TEC2014-53176-R HAVideo (2015-2017)

High Availability Video Analysis for People Behaviour Understanding

D2.2v2

CONTEXTUAL MODELLING AND EXTRACTION FOR PEOPLE BEHAVIOUR UNDERSTANDING

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

AUTHORS LIST

Marcos Escudero Viñolo Marcos.escudero@uam.es

HISTORY

Version	Date	Editor	Description
0.9	28/06/2016	Marcos Escudero Viñolo	Final working draft
1.0	30/06/2016	José M. Martínez	Editorial checking
1.1	09/03/2017	Marcos Escudero Viñolo	First draft (v.2)
1.9	26/03/2017	Marcos Escudero Viñolo	Final working draft (v.2)
2.0	30/03/2017	José M. Martínez	Editorial checking

CONTENTS:

1. Introduction

1.1. Motivation

In the design of analysis methods for people behaviour understanding one should discriminate between active and passive objects. The former are those which behaviour is aimed to be understood—in this project *the people*—whereas the latter is either the *affected* or the *receiver* of the behaviour—the objects with which *the people* interact while they *behave*—.

Bearing this in mind, we can define contextual information as every description of non-active and non-passive objects in the scene. Including the passive objects before the behaviour is *applied* on them. Let us define all these objects as contextual objects. This—intentionally coarse—definition allows to classify as context a broad set of diverse evidences. Contextual information may range from appearance and spatio-temporal descriptions of contextual objects, e.g. colour, texture and position on the video slice; to scene capture conditions, e.g. cameras internal and external calibration; and contextual object functionalities, e.g. potential uses and mobility.

In the core of this project lies the idea that the extraction of contextual information may be a key process to drive, support and constrain people behaviour understanding methods.

1.2. Objective

This document describes the work done during the last twenty-five months in the task of context modelling and extraction (T.2.2) and updates the work done in the task of context modelling and extraction (T.2.2) respect to the previous version of this deliverable [19] published in June 2016.

During the initial months, the group developed offline annotation tools that operate on video sequences and reduce user interaction by propagating initial user evidences along the video. Initial studies on online modelling of context were also carried out. During the last months, the group has presented three papers for evaluation and has advanced in the task of context-constrained point descriptions, camouflage handling and multi-camera context aggregation.

1.3. Document structure

The document is structured as follows:

Chapter 1: Introduction

Chapter 2: Context in mono-camera scenarios.

Chapter 3: Context in multi-camera scenarios.

Chapter 4: Conclusions and future work

2. Context in mono-camera scenarios.

2.1. Introduction

This chapter describes the approaches designed or being design by the group in the task of context extraction and maintenance in mono-camera recorded scenarios. Ongoing work is included here as this deliverable closes the task T.2.2 to which the works are related.

2.2. Offline context annotation.

2.2.1. Introduction: a contextual annotation tool

This chapter describes the enhancements made in the group on an existing contextual annotation tool [1]. The tool's operation can be sketched as follows: a frame of the video is first segmented into regions by a multi-coarse version of [2]. Then, the user is asked to group the regions—by reiterative merging—to shape regions of interest that delimitate the spatial position of objects. These regions on interest are entities of a higher semantic level (close to the object level). The so-obtained regions of interest are then user-labelled as members of a contextual class. Labels defining the classes are extracted from a simple predefined ontology.

The use of this tool is relatively simple and its design allows to describe contextual objects with tight-to-boundaries masks. The annotation of a single frame—e.g. the first in the video—may provide enough information for the characterization of fixed non-functional objects in static-camera recorded sequences (e.g. walls, ceiling, floor or tables). However, there are three main issues that need to be addressed:

- The tool is not designed to handle video, just images.
- The annotation process can be tedious if masks are required to be clipped to object boundaries.
- The annotation of dynamic objects or of static objects in moving camera scenarios is not considered.

Next sections briefly describe the works developed in the VPU to cope with these problematic.

2.2.2. Adapting to video inputs, automatic object detection and initial ideas on propagation

In [3], we enhance the base tool by including several functionalities. In particular, support for more image formats is added and reading video frames is included in the tool as a new feature. The most effective improvement for annotation efficiency is to add object detection algorithms to automate the annotation process. Moreover, frame by frame manual annotation propagation mechanism has been also added, which consists in moving an annotation while the frames are changing.

The supported image formats are provided by built-in Matlab functionality and they are: bmp, cur, gif, hdf, ico, jpeg, pbm, pcx, pgm, png, pnm, ppm, ras, tiff and xwd. The result of the annotation is: several files with the preprocessed data (region segmentation); a GIF image with

the label for each pixel, a txt file with the ID of each label and the corresponding text; an auxiliary file in MAT format to load data within Matlab.

The improvements introduced can be summarized as follows:

- Annotation for video files with various video codecs
- Automatic annotation supported by object detection
- Propagation of annotations across frames

The appearance of the annotation tool after including the improvements is presented in Figure 1, GUI feedback on provided results is depicted in Figure 2 and an example of the automatic annotations generated by the tool in Figure 3. The user can delete them or keep them for faster annotation.

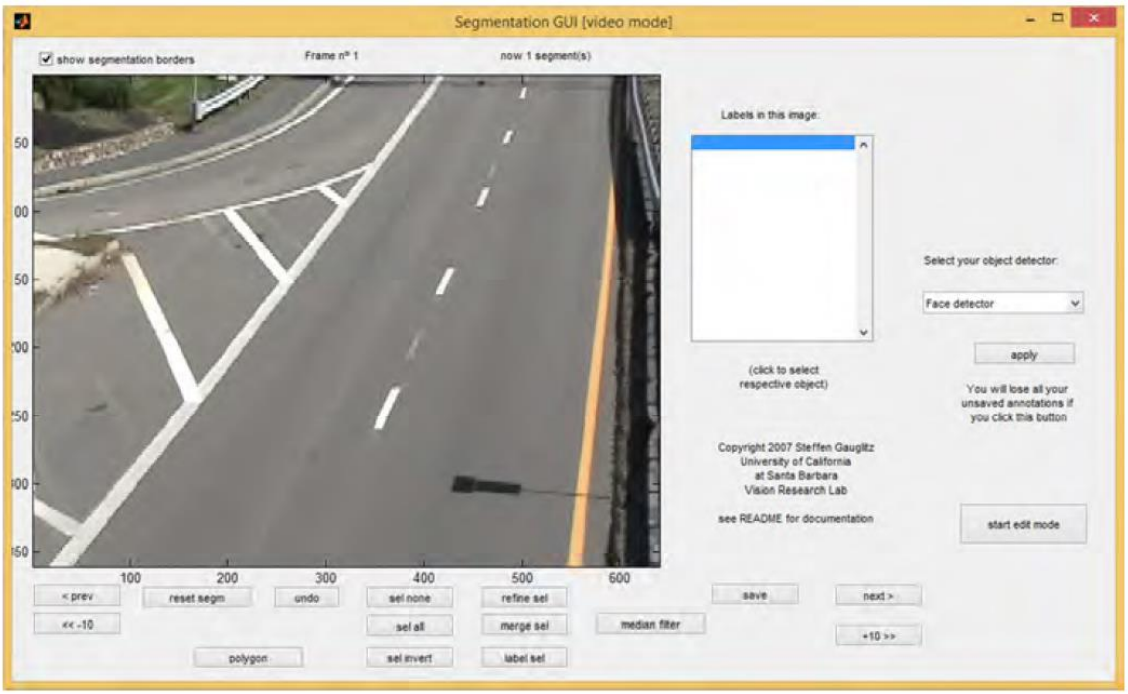

Figure 1. GUI of the improved context annotation tool

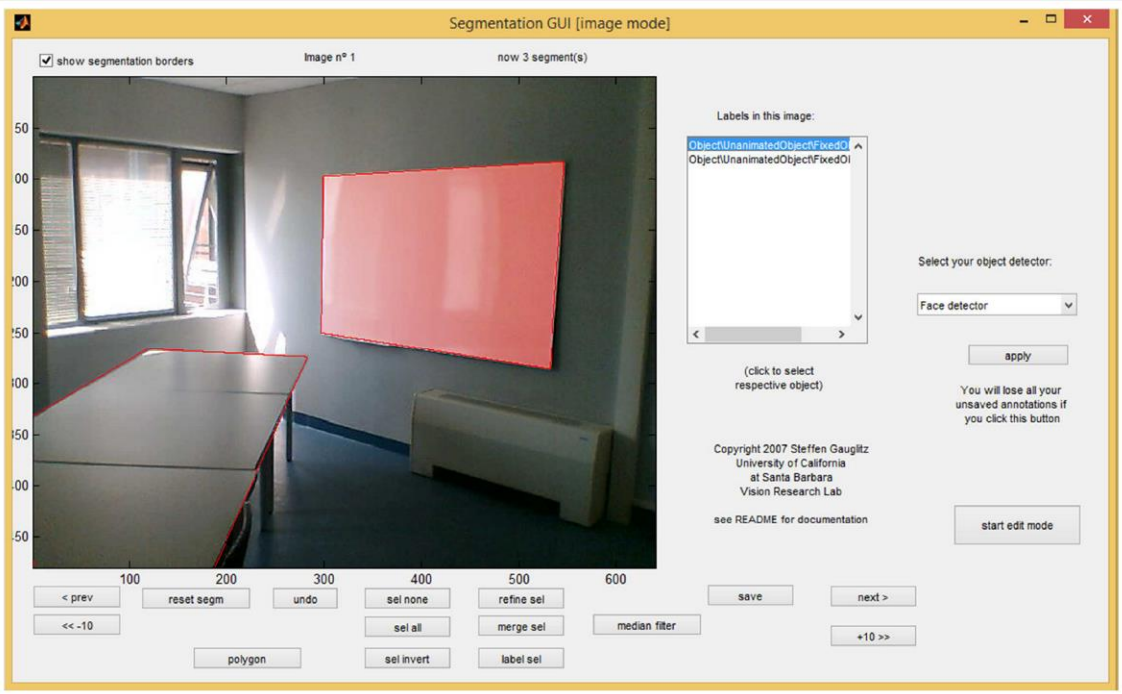

Figure 2. Snapshot of a context annotation tool in operation. A whiteboard and a table have been annotated as contextual objects.

