HAL
open science

FAST AND RELIABLE OBJECT CLASSIFICATION IN VIDEO BASED ON A 3D GENERIC MODEL

Marcos Zuniga, François Bremond, Monique Thonnat

To cite this version:

Marcos Zuniga, François Bremond, Monique Thonnat. FAST AND RELIABLE OBJECT CLASSIFICATION IN VIDEO BASED ON A 3D GENERIC MODEL. The 3rd International Conference on Visual Information Engineering (VIE 2006), Sep 2006, Bangalore, India. inria-00515615

HAL Id: inria-00515615

https://inria.hal.science/inria-00515615v1

Submitted on 7 Sep 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

FAST AND RELIABLE OBJECT CLASSIFICATION IN VIDEO BASED ON A 3D GENERIC MODEL

Marcos Zúñiga, François Brémond, Monique Thonnat

INRIA Sophia Antipolis, ORION group
2004, route des Lucioles, BP93
06902 Sophia Antipolis Cedex — France
(33) 4 9238 7657
firstname.surname@sophia.inria.fr

Keywords: video interpretation, object classification, 3D object model, reliability measure.

Abstract

We propose a new object classification approach for monocular video sequences, which allows to classify objects modelled independently from the camera position and object orientation. To achieve this independence, a simple 3D object model that represents an object as a parallelepiped is proposed. The approach is able to give good estimates of object dimensions and proposes visual reliability measures for the object dimensions. These measures give a representation of the visibility of the estimated dimension and are principally proposed to aid posterior phases of the video understanding process, as object tracking and event detection. The method obtains the 3D parallelepiped model estimation using a set of 2D moving regions (obtained in a segmentation phase), the perspective matrix transform (obtained from camera calibration using the pin-hole camera model) and predefined 3D models of expected objects in the scene. After classification, a merging step is performed to improve the classification performance by assembling 2D moving regions with better 3D model probability when together. This approach shows promising results on object classification, obtaining very high detection rates for complex situations and performing at video frame rate.

1 Introduction

Binocular visual perception allows human beings to perceive depth of their environment. At the same time, a person can shut one of his/her eyes and still preserve the depth sensation, without loosing too much of precision on depth estimation of the focused object. This capability is a consequence of the interpretation that the brain performs about the new visual information, by associating it to similar environments or objects previously observed, and then concluding on its nature and 3D shape. This means that the brain uses a priori knowledge to conclude about the attributes (e.g. position, dimensions) of an observed object.

Following this idea, we propose a new object classification approach for monocular video sequences using a simple 3D model of the expected objects in the scene. The proposed approach allows to classify objects of different nature in a way that is independent from the relative position between the object and the camera, considering a pin-hole camera model. For this purpose, we propose a simple 3D object model that represents an object as a parallelepiped. The model is described by the parallelepiped dimensions (width, length and height) and orientation in the ground plane of the scene. Also, visual reliability measures of the three estimated dimensions are proposed, which represent a measure of their visibility. These measures are intended to aid to fairly classify objects according to the more visually significant attributes. These measures have been principally proposed to aid posterior phases of the video understanding process, as dimensional estimation of tracked objects, multi-camera object fusion, and discrimination between visually reliable data from purely estimated data on event detection and learning.

Our approach tries to cope with several limitations imposed by 2D representations, but keeping their capability of being general models able to describe different objects and still being able to work in real-time. For more details, refer to section 2.

For implementing our approach a platform for image sequence understanding called VSIP (Video Surveillance Interpretation Platform) is used, which was developed at the research group ORION at INRIA (Institut National de Recherche en Informatique et en Automatique), Sophia Antipolis. VSIP is a generic environment for combining algorithms for processing and analysis of videos. This platform allows to flexibly combine and exchange various techniques at the different stages of the video understanding process. Moreover, VSIP is oriented to help developers describing their own scenarios and building systems capable of monitoring behaviours, dedicated to specific applications.

The platform corresponds to a two-level architecture. At the first level, VSIP extracts primitive geometric features like areas of motion. Based on them, objects are recognised and tracked.

