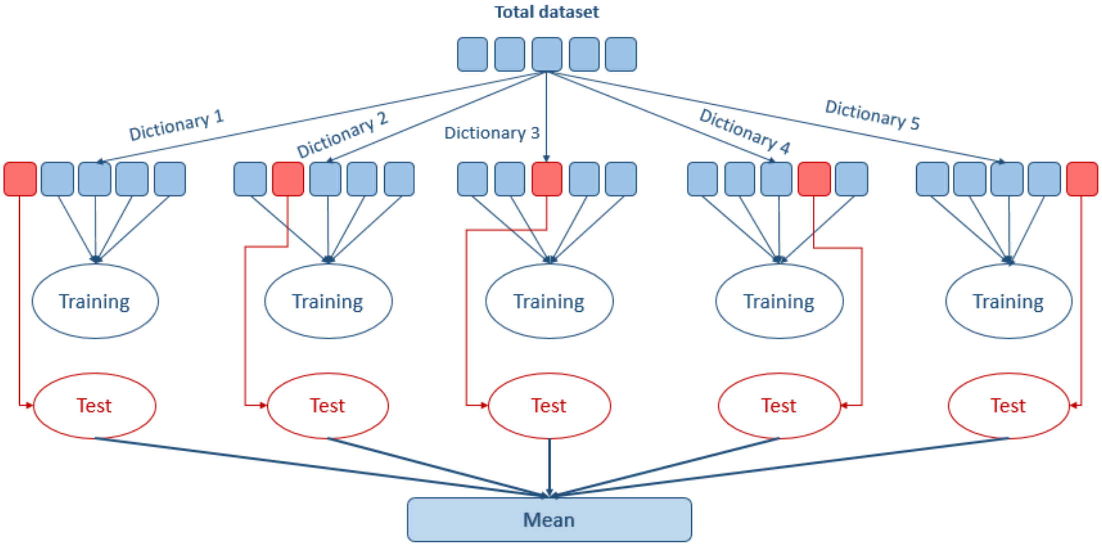


Figure 23: Cross-validation.

As for the conclusions, we have that SP-SIFT gets a better performance in the task of describing the appearance of the videos, but SIFT improves in the description of the movement. The Lagrangian measures are able to improve the results obtained by the optical flow fields because they are able to take into account more than two frames. The integration parameters of these measures - number of frames that are considered for the calculation of the motion - play a fundamental role in the performance of the classifier obtaining the best results when its value is of 5-6 frames.

6. Conclusions

In relation with segmentation, a long-term stationary object detection based on spatio-temporal change detection has been implemented and evaluated [7], we propose a block-wise approach to detect stationary objects based on spatio-temporal change detection without using background subtraction. In addition, an evaluation framework and a comparative analysis of state of the art local features detection and description algorithms has been developed [33].

In relation with people/object detection several approaches have been implemented and tested: people density estimation in crowded scenarios [8], people detection in groups [9], [10], people detection in residential and hospitalary environments [16], people detection using convolutional neural networks [37], a generic object-background segmentation based on detection [38], a platform for people detection algorithms evaluation [39] and a basic traffic signs detection and recognition approach [36]. A significant part of people density estimation approaches from the state of the art use background subtraction and extract features from foreground pixels. All of them use foreground-background segmentation getting good results for the studied scenarios. The proposed work [8], introduces the use of people-background segmentation for crowd density estimation. On the other side, [9] and [10] propose two different approaches in order to deal with people detection in crowded scenarios or in presence of groups of people. [16] proposes a sitting person model with the aim of completing a detector for a nursing home scenario and explore the possibility of creating synthetic images datasets reducing the amount of resources needed and save the cost of having to record sequences for a detector in this specific nursing home scenario. [37] presents a comparison between people detection approaches based in hand craft features versus those based on convolutional neural networks. [38] presents a generalization of a previous people-background segmentation approach [40] to any object model: a generic object-background segmentation. [39] presents a automatic benchmarking platform for people detection algorithms evaluation. Finally, [36] propose a basic traffic signs detection and recognition approach based in signs shapes and colour.

In relation with tracking, a video object tracker based on the point-of-interest TILDE has been implemented and evaluated [17]. TILDE-driven correspondences are used to spatially constrain the target position between consecutive frames. The final result is refined by means of a classic cross-correlation method. Also, a new visual attention model based on a joint perceptual space of both color and brightness for improved video tracking is proposed [23]. In addition, an enhanced version of a teacher tracking system in a classroom has been implemented and tested [27]. And finally, a fusion algorithm [32] has been implemented which examines the behavior of the combination of different tracker from the SoA under a long-term framework.

In relation with behaviour recognition, a configurable abandoned-stolen object detection system in security-video that integrates the most relevant techniques in each one of its stages is proposed [21]. Also, a fall detection video-based system to promote independent living for the elderly is proposed [25]. A unified prototype Graphical User Interface (GIU) for sport video-content analysis has been implemented (soccer [26] and basketball [31]). In addition, an automatic classification of videos using features descriptors (SIFT, SP-SIFT and Lagrangian) has been implemented [35].