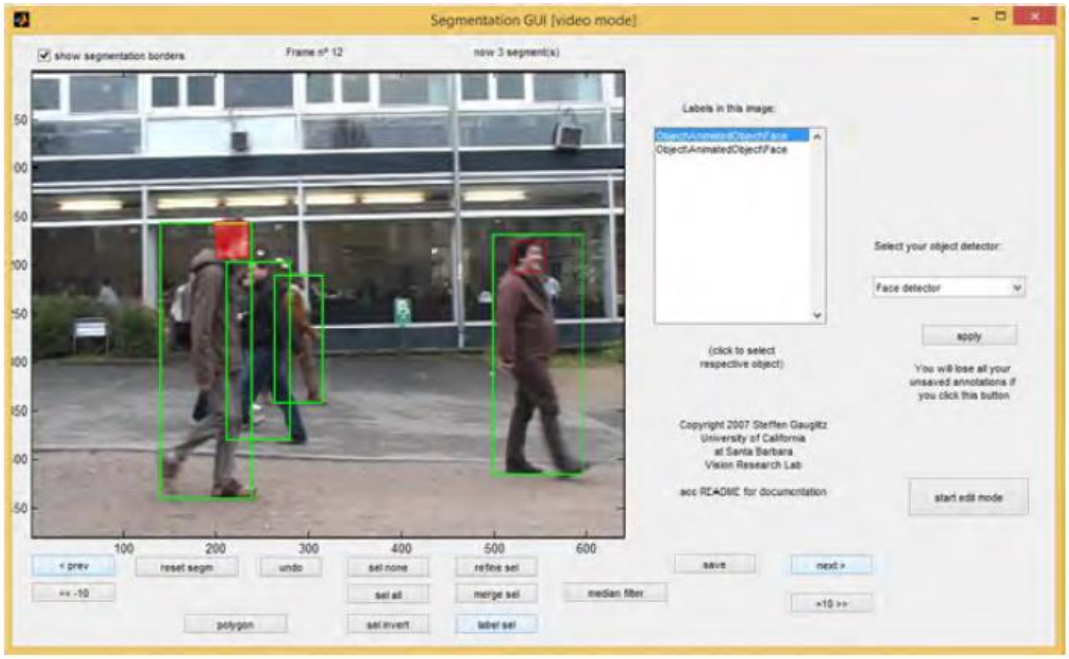

Figure 3. Automatic people detections annotated by the tool

Task	# base tool	# improved tool	base tool additional # to perform another annotation	improved tool additional # to perform another annotation
A	8	4	7	1
B	18	8	10	2

Table 1. Comparison between the base and the improved annotation tool in terms of required number of user interactions (#).

Task	base tool operation time	improved tool operation time
A	34	8
B	43	29
C	114	44

Table 2. Comparison between the base and the improved annotation tool in terms of time (measured in seconds).

Theoretically and experimentally, the improved tool requires a lower number of user interaction (Table 1) and time (Table 2) then the base tool when performing a set of predefined tasks (A, B & C).

2.2.3. Spatio-temporal propagation of contextual information and upgrading region-segmentation

In [4] the base tool is modified to minimize the number of required user interactions while maintaining the degree of usability of the annotations in similar terms to those obtained via the base tool. To this aim, we take advantage of the intrinsic characteristics of contextual objects: gradual and moderate temporal variation in terms of appearance, shape and position. In general terms, the aim is to propagate an initial user annotation in the first frame along the whole video. User-tool interaction strategies are also included to cope with system failures and new object appearances.

Obtained annotations are used to propagate the information on the next frame by automatically grouping regions non-requiring user interaction (see Figure 4). The so-obtained annotations are subsequently propagated to the next video frame and so on until the end of the video.

Through this naive strategy, the tool is potentially able to drastically reduce the number of required user interactions. Furthermore, due to the region-driven scheme, annotations are adapted to gradual changes of shape, appearance and position of the contextual objects. The region segmentation scheme is also upgraded to manage tighter-to-objects regions [5].

In order to evaluate the goodness of the designed tool respect to the base tool, two different comparison measures are used. The first one measures the number of user interactions required when using the improved tool respect to the base tool (*Si*)—the closer to one the lower the number of annotations, *Si=0* implies the same number of annotations—. The second evaluates the average quality of the propagated annotations respect to those obtained manually (*So*), i.e. the overlapping between the masks of the generated annotations—the closer to one the more similar, *So=0* implies complete dissimilarity of the annotation in every frame of the video—. Table 4 compares the base and the improved annotation tools in terms of the required number of user interactions, whereas Table 3 contains the *So* index. Finally, Figure 5 relates the measures. In the tables and the figure, results were obtained using two different region segmentation algorithms. In overall, obtained results support the design and development of the improved contextual annotation tool.

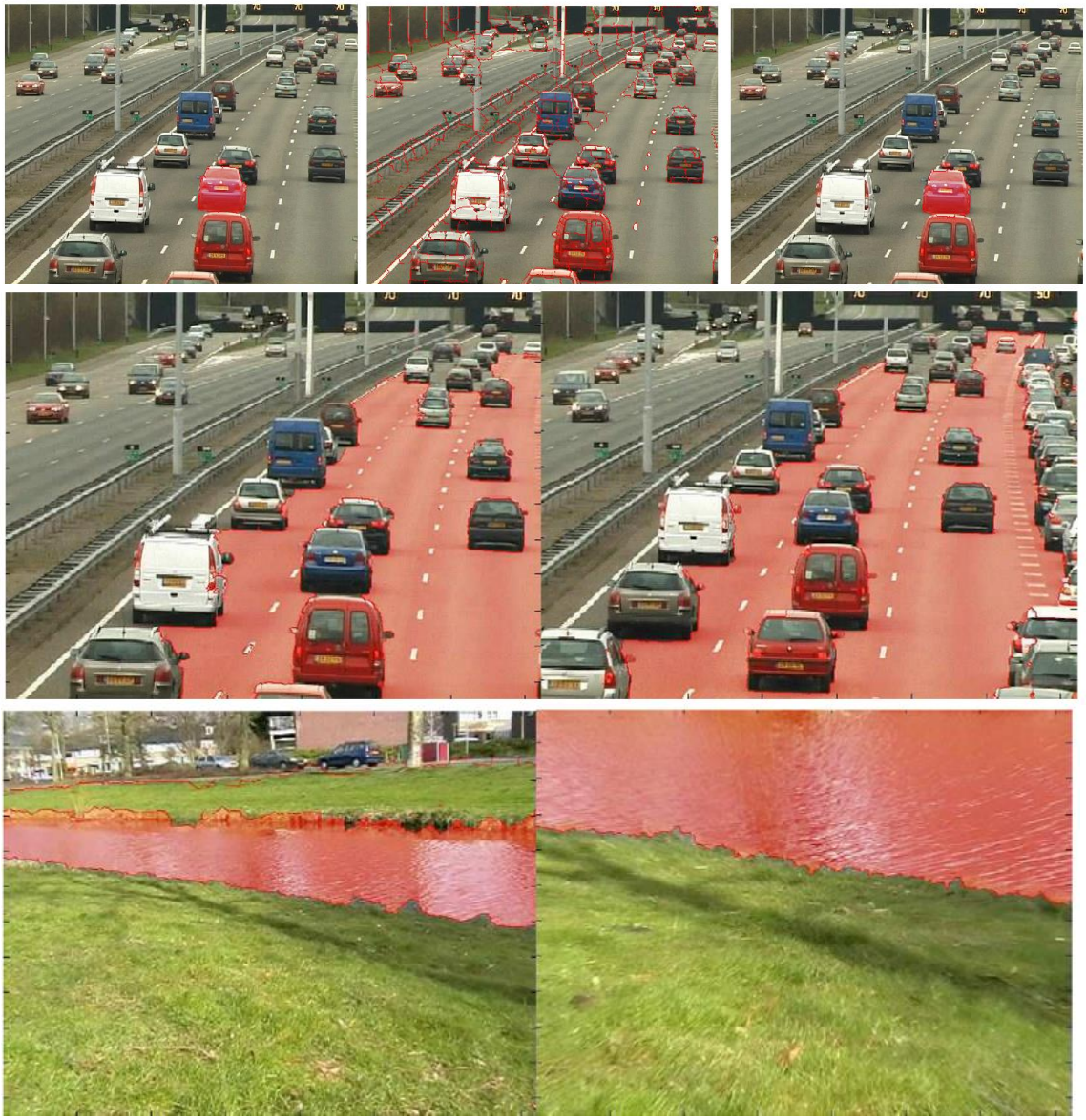

Figure 4. Spatio-temporal propagation of contextual information. Top row. Example for an active object (a car, for visualization purposes). Left. A frame with a RoI highlighted in red. Middle. Region segmentation in the following frame. Right. Regions significantly overlapping with the RoI are used to define the spatial extent of the RoI in the current frame. Middle and bottom rows. The same process used to extrapolate the information on contextual objects (a road and a river in a moving camera scenario).

	Sequence	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*	*(9)*	*(10)*	*(11)*	*(12)*	*(13)*	**Avg.**
So	**Superpixels [5]**	.94	.96	.87	.90	.92	.92	.94	.83	.90	.96	.74	.94	.92	.90
	Felzenswalb[2]	.84	.85	.90	.63	.92	.95	.89	.92	.97	.81	.92	.97	.96	.89

Table 3. Comparison between the base and the improved annotation tool in terms of annotations' overlapping and related *So* index.

Sequence	Superpixels [5]			Felzenswalb[2]		
	# base	# improved	*Si*	# base	# improved	*Si*
(1)Atasco	25	3	0.88	25	3	0.88
(1)Bandera	25	3	0.88	25	5	0.80
(3)Busy Boulevard	25	5	0.80	25	7	0.72
(4)Caldero	25	5	0.80	25	7	0.72
(5)Continuous PAN	25	4	0.84	25	4	0.84
(6)Fountain	25	3	0.88	25	3	0.88
(7)Molino	25	2	0.92	25	5	0.80
(8)Office	25	7	0.72	25	7	0.72
(9)Playa	25	2	0.92	25	2	0.92
(10)Rio	25	3	0.88	25	5	0.80
(11)Traffic	25	10	0.60	25	10	0.60
(12)Velas	25	4	0.84	25	7	0.72
(13)Wet Snow	25	5	0.80	25	5	0.80
Average	**25**	**4.31**	**0.83**	**25**	**5.38**	**0.78**

Table 4. Comparison between the base and the improved annotation tool in terms of the required number of user interactions (#) and related *Si* index.

