

# People and Luggage Recognition in Airport Surveillance Under Real-Time Constraints

V. Atienza-Vanacloig   J. Rosell-Ortega   G. Andreu-García   J. M. Valiente-González  
Grupo de Vision por Computador. DISCA. UPV. Spain  
{vatienna, gandreu, jvalient}@disca.upv.es, jarosell@doctor.upv.es

## Abstract

*This paper describes an approach to classify people, groups of people and luggage in the halls of an airport. The algorithm is included into a surveillance system which tracks and classifies objects and transmits this information to a higher computational level which fuses the information of several cameras covering overlapping areas. Two kind of features are used: foreground density features and features related to real-size of objects, obtained by applying a homographic model. A classification schema based on k-nn classifiers and a voting system makes the classification process highly robust. On-line and off-line experiments are introduced.*

a pyramidal schema. A homographic normalization is performed to homogenize blob sizes for computing size histograms in [1], but camera calibration and explicit normalization of blob sizes to obtain reliable real-world size measures for classification is seldom done.

In our application, airport surveillance, attention must be paid to individuals standing alone, groups of people and unattended luggage. The system consists of a set of specifically designed smart cameras running low-level vision algorithms. These units, are assigned the tasks of detecting objects in its assigned area, track them and perform a preliminary classification using only the information gathered through the camera and a limited local history, if any. This classification is then fed to a higher level which controls several cameras and fuses all the incoming data.

## 1. Introduction <sup>1</sup>

Visual surveillance, either indoors or outdoors, is an active research topic in computer vision and various systems have been proposed in recent years: [2], [6], [7]. The visual surveillance process may be divided into the following steps: environment modeling, motion detection, object classification, tracking, behavior understanding, human identification and data fusion.

Effective classification of objects in crowded scenarios is a challenging problem. Several examples of systems addressing this problem are available in the literature, for instance, in [2], simple features as symmetry and vertical projections of image blobs and others such as extremes of convex hull are used to discriminate groups from single people. In [3], dispersedness, image area and aspect ratio of blobs are used to classify humans, vehicles and groups, authors of [5] propose using linear combinations of simple rectangle filters in

Real-time computation of these features is crucial as much as having a high local rate of successful classifications; despite the fact that specific local errors in classification can be corrected at higher levels.

In this paper, we introduce a solution that integrates shape features, based on foreground density patterns, with features based on real size measures in order to obtain low confusion rates while maintaining reduced computational costs. We enhance shape features introduced in [4] by setting an unevenly distributed foreground density pattern designed to give different importance to some parts of the blobs over others, according to a priori information about objects shape. On the other hand, real-world size features as height and extent of visible area are computed by modeling the image formation process by means of a homography which, at the same time, allows to obtain position estimations for objects in the real world. We present a series of experiments which demonstrates the effectiveness of such approaches, as well as that of the overall system.

<sup>1</sup>Acknowledgments: This work is supported partially by the sixth framework programme priority IST 2.5.3 Embedded systems. Project 033279.

## 2 System outline

As said in the introduction, our system is designed to track and classify objects appearing in an airport.

We use a single frame with no targets as the starting background model and update it in predefined intervals of time; this is done so, due to hardware requirements. Being  $B_{t-1}$  the background model and  $F_t$  the current frame in time  $t$ , the updated model  $B(t)$  is calculated according to  $B(t) = \alpha \times B_{t-1} + (1 - \alpha) \times F_t$ , where  $\alpha$  is a chosen fixed value in the range  $[0, 1]$  (in our case,  $\alpha = 0.98$ ).

In order to detect regions of interest in frames, we use background subtraction together with temporal differences between  $F_t$  and two preceding frames  $F_{t-1}$  and  $F_{t-2}$ . Only those regions of interest whose area exceeds a threshold are considered to be blobs, discarding the rest.

Once blobs in current frame are detected, we find the relationships between blobs in current frame and blobs in previous frames. This is done by testing temporal overlappings between blobs. This takes us to face situations as blobs joining or splitting what, in the real world, translates into people joining into groups or groups of people splitting into separate people.

