ROBUST SEGMENTATION AND REPRESENTATION OF FOREGROUND KEY-REGIONS IN VIDEO SEQUENCES

Javier Ruiz Hidalgo and Philippe Salembier

Universitat Politecnica de Catalunya, Barcelona, Spain.
{jrh,philippe}@gps.tsc.upc.es

ABSTRACT

This paper deals with the extraction and characterization of foreground objects in video sequences. The algorithm first computes the mosaic image representing the background information and then extracts foreground objects. In this last step, the foreground objects are progressively extracted taking into account the reliability of the contour information. This extraction step is based on morphological tools. Finally, the foreground objects are characterized by their shape, texture and motion trajectory. Moreover, some information about the temporal evolution of non rigid objects is also extracted. This feature extraction algorithm is particularly suitable for the indexing, search and retrieval applications.

1. INTRODUCTION

The growing of multimedia applications, such as video-on-demand or digital library systems, have generated a strong interest in content-based analysis of video sequences. A general overview of the major techniques for video and image indexing can be reviewed in [1]. For video indexing applications, the initial phase generally consists in structuring the original content. A classical structuring approach consists in detecting shots in a video sequence and to group them into scenes (see [2, 3, 4] and the references herein). The description of shots is often based on key-frames. For instance in [5], several key-frames are used to represent a set of shots and to browse them. These key-frames are also indexed using standard techniques for still images. Other methods for shot representation involve a more complex analysis of the spatio-temporal content. In [6] and [7], for instance, the representation of a shot is composed of a set of layers. In this framework, mosaic images are often used to represent the background information over an entire shot [6]. Mobile foreground objects can then be superimposed to the mosaic representation in order to indicate their relative trajectories to the global background motion [8].

The shot representation technique proposed in this paper is based on a mosaic for the background and a set of foreground key-regions. One of the important feature of the proposed approach is that foreground key-regions are progressively extracted taking into account the reliability of the contour information. This extraction step is based on morphological tools. Moreover, foreground key-regions are characterized by their shape, texture and motion trajectory. Finally, some information about the temporal evolution of non rigid objects is also extracted allowing the object activity to be analyzed.

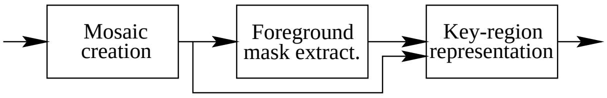

Fig. 1. Outline of the shot representation algorithm.

The paper is organized as follows. Section 2 gives an overview of the proposed algorithm. Section 3 presents the algorithm used for the computation of the background mosaic. Section 4 explains the foreground segmentation algorithm and the creation of key-regions. Finally, conclusions are drawn on section 5.

2. OVERVIEW OF THE APPROACH

The algorithm for mosaic and key-region extraction is based on three steps highlighted on Fig. 1. In the first step, the mosaic representing the background is created. The information resulting from this initial step is the background mosaic image and the set of warping parameters used to create the mosaic from the individual frames of the video sequence. In the second step, a foreground mask is computed for each frame by comparing the input frame with the appropriate part of the background mosaic image. Finally, key-regions are constructed and modeled in the third step by using the foreground masks collected on individual frames.

This approach results in a non-casual algorithm because the first step of the algorithm consist in computing the background mosaic for the entire shot. This step introduces a processing delay that is not compatible with real-time applications. However, this approach allows motion and background color information to be jointly used to obtain the foreground mask. This approach is different from classical techniques where only motion information is available [9, 10].

3. BACKGROUND MOSAIC CREATION

Fig. 2 illustrates the background mosaic creation algorithm. The approach is classical and involves four main blocks. The first one estimates the dominant motion $m_f(t)$ between two successive frames, $I(t)$ and $I(t-1)$, of the original sequence. The dominant motion is assumed to represent the motion of the background. The motion estimation is based on a 2D perspective motion model.

To allow a robust creation of the mosaic image, a gray level mask indicating whether pixels belong to the foreground or to the background is created in the second block (Foreground / Background mask computation). The mask image $W_f(t)$ assign high

The authors would like to thank the support from the European Commission and in particular, the Diceman ACTS Project.

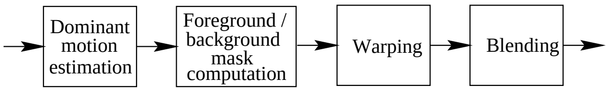
Dominant motion estimation
Foreground / background mask computation
Warping
Blending

Fig. 2. Mosaic creation algorithm.