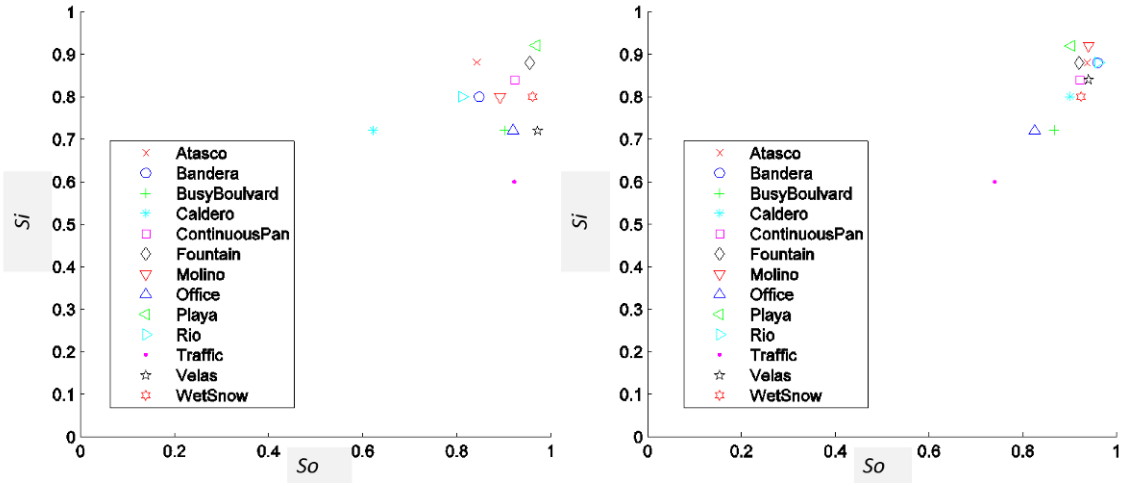

Figure 5. Relation between *Si* and *So* indexes for the set of sequences analysed when using [5] (left) and [2] (right) for region segmentation. In the figures, the top-right corner represents optimal performance.

2.3. Online context modelling.

2.3.1. Introduction: adapting to changes in contextual information

Video content is subjected to change significantly in space and time when analysing long sequences or scenarios recorded by moving cameras. Contextual information is part of this content and hence may also vary. The propagation schemes described in chapter 2.2 are not able to cope with scenarios in which contextual objects significantly change in appearance or position due to its interaction with active objects, climatological conditions or fast camera movements.

This chapter describes the developed strategies to use contextual information online, without the requirement of annotating it first (as in chapter 2.2). In this line, we propose: a strategy to use the appearance and motion of a point-of-interest (PoI) spatial context to constrain its description; a method to fight camouflage in background subtraction scenarios; an initial study in context transferring; and a method to online estimate the prior probabilities in road-scenarios.

2.3.2. Using spatial context to describe PoI

PoI are usually described according to their neighbourhoods, i.e. to their spatial context. Regions appear as an alternative to shape-fixed description supports on which to measure the distribution of spatial-based features. A region is prone to define a common entity at a—generally unknown—semantic level. The benefits of using regions as description supports are mainly summarized in a description-independence premise: as regions are prone to conserve object contours, the influence of contiguous objects in the description is eliminated. This might be especially useful when describing moving objects in videos. There, the object moves in a scene, while the objects around it remain static or move differently to it. Moreover, it might be also beneficial when, with independence of the object dynamics, the scene is captured from two different points of view. In this case, as the scene is 3−dimensional and the image plane is 2−dimensional, the captured configuration of the scene may vary substantially; changing completely the adjacency and occlusion relationships among the captured objects.

In [10], starting from a previous work [6], focused on a region-constraining scheme of the classical SIFT [7] descriptor, we developed regional versions of state-of-the-art 2-dimensional [8] and 3-dimensional descriptions [9]. We have explored two alternatives to perform the constraining. In the first one, the basic idea is to describe a point one-to-several times, one per region contained in the PoI's fixed support. Then, in the matching stage, correspondences are found by minimizing the distance between the PoI's descriptions (hard constraint). The second one, weights descriptions according to the resemblance of the described areas to the described PoI (soft constraint). See examples of hard-constrained versions of DAISY [8] , and SHOT [9] descriptors in Figure 6 and Figure 7. See Figure 8 for qualitative results of the method. The work here described is now under evaluation after revision [20].

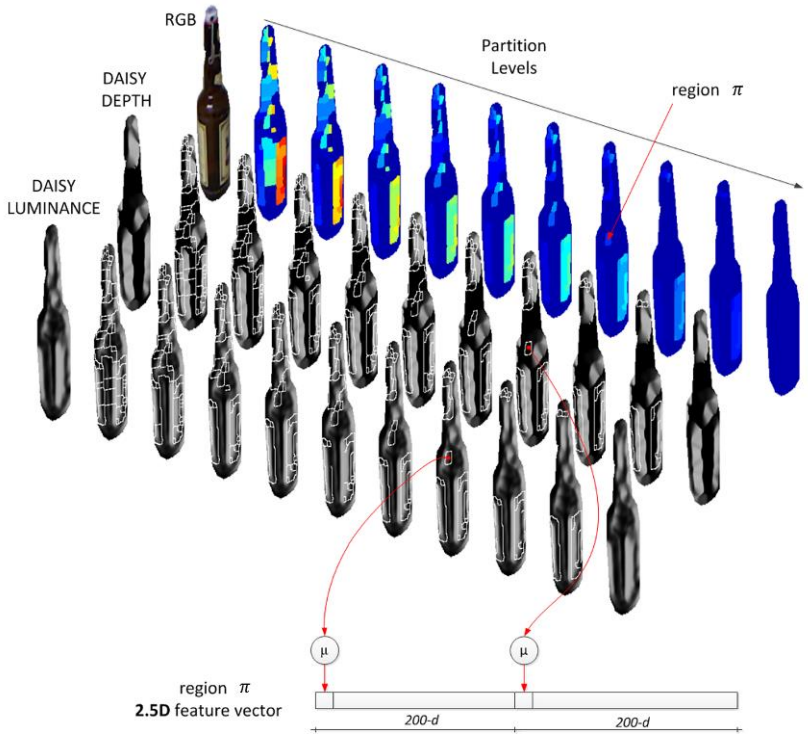

Figure 6. The R-DAISY descriptor. A RGB image is repeatedly partitioned into E coarseness-levels (E =10 in the example). Each region in a partition defines a region area over the luminance and depth DAISY [8] descriptions. The feature-vector for a region π is obtained by concatenating the median value, μ, of the descriptions inside the area defined by π. Extracted from [20].

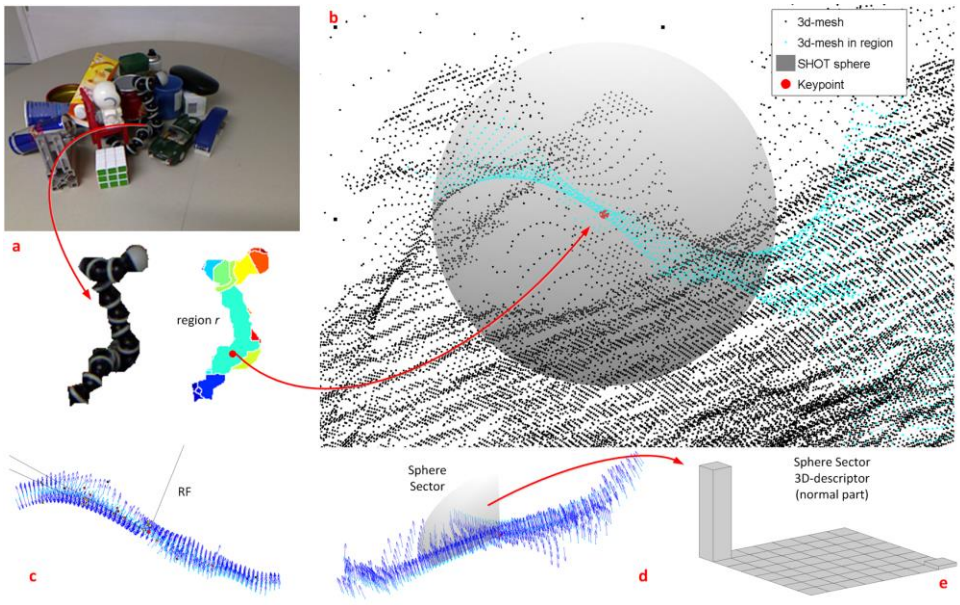

Figure 7. The R-SHOT descriptor. a. An object is partitioned into E coarseness-levels (only the 5th level is shown in the figure). b. For a given PoI (red point), its associated region in the partition defines a subset of 3D-coordinates (light blue points) over the whole image mesh (black points). Note that this masking strategy is able to avoid the use of points of a different object part in the SHOT description (some of the black points inside the sphere). c. PoI's neighbouring object-points (in the image these are plotted with associated normals) are used for the reference frame (RF) estimation. d. The RF is used for the computation of the local coordinates. e. Normals in each sphere sector are described by a two-dimensional sample-normalised histogram. From [20].

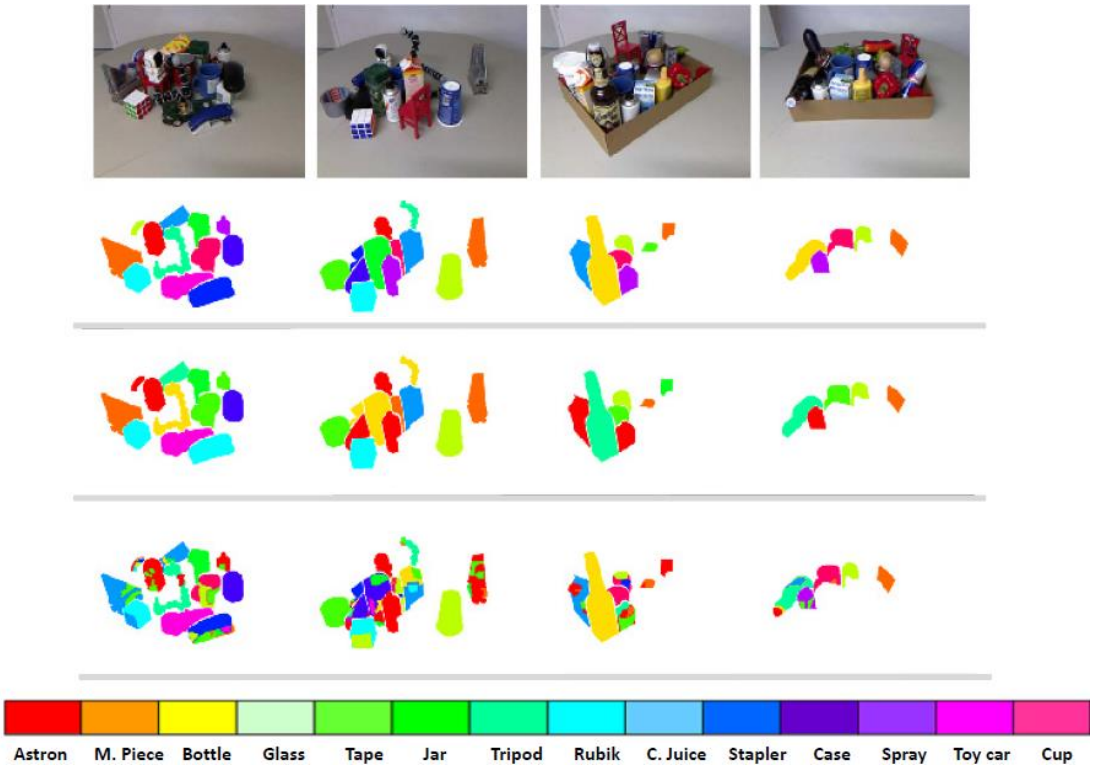

Figure 8. Regions-constrained descriptors for object recognition a. Test images (first row) and associated ground-truth (second row). Results for the SHOT descriptor (third row), and for the R-SHOT descriptor (fourth row). Associated labels are included in the bottom row of the Figure to ease visualization. Note that result images have been recomposed as each object is analysed isolated. From [20].