A classification schema based on a  $k$ -nn classifier and temporal information is used to classify objects into one of the following classes *person*, *group of people* and *luggage*. The temporal information consists of the history of classifications of the same object in order to achieve a more robust classification over time and is used to vote the most probable label that should be assigned to the object.

This classification is encapsulated into an XML file and sent to a higher computational level to be fused with information gathered from other cameras covering the same area from different points of view.

## 3 Feature sets

In this section, we introduce two different sets of features: foreground density features with granularity and real size features.

The first group derives from a previous work [4] where we proposed an easy schema to classify objects. We stated that either single persons, groups of people, or luggage show different patterns of occupancy. This leads us to think that the density of foreground pixels measured in a regional basis could be a good feature set for classifying the objects.

While scale invariance is a requisite in order to the  $k$ -nn classifier to succeed overcoming size changes of objects due to perspective effects, real size of blobs is

considered to be a very valuable information to be taken into account for classification purposes. Certainly, the size of a group of people in the image is expected to be greater than that of a single person and person size greater than that of luggage items. Moreover, real size considerations can help to achieve more reliable discrimination of noise.

### 3.1 Foreground density features with granularity

We proposed an schema consisting in dividing each bounding box always into the same number of regions, this way, the requirement of scale invariance is incorporated to the method. For each region  $R_i$ , the amount of foreground pixels ( $P_i$ ) and background pixels ( $B_i$ ) was calculated and the result of the division  $\frac{P_i}{P_i+B_i}$  was stored in a vector of values. This way, we have a set of values pointing out, for each object, where the majority of foreground pixels is likely to be. For instance, people are expected to have their maximum densities in the area of the body.

A step forward is giving more importance to some parts of the object over others. This can be done by dividing each object in three unequal regions, which we will call from now on, *head*, *body* and *legs*, being the head region the top of the object, legs region will correspond to the bottom of the object and body region will be the rest. In figure 1, striped lines separate the three different regions and straight lines delimit cells in regions. It is clear that head region may be more useful for distinguishing instances, for example, of classes *person* and *luggage*.

With this schema, we divide each region into a grid of  $n_r \times m_r$  regions, where  $r$  stands for the region (head, body or legs), being the grid size of each region independent of the rest. We then perform the same calculations for each region as before.

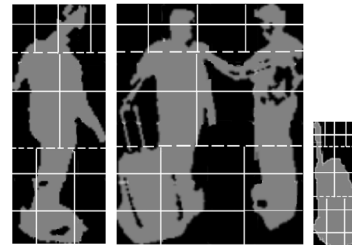


Figure 1: Sample of person, group of people, and luggage with different divisions in the 3 regions defined as head (granularity  $4 \times 2$ ), legs (granularity  $3 \times 3$ ) and body (granularity  $2 \times 2$ ).

### 3.2 Real size features

To estimate real-world measures of objects we model the projection of ground plane of the scene in the image by means of a homographic transformation. The *homography*  $\mathbf{H}$  is a  $3 \times 3$  matrix that relates a scene point  $\mathbf{x}_i$  with its corresponding image point  $\mathbf{u}_i$  by means of the equation  $\tilde{\mathbf{x}}_i = \mathbf{H}\tilde{\mathbf{u}}_i$ , where  $\tilde{\mathbf{x}}$  denotes point  $\mathbf{x}$  expressed in homogeneous coordinates. A calibration procedure must be performed to obtain homography parameters, this involves measuring real-world coordinates of four non-collinear points of the ground that are visible in the image. Figure 2 shows the calibration and transformation procedures. The rectified figure (right) is obtained from the original frame (left) by applying the estimated homography. The points used for calibration correspond to the base of four columns.

