A 3D MODEL BASED VISUAL SURVEILLANCE SYSTEM

Isabel Martins  
Instituto Superior de Engenharia do Porto  
INESC Porto  
e-mail: mmartins@inescn.pt  
Praça da República, 93 r/c, 4050-197 Porto, Portugal

Luís Corte-Real  
Faculdade de Engenharia da Univ. do Porto  
INESC Porto  
e-mail: lreal@inescn.pt

ABSTRACT
This paper presents a visual surveillance system tailored for remote surveillance of indoor environments. It uses a two layers video codec based on a 3D model of the background [2]. The objective is to achieve high data compression rates and to enable the support for applications that require the 3D knowledge of the scene. Examples of such applications are statistical applications (e.g., counting persons) and security applications (e.g., control of entrances in restricted areas). The system supports the ability to localize moving objects in the 3D space assuming that all objects are placed on the ground plane. This feature allows the localization of people on a top view map of the site. Results of some tests are presented in order to show the system ability to deal efficiently with the motion of the camera and to illustrate the two layers separation based on a background/foreground segmentation.

1 INTRODUCTION
Very often, in remote video surveillance applications, the remote site is a physically closed space that is a priori known and is being monitored by one or more video cameras. This is a typical multi-viewpoint application, where there are multiple perspectives of the same scene at the same moment. For this kind of applications the differences in the scene background for each view are due to the camera motion, the variation of the focal length of the lens or the variation of the lighting conditions. In that context, the use of a 3D model of the site and the application of rendering techniques to synthesize the background of the image seems to be a natural approach. The same model can be used to represent the backgrounds of the different views.

The use of a 3D representation of the background can have advantages not only from the point of view of data compression but also from the point of view of scene analysis. The knowledge of the 3D structure of the background scene can be used on the analysis of foreground objects, allowing, for instance, the automatic location of people or objects in the 3D space.

In the system presented in this paper, the representation of images is based on a two-layer model. The first layer is a 3D model of the background and the objects or persons over the background will be treated as a second layer. The objects identified in this second layer may be represented by 3D models previously known to the system or may be treated as 2D objects. The system is able to automatically detect, localize, and track multiple objects entering the scene. In the future, it should be able to classify the detected objects, namely, identify and recognize persons.

2 THE VISUAL SURVEILLANCE SYSTEM
The proposed visual surveillance system is based on an evolution of the codec described in [2]. The codec has two-layers: the first layer represents the background using a 3D model and the second layer represents the foreground using specific standardised techniques of MPEG-4. Presently, all the objects identified in this second layer are treated as 2D objects but, in the future, some of them may be represented by 3D models previously known to the system, like human bodies. This system has been evolving by incorporating new useful features for surveillance applications, taking advantage of its 3D approach.

The main functional blocks of the proposed system are the following:
1. estimation of camera parameters;
2. estimation of changes of illumination;
3. rendering of the synthetic background;
4. background/foreground segmentation;
5. foreground encoding;
6. individual foreground objects identification and tracking;
7. foreground objects classification;

Some of these blocks are already being developed, namely the blocks 2, 6 and 7. The performance of the
estimation of camera parameters, block 1, and the background/foreground segmentation, block 4, are also being studied, in order to improve the basic features of the system.

The 3D model of the background is assumed to be known. However, in the future, the progress on the automatic 3D model construction based on uncalibrated image sequences will allow the implementation of an automatic procedure for the installation of this system. For the tests presented on the section 3 a hand made model has been used.

2.1 Background/foreground segmentation

In most real-world applications, the scene is composed of multiple moving objects (a still background is an object moving with zero velocity). Therefore, it is necessary to segment the scene into the individual objects. So, motion-estimation techniques can be used to help segmentation [9].

In real-time applications, the segmentation has to be performed automatically. This is typically the most difficult case, not only because of time constraints but also because no user interaction is allowed [10]. Automatic methods based on color, texture and motion similarity often fail to capture semantically meaningful objects. In some applications, however, there is only the need to classify the pixels of a video sequence into two classes: moving objects (foreground) and still background. Different techniques have been proposed for this purpose [11][12].

In the present version of the system we are using an algorithm based on the differential images and on simple low-level processing. A binary mask is generated by a threshold operation. In order to simplify the obtained mask, a median filter is used followed by blow and shrink operations using morphological filters. As a final step of the foreground region detection, a binary connected component analysis is applied to the foreground pixels to assign an unique label to each foreground object. This mask is used to define the VOPs to be encoded using the techniques defined in MPEG-4 Video VM.

2.2 Background representation and coding

To represent the background image is necessary to estimate the projection and viewing transformations. For that purpose an automatic camera motion and lens zoom estimator has been implemented. Since it was assumed that the camera is mounted in fixed positions, only pan, tilt and zoom has been considered.

Video coding applications, in contrast to image analysis, require sending the motion parameters to the decoder as side information. In this framework, the cost of transmitting a dense motion field, like optical flow, is usually not affordable. So, the choice of the motion model is therefore restricted to fully parametric models [3]. The basic approach for motion/structure determination consists of two steps [4]: i) extract, match and determine locations of corresponding features (like edges, valleys and corners); ii) determine motion and structure parameters from the feature correspondences.