At the second level those events in which the detected objects participate, are recognised. Examples of this two-level architecture can be found in the works of [7] and [9].
We have used this platform at its first level (see figure 1), applying a background subtraction method for segmentation to obtain a set of 2D moving regions. Then, the classification phase uses the obtained moving regions, the perspective matrix of the scene, and predefined 3D parallelepiped models of expected objects on the scene, to find the most likely 3D model of the objects. Finally, a merging step is performed to improve the classification performance by assembling 2D moving regions showing a better 3D object likelihood when they are put together. The perspective matrix of the scene is previously obtained from an off-line camera calibration phase, considering the pin-hole camera model. Classification and merging processes will be described in detail on section 3.
This paper is organised as follows. Section 2 presents state-of-

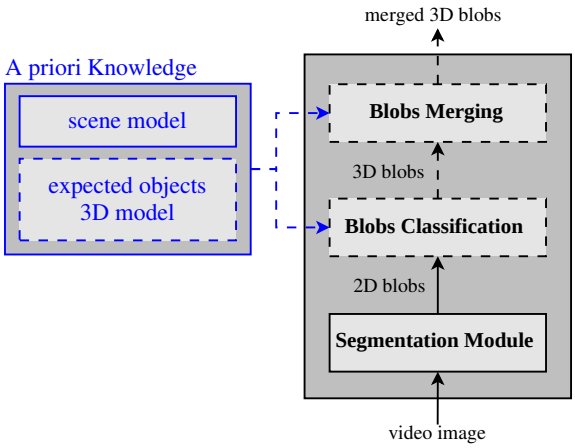

Figure 1: Proposed architecture of object classification approach. The steps depicted in the figure describe the data flow and processing modules during an object classification process. Dashed lines emphasise the main contributions of this work.

the-art on object classification in video. In section 3, we present a detailed description of the proposed 3D model and classifier. Section 4 presents and discusses the obtained results from 20 short sequences, extracted from two long duration videos.

2 Related Work

The objective of this section is to expose the principal advantages e inconveniences of using 2D object representations, but also the drawbacks of using very precise and detailed object representations to perform object detection in video.
2D representations have been used in several applications, with acceptable detection and classification performance. This representation has several advantages which justify its use. For certain applications, two dimensions are enough to describe the objects involved in the analysed scene, because: (a) The relative position between the camera and the observed object hides one dimension (e.g. tracking groups of people in a metro scene [6]), meaning that can be enough to model a 3D object with a 2D model. (b) The estimation of the other dimension is performed by merging information from different cameras (e.g. human posture detection [5], apron monitoring application on an airport [2]). (c) Object detection can be more interesting than classification for certain applications (e.g. detection of stopped vehicles in a highway [4]). Certainly, the processing time spent in calculating the attributes associated to 2D representations is inexpensive, allowing to cope with real-time constraint. These 2D models are sufficient to find the 3D position of an object, which is enough for certain applications.
Nevertheless, 2D representations present also several drawbacks, that make them useless for many applications. In situations where there are no multiple cameras or it does not exist an overlap between views on the zone of interest, the third dimension cannot be estimated by merging cameras information. If the 2D moving region considerably changes its appearance depending on its position relative to camera (see figure 2), dimensional estimation becomes unreliable. If the 2D representation considerably changes when the object rotates (see figure 3), dimensional estimation becomes also unreliable. For deformable objects (e.g. persons changing their posture), it would become a very hard task to define a 2D representation for each possible deformation of an object of this nature, considering that it can also change according to different positions relative to camera and different object orientations.
On the other extreme, different models have been proposed

(a) (b)

Figure 2: 2D moving region deformation by different positions of an object relative to the camera. Here, the same person (with same posture) is represented by very dissimilar 2D regions in the same video sequence. In figure (a) the person is far from the camera and it is possible to see his height, while in figure (b) the person is seen almost from top and almost nothing can be said about his height.

for specific objects (e.g. persons, vehicles), which are application and object dependent. Some authors use precise models of a specific object to perform detection. These models allow generally to obtain quite good detection rates and attribute estimations, but the computational cost associated to its utilisation is often too expensive to be real-time. [1] uses a 2D model of each body part of a human constrained by image motion parameters to perform tracking of walking persons and human gestures. [3] uses a very precise 3D model of human to detect

(a) (b)

Figure 3: 2D moving region deformation caused by change of orientation of the object. Here, the same car is represented by very dissimilar 2D regions in the same video sequence. In figure (a) a car is seen from its back part. Later, the car rotates to park and it is seen from its right part, as seen in figure (b).

postures. In this work, a human posture is described by a set of 23 parameters, subject to biomechanical constraints. This human model is used to generate silhouettes to be compared with the one detected for a person in the scene.
Other authors train classifiers with examples they expect to find in their applications. One of the precursors of this type of approach are [8]. The authors propose to train a system in the detection of basic features from an object, and to combine these basic features to construct a strong classifier, based on Adaboost algorithm. They present their method with an application to frontal view face detection, with high detection rates. A considerable number of works have taken this kind of approach. The problem of these methods is their dependence to a determined object rotation and camera position relative to the object position (e.g. in [8], the face of a person seen from one side would not be detected). Another problem is that the detection is restricted to objects similar to the training samples.
Our approach tries to cope with the majority of the limitations imposed by 2D models, but being general enough to be capable of modelling a large variety of objects and still preserving the capability to perform in real-time. Also, the utilisation of 2D moving regions as input allows our method to inexpensively cope with the problem of object translation, giving a good initial estimate of the position of detected objects. Next section explains how the method has been developed.

