ACCURATE VIDEO OBJECT SEGMENTATION THROUGH CHANGE DETECTION

Andrea Cavallaro and Touradj Ebrahimi
Signal Processing Institute (ITS)
Swiss Federal Institute of Technology (EPFL)
CH-1015 Lausanne, Switzerland

ABSTRACT

We propose an algorithm for the accurate extraction of video objects from color sequences. The semantics defining the video objects is motion, and the extraction algorithm is based on change detection. The color difference between frames is modeled so as to separate the contributions caused by sensor noise and illumination variations from those caused by meaningful objects. Sensor noise is eliminated by using a probability-based classification, and local illumination variations are removed using a knowledge-based approach that is formulated as a hypothesize-and-test scheme. Experimental results show that the proposed method provides accurate contours of multiple deformable objects, thus providing a reliable input to object-based applications such as those supported by the MPEG-4 and MPEG-7 standards.

1. INTRODUCTION

Object-based video coding and description provide the user with flexibility in content-based access and manipulation of multimedia data. To maximize the benefits of object-based representation, industry standards such as MPEG-4 and MPEG-7 need to be complemented with automatic techniques for extracting the video objects from video data.

The first step towards an automatic extraction of video objects is a clear characterization of the semantics defining the objects of interest. Unfortunately, since the semantics is a human abstraction and context dependent, a unique definition does not exist. In addition, since video objects cannot generally be characterized by simple homogeneity criteria (e.g., uniform color or uniform motion), their extraction is a difficult and sometimes loose task. A set of rules is therefore needed to automatically extract video objects. These rules are based either on special characteristics of the scene or on specific knowledge (a priori information). A typical example of methods based on a specific set-up of the scene is the blue screen approach (a.k.a. chroma-keying), whereas examples of methods based on a priori information are template matching, face detection, and moving object segmentation.

In this paper, motion information is consider as semantics for the extraction of video objects. Methods based on motion information rely on the assumption that the motion of a moving object is usually different from the motion of background and other objects. To automatically extract video objects either motion-based segmentation or motion detection can be used. Motion-based segmentation consists in classifying pixels of the image into clusters such that pixels in a given cluster have similar motion. It thus involves motion estimation followed by a clustering step. Motion detection simply identifies those pixels where apparent motion exists. Compared to motion-based segmentation, motion detection provides better contours and is computationally less expensive. For these reasons motion detection is often used to apply the human abstraction corresponding to moving objects. In particular, change detection techniques are widely used to extract video objects. A general model that allows us to compare different change detection techniques is proposed in Section 2. Furthermore a new technique for the accurate segmentation of moving objects is presented in Section 3. Experimental results are discussed in Section 4. Finally, in Section 5 we draw the conclusions.

2. MOTION AS SEMANTICS: CHANGE DETECTION

Change detection is a temporal segmentation tool aiming at identifying changes in image sets or image sequences at two different times. In the framework of semantic video object extraction, a change detection algorithm is ideally expected to extract the precise contours of objects moving in a video sequence (spatial accuracy). An accurate extraction is especially desired for applications such as video editing, where objects from one scene can be used to construct other artificial scenes, or computational visual surveillance, where the objects are analyzed to derive statistics about the scene.

Let us denote the image under test as \( f(x, y, n) \).\(^1\) The problem of change detection consists in finding, for each frame \( n \), a binary map \( c(x, y, n) \), defining the pixels in \( f(x, y, n) \).

\(^1\)In this paper, we represent a general sequence as function \( f(x, y, n) \) of three variables: two spatial, \( x \) and \( y \), and one temporal, \( n \).
which have changed with respect to the reference image \( f(x, y, r) \).\(^2\) The binary mask \( c(x, y, n) \), resulting from the change detection analysis, is defined as

\[
c(x, y, n) = \begin{cases} 
1 & \text{if a change occurred at } (x,y) \text{ at time } n, \\
0 & \text{otherwise.}
\end{cases}
\]

