# A Framework for the Detection and Interaction with Pedestrian and Objects in an Unknown Environment

A. Moro, K. Terabayashi, K. Umeda, and E. Mumolo *Member, IEEE*

*Abstract*— **In this paper, we present a visual scene description and interaction framework for pedestrian and mobile objects detection and tracking applications. The framework is built upon a previously developed stereo vision system. The proposed algorithms raise up the information level in order to allow to query about the scene using natural language or semantic operators and give a simpler interface with other technologies.**

*Index Terms*— **Scene Analysis, Classification, Tracking, Natural Language**

## I. Introduction

For reliable video stream applications, detections and tracking, suitable static and dynamic information on the visual scene are required. Robust detection and tracking of humans or objects is a key enabling technology for many applications. It is key for knowing who is where in a scene and what their actions have been. It potentially allows other layers in an application framework to infer beliefs about those people. However, depending upon the end-user application, a variety of event detection approaches have been developed. Recently, Ess et al. [4] showed a practical vision-based sensing application which demonstrates that vision-based algorithms have progressed sufficiently for appropriate utilization both in static and dynamic environments. It is becoming popular also the introduction of Natural Language (NL) interfaces in vision systems. In these systems, human behavior is represented by scenarios, i.e. predefined sequences of events [15]. The scenario is evaluated and automatically translated into text. In this paper, we propose a purely vision-based real-time framework which raises up the level of abstraction of scene understanding and allows a simple natural language interaction. Our proposed system uses as input stereo images from a forward-looking fixed camera. The analysis of the image and the absence of training phase allows the system to be adaptable and usable in several indoor and outdoor contexts. The visual system elaborates the input image and combines detection, tracking and classification capabilities with a 3D mapping based on stereo depth data. Its results can be used directly as additional input for an integrated network sensor system.

The work reported in this paper was motivated by the need to develop such a framework to solve a real problem. The implementation of this system is in fact related to the visual component of OSOITE project [16]. The task is to realize a system which detects and counts pedestrians using video data. A natural language interface has been introduced as a method to query the system by who is immersed in the system itself or supervises it. In our application it is a problem of counting the pedestrians moving in a defined direction (i.e. on a street, one of the two possible directions). The framework described here allows to perform complex measures on the detected mobile objects and pedestrians. A method based on the occupied surface, defined by detected blob, permits to count how many people move in a given direction in a time period. A supervisor may query the system to know how many people are on the scene, how many move in a direction or, i.e., if a car has been classified, how many people are near a car. Moreover, the described system works in real-time and can be used for video surveillance of complex systems.

The paper is structured as follow: the section II reviews previous work. Section III gives a description of the proposed framework with a focus on the main components. Section IV contains the results obtained in several environments and a description of the system configuration. In section V, conclusion and future works are reported.

## II. Related Works

The detection of moving objects and people from a streaming video is a basic and fundamental problem of a large amount of vision systems including robotics applications, objects/human detection, tracking, traffic control, and semantic annotation. Reliable object and pedestrian recognition, pose estimation and tracking are critical tasks in reality applications [7]. In general, accurate detection of objects or pedestrians in the visual scene can be obtained by applying category specific models, either directly on the camera image [8], on 3D depth information [2], or combination of both [14]. Tracking detected objects over time presents additional challenges due to the complexity of data association in crowded scenes, or prediction of trajectories. Targets are typically followed using classic tracking approaches such as Extended Kalman Filters (EKF) [3] or Joint Probabilistic Data Association Filters (JPDAF) [6]. Recently, Rodriguez et al. [13] described a tracking method for unstructured crowded scenes where the motion of the crow

A. Moro is with the Department of Electronics, Electrical Eng. and Computer Science, Univ. of Trieste / CREST, JST, P.le Europa 1, 34127 Trieste, Italy
K. Terabayashi and K. Umeda are with Department of Precision Mechanics, Faculty of Science and Engineering, Chuo University / CREST, JST, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
E. Mumolo is with Department of Electronics, Electrical Eng. and Computer Science, University. of Trieste, P.le Europa 1, 34127 Trieste, Italy