3 The 3D Object Classification

In this section we first describe the proposed 3D model and its attributes (section 3.1) and the proposed visual reliability measures. Then, in section 3.2, we present the classification approach and explain how the merging process works.
The proposed method applies a background subtraction method to perform segmentation. This method consists on a set of tests in different colour spaces applied to each pixel for discrimination between foreground and background pixels. Then, moving regions are detected and represented as a set of 2D bounding boxes (i.e. 2D blobs). Next, the classification phase uses the obtained 2D blobs, the perspective matrix of the scene and predefined 3D parallelepiped models of the expected objects on the scene, to find the most likely 3D model of each object. Finally a merging phase is performed to improve the classification performance by assembling 2D blobs showing a better 3D object likelihood when they are put together. The perspective matrix of the scene is previously obtained from an off-line camera calibration phase, considering the pin-hole camera model. In the following section we present the mentioned 3D parallelepiped model.

3.1 The 3D Parallelepiped Model

A large variety of objects can be modelled (or, at least, enclosed) by a parallelepiped. The proposed model is defined as a parallelepiped perpendicular to the ground plane of the analysed scene. Starting from the basis that a moving object will be detected as a 2D blob b with 2D limits $(X_{left}, Y_{bottom}, X_{right}, Y_{top})$, 3D dimensions can be estimated based on the information given by predefined 3D parallelepiped models of the expected objects on the scene.
An attribute model $\tilde{\mathbf{q}}$, for an attribute q can be defined as:

$$\tilde{\mathbf{q}} = (Pr_q(\mu_q, \sigma_q), q_{min}, q_{max}) \tag{1}$$

where Pr_q is a probability distribution described by mean μ_q and standard deviation σ_q, where $q \sim Pr_q(\mu_q, \sigma_q)$. q_{min} and q_{max} represent the minimal and maximal values for the attribute q, respectively.
Then, a predefined 3D parallelepiped model $Q_{\mathbf{C}}$ for an object class $\mathbf{C}$ can be defined as:

$$Q_{\mathbf{C}} = (\tilde{\mathbf{w}}, \tilde{\mathbf{l}}, \tilde{\mathbf{h}}) \tag{2}$$

where $\tilde{\mathbf{w}}$, $\tilde{\mathbf{l}}$, and $\tilde{\mathbf{h}}$ represent the attribute models for the 3D attributes width, length and height, respectively.
For the applications presented in this work, attributes w, l and h have been modelled as Gaussian probability distributions with parameters (μ_w, σ_w), (μ_l, σ_l) , and (μ_h, σ_h), respectively.
A 3D parallelepiped model $S_{\mathbf{O}}$ for an object $\mathbf{O}$ detected in the scene (see figure 4) is described by:

$$S_{\mathbf{O}} = (\alpha, (w, R_w), (l, R_l), (h, R_h)) \tag{3}$$

where α represents the parallelepiped orientation angle (figure 4(b)), defined as the angle between the direction of length 3D dimension and x axis of the world referential of the scene. w, l and h represent the 3D values for width, length and height of the parallelepiped, respectively. l is defined as the 3D dimension which direction is parallel to the orientation of the object. w is the 3D dimension which direction is perpendicular to the orientation. h is the 3D dimension parallel to the z axis of the world referential of the scene. R_w, R_l and R_h are 3D visual reliability measures for each dimension. These measures represent the confidence on the visibility of each dimension of the parallelepiped and are described below.
For obtaining the dimensions of the 3D model we need to calculate the 3D position of the vertexes of the parallelepiped in