Different strategies \( \mathcal{M} \) can be adopted to compute \( c(x, y, n) \) as a function of \( f(x, y, n) \) and \( f(x, y, r) \). We propose a decomposition of \( \mathcal{M} \) into four major steps, namely feature extraction, feature analysis, classification, and post-processing. The block diagram of the proposed scheme is shown in Figure 1. The choices related to the algorithm \( \mathcal{M} \) comprise the features to extract from \( f(x, y, n) \) and \( f(x, y, r) \), the distance metric to quantify changes, and the classification strategy to detect changes.

a) Feature extraction. The first step towards change detection consists in transforming each input frame of the image sequence \( f(x, y, n) \) into the most appropriate feature space, \( g(x, y, n) \). The choice of the feature space depends on the application. The signal \( g(x, y, n) \) may represent the luminance [1, 3, 4], the color components [5], or more complex features. More complex features are usually employed in applications where the illumination of the scene cannot be easily modeled [6, 7]. Finally, the parameters of a region-based model of the input image can be used [2].

b) Feature analysis. The result of this first step is a sequence \( g(x, y, n) \), representing the signal on which the change detection operation will be performed. A feature analysis step follows the feature extraction step in order to derive an indicator of the level of activity in the frame under consideration, resulting in an activity index. The activity index is computed by comparing \( g(x, y, n) \) with the reference image, \( g(x, y, r) \). The reference image can be either a frame representing the background of the scene, a previous frame, or any other appropriate entity. A distance operator provides a pixel level feature distance and can be implemented as pixel-wise difference [1, 6], image ratioing [2], or a difference metric based on second order statistics [8].

c) Classification. The indicator of the level of activity, \( t(x, y, n) \), has to be classified in one of the two classes: changed or unchanged. To obtain the classification, \( t(x, y, n) \) is binarized by thresholding. The threshold can be set empirically [2, 8] or computed adaptively [1, 3, 4, 7]. In the former case, the threshold is fixed for all pixels in frame and all the frames in the sequence. Its value is usually determined experimentally based on a large database. In the latter case, the threshold is adapted according to some rules, as described in Section 3.

d) Post-processing. The result of the classification step is affected by noise coming from different sources. Apart from camera noise, imperfections may arise from approximations of the model used in the change detection algorithm. Various sources of noise could be responsible for false alarms in the change detection mask \( c(x, y, n) \). To reduce the probability of false alarms, post-processing strategies have been proposed in the literature. These strategies are applied either to the binary image result of the classification only [1, 6], or to both the binary images and the original frame [3, 4].

3. PROPOSED METHOD

The temporal changes identified by the change detection process are used to segment the video objects from the background. However, temporal changes may be generated not only by the objects, but also by noise components. The main sources of noise in the feature analysis step are illumination variations, and camera noise. We propose to eliminate the former in the classification step and the latter in the post-processing step.

3.1. Classification

The classification step decides whether the foreground signal corresponding to an object is present in each pixel position, while discarding the effect of the camera noise. To discount the residual effect of the camera noise, the activity index is binarized by thresholding. Early change detection techniques fixed the threshold empirically. They performed sufficiently well on sequences where moving objects are well contrasted from the background. However, thresholds have to be tuned manually according to the characteristics of the sequence and often need to be updated from one frame to another along the sequence itself. These major drawbacks limit this approach for fully automatic applications. To overcome these problems and to obtain a much more flexible procedure, we employ a locally adaptive threshold. This dynamic thresholding strategy models the noise statistics and applies a significance test [3]. The noise model is defined based on the following assumptions: all pixels in the neighborhood have changed only because of noise (hypothesis \( H_0 \)), and each frame of the sequence is affected by an additive Gaussian noise with a certain mean and variance. Under these hypotheses the noise model is described by a \( \chi^2 \) distribution, whose properties depend on the number of pixels in the neighborhood and on the variance of the Gaussian noise affecting each frame of the sequence. Given the \( \chi^2 \) distribution and a significance level \( \alpha \), the adaptive value of the threshold \( \tau_n \) can be computed through

\[
\alpha = P\{\delta > \tau_n | H_0\},
\]

where \( H_0 \) is the hypothesis that no change occurred in the pixel position under test. The significance level \( \alpha \) is a stable parameter that does not need manual tuning.