978-1-4244-7910-8/10/$26.00 ©2010 IEEE

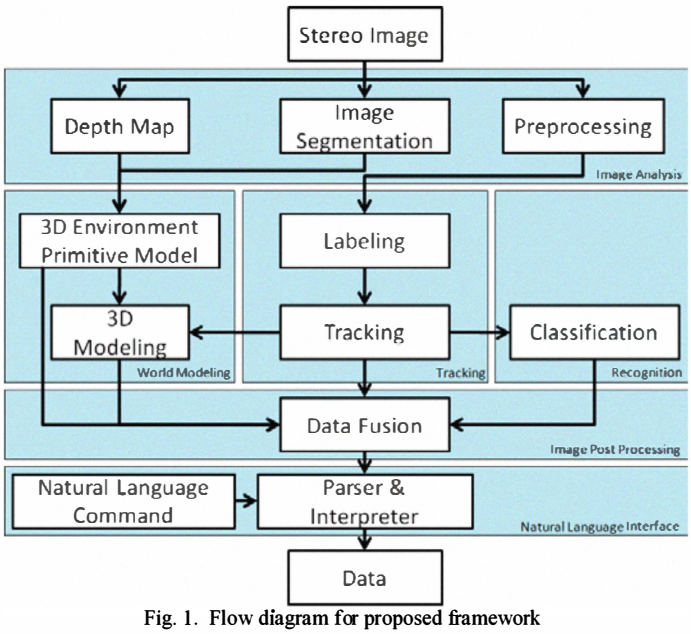


Fig. 1. Flow diagram for proposed framework

appears to be random with different participants moving in different directions. General solution to the problem of object class detection and abstraction are still far to be solved [19]. The early methods apply several single-view detectors independently and combine their responses via some arbitration logic. Thomas *et al.* [17] proposed a single integrated multi-view detector that accumulates evidence from different training views. Yan *et al.* [20] take in consideration the 3D shape of the model to recognize and the spatial connection for relating multiple 2D views.

## III. The Proposed Framework

The proposed framework is designed for a fixed stereo camera system. In Fig. 1 the main components of the system are reported. For each input frame, the blocks are executed the points belonging to moving objects. The image is also segmented in order to identify the geometrical structure of the scene. Then the blob regions are detected and the tracking is applied. Combined with the tracked information and 3D round reconstruction, a 3D model box is created around the tracked objects. In order to recognize what is moved on the scene, each region is classified in some objects of interest. The outputs of these stages are collected together and represent a high level description of the scene.

As a final step, a parser and interpreter is used in order to query the desired requested information from the scenes using NL sentences.

### A. Preprocessing

The image preprocessing components are based on the system described in [10]. For the sake of completeness, in the following we summarize the main components. Given the image frame and a rectangular search window W, thresholds values into the entire possible window W are computed as defined in [10].

The changes on the images can be due not only to the movements in the scene, but because the illumination change or because the moving object or pedestrian generate a shadow. Shadow may be cause of false detection or merging of blobs. A shadow detector is used, in order to increase the vision system performances.

### B. Image Segmentation

Input image generally contains information both of dynamic regions, i.e. pedestrians, and static regions, i.e. ground or buildings. To know how the environment is structured it is useful to better describe the scene. However, because the environment can change and in order to maintain the system more flexible possible, we adopted the automatic image segmentation approach described in [5].

### C. 3D Environment Primitive model

Objects and pedestrians are supposed to be on the same plane and this preliminary condition allows the correction of possible errors generated in the change detection and tracking phases. For detailed information please see [12]. All the transformations are performed in the camera space. Due to the errors in the depth map, just the ground plane is searched. Given the segmented regions, the ground is identified as the largest bottom region in the scene which has enough points to calculate the ground plane. Defined the regions of the image, the ground plane equation is obtained by plane fitting of the points using orthogonal regression.

### D. 3D Modeling

A detected rectangular region in the scene, surrounding the detected changed pixel, is generated. This region lacks of the information about its occupancy. Using a stereo camera, it is possible to obtain a 3D description of the scene, but not the depth of the objects in the scene. A 3D bounding box is then created for each object.

### E. Classification

Detectors based, for instance, on SIFT features [11] are generally able to overcome cluttered scene and recognize a specific object. The choice of recognition algorithm may impose a computational burden. The category generalization can be performed using HAAR based approaches [18], which requires a long training phase. Instead, in this work we used the Pseudo2D Hidden Markov Model stochastic classifier, described in [9]. It is easy to train, it has low computational cost and it does not require particular features to obtain a good rate of classification.

### F. Data Fusion

All the acquired data is collected and indexed properly to be accessible to the NL interpreter. Detected Regions are indexed by class type and labeled by color. The data from bounding box and position are gathered together to the position of the object detected. If the equation of the floor is available, a correction of the bounding box region is applied at this level.