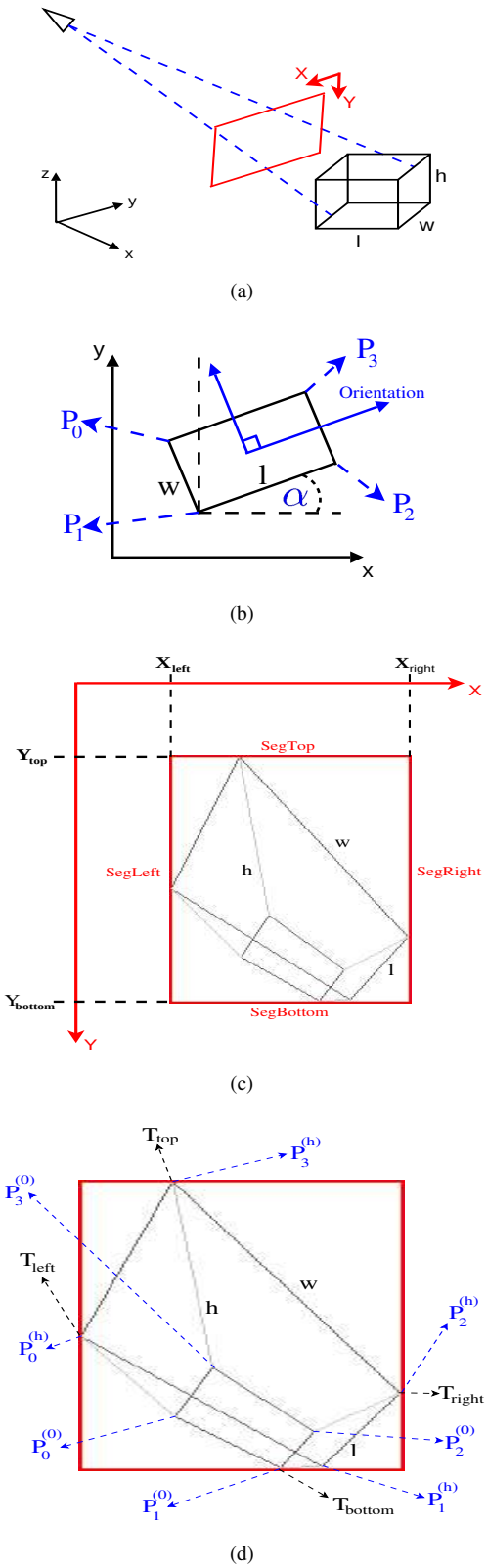

Figure 4: 3D parallelepiped model for detected objects. (a) 3D view of the scene. (b) Top view of the scene. (c) Point of view from the camera explaining image 2D referential variables. (d) Point of view from the camera explaining world 3D referential variables.

the world referential of the scene. We define $P_i^z(x_i, y_i)$, with $i \in \{0,1,2,3\}$ and $z \in \{0,h\}$, as the eight 3D points (x_i, y_i, z) that define the parallelepiped vertexes, with $P_i^{(0)}$ corresponding to the i-th base point and $P_i^{(h)}$ corresponding to the i-th vertex on height h, as shown in figure 4(d). We also define P_i, with $i \in \{0,1,2,3\}$, as the coordinates for each vertical edge of the parallelepiped on plane xy of the world referential of the scene, as depicted in figure 4(b). Then, w and l are defined below.

$$\begin{aligned} w &= \mathbf{d}(P_0, P_1) = \mathbf{d}(P_2, P_3) \\ l &= \mathbf{d}(P_1, P_2) = \mathbf{d}(P_3, P_0) \end{aligned} \tag{4}$$

where $\mathbf{d}(\cdot,\cdot)$ is the euclidean distance function.
We want to find a parallelepiped bounded by the limits of the 2D blob b. For this purpose four line segments are defined, as depicted in figure 4(c):
SegLeft: Defined by points $[(X_{left}, Y_{top}); (X_{left}, Y_{bottom})]$.
SegBottom: Defined by points $[(X_{left}, Y_{bottom}); (X_{right}, Y_{bottom})]$.
SegRight: Defined by points $[(X_{right}, Y_{top}); (X_{right}, Y_{bottom})]$.
SegTop: Defined by points $[(X_{left}, Y_{top}); (X_{right}, Y_{top})]$.
Then, we define points $(T_{left}, T_{right}, T_{top}, T_{bottom}) \in \{P_i^z | i \in \{0,1,2,3\}, z \in \{0,h\}\}$ as the vertexes that comply with equations (5).

$$\begin{aligned} \textbf{ImageProjection}(T_{left}) &\in SegLeft \\ \textbf{ImageProjection}(T_{right}) &\in SegRight \\ \textbf{ImageProjection}(T_{top}) &\in SegTop \\ \textbf{ImageProjection}(T_{bottom}) &\in SegBottom \end{aligned} \tag{5}$$

where **ImageProjection**$(\cdot)$ is a function that projects a point from the world referential of the scene onto the image plane.
Considering h as a parameter, the points P_i, with $i \in \{0,1,2,3\}$ define eight variables to be solved. Therefore, at least eight equations are required to find the solutions for these variables. Using the pin-hole camera model equation (6), with $\mathbf{M}$ the calibrated perspective matrix, and the four relations of equation (5), four linear equations can be defined between each pair of variables (x_i, y_i) from each point P_i, with $i \in \{0,1,2,3\}$, as depicted in figures (4(c)) and (4(d)).