7. References

[1] D. Ortego, J. C.s San Miguel, J. M. Martínez, “Long-Term Stationary Object Detection Based on Spatio-Temporal Change Detection”, IEEE Signal Processing Letters, 22(12), 2368-2372, Dec. 2015.

[2] Q. Fan, P. Gabbur, and S. Pankanti, “Relative attributes for large-scale abandoned object detection”, in Proc. IEEE Int. Conf. Computer Vision (ICCV), Dec. 2013, pp. 2736–2743.

[3] K. Lin, S. Chen, C. Chen, D. Lin, and Y. Hung, “Abandoned object detection via temporal consistency modeling and back-tracing verification for visual surveillance”, IEEE Trans. Inf. Forensics Secur., 2015.

[4] A. Albiol, L. Sanchis, A. Albiol, and J. Mossi, “Detection of parked vehicles using spatiotemporal maps”, IEEE Trans. Intell. Transp. Syst., vol. 12, no. 4, pp. 1277–1291, 2011.

[5] A. Bayona, J. SanMiguel, and J. Martinez, “Comparative evaluation of stationary foreground object detection algorithms based on background subtraction techniques”, in Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance (AVSS), 2009, pp. 25–30.

[6] T. Bouwmans, “Traditional and recent approaches in background modelling for foreground detection: An overview”, Comput. Sci. Rev., vol. 11–12, pp. 31–66, 2014.

[7] D. Ortego, J.C. SanMiguel, J.M. Martinez, "Long-Term Stationary Object Detection Based on Spatio-Temporal Change Detection", IEEE Signal Processing Letters, vol. 22, no. 12, pp. 2368-2372, 2015.

[8] Estimación de la densidad de personas en entornos densamente poblados (People density estimation in crowded environments), Rosely Sánchez (advisor: Álvaro García-Martín), Proyecto Fin de Carrera (Master Thesis), Ingeniería de Telecomunicación, Escuela Politécnica Superior, Univ. Autónoma de Madrid, May 2015.

[9] Detección de Personas en Grupos (People detection in groups), Marta Villanueva Torres (advisor: Álvaro García), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Jul. 2015.

[10] Detección de personas en presencia de grupos (People detection in presence of groups), Sergio Merino Martínez, (advisor: Álvaro García Martín), Trabajo Fin de Máster (Master Thesis), Master en Investigación e Innovación en TIC, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Sep. 2015.

[11] A.N. Marana, L.F. Costa, R.A. Lotufo, and S.A. Velastin. On the e-cacy of texture analysis for crowd monitoring. In Computer Graphics, Image Processing, and Vision, 1998. Proceedings. SIBGRAPI '98. International Symposium on, pages 354-361, 1998.

[12] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L. Xu. Crowd analysis: a survey. Machine Vision and Applications, 19(5-6):345-357, 2008.

[13] A. Garcia-Martin, A. Cavallaro, and J.M. Martinez. People-background segmentation with unequal error cost. In Image Processing (ICIP), 2012 19th IEEE In-ternational Conference on, pages 157-160, 2012.

[14]A. Garcia-Martin, R. Heras Evangelio and T. Sikora. A Multi-configuration Part-based Person Detector. In Proceedings of the 11th International Conference on Signal Processing and Multimedia Applications (ICETE 2014).

[15]P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32(9), pp. 1627-1645, 2010.

[16]Detección de personas en entornos residenciales y hospitalarios, Jesús Molina Merchán, Trabajo Fin de Grado, Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Universidad Autónoma de Madrid, Escuela Politécnica Superior, May 2016.

[17]Seguimiento de objetos mediante constricción espacial de puntos TILDE. Cesar Betancur Garcia (tutor: Marcos Escudero-Viñolo), Trabajo Fin de Grado, Escuela Politécnica Superior, Univ. Autónoma de Madrid, Julio 2016.

[18]Verdie, Y., Yi, K., Fua, P., & Lepetit, V. (2015). TILDE: a temporally invariant learned DEtector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5279-5288).

[19]Nam, H., & Han, B. (2016). Learning multi-domain convolutional neural networks for visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4293-4302).

[20]Briechle, K., & Hanebeck, U. D. (2001, March). Template matching using fast normalized cross correlation. In Aerospace/Defense Sensing, Simulation, and Controls (pp. 95-102). International Society for Optics and Photonics.

[21]Integración y evaluación de sistemas de robo-abandono de objetos en video-seguridad, Jorge Gómez Vicente, Proyecto Fin de Carrera, Ing. Telecomunicación, Univ. Autónoma de Madrid en curso de realización. (tutor: Juan C. SanMiguel), Julio 2016.

[22]Álvaro García-Martin, Rosely Sánchez Ricardo, Jose M. Martinez: "Estimación densidad de personas basada en segmentación persona-fondo (People density estimation based on people-background segmentation)", Actas del XXXI Simposium Nacional de la Unión Científica Internacional de Radio - URSI 2016, Madrid, Spain, Sept. 2016.