### G. Parser & Interpreter

An operator or a human in the environment may desire to interact with the scene using a natural human-computer interface. It is possible to use a generative grammar to give a description of the scene using sentences. This technique simplifies a semantic analysis but reduces the freedom of the queries. For example an operator watching the scene from his location may say: "'Move robot A to the near human and give

the position of human red"'.

An analytical irregular grammar is used to convert an arbitrary input obtained by an operator or other systems. For this purpose, a simple Left to Right (LR) parser [1] able to extract imperative commands has been implemented. By defining the grammar as:

$$\mathcal{G} = \langle V, \Sigma, R, S \rangle$$

Where $V = \{C_1, C_2, \breve{A}\}$,$\Sigma = \{\breve{b}\}$ and R is described as follow:

$S \rightarrow C_1\breve{A}\breve{b} \mid C_1C_2$

$C_2 \rightarrow C_2\breve{A}\breve{b} \mid \varepsilon$

simple sentences can be parsed and interpreted. $\breve{A}$ is the subset of classified objects at a certain instant and $\breve{b}$ is the subset of properties filled by module.

## IV. Results

In order to evaluate the proposed framework, several sequences have been captured and analyzed. The sequences was taken in indoor and outdoor environments considering different moments of the day. It has been used a Point Gray Research Bumblebee camera with 640 x 480 pixel resolution. The results have been obtained using an Intel Core2 Quad CPU, 2.83 GHz with 4 GB ram. The videos are obtained at a frame rate of 30 frames/sec. For indoor environment movies are taken in static background where the light source is controlled. However, the outdoor environment suffers of light change. For a quantitative evaluation, the detection performance, tracking performance, classification performance have been evaluated. The detection and preprocessing phase is the core component of the framework. Example of detection and preprocessing results

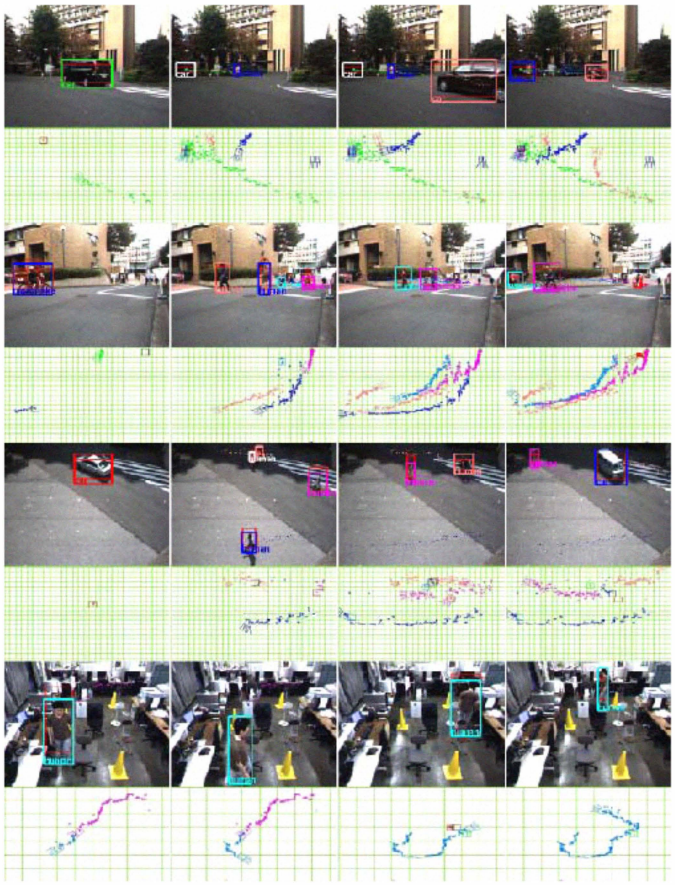

Fig. 2. Example of detection, tracking, classification

TABLE I
SHADOW DETECTION RATE

| | Accuracy |
|---|---|
| Blob Detection Rate | 91.8% |
| Blob Accuracy | 30.3% |
| Shadow Detection Rate | 83.3% |
| Shadow False Alarm Rate | 8.23% |

Detection of changed pixel and shadow detection. Blob accuracy is related to the number of pixel belong to the interesting objects.