$$\begin{pmatrix} X_t \\ Y_t \\ t \end{pmatrix} = M \cdot \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{6}$$

Considering α also as parameter, the following equations must also be solved in order to obtain the values for all vertexes:

$$\begin{aligned} (y_2 - y_1)/(x_2 - x_1) &= \tan(\alpha) \\ (y_3 - y_2)/(x_3 - x_2) &= \tan(\alpha + \pi/2) \\ (y_0 - y_3)/(x_0 - x_3) &= \tan(\alpha + \pi) \\ (y_1 - y_0)/(x_1 - x_0) &= \tan(\alpha + 3\pi/2) \end{aligned} \tag{7}$$

These equations are derived from the equation of the angle that must be formed between each parallelepiped base line and the x axis of the world referential of the scene, given the orientation α. It is also considered that, for each line formed by the base points, the next line must be rotated on $\pi/2$ degrees to form the

rectangular base of the parallelepiped.
It is now possible to calculate the values for all vertexes P_i^z, with $i \in \{0,1,2,3\}$ and $z \in \{0,h\}$. Then, w and l can be determined by calculating the distance between the obtained 3D points (see equation (4)). Then, considering perspective matrix M and 2D blob $b = (X_{left}, Y_{bottom}, X_{right}, Y_{top})$, a parallelepiped model $S_{\mathbf{O}}$ for a detected object $\mathbf{O}$ can be completely defined as a function f:

$$S_{\mathbf{O}} = f(\alpha, h, M, b) \tag{8}$$

Equation (8) defines how to obtain the three dimensions of a parallelepiped starting from its height h, orientation α, 2D blob b limits and calibration matrix M.
The visual reliability measures remain to be determined and are described below.

Visual Reliability Measures

A reliability measure R_q for a dimension $q \in \{w,l,h\}$ is intended to quantify the visual evidence for the estimated dimension, by visually analysing how much of the dimension can be seen from the camera point of view. The objective is to find a measure that give a minimal value (e.g. 0) when attribute is not visible, and a maximal value (e.g. 1) when the dimension is totally visible. We have chosen to find a function $R_q(S_{\mathbf{O}}) \rightarrow [0,1]$, where visual reliability of the attribute is 0 if the attribute is not visible and 1 if is completely visible.
Considering $P_j \in \{P_i, i \in \{0,1,2,3\}\}$ as the nearest point to (x_c, y_c), where (x_c, y_c, z_c) is the 3D position of the focal point of the camera, then point (x_j, y_j) is considered as the point having the best visibility for height 3D dimension. Defining $X_{2D}(P_k^{(z_k)})$ as the X coordinate and $Y_{2D}(P_k^{(z_k)})$ as the Y coordinate of the image projection of a 3D point $P_k^{(z_k)}$ of height z_k onto the image plane, we can define R_h as shown in equation (9).

$$R_h = \max\left(\frac{dY_h}{||SegLeft||}; \frac{dX_h}{||SegBottom||}\right) \tag{9}$$

where $dY_h = |Y_{2D}(P_j^{(0)}) - Y_{2D}(P_j^{(h)})|$ and $dX_h = |X_{2D}(P_j^{(0)}) - X_{2D}(P_j^{(h)})|$, and operator $||\cdot||$ determines the magnitude of its argument.
In the same way, we can define reliability measures for R_l and R_w as shown in equations (10) and (11), respectively. These measures represent visual reliability as the maximal magnitude of projection of a 3D dimension onto the image plane, in proportion with the magnitude of each 2D blob limiting segment. Thus, the maximal value 1 is achieved if the image projection of a 3D dimension has the same magnitude compared with one of the 2D blob segments.

$$R_l = \max\left(\frac{dY_l}{||SegLeft||}; \frac{dX_l}{||SegBottom||}\right) \tag{10}$$

where $dY_l = |Y_{2D}(P_1^{(h)}) - Y_{2D}(P_2^{(h)})|$ and $dX_l = |X_{2D}(P_1^{(h)}) - X_{2D}(P_2^{(h)})|$.

$$R_w = \max\left(\frac{dY_w}{||SegLeft||}; \frac{dX_w}{||SegBottom||}\right) \tag{11}$$

where $dY_w = |Y_{2D}(P_0^{(h)}) - Y_{2D}(P_1^{(h)})|$ and $dX_w = |X_{2D}(P_0^{(h)}) - X_{2D}(P_1^{(h)})|$.
The following section describes how the method performs classification.