2.3.3. Using spatial and motion context to describe PoI

An ongoing work in the group [21] is extending this idea to the use also of motion information. Motion information has become a key task in violence and action recognition so, following [22], in this work we use the accumulated direction Lagrangian measures [23] for the task of violent-video detection. These measures provide similar fields to the optical flow fields and represent the dynamic patterns in the scene. We propose to extend [22] by using the Super-Pixel-based isolation of SIFT (SP-SIFT) [24] to encode spatio-temporal patterns instead of the traditional SIFT descriptions there used. The SP-SIFT descriptor is able to abstract the main object description from the background, which can be advantageous in tasks such as recognition of objects where background information is not important.

Figure 9 shows a block diagram of the proposed approach. Approaches based on feature descriptors typically use the bag-of-words model to represent each video sequence as a histogram over a set of visual words created by the k-means algorithm to generate a fixed-dimensional encoding. In this work, these histograms are classified using a Support Vector Machine (SVM).

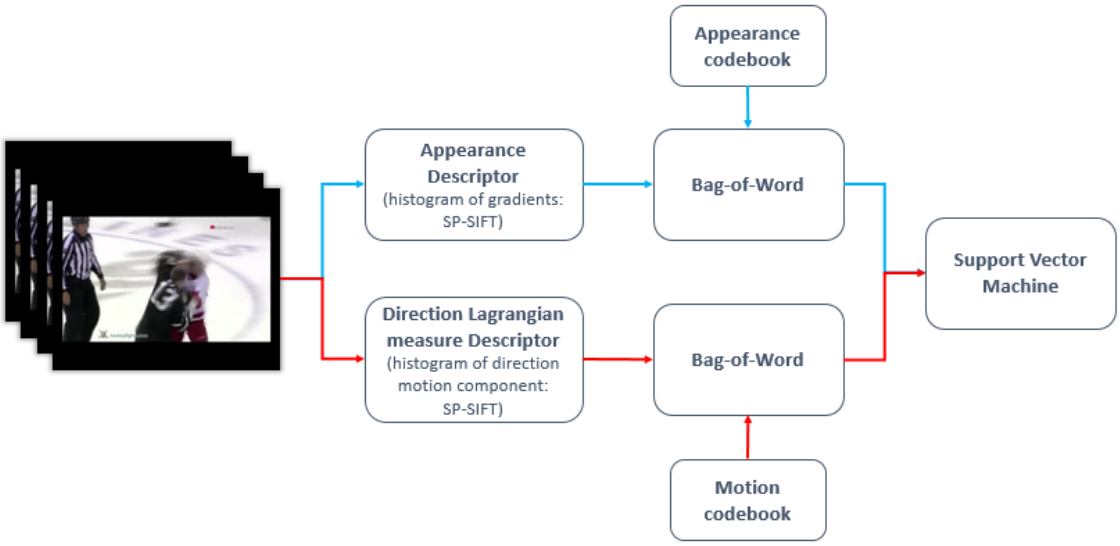

Figure 9. Block diagram of the proposed approach Motion and spatial constrained point of interest description for violent-video detection. Extracted from [21].

2.3.4. Fighting camouflage

The challenge of discriminating highly similar background and foreground samples is usually known as camouflage. Camouflage—being a problem of foreground-to-background resemblance—is caused by a combination of at least three factors: the feature(s) used for characterisation; the metric used to compare new instances with a background model—usually linked to the features defining a metric space—; and the model maintenance mechanisms, that determine the background model at each time instant. We believe that the effect of foreground camouflage in foreground segregation schemes (e.g. background subtraction BS) has been generally underestimated. The effects of camouflage situations vary from small holes in the detected foreground objects, usually solved via post-processing, to missing notable parts or even the whole foreground object. This last situation may affect the correct operation of segregation-dependent tasks as object tracking, people detection or video event detection. Camouflage is still stated as an unsolved context-related problem in recent approaches [25] and it is considered the main focus for improvement in international BS benchmark evaluations [26] [27]. To fight camouflage, we propose to derive a new characterisation scheme based on the widely-known Discrete Cosine Transform (DCT): the *WRAC* and its spatially-smoothed version, the *SWRAC*.

In particular, we use the DCT to characterise the pixel neighbourhood. In the literature, a common approach for different applications is to select a subset of these basis-functions, which is fixed for every analysed signal. In this paper, we question this approach and propose a novel method to select a signal-dependent subset of basis-functions. As we propose to describe each pixel with a different set of DCT coefficients, each associated to a particular basis-function, in order to compare any two so-obtained descriptions a distance function is required: we propose a novel metric to cope with this situation.

Experimental results prove, first, that the subset of relevant basis-functions varies for every signal, in this case for every pixel neighbourhood. The presented experiments over the Change Detection Data-set show, second, that the proposed results show that the proposed scheme notable reduces the likelihood of camouflage respect to other popular features: 92% respect to the pixel luminance, 82% respect to the RGB values, and 65% respect to the best performing LBP configuration evaluated.

This work is currently under revision [28]. Figure 10 illustrates some examples of the created camouflage data-set whereas Figure 11 depicts obtained results when compared with state-of-the-art alternative descriptors.

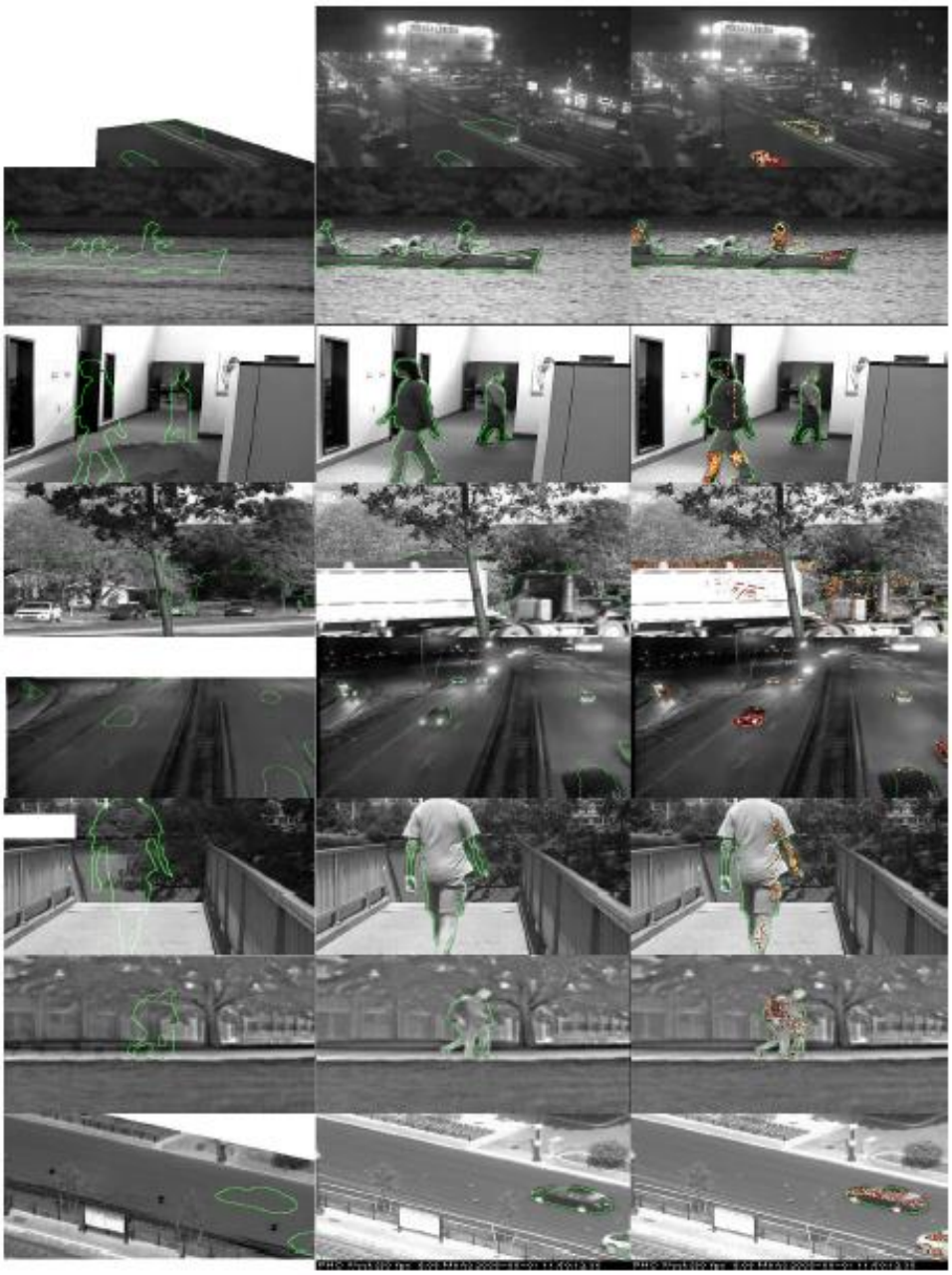

Figure 10. Examples of the luminance based camouflage reference for selected frames of some video sequences from the data-set. Background Model (left column)—white areas are out of ROI samples—; selected frame (middle column); and camouflage likelihood (right column)—the redder the more likely. To ease visualisation, the contour of the foreground mask is over-imposed in green. Extracted from [28].