TABLE II
COMPUTATION TIME

| Component | Time |
|---|---|
| Preprocessing | 0.01s |
| Segmentation | 0.117s |
| Depth Map | 0.03s |
| Labeling | 0.001s |
| Tracking | 0.1s |
| 3D Environment & Modeling | 0.028s |
| Classification | 0.004s x model |
| Parser & Interpreter | 0.029s |

The elaboration time in this table has to be considered an average of the current performance. Several components (i.e. preprocessing, classification.) are strictly connected to the number of points or regions found.

are shown in Fig. 2. Below each image, we report the tracked trajectories in a map.

In table I, we report average blob and shadow detection rates. These measures have been obtained comparing the elaboration results with a ground truth hand segmented. To assess the suitability of the proposed system for scene position and description, the precision of the pedestrian and object position has been investigated.

Classification algorithm performance is evaluated on the tracked object, comparing the classified objects with expected objects. In these evaluations we used the same trained models for different points of view and environments. The objective was to estimate the flexibility of the classification method in different conditions. To train the models we used images of the CALTECH101 databases (see Fig. 3). The number of images for each category is limited from a minimum of 15 to a maximum of 25. We used the same six objects categories in indoor and outdoor environments. Recognition rate can be increased by extending the training phase.

In Fig. 4 we report a confusion matrix related to the

Fig. 3. Example of images used in training phase. The training set are shown by groups of nine. From top left to bottom right: Cars, Door, Motorbike, Human, Chairs, No Classified.

classification with 2DHmm of five categories, namely Human, Car, Motorbike, Door, Chair.

The bounding boxes are colored to make evident the tracking output (due to the limited palette, some colors repeat). The computation time of the system can be seen in table II. The entire system is implemented in C++. Actually, the bottleneck is represented by the scene segmentation and tracking phases.

The obtained results can be sent to a human operator or another system which query the proposed framework. The current admitted words are "move, go, give, position, near", the six objects categories and the name of the colors as terminal elements. If a sentence like:

"give the position of the human blue"

is given, the parser will decompose the phrase and create the following command: - "give position human blue".

In Fig. 5 we report four examples of natural language enquiry to get the information about the position of humans. In the sequence of Fig. 5, from left to right and from top to bottom, it is possible to see four different situations. In the first two cases, at the request:

- "give position of a human"

no position will be returned because there are not human beings in the scene. In the third picture, due to ambiguity, other information are necessary:

- "give position of human green"
- "give position of human near the chair"

These commands will return the position of sitting human and standing up human. In the last image, instead, a simple request like:

- "give position of human"

will return the position of sitting human.

## V. Conclusions

In this paper, we present a computer vision framework for the detection and classification of objects in the visual scene and a natural language interface to easily interact with the system. Our system relies on a flexible detecting algorithm which is easily adaptable in a various environments. Classification of objects and pedestrian increase the quantity of information which can be obtained from the scene. The inferred predictions can then be used by other systems applying a sensor fusion support. In the future work, we plan to optimize individual system components further with respect to run-time and performance. Actually single thread implementation works at 3- 4 fps, but additional improvements are necessary for a true real-time performance. A scene understanding algorithm will be applied in order to increase the scene description quality. Moreover, it will be realized both top-down and bottom-up natural language support to better interface with other systems.

## VI. References


[1] A. Aho, M. S. Lam, R. Sethi, J. Ullman, "Compilers: Principles, Techniques, and Tools", ISBN 0-321-48681-1, 2006.


| | Human | Car | Motorbike | Door | Chair | Not Classified |
|---|---|---|---|---|---|---|
| Human Indoor / Outdoor | 69.7 / 58.3 | 17.7 / 19.4 | 0.6 / 1.5 | 0.2 / 13.9 | 11.2 / 6.4 | 0.6 / 0.5 |
| Car | 14.6 | 63.8 | 1 | 0.5 | 20.1 | 0 |
| Motorbike | 30.2 | 25.4 | 42.8 | 0 | 1.6 | 0 |
| Door (*) | 11.4 | 1.1 | 6.2 | 49 | 3.1 | 29.2 |
| Chair (*) | 25.1 | 10.2 | 13.8 | 7.4 | 41.3 | 2.2 |

Fig. 4. On the top, the graph shows the classification performance for both indoor and outdoor scenes. The marked objects (door and chair) are special cases and the sequences where these objects were available were limited.