3.2 The 3D Object Classifier

The method searches the optimal fit for each predefined parallelepiped model, scanning different values of h and α. After finding optima for each class based in the probabilistic measure PM (defined in equation (12)), the method decides the class of the analysed blob by using the reliability measure RM, defined in equation (13). This operation is performed for each blob on the current frame.

$$PM(S_{\mathbf{O}}, C) = \prod_{q \in \{w,l,h\}} Pr_q(q|\mu_q, \sigma_q) \tag{12}$$

$$RM(S_{\mathbf{O}}, C) = \frac{\sum_{q \in \{w,l,h\}} R_q \times Pr_q(q|\mu_q, \sigma_q)}{\sum_{q \in \{w,l,h\}} R_q \times Pr_q(\mu_q|\mu_q, \sigma_q)} \tag{13}$$

Given a perspective matrix $\mathbf{M}$, object classification is performed for each blob b from the current frame in the following way:

For each class C of predefined models
  For all valid pairs (h, α)
    $S_O \leftarrow F(\alpha, h, M, b)$;
    if $PM(S_O, C)$ improves best current fit $S_O^{(C)}$ for C,
    then update optimal $S_O^{(C)}$ for C;
Class$(b) = argmax_C(RM(S_O^{(C)}, C))$;

Equation (12) presents the PM criteria for comparing between different configurations for a same class. The idea is to first find the most probable configuration, regardless the visual reliability of its attributes. Equation 13 presents the criteria of comparison between classes. The justification of this measure is that dimensional attributes that are not visually reliable will not be considered (or will be less considered) for comparison between classes. This measure makes more fair comparisons between classes in the sense that classes with very good dimensional estimates (very likely) of not visible attributes will not be rewarded on their evaluation, because there is no visual evidence of those attributes. We search the most probable class in terms of the high visibility of the attributes.
Finally, after classification, a merging module is applied. This module searches for blob pairs that are near in distance and where the resulting blob obtained from merging them gives a higher likelihood for a class, compared with the results obtained in the classification phase.
Next section presents the obtained results from 20 short sequences, extracted from two long duration videos of different nature.

(a) (b) (c) (d)

Figure 5: Results for different frames of the parking video. Figures (a), (b) and (c) correspond to zoomed versions of captured frames. Figure (d) shows the real view from the camera.

Name	Description	TP	FP	FN	CM[fr/s]	Total[fr/s]	CM[bl/s]	Total[bl/s]	Precision	Sensitivity
Borel 1	1 car parking right	20	0	0	156.27	44.85	164.08	47.09	1.0	1.0
Borel 2	1 person to bottom	20	0	0	161.32	52.36	161.32	52.36	1.0	1.0
Borel 3	1 car going far	20	0	0	49.15	30.82	93.38	58.56	1.0	1.0
Borel 4	1 car parking left	20	3	0	5.55	5.07	34.7	31.67	0.8596	1.0
Borel 5	1 car and 1 person	39	3	1	5.22	4.84	40.2	37.28	0.9286	0.9750
Borel 6	2 cars	40	0	0	17.42	8.83	94.96	48.11	1.0	1.0
Borel 7	2 persons bottom	40	0	0	47.97	26.01	100.74	54.63	1.0	1.0
Borel 8	1 person very far	20	0	0	224.76	39.89	224.76	39.89	1.0	1.0
Borel 9	2 persons walking	40	0	0	16.29	10.22	38.28	24.01	1.0	1.0
Borel 10	2 cars very near	40	0	0	18.69	12.91	83.19	57.47	1.0	1.0
	Mean Values	29.9	0.6	0.1	70.26	23.58	103.56	45.11	0.9798	0.9975

Table 1: Obtained results for parking video.