\(^2\)The reference frame, \( f(x, y, r) \), does not necessarily correspond to a given frame of the sequence.
3.2. Post-processing

The post-processing step makes use of both the binary mask resulting from the classification, \( c(x, y, n) \), and the original images, \( f(x, y, n) \). The proposed approach is based on the evaluation of heuristic rules which derive from the domain specific knowledge of the problem. The physical knowledge about the spectral and geometrical properties of shadows are used to define explicit criteria which are encoded in the form of rules. A bottom-up analysis organized in three levels is performed. This hierarchical control structure uses the hypothesize-and-test scheme.

a) Hypothesis generation. The presence of a shadow is first hypothesized based on some initial evidence. A candidate shadow region is assumed to correspond to a darker region than the corresponding illuminated region (the same area without the shadow). The color intensity of each pixel is compared to the color intensity of the corresponding pixel in the reference image. A pixel \((x, y, n)\) becomes a candidate shadow pixel if all color components are smaller than the corresponding pixel in the reference frame.

b) Accumulation of evidence. The hypothesized shadow region is then verified by checking its consistency with other additional hypotheses. The presence of a shadow does not alter the value of invariant color features. However, a material change is highly likely to modify their value. For this reason, the changes in the invariant color features \( c_1c_2c_3 \) [10] are analyzed to detect the presence of shadows. The second additional evidence about the existence of a shadow is derived from geometrical properties. This analysis is based on the position of the hypothesized shadows with respect to objects. The existence of the line separating the shadow pixels from the background pixels (the shadow line) is checked when the shadow is not detached, that is, an object is not floating, or the shadow is not projected on a wall. If a shadow is completely detached, the second hypothesis is not tested. In case a hypothesized shadow is fully included in an object, the shadow line is not present, and the hypothesis is then discarded.

c) Information integration. Finally, all the pieces of information are integrated to determine whether to reject the initial hypothesis.

The post-processing step results in a spatio-temporal regularization of the classification results.

4. RESULTS

The results of the proposed change detection before and after the post-processing step are compared. The same set of parameters has been used for all the sequences. Test sequences from the MPEG-4 and MPEG-7 data set are used, as well as test sequences from the test set of the European project art.live.3 All sequences are in CIF format (288 × 352 pixels) and the frame rate is 25 Hz. The results of the segmentation are visualized as follows. Two columns are displayed: on the left hand side, sample results of the probability-based classification are shown. The results of the change detection are visualized by superposing the change detection mask over the original sequence. On the right hand side, the post-processed results of the corresponding frame in the same row are shown.

In Figure 2 the sequence Hall Monitor is considered. As can be seen, the contours of the extracted objects are correctly defined and they are stable over time. The sequence Group is considered in Figure 3. By applying the probability-based test, objects are segmented, but spurious areas corresponding to shadows are also detected as (Figure 2 (a) and Figure 3(a)). The results shown in Figure 2 (a) and Figure 3(a) differ from those of an ideal object extractor for two aspects. The first aspect is the low-pass filter effect introduced by the use a windowing, whose dimensions are a trade-off between robustness to noise and accuracy of detection. For this reason, the extracted contours are slightly larger than the real ones. This error can be easily corrected by a further post-processing module if the application requires contours exactly fitting the objects. The second deviation from an ideal extraction is the presence of shadows in the change detection mask. Shadows are in fact detected as moving objects since they possess the same characteristics. From a theoretic point of view, this is a correct decision since the statistical change detection is expected to detect changes that are not due to noise. However, in object-based applications, the detection of shadows is annoying, as it modifies the coherence of the shape of the object. This problem is overcome by the proposed post-processing (Figure 2 (b) and Figure 3(b)).

3http://www.tele.ucl.ac.be/PROJECTS/art.live/
5. CONCLUSIONS

An adaptive change detection algorithm for the extraction of multiple moving objects has been presented. The algorithm is designed based on a color differentiation between frames, to isolate the error contributions caused by sensor noise and illumination variations. Sensor noise is eliminated in the classification step of the change detector by taking into account its statistics, thus adapting the detection threshold to local information. Local illumination variations are eliminated in the post-processing stage of the change detector, by using a knowledge-based approach organized in a hypothesize-and-test scheme. Experimental results show that changed regions are detected with good spatial precision in different test sequences.

6. REFERENCES