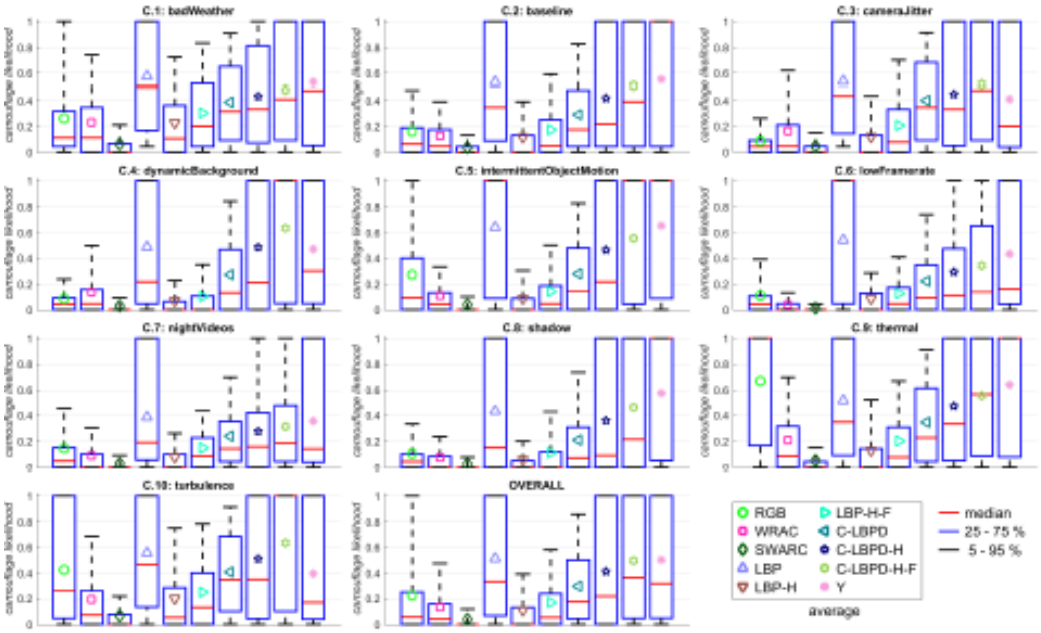

Figure 11. Performance comparison of the ten evaluated feature-metric combinations on the sampled camouflaged data-set. Extracted from [28].

2.3.5. Contextual transferring.

The idea of this project is to transfer existing annotations of contextual objects—*the knowledge*—to an unobserved new scene which contains objects which can be semantically understood of being of the same contextual categories as those in *the knowledge*.

The project starts from an existing approach to transfer masks [29] and adapts it to the scope of contextual transferring. To this aim, *the knowledge* is composed of widely-used human-annotated data-sets recorded at varied scenarios, including the Label-Me [30] the AD20K [31] and the CityScapes [32] data-sets. The project's end is scheduled to July 2017.

2.3.6. Learning contextual priors to detect vehicles

The goal of [15] is the implementation and analysis of a system able to segment a video flow generated by a forward-looking camera placed in the frontal of a vehicle, with the aim of detecting vehicles, pavement and lane marks. The system handles contextual information of the scene to generate statistical models of the regions of interest, which are optimized through the Expectation-Maximization algorithm to classify the images using the Bayesian decision theory.

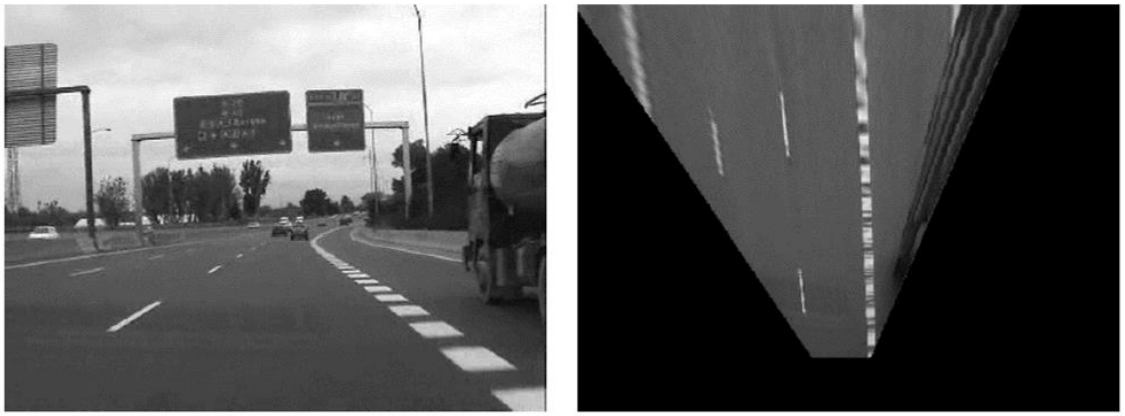

Figure 12. Inverse Perspective Transform

Figure 13. Left. Original image after IPM. Middle. Pavement detection. Right. Line detection.

In particular, the method operates on IPM (Inverse Perspective Mapping) images (see Figure 12) and assumes that the only visible classes are pavement, lines and vehicles. Additionally, an unknown class is also included to cope with road vegetation and the rest of non-accounted objects.

The basic idea is to estimate the spatial coverage of pavement and lines and extract data distributions in these areas (see Figure 13). These distributions are then used as priors to improve the detection of vehicles.

All the stages of the system have been studied and implemented, contributing also with proposals to improve the performance of some of them. Subsequently, to evaluate the performance of the algorithm and the validity of the improvements, we have designed a group of specific tests for each class, where three different versions of the systems are compared against a ground-truth created specifically for this purpose. Preliminary results partially support the designed method (see Figure 14).

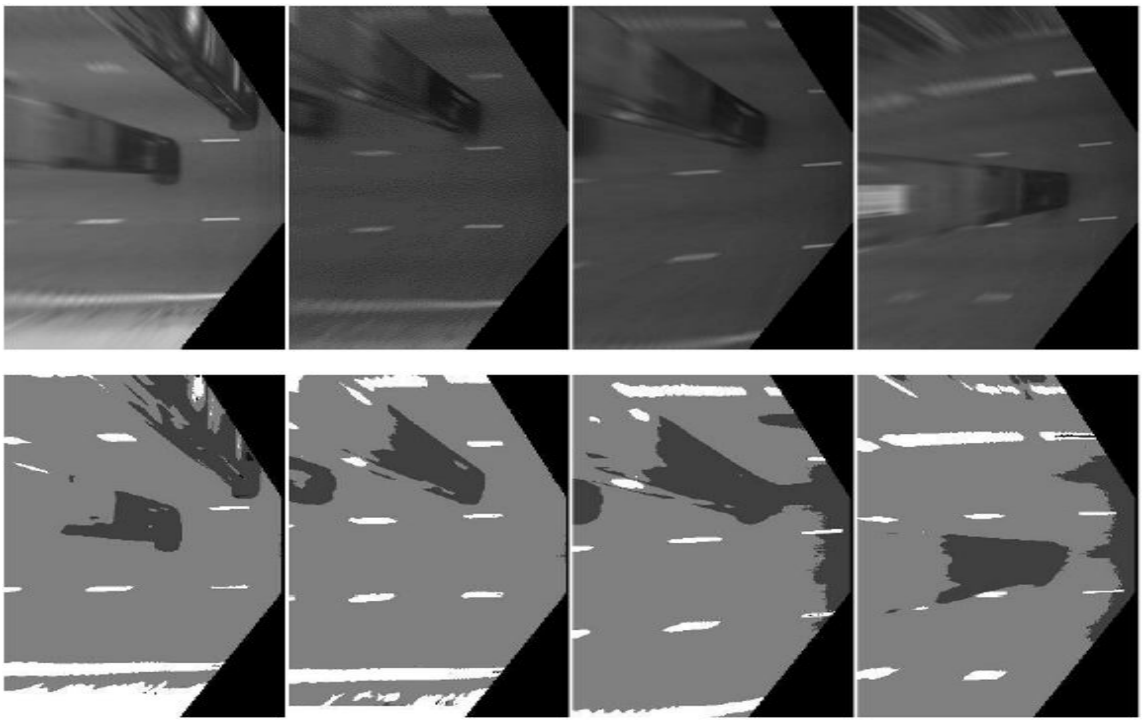

Figure 14. Top row. Original frame. Bottom row. Obtained classification, a different grey value is assigned to each class.

3. Context in multi-camera scenarios.

3.1. Introduction

This chapter describes the approaches designed or being design by the group in the task of context extraction and maintenance in multi-camera recorded scenarios. The chapter describes: an initial approach to estimate the relative position of two widely-separated cameras recording the same scene; a method to create a panoramic view of a scene recorded via a moving camera; a scheme to relate people detection between two-cameras; and a strategy to combine contextual information in a multi-camera scenario. Ongoing work is included here as this deliverable closes the task T.2.2 to which the works are related.

3.2. Relating cameras in wide-baseline scenarios

Knowing the relative position between any two cameras (the epipolar geometry) recording the same scene is essential to propagate and combine the individual analysis results obtained for each camera view. When the cameras are shortly separated, the relations between images points can be easily derived, e.g. it is viable to impose a set of spatial constraints to define narrow image areas on which the corresponding point to a searched one is prone to be found. However, when the cameras are widely separated these constraints are hardly to set. Additionally, objects' appearance may change severely due to different illumination and capture conditions and to projective distortions.

The work done in [11] aimed to design and develop a set of algorithms able to cope with the problem of modelling the epipolar geometry between two widely cameras capturing the same scene. The ultimate goal of this process is the automatic estimation of the relative position between both cameras. In fact, whereas the epipolar geometry is fully explained through the fundamental matrix, the relative position may be extracted, up to a scene homography, from the fundamental matrix. To fulfil the faced task, a study of the image capture and camera calibration processes was first required. Automatic calibration strategies usually rely on a previous set of inter-image correspondences. To this aim, we explore the use of Hessian-based detectors on the scale-space [12] for the feature detection stage, SIFT [7] and DAISY [8] descriptors were assessed to describe the extracted features, whereas the NRDC [13] algorithm was evaluated as an alternative for feature matching.