4 Results

Two videos have been tested for validating this approach. The first video corresponds to a parking sequence where cars and persons interact. Two object models are used for this sequence. The evaluation objective of this video is to validate the capability of the approach for coping with the problem of object rotation and relative position to camera. The second video corresponds to a lock chamber from a bank camera, with high deformation of the detected blobs because of the proximity of persons to the camera. For this video, three models representing one person and groups of two and three persons are defined (the space of the chamber allows a maximum of three persons at the same moment). The lock chamber video is used to validate the approach capability to detect highly deformed objects and to differentiate between very similar classes. Ten short sequences of 20 frames have been selected from each video, giving a total of 400 analysed frames. The selected sequences avoid occlusion situations (because it is not the scope of this work and this feature is not yet considered on the approach), but consider situations of different distances between objects and the camera focal point, and different object orientations. A computer Intel Pentium IV, Xeon 3000 Mhz, have been used for performing the tests. For each sequence, we have counted True Positive (TP) as the objects which class corresponds with the ground truth, False Positive (FP) as the classification of an object that is not in the ground truth, and False Negative (FN) as the misclassification of an object that is in the ground truth. This means that classifying an object with a class different from ground truth is considered as two errors at the same time (one FP and one FN), while not classifying it at all meant just a FN. We have also calculated $precision = TP/(TP+FP)$ and $sensitivity = TP/(TP+FN)$. For real-time capability measurement we have calculated $[frames/second]$ ($[\mathbf{fr}/\mathbf{s}]$) and $[blobs/second]$ ($[\mathbf{bl}/\mathbf{s}]$) rates for Total procedure time (Segmentation, Classification and Merging) and for the time spent just in Classification and Merging (referred as **CM** in tables 1 and 2). In figures 5 and 6, each detected object is enclosed by a 2D bounding box and by the corresponding 3D parallelepiped. The base of parallelepiped is represented by blue lines, while projected lines in height h are represented by green lines. 2D bounding boxes take different colours according to the classified object (person: red, 2 persons: green, 3 persons: blue, car: brown). Cars in parking sequence that seem not detected are considered as part of the background of the scene.

For the parking sequence, 3D models for persons and cars were predefined. The results for this sequence are shown in table 1 and images of these results are shown in figure 5. Parking results show a very good performance, obtaining a global precision of 0.98. The encountered errors have been principally caused by poor segmentation in some frames because of illumination changes. The method have been able to discriminate

(a) (b) (c) (d)

Figure 6: Results for different frames of the bank locked chamber video. Four frames for the selected sequences are shown.

Name	Description	TP	FP	FN	CM[fr/s]	Total[fr/s]	CM[bl/s]	Total[bl/s]	Precision	Sensitivity
Sas 1	1 p. with folder	20	0	0	78.75	21.01	129.94	34.67	1.0	1.0
Sas 2	1 mean height p.	20	0	0	50.64	20.54	101.28	41.07	1.0	1.0
Sas 3	1 tall p.	17	3	3	37.46	15.20	69.30	28.12	0.8500	0.8500
Sas 4	2 p. semi-ext. arms	20	0	0	45.36	13.33	77.11	22.65	1.0	1.0
Sas 5	2 p. not aligned	18	2	2	37.11	15.36	81.65	33.80	0.9000	0.9000
Sas 6	2 p. aligned	20	0	0	109.90	26.89	109.90	26.89	1.0	1.0
Sas 7	2 p. extended arms	15	5	5	77.53	23.45	93.04	28.14	0.7500	0.7500
Sas 8	3 p. 1	20	0	0	81.65	19.53	126.55	30.28	1.0	1.0
Sas 9	3 p. 2	19	1	1	26.92	13.77	63.27	32.35	0.9500	0.9500
Sas 10	3 p. 3	20	0	0	74.36	21.23	96.67	27.60	1.0	1.0
	Mean Values	18.9	1.1	1.1	61.97	19.03	94.87	30.56	0.9450	0.9450

Table 2: Obtained results for bank locked chamber video.

objects at different orientations and positions relative to the camera. For instance, figure 5(a) shows the same person in two different frames detected as a person, showing the method capability for coping with different positions relative to the camera. In figure 5(b) a very difficult to detect person, because of its distance to the camera (left image), is successfully detected in the classification phase (right image). Figures 5(c) and 5(d) show the capability of the method for coping with different positions and orientations of cars and for coping with more than one object class at the same frame.

For the bank locked chamber sequence, models for one, two and three persons have been defined, where the model of one person is identical to the person model used in the parking video. The results for the bank locked chamber sequence are shown in table 2 and images of these results are shown in figure 6. Locked chamber results show a very good performance, obtaining a global precision of 0.95. The encountered errors have been principally caused by the proximity between predefined models. The obtained results for some sequences are sometimes very similar with the next class (one person similar with two persons, or two persons similar with three) because of some postures and configurations of persons, that lead to some misclassification. However, in terms of results, the method shows the different configurations with similar likelihood that could occur, which could be a beneficial situation for other purposes. Figures 6(a), 6(c), and 6(d) show examples of classification for the three different classes. Figure 6(b) shows the case of a tall person, who has been sometimes misclassified as two persons.

