Shape Recognition Based on a Video and Multi-Sensor System

Huy-Binh BUI NGOC*, François BREMOND**, Monique THONNAT**, Jean-Claude FAURE*
{binh.bui,jean-claude.faure}@ratp.fr, {francois.bremond, monique.thonnat}@sophia.inria.fr

*RATP - SIT/SVO/BRS/CTSC - NOISY-LE-GRAND - France
**ORION Research Team, INRIA Sophia-Antipolis, France

Abstract

We present in this paper a real-time system for shape recognition. The proposed system is a video and multi-sensor platform that is able to classify the mobile objects evolving in the scene into several expected categories. The key of the recognition method is to compute mobile object properties thanks to the camera and sensors and then to use Bayesian classifiers. A learning phase based on ground truth data is used to train the Bayesian classifiers.

Our recognition method has been integrated into an existing access control device used in public transportation (subway) at RATP to improve safety and comfort, to prevent fraud and to count people for statistical matters. The expected categories in this case are mainly "adult", "child", "suitcase" and "two adults close to each other".

Keywords: shape recognition, Bayesian classifiers, supervised learning.

1. Introduction

We propose in this paper a new system for shape recognition based on a video and multi-sensor system. Our goal is to design a system with very high recognition rate complying with real-time constraint. To achieve this goal, we have conceived a device combining a static camera and a set of lateral sensors. Cameras are often static in visual surveillance network to get a robust low-level detection of mobile objects. The lateral sensors are very useful to separate people entering the access control site. The real-time constraint is very challenging as it implies that the solutions should be kept with a maximal computing time.

After an overview of the system in section 3, we give a detailed description for the main tasks of the interpretation process in section 4. The performance of the system is illustrated by the experimental results described in section 5. The paper concludes with the current limitations and the future work for enhancing the robustness of such video understanding system.

2. Related Work

In recent years, many video interpretation systems have been developed in the computer vision community. These systems are usually composed of algorithms for (a) detecting and tracking mobile objects and (b) recognizing mobile object behaviors and related scenarios. In [3], N. Moenne-Locoz and al. use a Recurrent Bayesian Network to model the temporal evolution of the visual features characterizing human behaviors and to infer the occurrences whatever the time-scale. In [6], Haritaoglu and al. have developed techniques for shape analysis and tracking to locate people and their parts (head, feet, etc). In [2], Zhao and Nevatia have used just one camera in realistic situations and an articulated dynamic human model to recognize postures of a walking and running person. However, few systems have been successfully applied to real word applications due to a large variety of video interpretation issues (e.g. motion detection and tracking are often uncertain and incomplete) and due to strong requirements to obtain a real-time, efficient and robust system. Moreover, most of existing systems focus only on mono camera processing therefore they cannot take advantages of other information sources.

Thus, we propose a system that is able to detect and classify people and objects with very high recognition rate and with real-time constraint. Our approach consists in applying Bayesian classifiers for shape recognition to handle the uncertainty accurately.

3. System Overview

Our goal is to have as much information as possible on the scene to understand precisely who is entering the site. To reach this goal, a fixed camera is placed above, at the height of about 2.5m, while a set of lateral sensors is placed on the side as shown in figure 1. The camera observes the mobile objects from the top to detect and locate them. The lateral sensors observe the side of mobile objects, help to separate the detected mobile objects and provide information on their lateral shape.

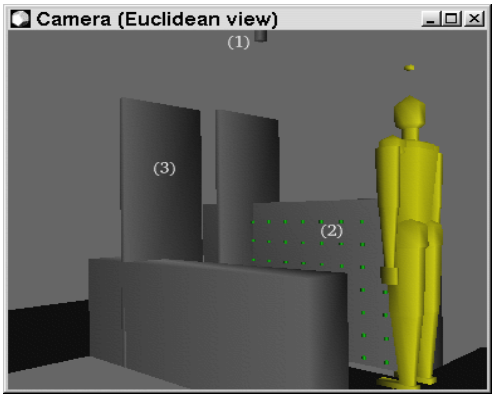


Figure 1. *The access control site contains (1) a top camera, (2) a set of lateral sensors and (3) an access door.*

The interpretation process is composed of four main tasks as shown in figure 2. First a motion detector detects mobile objects evolving in the scene thanks to the top camera. Second, the mobile objects detected as one moving region can be separated thanks to the computation of the vertical projections of pixels or thanks to the lateral sensors. Third, the mobile objects are classified into several mobile object categories (e.g. adult, child, suitcase, two adults close to each other) using Bayesian classifiers. Finally, the mobile objects are tracked to improve the reliability of the recognition process.