Fig. 3. Two frames of the *nhkvideo7* sequence.

values to pixels following the dominant motion (assumed to belong to the background) while it assigns low values to pixels that do not follow the dominant motion $m_f(t)$. The mask image is computed with the morphological motion operators described in [10].

The third step is the warping. The dominant motion parameters $m_f(t)$ are accumulated in order to align all video frames with a common spatial reference. To prevent the propagation of possible errors during the dominant motion estimation, the accumulated motion parameters at time t are recomputed during the warping step using the previously computed mosaic image $M(t-1)$. The new motion parameters, called warping parameters, $m_m(t)$, relate the mosaic image $M(t)$ with the input frames at time t. They are used to re-align the final mosaic image with the input frames.

Finally, the fourth block is the blending that updates the current mosaic image $M(t)$ using the computed warping parameters $m_m(t)$, the current input image $I(t)$ and the mask image $W_f(t)$. A mosaic of the MPEG-7 *nhkvideo7* test sequence (Fig. 3 shows two frames of the sequence) can be seen in Fig. 8.

4. KEY-REGION EXTRACTION AND REPRESENTATION

4.1. Overview of the algorithm

The key-region extraction algorithm can be seen in Fig. 4. The approach is quite similar to the one used for the background mosaic creation. First, the mosaic image and the corresponding warping parameters are un-warped to obtain a background image aligned in time with the current input frame. Then a foreground mask is extracted for each input video frame. This mask is obtained by comparing the background image with the current video frame. The contours of foreground regions are obtained with a watershed algorithm [11, 12]. Finally, the two last blocks track the resulting foreground masks in time, combine them and model the key-regions for the entire shot.

4.2. Foreground mask extraction

The mosaic alignment uses the previously computed mosaic image and the warping parameters $m_m(t)$ to align in time the mosaic image with the input original frame $I(t)$. The resulting motion compensated background image ($B(t)$ in Fig. 5) is used to obtain a difference image, $D(t) = I(t) - B(t)$, where foreground regions are highlighted from the background (middle-left of Fig. 5).

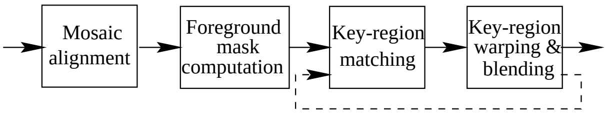

Fig. 4. Key-region extraction algorithm.

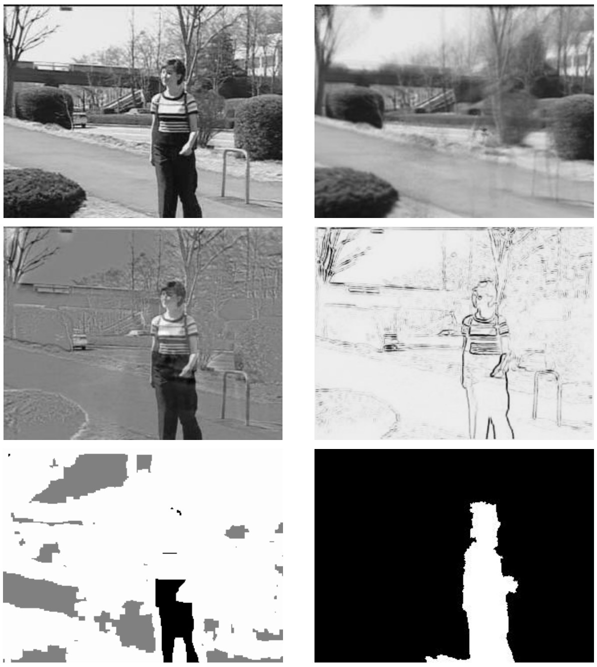

Fig. 5. Example of foreground mask extraction. Top row: input image $I(t)$ and aligned background image $B(t)$. Middle row: difference image $D(t) = I(t) - B(t)$ and gradient image $G(t)$. Bottom row: background (grey) and foreground (dark) markers, and final foreground mask.