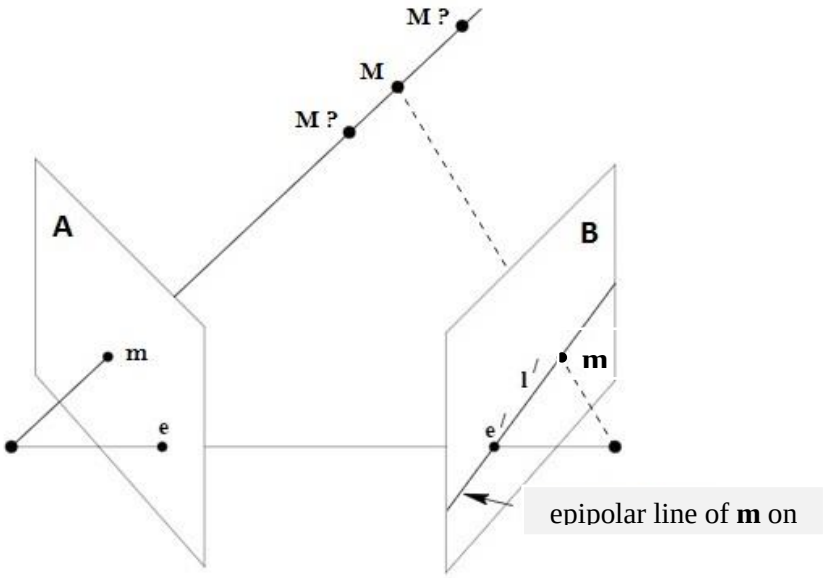

Figure 15. The epipolar constraint.

The use of these feature-matching pipeline leads—in conjunction with a RANSAC estimator—to an initial estimation of the fundamental matrix. Afterwards, such matrix is refined by three strategies designed and developed during this project. The three strategies are based on the classical epipolar constraint (see Figure 15):

Given a scene point M which projection on image A is m; the corresponding point, m' in B must lie on the epipolar line of m, l' on B and vice versa.

The first strategy consists on a double check matching. For a given point *m* in A and a given estimation of the fundamental matrix, we search for the most-similar point in the other image falling on *l'* (projection). Then, for the found point *m'* the inverse process is applied (retro-projection), i.e. we search in A the most similar point to *m'* falling on *l*. If the found point, let say *m''*, coincides with *m*, a matching is declared. See first and second rows in Figure 16. In the first strategy, we have assumed that the initial estimation fundamental matrix is at least partially correct. To cope with bad estimations, we derive the second strategy: a mask-based exhaustive searching. In this searching, instead of searching only on points on the epipolar line, we wide the candidate area to a rectangle aligned with the epipolar line. In particular, the line serves as the central axis of the rectangle. See third row in Figure 16. Finally, the third strategy consist on combining the double check and the mask-based exhaustive searching.

Moreover, a Graphical User Interface (GUI) which eases the manually introduction of reliable correspondences has been also designed and developed. The aim of this GUI is to face scenarios where the feature-matching pipeline is not able to produce an initial estimation of enough quality (See Figure 17.). Finally, both the initial estimates of the fundamental matrix and its subsequent refinements have been evaluated by means of different experiments and error measurements. Achieved results are still unable to produce a reliable fundamental matrix, but improve previous approaches and pave the road to future enhancements of the method.

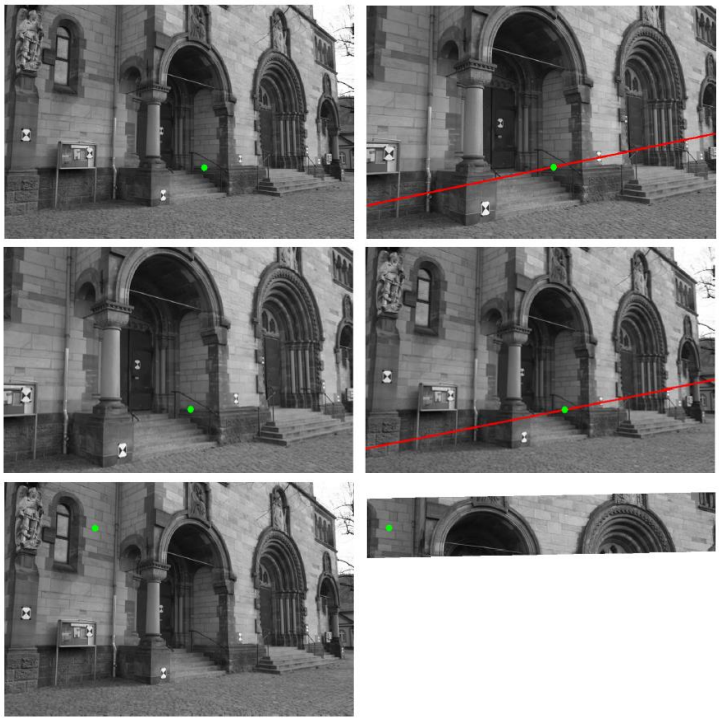

Figure 16. Searching strategies. Projection with epipolar constraining (1st row). Retro-projection with epipolar constraining (2nd row). Mask-based exhaustive searching (3rd row).

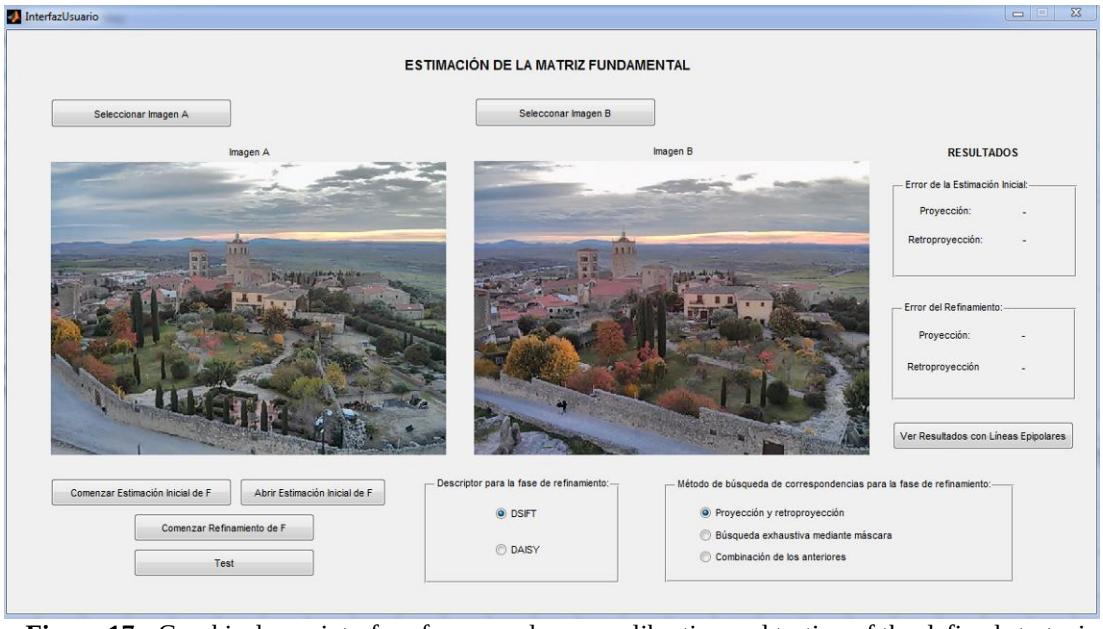

Figure 17. Graphical user interface for manual scene calibration and testing of the defined strategies.

3.3. Scene-reconstruction based on PoI

Scene reconstruction techniques are usually designed to work on static cameras. The capture conditions provided by these systems are far better than the capture conditions on pan-tilt-zoom (PTZ) cameras, or in full mobile cameras. In order to increase the area covered by a single static camera without adversely affecting the results of the algorithms, PTZ cameras are a good approach. They offer a good trade-off between area covered and capture conditions.

Background initialization on PTZ cameras is a process similar to a panoramic generation. However, it faces the same problems that background initialization on static cameras. Those problems are most of the times related to foreground objects present in the scene while the background is being captured. To cope with these situations in PTZ scenarios, a two-stage solution is proposed [14]. It is designed to operate on a basic system of panoramic generation based on local features. As shown in Figure 18, the first stage (blocks in green) is focused on error detection, and the second stage (blocks in blue) deals with error correction strategies.

The basic system receives an input image, captured by the PTZ while scanning the scene. SURF PoIs are detected on the input frame, and matched with previously detected points in the panoramic generated till that moment. Using the matching information, a homography between the input frame and the panoramic is estimated. The final stage is the mapping of the input frame to the panoramic through the estimated homography. Once the input frame is transformed, the panoramic is updated.

In order to detect errors, we design an error detection method, which is composed of two modules: correspondence and frame-insertion evaluation. Correspondence evaluation compares the number of correct correspondences obtained in the matching stage against an adaptive threshold. If the number of correspondences is under that threshold an error is considered. A low number of correspondences can be caused by the presence of foreground objects occluding the background, but also to alternative challenges as blurring or illumination changes. In any case, the estimated homography may not be reliable for the insertion. Differently, if the number of correspondence is high enough, the second error detection stage consists on comparing the input frame and the panoramic generated till that moment in the overlapping area. If colour differences are high, foreground objects are supposed to be in the scene while the last frame is captured, and an error alert is activated.

In order to cope with these errors, two different strategies are proposed. The first one, named stability correction, relies on the assumption that the background is the most repetitive representation in a sequence. Bearing this in mind, each time any of the two errors is detected, the frame is discarded. Additionally, the method defines an inactive time on which no analysis is performed. The second strategy, named multi-panoramic correction (MP in Figure 19), relies on the assumption that each error will generate a new panoramic, and at the end of the swept, the algorithm will be able keep and merge the correct ones, and avoid the others.

Quantitative and qualitative results presented in [14] suggest that the proposal performs robustly under challenging situations for background initialization. An example is shown in Figure 20, where left image is captured with the basic system, and mistakes in the reconstruction are easily noticeable. The obtained panoramic constitutes a reliable holistic representation of the scene on which to annotate context objects and intrinsically defines camera view range and capture position.

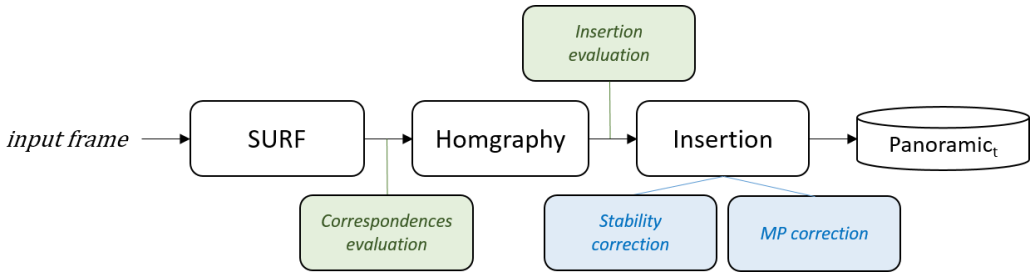

Figure 18. Proposed scheme for panoramic generation. In white, basic system. In green, error detection modules. In blue, error correction modules.