Moreover, we use a 3D model of the empty scene as a priori contextual knowledge of the observed environment. We define in the 3D scene model the 3D positions and dimensions of the equipment (e.g. the access door), the zones of interest (e.g. the entrance/exit zone) and the expected objects in the scene (adult, child, suitcase, two adults close to each other). Using context is essential for object recognition and for establishing the confidence in the whole interpretation system.

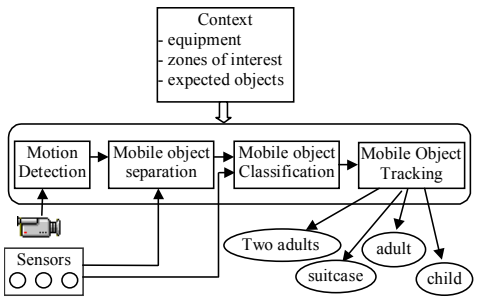


Figure 2. *The interpretation process takes as input a video stream and sensor information and outputs the recognized shape of mobile objects.*

4. Motion Detector

The goal of the Motion Detector is to detect for each frame the moving regions in the scene. In our approach, motion is detected mainly by thresholding the difference of the current image I_c with respect to a reference image I_r. Then for each pixel we compute the absolute difference between its intensity (grey or color) and the intensity of the corresponding pixel in the reference image. If this difference is greater than a certain threshold δ, the pixel is marked as moving and otherwise it is marked as stationary (cf. equation 1).

$$Pixel(x,y)_{moving} = (Abs(Difference(I_r(x, y), I_c(x, y)))) > \delta$$

Equation 1. *Test for a moving pixel*

We then update the reference image with information from the current image. For each pixel marked as stationary, we integrate a significant portion of the current image to the reference image according to the following equation:

$$I_r = (1-\alpha)I_r + \alpha I_c$$

Equation 2. *Update the reference image for stationary pixels.*

Moreover, to correct detection errors (e.g. to integrate noise to the reference image), for each pixel marked as moving, we also integrate a small portion of the current image according to the following equation:

$$I_r = (1-\beta)I_r + \beta I_c$$

Equation 3. *Update the reference image for moving pixels.*

A typical challenge of motion detection is to handle shadows. To remove the shadows, we have installed a light on the floor of the site.

The motion detection takes the majority of the whole interpretation process time. To have faster motion detection (complying with the real-time constraint), instead of testing all the pixels of image, we only test the pixels at a regular step (every η pixels, with η a parameter of motion detector). If the pixels between two consecutive steps have the same label (e.g. “moving” or “non moving”) then we consider that all the intermediate pixels (between the two pixels tested previously) have also this label. If it is not the case, we have to go back and test recursively the intermediate pixels. As statistically fewer pixels are "moving" in each frame, only a small number of pixels need to be tested. So, the motion detector can save time. We call such motion detector a “RLE (Run Length Encoding) Motion Detector”. In our experimentation, with $\eta = 4$, the motion detector takes

only 30% of the whole interpretation process time instead of 75% with $\eta = 1$.

5. Mobile Object Separation

A common error of motion detection is to detect several mobile objects (people walking closely to each other or person carrying a suitcase) as only one moving region (cf. figure 3). The mobile object separation task consists in separating the moving regions that could correspond to several individuals into distinct moving regions. To accomplish this task, two techniques are combined together: computation of pixel projections and utilisation of lateral sensors.

5.1. Using Vertical Projections of Pixels

For each moving region, we calculate the potential points of separation (called separators) corresponding to potential borders between two persons. For that, we calculate the vertical projections of the moving region pixels as shown in figure 3. When a "valley" is detected between two "peaks", we regard this valley as a potential separator between two distinct persons and the peaks as the gravity centers of these persons. If the size (the 3D length and the 3D width) of both distinct persons matches the dimension of a real person then this separator is valid.