We are using an estimation method based upon the one proposed in [3], where the binary matching of the edges in successive frames is the salient feature. It is assumed that the space coordinates \((x, y, z)\) of a point in the 3D space and its coordinates in the image plane \((X, Y)\) are related by the perspective transformation

\[
X = F^x_t \quad Y = F^y_t
\]

Combining zoom, pan and rotation models, it arrives at a model that uses 4 parameters to describe the motion of the camera in situations where there is no camera translation:

\[
\begin{pmatrix}
X_2 \\
Y_2
\end{pmatrix} = \begin{pmatrix}
c_1 & c_2 \\
-c_2 & c_1
\end{pmatrix} \begin{pmatrix}
X_1 \\
Y_1
\end{pmatrix} + \begin{pmatrix}
c_3 \\
c_4
\end{pmatrix}
\]

where:

\[
c_1 = F_1/F_2; \\
c_2 = t_z * c_1; \\
c_3 = F_2 * t_y + t_z * F_2 * t_x; \\
c_4 = F_2 * t_x - t_z * F_2 * t_y;
\]

being \(F_1\) and \(F_3\) the focal length before and after the zoom, \(t_x, t_y\) and \(t_z\) the rotation angles around the \(x, y\) and \(z\) axes respectively. It is assumed that these rotation angles are small enough so that \(\cos(t_x) \approx 1, \cos(t_y) \approx 1, \cos(t_z) \approx 1, \sin(t_x) \approx t_x, \sin(t_y) \approx t_y\) and \(\sin(t_z) \approx t_z\). It is also assumed that the rotation is small enough so that \(x_{1,2}t_z \ll z\) and \(y_{1,2}t_y \ll z\).

In our application, where it is assumed that there is no rotation around the camera axis, \(t_z = 0\), this model was simplified to 3 parameters:

\[
\begin{pmatrix}
X_2 \\
Y_2
\end{pmatrix} = \begin{pmatrix}
c_1 & 0 \\
0 & c_1
\end{pmatrix} \begin{pmatrix}
X_1 \\
Y_1
\end{pmatrix} + \begin{pmatrix}
c_3 \\
c_4
\end{pmatrix}
\]

where:

\[
c_1 = F_1/F_2; \\
c_3 = F_2 * t_y; \\
c_4 = F_2 * t_x.
\]

To find the edges we are using an efficient recursive algorithm proposed in [6], that computes the directional derivatives of the smoothed image, from where the gradient magnitude image is computed. It is then non-maxima supressed in the exact gradient direction and thresholded with hysteresis.

Other methods for the estimation of camera parameters from image sequences without any a priori information on the camera system are being tested [7].

It should be noted that with this codec the number of parameters (bits) needed to encode each background image is fixed and independent of the achieved quality.
The quality of the synthetized background image depends only upon the quality of the model and the renderer, and the accuracy of the estimation of the camera parameters.

Since the rendering is a complex task, it is important that different decoders with different computational power, be able to decode the same bitstream. The proposed codec has this feature, as the information carried by the bitstream is independent of the decoder rendering module.

2.3 Foreground coding

The foreground layer is coded using the techniques defined in the MPEG-4 video VM [8]. The MPEG4 video coder uses a decomposition of the scene on Video Objects (VOs). Each VO is characterized by a shape, a motion and a texture. Since the procedure for VO identification is not standardized, we use the mask generated by background/foreground segmentation as input to the MPEG-4 video coder.

In the present version of the codec, each view of the scene is encoded independently. In the future we will explore the multiview nature of the application in order to improve the coding efficiency.

2.4 Features for surveillance applications

The system is able to automatically detect and track the different moving objects present in the scene, based on the the background/foreground segmentation and on the local motion. With these information the system supports the ability to localize objects in the 3D space assuming a ground plane hypothesis to regularize the 2D-into-3D transformation: all objects are considered placed on the ground plane [13]. Presently, the system is able to show the localization of the detected objects on a top-view map of the surveilled site. This allows to follow the movements of persons in the map, the detection of anomalous presence of persons in forbidden areas, etc..

3 RESULTS

Figure 1 shows the results of applying the proposed codec to the “zoom” sequence, composed of 300 frames in CIF format, and that contains only camera zoom operation.

Figure 2 shows the evolution of the estimated camera parameter $c_1$, the zoom parameter, for this “zoom” sequence, that starts with a zoom in and ends with a zoom out.

Presently, 16 bits are used to encode each of the camera parameters. However, the best way to encode these parameters is still a topic for further study.

As stated before, the quality of the synthetized background image is independent of the number of bits generated and depends only upon the quality of the model

and the renderer, and the accuracy of the estimated camera motion or zoom.

The gain in bitrate saving using the proposed codec compared with an MPEG-4 VM based codec that encodes both foreground and background is evident, particularly for sequences containing camera motion or zoom, as reported in [1].

Figure 3 shows an image from a test sequence with

Figure 1: Images from the “zoom” sequence: left) original sequence; right) synthetic sequence.

Figure 2: Evolution of the estimated camera zoom parameter $c_1$ for the “zoom” sequence.
people crossing the scene. The original image 3(a) is divided in background and foreground. The foreground/background segmentation is performed automatically as described before. The background is represented by the synthetic image 3(b), while the foreground is encoded with the object-based techniques 3(c). The resulting encoded image 3(d) is generated composing the synthetic background with the encoded foreground.

Figure 3: Images from a test sequence: a) Original image; b) first layer (synthesized background); c) second layer (foreground object); d) encoded image.

4 CONCLUSIONS

A visual surveillance system tailored for remote surveillance of indoor environments has been presented. It uses a two layers video codec based on a 3D model of the background. The objective is to achieve high data compression rates and to enable the support for applications that require the 3D knowledge of the scene.

Although the first results are promising, further work will be necessary in order to improve the automatic estimation of camera parameters and the automatic background/foreground segmentation.

Techniques to support foreground object classification, like identification and recognition of people [14] are not yet implemented but will be incorporated in the future.

With the incorporation of 3D models of foreground objects, like human bodies, the system will be able to include these objects in the 3D model of the site, allowing the observation of the scene from different viewpoints and not only the camera viewpoint.

References