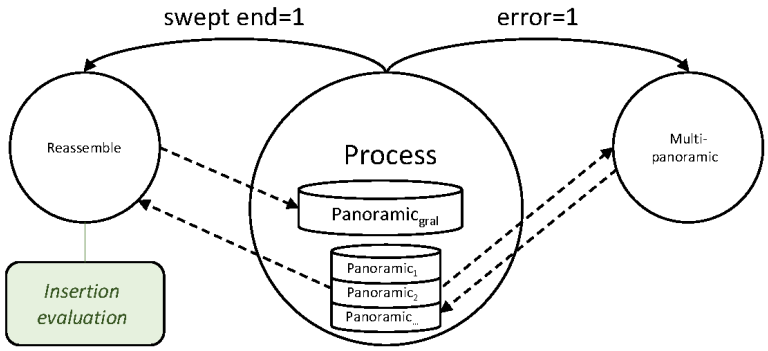

Figure 19. State diagram for MP correction. A new panoramic is generated with each error. An insertion evaluation module is included in the final reassembling process.

Figure 20. Qualitative comparative. Left image is obtained with basic system. Right image is obtained after MP correction

3.4. Using scene context to relate people detections across cameras

The work first described in [16] was focused on the use of scene context for the inter-camera combination of people detections in scenarios recorded with at least two cameras. The aim was to cope with situations that hinder people detection algorithms: occlusions, illumination changes, perspective change, etc. An extension of this work is now under evaluation [34]. The context—in the shape of inter-camera extrinsic calibration—is here used by transferring people detection projections between cameras. The idea is that a location on a camera view—see Figure 21 (a) and blue solid line in Figure 21 (b)—has to be adapted to the other camera view before projection—see red solid line in Figure 21 (b)—.

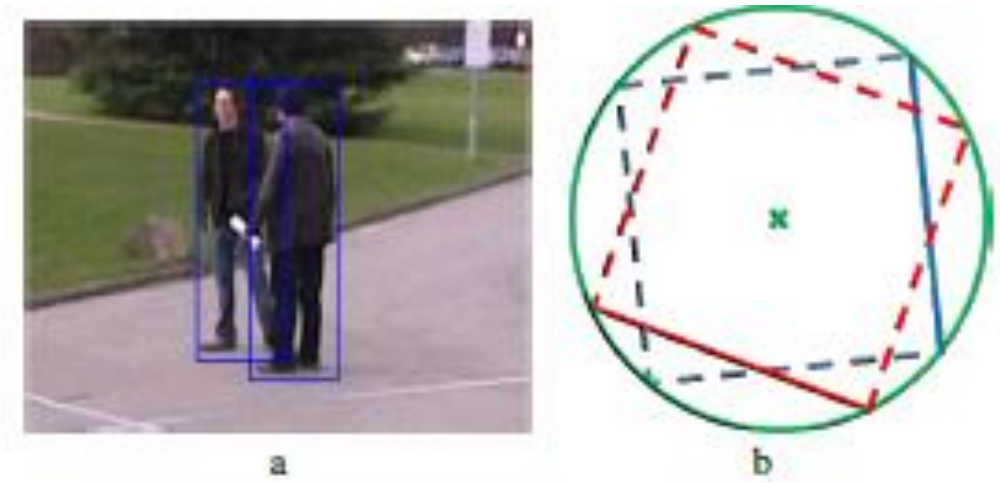

Figure 21. Inter-camera transferring of people detections using multi-camera contextual information. Extracted from [34].

Figure 22 depicts the faced scenario. In the proposed approach, the floor plane is used to derive an IPM mapping on which to relate people detections. To evaluate system's performance with independence of scene calibration, the mapping is done manually. Therefore, assuming that cameras remain static along the sequence. The algorithm starts by detecting people in each camera. The used method is trained to detect both standing people and people in wheelchairs. Obtained detections are projected on to a common floor plane and propagated to the other camera by a height estimation method. Detections from different cameras are then combined by overlapping. Obtained results for different overlapping hypotheses validate the designed propagation strategy (see Table 5).

Figure 22. Left. Map of the scenario. Right. Captured views for each camera. Red dots are used to calibrate the floor plane.

Figure 23. People detections are projected under a cylinder-based assumption on the floor. Yellow (blue) ellipse is the projection of the circle inscribed in (circumscribing) the bounding-box projection on the floor plane. Individual people detection algorithms are trained to handle wheelchair situations.

	Area under curve		
	1st Camera	**2nd Camera**	**Total**
Detection in the camera	0.772	0.694	0.733
Detections in the other camera	0.703	0.733	0.717
Combined detecting with 0.999 Overlapping.	0.621	0.609	0.615
Combined detecting with 0.9 Overlapping.	0.625	0.611	0.618
Combined detecting with 0.75 Overlapping.	0.690	0.663	0.676
Combined detecting with 0.5 Overlapping.	0.789	0.752	0.770

Table 5. Area under curve for individual, propagated and combined detections.

3.5. Semantic segmentation in multi-camera scenarios

The goal of this ongoing project ([35]) is to combine contextual annotations obtained for different views and to temporally and spatially refine such annotations by using video temporal coherency. Furthermore, we also aim to hallucinate contextual annotations for virtual views in-between the real camera views.

To this aim, we have defined three main stages in the algorithm:

1- Extraction of contextual information per view. We use a CNN called PSP-Net [36] which has been trained using more than a hundred different classes. Image areas are automatically classified into non-mobile scene-related contextual objects, including: corridors, floor, paths, walls, columns, doors and counters.

2- Projection on to a common plane. A set of plane induced homographies tailored to a set of sampled positions for each of the three cameras recording the scene is used to relate captured pixels with the zenith plane. Through these homographies semantic maps of each camera sampled positions are projected onto the ground floor (see Figure 24).

3- To also cope with non-sampled camera positions—as in PTZ-scenarios—we define a perspective correction algorithm that relates a camera view with its closest fixed camera position in a codebook. To do so, matching between AKAZE [37] point-of-interests is used to define the best candidate. Semantic maps is then corrected to align with the found position (see 0 for an example).

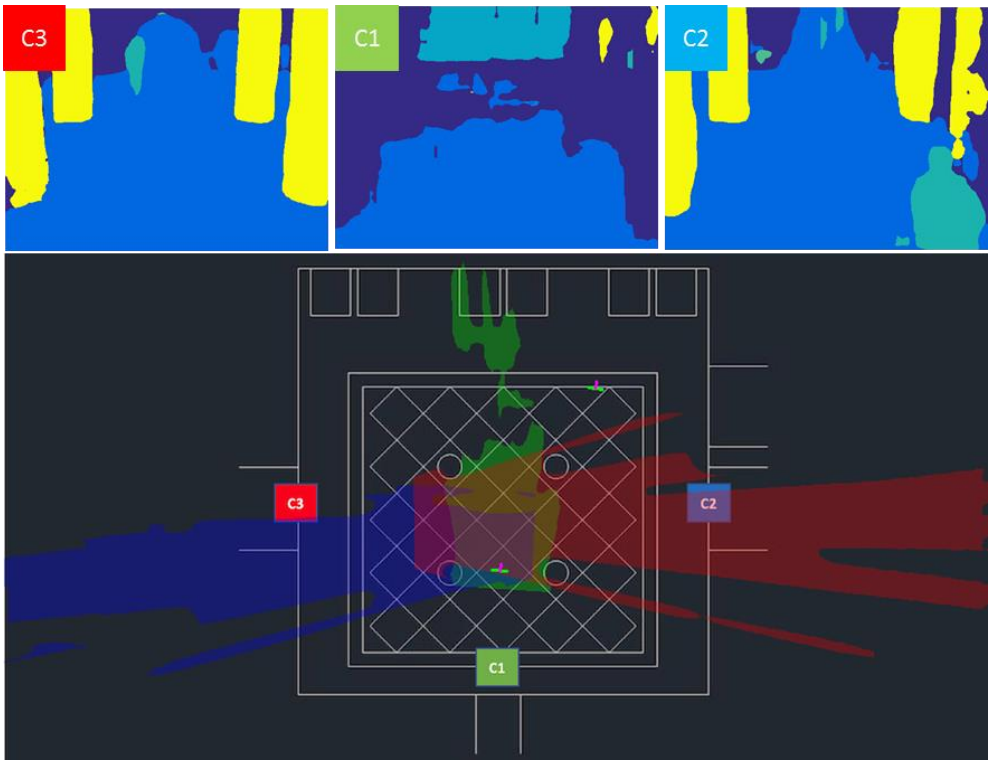

Figure 24. Top row, semantic maps obtained for a fixed pose of each camera. Blue pixels represent floor areas; yellow pixels define columns; dark turquoise pixels indicate windowed areas and light turquoise pixels are representative of people regions. Bottom row, projection of floor areas onto the floor plane (a different colour for each camera).

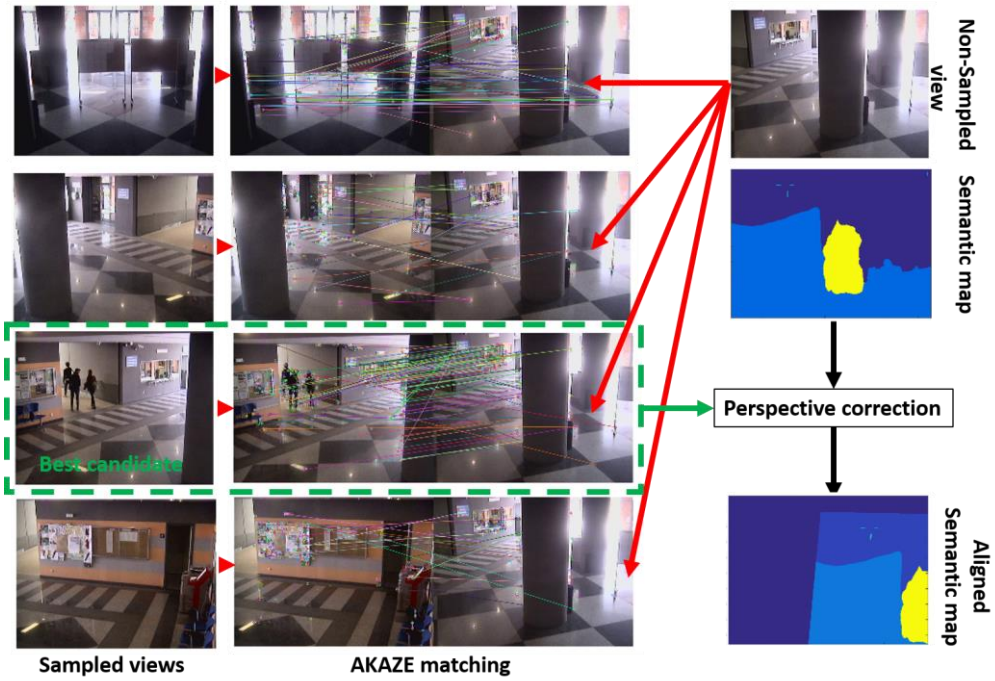

Figure 25. Left: Set of sampled views for a given camera. Centre, AKAZE matching between sampled views and a given non-sampled view—the more lines the bigger the number of matchings—. Right, top: non-sampled view and associated semantic map; bottom, semantic map aligned with best candidate from the set of sampled views.