A watershed algorithm is used to segment the foreground regions. The watershed algorithm is applied on a gradient image and relies on markers indicating roughly the interior foreground and background regions. The gradient image should highlight the contours of foreground regions. It is primarily constructed on the basis of the gradient of the difference image: $\mathcal{S}\{D(t)\}$, where $\mathcal{S}\{\cdot\}$ represents the gradient operator. This gradient not only highlights the contours of foreground regions but also all textured areas. To improve the robustness of the algorithm, the gradient is weighted by the temporal gradient: $\mathcal{S}\{|I(t) - I(t-1)|\}$. Therefore, the final gradient is :

$$G(t) = \mathcal{S}\{D(t)\} \cdot \mathcal{S}\{|I(t) - I(t-1)|\} \quad (1)$$

The markers are simply computed as follows: All connected components of the space where the difference image is below (above) a given threshold, T_1 (T_2), are used as background (foreground) markers. Fig. 5 (bottom-left) shows an example of the extracted markers for the *nhkvideo7* sequence. The bottom-right row of Fig. 5 shows the resulting foreground mask. In this case, the girl is successfully segmented from the background.

4.3. Key-region matching and modeling

The foreground masks, $M_f(t)$, extracted in the previous section for each individual input frame are matched with key-regions over the entire shot, combined and modeled. This last step is not only necessary to have a global model for each key-region valid for the entire shot but also to improve the quality of the estimation obtained on a frame basis. The matching is based on a simple overlapping algorithm: a connected component of the foreground mask is assigned to an existing key-region if the overlap with the last assigned foreground mask of the corresponding key-region is sufficient. This works well on real scenes where changes between frames (at 25/30 fps) are usually small. Then the current foreground mask and the existing key-region are aligned using a perspective model. If the connected component of the foreground mask does not correspond to any known key-region, a new key-region is created.

The following step consists in updating the existing key-regions with the information from the current foreground masks. The first update addresses the key-region shape and is based on the reliability of the contours. Assume that I is an image and M a mask, let $\mathcal{C}\{I, M\}$ denote an image equal to zero except on the contours of the mask M where it takes the values of I.

$$\mathcal{C}\{I, M\} = \begin{cases} I, & \text{if } \mathcal{S}\{M\} \neq 0 \\ 0, & \text{if } \mathcal{S}\{M\} = 0 \end{cases} \tag{2}$$

The contour reliability of the foreground mask $M_f(t)$ is given by:

$$C_f(t) = \mathcal{C}\left\{\mathcal{S}\{I(t)\}, M_f(t)\right\} \tag{3}$$

$C_f(t)$ is a confidence value because low values imply that the contour does not correspond to contrasted edges (this can occur, for instance, when the foreground occludes a background of the same color). By contrast, high values of $C_f(t)$ correspond to strong edges on the original image and therefore to reliable contours.

Fig. 6 illustrates the use of the reliability information to update the key-region shape. In this example, the foreground mask extracted at frame 2039 of the *nhkvideo7* sequence is of poor quality because the contrast between the right part of the girl and the background area was low at that time instant. The two images on the left of Fig. 6 show the extracted foreground mask $M_f(t)$ and the corresponding contour reliability $C_f(t)$. The reliability of the current extracted mask is compared with the reliability of the key-region contours, $C_k(t-1)$, computed with past foreground masks. The updated contour of the key-region are obtained by applying a watershed algorithm on $\{aC_f(t) \vee (1-a)C_k(t-1)\}$, where $\vee$ denotes the maximum (the markers for the watershed are one point outside the mask and one in the center of the mask). The watershed extracts the crest line of highest value between $aC_f(t)$ and $(1-a)C_k(t-1)$. The watershed algorithm automatically combines both contour information keeping the most reliable part of them. Let us denote by $\hat{C}_m(t)$, this combined confidence value.

The parameter $a \in [0, 1]$ controls the memory of the allowed modifications to the shape of extracted foreground masks. If $a \simeq 0$, the previously computed key-region contours are more trusted than current contours coming from the foreground mask. In this case, errors in the foreground mask are easier to fix but tracking non-rigid foreground regions becomes more difficult. On the other side, if $a \simeq 1$ non-rigid regions are easier to track but errors in the foreground mask are also more difficult to correct. In our case, a value of $a = 0.5$ has been used for all examples.

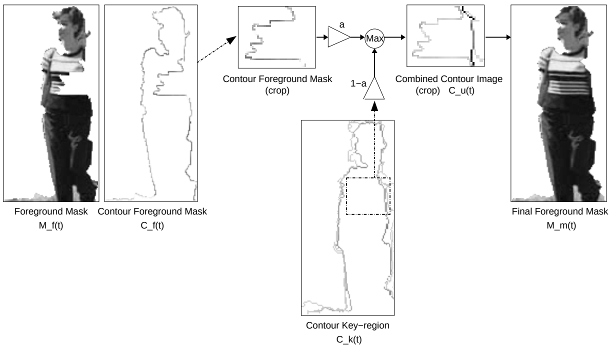

Fig. 6. Re-estimation of the key-region shape using the reliability of contours.