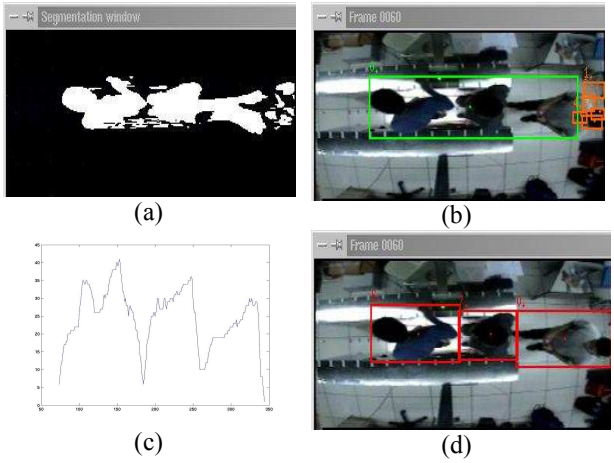


Figure 3. *Separation using the vertical projections of pixels: images (a), (b) illustrated three persons detected as one moving region; image (c) illustrated the vertical projections of pixels and image (d) shows that the moving region has been separated into three distinct moving regions.*

5.2. Using the Lateral Sensors

The separation method using the vertical projections of pixels depends on the position of the persons relatively to the camera. This method cannot separate two adults walking closely to each other or far from the camera (cf. figure 4). In this case, we use lateral sensors to detect the point of separation. For example, we can detect the non-occluded sensors announcing a space between two adults. More exactly, a separator is a non-occluded sensor found between two bands of occluded sensors (cf. figure 4).

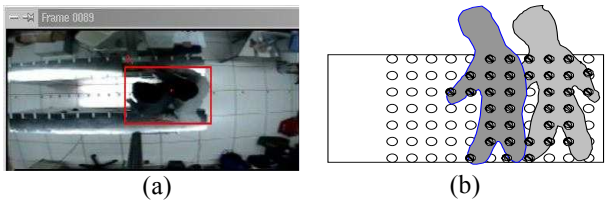


Figure 4. *Separation using lateral sensors. Image (a) shows two adults walking closely to each other and detected as one region mobile; drawing (b) shows three non-occluded sensors detected between two adults. These three sensors form a separator.*

In addition to help to separate two adults walking closely to each other, lateral sensors provide also clues to separate the objects associated to the persons such as bags, suitcases and in certain cases children. To separate objects, we define a separator as a column of sensors having a large majority of non-occluded sensors. These separators enable to separate two consecutive suitcases and a suitcase or a child from the adult if the distance between them is big enough (cf. figure 5).

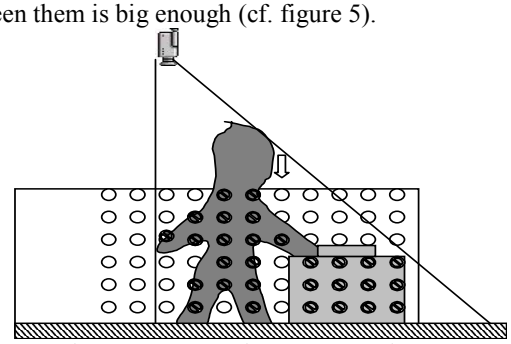


Figure 5. *The top camera does not see the suitcase but lateral sensors help to separate it from the person. The separator (column of sensors having a large majority of non-occluded sensors) is indicated by an arrow.*

6. Mobile Object Classification

6.1. Mobile Object Models

Initially, we have to build a model (class) for different mobile objects such as "adult", "child", "suitcase" and "two adults close to each other". We are supposed to have the model for two adults close to each other because in cases where two adults are walking very closely to each other, neither the vertical projections of pixels nor the lateral sensors can separate them. The model for a mobile

object is built from its characteristics obtained by the top camera and the lateral sensors. The top camera provides information on its 3D length L_t and its 3D width W_t. For lateral sensors, we divide the zone of sensors at the mobile object position into n sub-zones (cf. figure 6). Then, for each sub-zone i, we calculate the density S_i of the occluded sensors and we use this density as a characteristic of the mobile object.

The number of sub-zones, their dimension and their position should be chosen intelligently according to the properties and the people body parts (e.g. the legs is one of the sensitive body part for a person). In our experimental test, to simplify the calculation, we divide the zone of sensors into 9 sub-zones. The dimension of each sub-zone is defined proportionally with the dimension of the zone as shown in figure 6.