4. Conclusions and future work

In this document, we have briefly described the contributions made in the scope of the HAvideo project for the tasks of use and extraction of contextual information for video-based human behaviour understanding. This document updates a previous version of the deliverable [19].

We start by motivating the use of context, providing a contextual definition that covers essential information in video analysis. Then, we present two methods to upgrade and existing annotation tool [19]. The first one [3], adapts the tool to video inputs and integrates object detection methods in the tool's operation. The second one [4] proposes a naive strategy to propagate user annotations along the video. Designed strategies wide the potential uses of the tool and improve its usability.

Following, we describe five methods to automatically extract contextual information from the video content in mono-camera scenarios. The first one [10], [20]—operating at a low semantic level—, uses the spatial context of points-of-interest to constraint their descriptions. The second one [21] extends [20] by incorporating motion information. The third one [28] proposes a new descriptor able to fight camouflage situations. The fourth one, [33], describes a preliminary work on contextual transferring from a knowledge data-set to a given unobserved image. The fifth one [15] generates probabilistic models of context to improve object detection. Designed strategies pave the road to automatic context extraction, but also illustrate the challenging nature of this task. Future work will be mainly focused on improving these approaches and on proposing new methods for online context extraction.

In multi-camera scenarios, we describe four methods to generate, use and transfer context. The first one [11], aims to relate the position of widely-separate cameras recording the same scene. The second one [14], generates a panoramic view of a PTZ-recorded scene given the scene context. In the third one, we formalise a transferring approach to relate people detection across cameras based on scene off-line obtained context—camera calibration and plane-induced homographies—. Finally, in the fourth one, we define an approach to transfer and adapt semantic maps—information of contextual objects in the scene—in a PTZ-scenario recorded by three moving cameras.

Our future work will be focused on the design of strategies to incorporate contextual information into diverse analysis approaches. Updates on ongoing approaches will be provided in a future deliverable [38].

5. References

[1]. Elisa Drelie Gelasca, Joriz De Guzman, Steffen Gauglitz, Pratim Ghosh, JieJun Xu, Emily Moxley, Amir M. Rahimi, Zhiqiang Bi and B. S. Manjunath, "CORTINA: Searching a 10 Million + Images Database", VRL, ECE, University of California, Santa Barbara, Sep. 2007.

[2]. Felzenszwalb, P. F., & Huttenlocher, D. P. Efficient graph-based image segmentation. International Journal of Computer Vision, Volume 59, Issue 2, p. 167-181. 2004.

[3]. Desarrollo de herramienta para la anotación manual de secuencias de vídeo, Yoel Witmaar García (advisor Juan Carlos), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Julio 2015,

[4]. Diseño y desarrollo de una herramienta para la extracción semi-supervisada de la información de contexto, Raúl García Jiménez (advisor: Marcos Escudero Viñolo), Trabajo Fin de Grado, Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Jul. 2016.

[5]. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence, Volume 34, Issue 11, p. 2274-2282. 2012.

[6]. Navarro, F., Escudero-Viñolo, M., & Bescós, J. Sp-sift: enhancing sift discrimination via super-pixel-based foreground-background segregation. Electronics Letters, Volume 50, Issue 4, p. 272-274. 2014.

[7]. Lowe, D. G. Object recognition from local scale-invariant features. In Computer vision, 1999. The proceedings of the seventh IEEE international conference on Volume 2, p. 1150-1157). 1999.

[8]. Tola, E., Lepetit, V., & Fua, P. Daisy: An efficient dense descriptor applied to wide-baseline stereo. IEEE transactions on pattern analysis and machine intelligence, Volume 3, Issue 5, p. 815-830. 2010.

[9]. Salti, S., Tombari, F., & Di Stefano, L. SHOT: unique signatures of histograms for surface and texture description. Computer Vision and Image Understanding, Volume 125, p. 251-264. 2014.

[10]. Contributions to region-based image and video analysis: feature aggregation, background subtraction and description constraining, Marcos Escudero Viñolo (advisor: Jesús Bescós Cano), Escuela Politécnica Superior, Universidad Autónoma de Madrid, Enero. 2016.

[11]. Detección de la posición relativa de una cámara en situaciones de grandes desplazamientos (Detection of camera relative position in wide-baseline scenarios), María Narváez Encinal (advisor: Marcos Escudero Viñolo), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Jun. 2015.

[12]. Lindeberg, T. Scale selection properties of generalized scale-space interest point detectors. Journal of Mathematical Imaging and Vision, Volume 46, Issue 2, p. 177-210. 2013.

[13]. HaCohen, Y., Shechtman, E., Goldman, D. B., & Lischinski, D. Non-rigid dense correspondence with applications for image enhancement. ACM transactions on graphics (TOG), Volume 30. Issue 4, p. 70. 2011.

[14]. Control de cámaras PTZ para la reconstrucción de escena basada en puntos de interés, Elisa Martín Pérez (advisor: Fulgencio Navarro Fajardo), Trabajo Fin de Grado, en Ingeniería de Tecnologías y Servicios de Telecomunicación, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Jul. 2016.

[15]. Detección de vehículos mediante el análisis de imágenes (Image-based vehicle detection), Juan Ignacio Bravo (advisor: Luis Salgado), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Jun. 2015.

[16]. Detección de personas en entornos multicámara utilizando información contextual, Alejandro Miguélez Sierra (advisor: Rafael Martín), Trabajo Fin de Grado, Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Ene. 2016.

[17]. C. Wolf, J. Mille, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandrea, C.-E. Bichot, C. Garcia, and B. Sankur, _Evaluation of video activity localizations integrating quality and quantity measurements,_ vol. 127, no. 1, pp. 14_30, 2014.
[18]. 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, no. 9, pp. 1627_1645, 2010.
[19]. D2.2/TR2.2 CONTEXTUAL MODELLING AND EXTRACTION FOR PEOPLE BEHAVIOUR UNDERSTANDING. Version 1. June 2016. http://www-vpu.eps.uam.es/HAVideo/public_docs/D2.2v1(HAVideo)Contextual%20modelling%20and%20extraction%20for%20people%20behaviour%20understanding.pdf
[20]. Marcos Escudero-Viñolo, Jesús Bescós, "Severe-occluded 3D object identification via region-based descriptions". Resubmitted after major revision to Signal Processing: Image Communication. April, 2017.
[21]. Automatic video classification using feature descriptors, María Narváez Encinal (advisor Álvaro García), Trabajo Fin de Máster (Master Thesis), Máster Universitario en Ingeniería de Telecomunicación, Escuela Politécnica Superior, Universidad Autónoma de Madrid, scheduled to June/July 2017.
[22]. T. Senst, V. Eiselein, and T. Sikora, A local feature based on lagrangian measures for violent video classification, in 6th International Conference on Imaging for Crime Prevention and Detection (ICDP-15), pp. 1-6, IET, 2015.
[23]. A. Kuhn, T. Senst, I. Keller, T. Sikora, and H. Theisel, A lagrangian framework for video analytics, in Multimedia Signal Processing (MMSP), 2012 IEEE 14th International Workshop on, pp. 387-392, IEEE, 2012.
[24]. F. Navarro, M. Escudero-Vinolo, and J. Bescós, Sp-sift: enhancing sift discrimination via super-pixel-based foreground-background segregation, Electronics Letters, vol. 50, no. 4, pp. 272-274, 2014.
[25]. Hu, W., Yang, Y., Zhang, W., Xie, Y., Feb 2017. Moving object detection using tensor-based low-rank and saliently fused-sparse decomposition. IEEE Transactions on Image Processing 26 (2), 724–737.
[26]. Goyette, N., Jodoin, P.-M., Porikli, F., Konrad, J., Ishwar, P., 2012. Changedetection.net: A new change detection benchmark dataset. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society 541Conference on. IEEE, pp. 1–8.
[27]. Wang, Y., Jodoin, P.-M., Porikli, F., Konrad, J., Benezeth, Y., Ishwar, P., 2014. Cdnet 2014: An expanded change detection benchmark dataset. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on. IEEE, pp. 393–400.
[28]. Marcos Escudero-Viñolo, Jesús Bescós, "Squeezing the DCT to fight camouflage". Submitted to Image and Vision computing. April, 2017.
[29]. Kuettel, Daniel, and Vittorio Ferrari. "Figure-ground segmentation by transferring window masks." Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.
[30]. B. Russell, A. Torralba, K. Murphy, W. T. Freeman.LabelMe: a database and web-based tool for image annotation . International Journal of Computer Vision, 2007.
[31]. Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso and Antonio Torralba. Scene Parsing through ADE20K Dataset. Computer Vision and Pattern Recognition (CVPR), 2017.
[32]. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[33]. Spatio-temporal segmentation of contextual objects in video, Sergio Serra (advisor Marcos Escudero-Viñlo), Trabajo Fin de Grado (Graduate Thesis), Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación, Escuela Politécnica Superior, Universidad Autónoma de Madrid, scheduled July 2017.
[34]. R. Martín-Nieto, A. Miguelez-Sierra, A. García-Martín and José M.Martinez, "Improving Multi-camera detection using contextual information". Submitted to International Conference on Image Processing, ICIP 2017. January 2017.
[35]. Online contextual updating in multi-camera scenarios, Alejandro López Cifuentes (advisor Marcos Escudero-Viñolo), Trabajo Fin de Máster (Master Thesis), International Master-level Program in Image Processing and Computer Vision, Escuela Politécnica Superior, Universidad Autónoma de Madrid, scheduled to June/July 2017.

[36]. Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2016). Pyramid Scene Parsing Network. arXiv preprint arXiv:1612.01105.
[37]. Alcantarilla, P. F., & Solutions, T. (2011). Fast explicit diffusion for accelerated features in nonlinear scale spaces. IEEE Trans. Patt. Anal. Mach. Intell, 34(7), 1281-1298.
[38]. Online adaptive people behaviour understanding based on contextual and quality information (version 3). Scheduled to December 2017.