[23]Víctor Fernández-Carbajales, Miguel Ángel García, José M. Martínez, “Visual Attention Based on a Joint Perceptual Space of Color and Brightness for Improved Video Tracking”, Pattern Recognition, 60:571-584, Dec. 2016 (online June 2016), Elsevier, ISSN 0031-3203 (DOI 10.1016/j.patcog.2016.06.007).

[24]C. Izmailov, E. Sokolov, “Spherical model of color and brightness discrimination”, Psychol. Sci., 2 (1991), pp. 249–259.

[25]Detección de caídas mediante vídeo-monitorización (Fall detection using video), David Dean Pulido (advisor: José M. Martínez), Proyecto Fin de Carrera (Master Thesis), Ingeniería de Telecomunicación, Universidad Autónoma de Madrid, Escuela Politécnica Superior, Mar. 2016.

[26]Análisis de actividad en vídeos deportivos multicámara (Activity analysis in multicamera sports videos), Ángel Mora Sánchez (advisor: Rafael Martín), Proyecto Fin de Carrera (Master Thesis), Ingeniería de Telecomunicación, Universidad Autónoma de Madrid, Escuela Politécnica Superior, Mar. 2016.

[27]Automatización de funciones en el seguimiento del profesor para la emisión de clases presenciales (Automatization of functions for teacher tracking in lectures broadcasting), Alberto Palero Almazán, (advisor: Jesús Bescós Cano), Trabajo Fin de

Máster (Master Thesis), Master en Ingeniería de Telecomunicación, Univ. Autónoma de Madrid, Jul. 2016.

[28]Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. Journal of basic Engineering, 82(1), 35-45.

[29]Sergio López, Diego Ortego, Juan Carlos Sanmiguel, Jose M. Martinez, "Abandoned Object Detection under Sudden Illumination Changes", Actas del XXXI Simposium Nacional de la Unión Científica Internacional de Radio - URSI 2016, Madrid, Spain, Sept. 2016.

[30]Detección de objetos abandonados para vídeo-vigilancia a largo plazo (Abandoned object detection for long-term video-surveillance), Sergio López Álvarez (advisor: Diego Ortego), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Univ. Autónoma de Madrid, Jun. 2016.

[31]Análisis de actividad en vídeos de baloncesto (Activity analysis in basketball videos), Rubén García García (advisor: Rafael Martín), Proyecto Fin de Carrera (Master Thesis), Ingeniería de Telecomunicación, Univ. Autónoma de Madrid, Jul. 2016.

[32] Seguimiento de objetos en vídeo a largo plazo (Long-term objects tracking in video sequences), Borja Maza Vargas (advisor: Juan Carlos San Miguel), Proyecto fin de Carrera (Master Thesis), Ingeniería de Telecomunicación, Univ. Autónoma de Madrid, Sept. 2016.

[33] K. Mikolajczyk and C. Schmid, _A performance evaluation of local descriptors,_ IEEE transactions on pattern analysis and machine intelligence, vol. 27, no. 10, pp. 1615_1630, 2005.

[34]Evaluación comparativa de técnicas de detección y descripción de puntos de interés en imágenes (Comparative evaluation of Points of Interest techniques for detection and description in images), Miguel Martín Redondo, (advisor: Fulgencio Navarro Fajardo), Proyecto fin de Carrera (Master Thesis), Ingeniería de Telecomunicación, Univ. Autónoma de Madrid, Jul. 2016.

[35]Clasificación automática de vídeos utilizando descriptores de características (Automatic classification of videos using features descriptors), María Narváez (advisors: Álvaro García-Martín, Tobias Senst), Trabajo Fin de Máster (Master Thesis), Master en Ingeniería de Telecomunicación, Univ. Autónoma de Madrid, Jul. 2017.

[36]Detección y Reconocimiento de Señales Viales (Traffic Signs Detection and Recognition), José Manuel Esteve de Prada (advisor: Álvaro García Martín), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería Informática, Univ. Autónoma de Madrid, Jul. 2017.

[37]Detección de Personas mediante Redes Convolucionales (People Detection using Convolutional Networds), Esther Sánchez Atienza, (advisor: Álvaro García-Martín), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Univ. Autónoma de Madrid, Jul. 2017.

[38]Segmentación Objeto-Fondo (Object-Background Segmentation), Paula Moral de Eusebio, (advisor: Álvaro García-Martín), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Univ. Autónoma de Madrid, Jul. 2017.

[39]Plataforma de Evaluación de Algoritmos de Detección de Personas (Platform for People Detection Algorithms Evaluation), Anthony Bryan Santiago Mendieta, (advisor: Álvaro García-Martín), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Univ. Autónoma de Madrid, Jul. 2017.

[40]Á. García-Martín, A. Cavallaro and J. M. Martínez, "People-background segmentation with unequal error cost," 2012 19th IEEE International Conference on Image Processing, Orlando, FL, 2012, pp. 157-160.