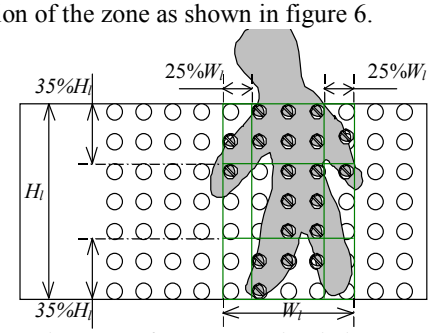


Figure 6. *The zone of sensors is divided into 9 sub-zones and the density of occluded sensors in each sub-zone is used as characteristic of mobile object shape.*

To reinforce the mobile object model, we also consider the lateral 3D width W_l and the lateral 3D height H_l of the zone as characteristics of the mobile object.

In conclusion, in our implementation, a mobile object model is a set of 13 characteristics: L_t, W_t, W_l, H_l and S_i, i=1..9. However, we can add other characteristics of mobile object to enrich the model.

6.2. Training Bayesian Classifiers

For each class of mobile object, we use about 250 instances representative of the class to train a dedicated classifier. For each frame, we compute and record the values of mobile object characteristics (i.e. L_t, W_t, W_l, H_l, S_i, i=1..9). We count the number of mobile objects having the same value c for the characteristic C. So we obtain the frequency for a given mobile object class to have the value *c* for the characteristic C. In other words, we obtain the conditional probability P(c|F) that a mobile object has the value *c* for the characteristic C knowing that this mobile object belongs to class F.

By counting the number of mobile objects of other classes (i.e. all classes excluding F) having the same value c for the characteristic C, we also obtain the conditional probability P(c|¬F) that a mobile object has the value c for the characteristic C knowing that this mobile object belong to another class (¬F corresponds to all classes excluding class *F*).

We have developed a tool permitting a user to annotate a frame with information for the learning task. Once the user has chosen a video sequence, the tool visualizes each frame of the sequence and asks the user to delimit the mobile object seen in the frame and to give its class (e.g. adult, child, suitcase, two adults). The tool then, for each mobile object, calculates automatically the values (l_t, w_t, w_l, h_l, s_i, i=1..9) of its characteristics (corresponding to ground truth) and records them into 13 files. These files are used latter as the training data. They are useful also for evaluating the recognition method. For example, we compare the output of the recognition module with the ground truth data in these files.

6.3. Mobile Object Classification

To classify mobile objects into the expected classes, we compare, in each frame, the characteristics of the mobile object with the characteristics of the mobile object classes using the Bayes rule.

For each frame and for each mobile object o, we build a vector containing its ***degrees of membership*** D(o∈F) for all classes F. The degree of membership is the ratio of the probability P(o∈F) that the mobile object *o* belongs to the class F divided by the probability P(o∈(¬F)) that the mobile object o belongs to another class (¬F corresponds to any class excluding class F) as shown by equation 4.

$$D(o \in F) = \frac{P(o \in F)}{P(o \in (\neg F))}$$

Equation 4. *The degree of membership is defined as the ratio of the probability that the mobile object o belongs to the class F.*

By using Bayes rule and by replacing the mobile object by its characteristic set, we obtain:

$$D(o \in F) = \frac{P(F | l_t, w_t, w_l, h_l, s_1, s_2, ..., s_9)}{P(\neg F | l_t, w_t, w_l, h_l, s_1, s_2, ..., s_9)}$$

Equation 5. *The degree of membership is computed as the ratio of the conditional probability that the mobile object characteristic corresponds to the class F.*

Where P(F |l_t, w_t, h_l, w_l, s_1,…, s_9) is the conditional probability that the mobile object characteristic correspond to class F knowing the value set (l_t, w_t, h_l, w_l, s_1,…,s_9).

After checking that all characteristics are independent we simplify equation 5:

$$D(o \in F) = \frac{P(l_t|F) \times ... \times P(s_8|F) \times P(s_9|F)}{P(l_t|\neg F) \times ... \times P(s_8|\neg F) \times P(s_9|\neg F)}$$

Equation 6. *the degree of membership is computed as the ratio of the conditional probability of each characteristic corresponding to the class F.*

Where P(c|F) (respectively P(c|¬F)) is the conditional probability that a mobile object characteristic C has the value c knowing that this mobile object belongs to class F (respectively to another class). These conditional probabilities are obtained from the ground truth data as discussed in the precedent section.