This transformation serves two purposes: i) to estimate the real-world coordinates of each object in the scene and ii) to estimate height (in meters) and extension of visible area ( $\text{m}^2$ ), two measures of interest for object classification. To achieve the first aim we apply the transformation to the central lowest point of the object in the image, which is considered to lie on the floor. To measure real height and size of the visible area of object, we assume that all its points are nearly equidistant to the camera (*weak perspective* condition). This will be reasonably fulfilled by isolated standing people and luggage, for which height and area measures will be reliably obtained. With relation to groups of people, which can extend through large areas of the ground, the weak perspective assumption will be often largely violated. This will lead to obtaining large values for height and area, due to the raised position of the camera over the ground. This fact will be positive to discriminate groups from single people, as long as these values will

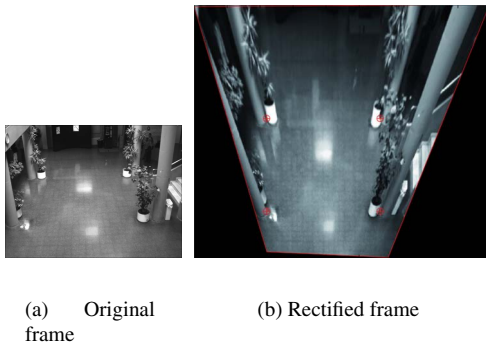


Figure 2: Homographic transformation example. The base points of four columns are used for calibration.

exceed suitable thresholds for a single person.

Based on the weak perspective assumption for the points of a single object, the area and height of its visible surface can be effectively approximated by considering that all of them are equally distant from the camera than the lowest point. As this lowest point lies on the ground plane, its scale ratio  $S = \text{real\_size}/\text{image\_size}$ , applicable both in the horizontal and vertical directions, can be easily computed from the homographic transformation. Be  $\mathbf{u}_l$  and  $\mathbf{u}_r$  the image coordinates of the bottom-left and bottom-right corners respectively of the bounding box of an object. The corresponding real-world points are computed as  $\tilde{\mathbf{x}}_l = \mathbf{H}\tilde{\mathbf{u}}_l$ ,  $\tilde{\mathbf{x}}_r = \mathbf{H}\tilde{\mathbf{u}}_r$ . The real-world position of the object is computed as  $\mathbf{x}_c = (\mathbf{x}_l + \mathbf{x}_r)/2$ . Then, the scale ratio is obtained by  $S = (||\mathbf{x}_r - \mathbf{x}_l||) / (\mathbf{u}_r^x - \mathbf{u}_l^x)$ , where  $\mathbf{u}^x$  denotes the x-component of the image point  $\mathbf{u}$ . Figure 3 shows the capacity of this procedure to determine the real-world position of objects on a rectified image of the ground and the normalization of objects sizes by applying the computed scale ratios. It can be observed how the normalized figures maintain a common scale in spite of their different distances to the camera.

## 4 Experiments

Off-line experiments were made in order to test the validity of having more granularity in some parts of the blob over others. These experiments were done using blobs extracted from videos recorded by ourselves. We extracted and labeled manually 2563 objects, corresponding to 1344 images of class *single people*, 229 of class *group of people* and 990 images of class *luggage*. The segmentation criteria was the same as in paper [4].

Previously, in separated tests, we decided that the top 20% blob would correspond to the head region, the bot-

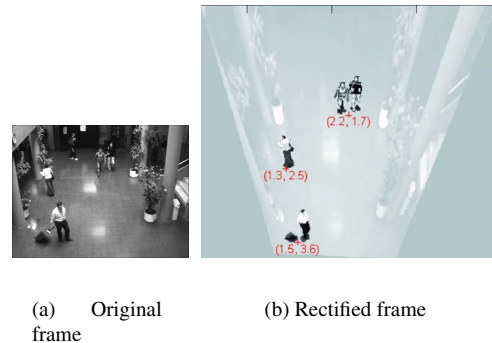


Figure 3: (a) Original frame. (b) Objects with real-world coordinates and normalized sizes.

Table 1: Confusion tables for different division granularities with  $k = 5$ .