An application for the bank locked chamber sequence consists in generating alarms if more than one person is at the same time in the locked chamber. In this case, a TP corresponds to the detection of more than one person when more than one person is present on the scene, a True Negative (TN) corresponds to the detection of one or zero persons when one or zero persons are in the scene, a FP corresponds to the detection of more than one person when one or zero persons are present in the scene, and FN corresponds to the detection of one or zero persons when more than one person is in the scene. Here, 140 TP, 57 TN, 3 FP and 0 FN were found, giving a precision of 0.98 and a sensitivity of 1. Table 3 shows the classification results. Each row represents ground truth and each column represents the detected object. Notice that committed errors were always associated with the detection of more or less one person, compared with the real number of persons.

In both cases, results are obtained in real-time. Notice that the performance of the complete method (considering Segmentation, Classification and Merging phases) is about $20[frames/sec]$, while the method performance considering only Classification and Merging phases is about $65[frames/sec]$, showing that the time spent in the classification and merging phases is inexpensive compared with the time spent in the segmentation phase.

	1p	**2p**	**3p**
1p	57	3	0
2p	0	73	7
3p	0	1	59

Table 3: Classification results for bank locked chamber video, for objects one-person (**1p**), two-persons (**2p**) and three-persons (**3p**).

5 Conclusion

The proposed classification method has shown promising results in object classification. First, the proposed approach has been able to cope principally with the problems of object position relative to the camera position, object orientation and dimensional deformation caused by camera proximity, with high classification rates. Second, the method has shown its capability of performing at video frame rate.

One of the principal limitations of the method is its inability of discrimination between more than one situation geometrically plausible, because it does not use pixel-level information of moving regions. However, sometimes it is more appropriate to postpone the decision of classification to later phases when more information is available.

Visual reliability measures are intended to be used by further video interpretation phases, as data integration in multi-camera data fusion and discrimination between visually plausible data from pure data estimation to aid in the detection and learning of primitive states and events.

Future work comprises the integration of this method with object tracking techniques, the capability of coping with occlusion situations and the improvement of the computation time. The problem of static occlusion (moving object occluded by a static object or image borders) can be treated with the information obtained from the proposed classification approach, but dynamic occlusion (a moving object occluded by another moving object) will require more information that can be obtained from tracking techniques.

Acknowledgements

This research has been supported in part by the Science and Technology Research Council of Chile (CONICYT) in the framework of INRIA (Sophia-Antipolis) and CONICYT cooperation agreement.

References

[1] M. Black, Y. Yacoob, X. Ju. "Recognizing human motion using parameterized models of optical flow", in Mubarak Shah, Ramesh Jain, editors, *Motion-Based Recognition*, pp. 245–269. Kluwer Academic Publishers, Boston, (1997).

[2] M. Borg, D. Thirde, J. Ferryman, F. Fusier, V. Valentin, F. Brémond, M. Thonnat. "A real-time scene understanding system for airport apron monitoring", in *Proceedings of 2006 IEEE International Conference on Computer Vision Systems (ICVS 2006)*, New York, USA, (January 5-7 2006). IEEE Computer Society.

[3] B. Boulay, F. Brémond, M. Thonnat. "Applying 3d human model in a posture recognition system", in *Pattern Recognition Letter, Special Issue on vision for Crime Detection and Prevention*, (2006).

[4] R. Cucchiara, R. Melli, A. Prati, L. De Cock. "Predictive and probabilistic tracking to detect stopped vehicles", in *Proceedings of Workshop on Applications of Computer Vision (WACV)*, pp. 388–393, Breckenridge, USA, (4-7 January 2005).

[5] R. Cucchiara, A. Prati, R. Vezzani. "Posture classification in a multi-camera indoor environment", in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, **1**, pp. 725–728, Genova, Italy, (11-14 September 2005).

[6] F. Cupillard, F. Brémond, M. Thonnat. "Tracking groups of people for video surveillance", in *Proceedings of the European Workshop on Advanced Video Based Surveillance Systems (AVBSS01), Kingston, United Kingdom*, (September 2001).

[7] Y. Ivanov, A. Bobick. "Recognition of visual activities and interactions by stochastic parsing", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), pp. 852–872, (2000).

[8] P. Viola, M. Jones. "Rapid object detection using a boosted cascade of simple features", in *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, **1**, pp. 511–518, Kauai, HI, USA, (8-14 December 2001).

[9] T. Vu, F. Brémond, M. Thonnat. "Automatic video interpretation: a novel algorithm for temporal scenario recognition", in *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI03), Acapulco, Mexico*, (August 2003).