The mobile object o is then classified into the class with the biggest degree of membership.

7. Mobile Object Tracking

The Bayesian classifiers sometimes miss recognize or do not recognize the class for a mobile object due to the large variety of lateral shapes (i.e. due to lack of training data). To increase the recognition reliability, we track mobile object when they evolve through the scene. This tracking stage enables, on one hand, to correct potential frame to frame classification errors and on the other hand, can help latter to recognize human behaviors and scenarios.

7.1. Mobile Object Matching

The mobile object matching stage consists of matching the mobile objects previously detected at time t-1 with new ones detected at time t. To calculate these correspondences, we currently use three different criteria: their compatibility of lateral shape, their 3D distance and the overlap between their bounding boxes. The decision to match or not two mobile objects is made based on a thresholding of the weighted sum of these criteria.

For each criterion k, we construct a binary matrix $n_o \times n_n$ M_k (n_o and n_n are the number of mobile objects detected at t-1 and the number of new mobile objects detected at t) containing the correspondences for the criteria k. The final decision matrix M is computed as the weighted sum of the matrices M_k as shown by equation 7.

$$M = \frac{\sum_{k=1}^{3} w_k M_k}{\sum_{k=1}^{3} w_k}$$

Equation 7. *The matrix M combines the three matrices of correspondences between mobile objects.*

If M(i, j), ($i=1..n_o$; $j=1..n_n$) is greater than a certain threshold ∂ then the mobile objects o_i at t-1 and o_j at t are matched. If an old mobile object at t-1 matches with several new mobile objects at t, the mobile object having the best correspondence (i.e. the greatest correspondence) is chosen.

7.2. Recognition Refinement

To increase the reliability of the shape recognition algorithm (i.e. to correct potential classification errors), we maintain the temporal coherency of the membership degree vector D composed of the membership degrees for all classes. For each previously detected mobile object at t-1, we update this vector D with the temporary membership degree D_t detected at time t as shown by equation 8.

$$D = D + \omega_t D_t$$

Equation 8. *Update of the membership degree vector D.*

Where ω_t, $0 \leq \omega_t \leq 1$ is the ***confidence weight*** at t of the Bayesian classifiers. This confidence weight is chosen according to the lateral sensor density where the mobile object is the detected. For example, a high confidence weight is chosen if the mobile object at time t is found in a zone where there are many lateral sensors (i.e. where we obtain a more precise shape of mobile object).

The final class of mobile object is chosen according to the biggest value in the new vector of membership degrees.

8. Results

The recognition module has been tested in two stages: a stand-alone experimentation on recorded image sequences (i.e. test offline) and an experimentation in live in interaction with the kernel of an existing access control device used in subways at RATP.

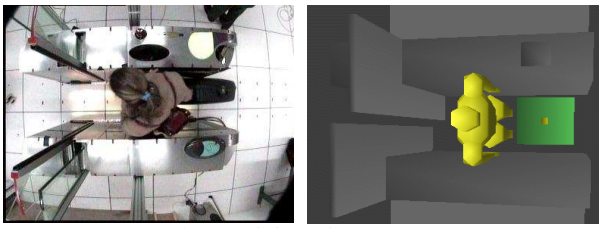
To train the Bayesian classifiers, for each class "adult", "child", "suitcase", we used about 300 frames as training data and about 1000 frames for testing. For the class "two adults close to each other", at the present time, we have only 32 frames in total to represent this class. For this class, we used 15 frames as training data and 17 frames for testing.

In the stand-alone stage, the results are very promising. A large majority of mobile objects have been correctly recognized with a high degree of membership (cf. figure 7 and table 1). More than 94% of adults, children and suitcases are correctly recognized. More precisely, for the adult class, the true positive is 98%, the false positive is 1% and the false negative is 2%. The true

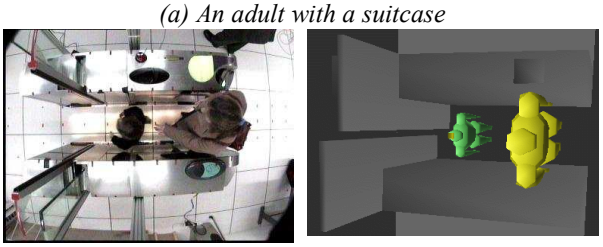
positive for "two adults close to each other" is about 73% due to the lack of training data in the learning phase.