$3 \times 4, 7 \times 8, 5 \times 5$	<i>P</i>	<i>G</i>	<i>L</i>
<i>P</i>	0.98	0.04	0
<i>G</i>	0.02	0.96	0
<i>L</i>	0	0	1
$5 \times 2, 8 \times 7, 5 \times 7$	<i>P</i>	<i>G</i>	<i>L</i>
<i>P</i>	0.98	0.04	0
<i>G</i>	0.02	0.94	0
<i>L</i>	0	0.02	1
$3 \times 1, 4 \times 3, 4 \times 3$	<i>P</i>	<i>G</i>	<i>L</i>
<i>P</i>	0.99	0.07	0.01
<i>G</i>	0.01	0.93	0
<i>L</i>	0	0	0.99

tom 40% blob would correspond to legs, and the rest to the body. Then we made experiments varying the values of  $n_{head}$ ,  $m_{head}$ ,  $n_{body}$ ,  $m_{body}$ ,  $n_{legs}$  and  $m_{legs}$  in the range 1..20.

We trained a  $k$ -nn classifier using 80% of the database and the test set was composed of the other 20%. The experiments were repeated 100 times to ensure statistical independence of the selected samples. In table 1 we show confusion tables for different values of  $n_r$ ,  $m_r$ , and  $k = 5$ . Results point out that not always a bigger granularity in any of the regions yields better results.

The main issue we took into account when selecting the best arrangement, was the performance with class *group of people*. As we learned in the previous work, objects belonging to this class are easily confused with instances of class *person*, and it is easy to see that, depending on how people is arranged inside the group (for instance, with occlusions), a group of people is quite similar to a single person. The arrangement of divisions ( $n_{head} = 3, m_{head} = 4, n_{body} = 7, m_{body} = 8, n_{legs} = 5, m_{legs} = 5$ ) was the chosen one.

Confusion between *luggage* and *group of people* is nearly reduced to zero with this new schema and confusion of classes *person* and *group of people* is also very low; improving results obtained without forcing different granularities in different areas of the object.

In another set of on-line experiments, we tested the performance of size features, foreground density features and the convenience of combining classifiers based on them both. On-line experiments were done with a 900-frame video different from those used for training the classifier.

Table 2 shows individual results for each set of features and a combination of both of them by means of a voting schema. Optimum results are obtained when combining both sets of features; classes *person* and *luggage* are perfectly separated, still some instances belonging to class *group of people* are confused with *per-*

Table 2: Comparison of confusion tables yielded by the different features used and their combination.

classes	Density features			Size features			Combined		
	<i>P</i>	<i>G</i>	<i>L</i>	<i>P</i>	<i>G</i>	<i>L</i>	<i>P</i>	<i>G</i>	<i>L</i>
<i>P</i>	0.96	0.17	0	0.98	0.02	0	1	0.03	0
<i>G</i>	0.04	0.83	0	0.01	0.98	0	0	0.97	0
<i>L</i>	0	0	1	0.01	0	1	0	0	1

*son*, this is normal if we take into account that people in groups may occlude one to each other and appear as a person.

## 5 Conclusions

In this paper we introduced a classification approach based on combining two different sets of features together with a voting schema that considers the history of classification of objects.

Experiments of both set of features, on their own and their combination, show that the combination of two classifiers, one based on each set, yields better results reducing interclass confusion to residual values. Despite the difficulty of the issue, results are very satisfactory.

Future works involve applying clustering algorithms in order to reduce the information needed to classify objects in order to be able to use this data in the set of cameras used in our project.

## References

- [1] D. G. D. Kong and H. Tao. A viewpoint invariant approach for crowd counting. *International Conf. on Patterns Recognition*, pages 1187 – 1190, 2006.
- [2] I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 809 – 830, 2000.
- [3] A. J. L. R. T. Collins and T. Kanade. A system for video surveillance and monitoring. *Carnegie Mellon Univ., Pittsburgh, PA, Tech. Report, CMU-RI-TR-00-12*, 2000.
- [4] J. A. Rosell, G. Andreu, A. Rodas, V. Atienza, and J. M. Valiente. Feature sets for people and luggage recognition in airport surveillance under real-time constraints. *VISI-GRAPP08*, pages 662 – 665, 2008.
- [5] P. Viola. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, pages 153 – 161, 2005.
- [6] T. T. W. Hu, L. Wang, , and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems Man, and Cybernetics-part C: Applications and Reviews*, pages 334 – 351, 2004.
- [7] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfunder: Real-time tracking of the human body. *IEEE Transactions on PAMI*, pages 780 – 785, 1997.