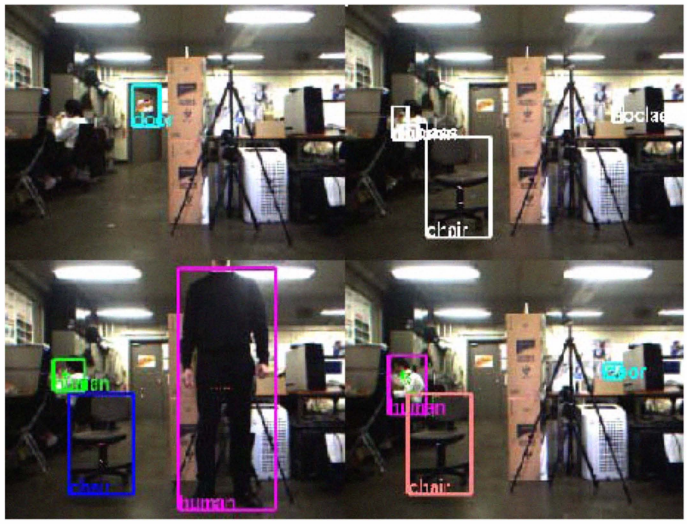


Fig. 5. Example detection and classification in indoor environment. At the request of "Give the position of human purple", the Parser and Interpreter return the coordinate just for the instant represented with the third and fourth picture. The third picture represents also a case of ambiguity for a simple request like "Give the position of the human".


[2] K. O. Arras, O. M. Mozos, and W. Burgard. Using boosted features for the detection of people in 2d range data. In ICRA, 2007

[3] I. J. Cox. A review of statistical data association techniques for motion correspondence. IJCV, 10(1):53–66, 1993

[4] A. Ess, B. Leibe, K. Schindler, L. Van Gool, "Moving Obstacle Detection in Highly Dynamic Scenes", 2009

[5] P.F. Felzenswalb, D.P. Huttenlocher, "Efficient Graph-Based Image Segmentation", IJCV, Vol 59(2), 2004.

[6] T. E. Fortmann, Y. Bar Shalom, and M. Scheffe. "Sonar tracking of multiple targets using joint probabilistic data association", IEEE J. Oceanic Engineering, 8(3):173–184, 1983

[7] I. Gordon, D. G. Lowe, "What and where: 3d object recognition with accurate pose", CLOR06, pp. 67-82, 2006

[8] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. International Journal of Computer Vision, 77(1-3):259–289, May 2008

[9] A.Moro、E. Mumolo, M. Nolich, "Visual Scene Analysis Using Relaxation Labeling and Embedded Hidden Markov Models for Map-Based Robot Navigation", ITI, 2008.

[10] A.Moro, K. Terabayashi, K. Umeda, E.Mumolo, "Auto-adaptive threshold and shadow detection approaches for pedestrians detection", *AWSVCI*, pp.9-12, 2009

[11] J. Mutch, D.G. Lowe, "Multiclass Object Recognition with Sparse, Localized Features", CVPR, pp. 11-18, 2006

[12] D. Nister, O. Naroditsky, and J.R.Bergen, "Visual Odometry", CVPR, 2004.

[13] M. Rodriguez, S. Ali, T. Kanade, "Tracking in Unstructured Crowded Scenes", ICCV, 2009

[14] L. Spinello, R. Triebel, and R. Siegwart, "Multimodal people detection and tracking in crowded scenes", In Proc. of The AAA I Conference on Artificial Intelligence (Physically Grounded AI Track), July 2008

[15] C.F. Tena, P.Baiget, X. Roca, J. Gonzalez, "Natural Language Description of Human Behavior from Video Sequence", KI '07: Proceedings of the 30th annual German conference on Advance in Artificial Intelligence", 2007.

[16] N. Thepvilojanapong, et al., "OSOITE: Towards Real-World Search", The 17th IASTED International Conference on Applied Simulation and Modelling (ASM 2008), 2008.

[17] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. Van Gool, "Towards multi-view object class detection", CVPR, volume 2, pp. 1589-596, 2006.

[18] P.Viola, et. al. "Rapid object detection using a boosted cascade of simple features", IEEE CVPR, 2001

[19] G. Wang, Y.Z.Li Fei-fei, "Using dependent regions for object categorization in a generative framework", CVPR, pp.1597-1604, 2006

[20] P. Yan, S. M. Khan, M. Shah, "3D Model based Object Class Detection in An Arbitrary View"