Mobile Object	True Positive	False Positive	False Negative	Frames used for testing	Frames used as training data
Adult	98%	1%	2%	1102	327
Child	94%	3%	6%	1050	295
Suitcase	95%	2%	5%	1008	305
Two adults close to each other	73%	0%	27%	17	15

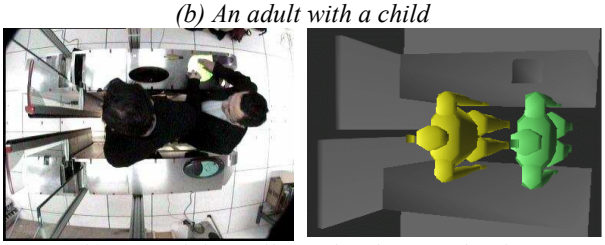
Table 1. *More than 94% of adults, children and suitcases are correctly recognized.*



(a) An adult with a suitcase



(b) An adult with a child



(b) Two adults walking closely to each other

Figure 7. *The images on the right show the recognition result in a 3D animation from the processing of images on the left.*

The recognition module sometimes miss classifies a child with a small suitcase and vice versa due to the similarity of appearance. Almost all the potential errors in the frame to frame classification have been corrected by the frame to frame tracking.

The recognition result depends on training videos used in the learning phase. To obtain better results, we should enrich the training data for each class. For example, for the class "adult", the training data should include large variety of persons (fat, thin, tall, short, adult in summer/winter clothes, etc).

In the live experimentation, the recognition module runs on a PC (Pentium IV 2.8 GHz, 1GB memory, Linux) and receives a video stream at 25 images per second. The maximal time for processing one image is inferior to 35ms. The real-time constraint is then satisfied.

9. Conclusion and Future Work

We have described in this paper a video and multi-sensor interpretation process for shape recognition. The key of recognition method is to compute mobile object properties thanks to a camera and a set of lateral sensors and then to use Bayesian classifiers. The system has been tested offline and in live and gives very promising results.

The recognition, as previously said, depends on training videos and sensor data. Since realistic training data cannot include all varieties of mobile object classes and shapes, the first next step will consist in studying supervised and non-supervised machine learning techniques in order to learn dynamically new classes of mobile objects. Moreover, with the objective of helping the system to control the access safely and comfortably while preventing from fraud, the second next step should consist in human behavior and scenario recognition in order to understand and anticipate the evolutions of mobile objects. For this, we can adapt methods proposed in [3] and [5]. Finally, for the system to be more robust, the third next step will consist in studying the system autonomy. For example, the system should be able to detect failures (sensors or camera breakdown, change of light) and set up a degraded operation mode according to the resource available. The objective will be to have a system that can reconfigure itself dynamically and autonomously.

References

[1] S. Hongeng, F. Brémond, and R. Nevatia. *Bayesian framework for video surveillance application*. 15th International Conference on Pattern Recognition, 2000.

[2] T. Zhao, R. Nevatia, 2004, "*Tracking Multiple Humans in Complex Situations*," *PAM*I, pp. 1208-1221.

[3] N. Moenne-Locoz, F. Brémond and M. Thonnat. *Recurrent Bayesian Network for the Recognition of Human Behaviors from Video*. Third International Conference on Computer Vision Systems (ICVS 2003), 2003. Proceedings, pp. 68 – 77.

[4] P. Viola and M. Jones. *Robust Real-time Object Detection.* Second International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing and Sampling, Canada, July 13, 2001.

[5] V.T. Vu, F Brémond and M. Thonnat. *Automatic Video Interpretation: A novel algorithm for temporal scenario recognition*. Eighteenth International Joint Conference on Artificial Intelligence - IJCAI 2003. Acapulco Mexico.

[6] Haritaoglu I, Harwood D and Davis L, 2000, "*W4: real-time surveillance of people and their activities*", IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 809-830.