The resulting mask $M_m(t)$ is shown on the right side of Fig. 6. As can be seen, the error in the shape of the foreground mask extracted at time t has no real influence because its reliability is low whereas the reliability of the corresponding key-region contour is rather high. In general, this procedure allows us to progressively improve the estimation of the key-regions contour taking into account the reliability of what is observed in time.

The final step of the algorithm updates the key-regions template. It corresponds to the Key-region Warping and Blending block of Fig. 4. The key-region template is composed of three images: An appearance image, a contour image and a texture image. The appearance image $A_k(t)$ shows the frequency with which a pixel has been segmented as foreground and assigned to the key-region k. If the mask $\hat{M}_m(t) = 1$ denotes pixels that have been extracted and assigned to key-region k, the updating of the appearance image can be done with $A_k(t) = A_k(t-1) + \hat{M}_m(t)$. The contour image stores the confidence of the key-region contours. The texture image represents the overall texture of the key-region and is updated using the compensated mask and the input image. The updating equations of contours and texture image are illustrated in equation 4.

$$\begin{aligned} T_k(t) &= \left(A_k(t-1)T_k(t-1) + \hat{I}(t)\right)/A_k(t-1) \\ C_k(t) &= \left(A_k(t-1)C_k(t-1) + \hat{C}_m(t)\right)/A_k(t-1) \end{aligned} \tag{4}$$

where $\hat{I}(t)$ is the compensated image. Note that only pixels included in the foreground segmented mask $\hat{M}_m(t)$ are updated. Fig. 7 shows the key-region template from a shot where a person walks in front of the camera. From left to right, the appearance, contour and texture template image of the key-region corresponding to the walking person are presented. The appearance and texture image contain information about the activity of the key-region. In this case, higher body parts (body, chest) show no relative movement while lower parts (legs) show a considerable amount of relative motion. This representation is particularly attractive to analyze the activity of non rigid objects.

Fig. 8 shows a complete shot representation of the *nhkvideo7*

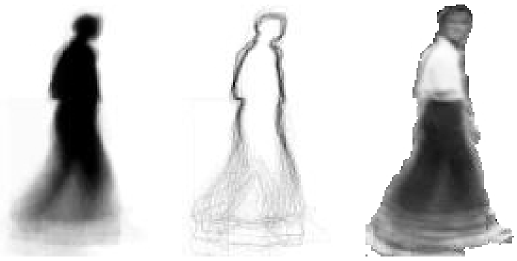

Fig. 7. Modeling of key-regions. The template of a key-region is composed, from left to right, of an appearance image $A_k(t)$, a contour image $C_k(t)$ and a texture image $T_k(t)$.

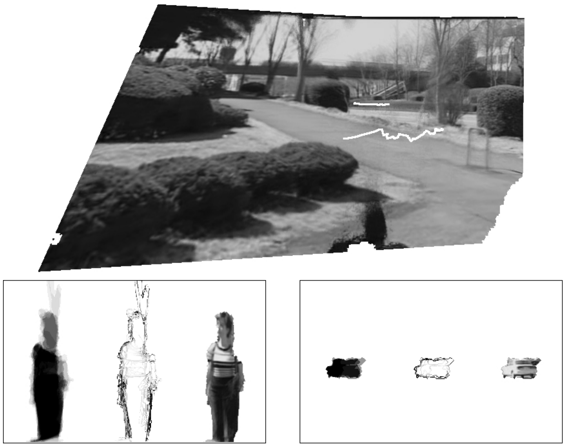

Fig. 8. Final result for the *nhkvideo7* sequence. A background mosaic image and two key-regions are found. Key-regions are represented from left to right by an appearance image A, a contour image C and a texture image T.

sequence. The background information is separated from the foreground regions of the scene. In the original sequence, the camera follows the walking girl while a car crosses the road in the background. Two key-regions have been extracted corresponding to the girl and the car. Bottom images show the corresponding texture images of the two key-regions. Superimposed to the final mosaic image, the relative motion respect to the camera is drawn providing a fast visual representation of the motion followed by each key-region. The top white line shows the motion followed by the car while the one on the bottom shows the path followed by the girl during the analyzed sequence.

5. CONCLUSION AND FUTURE WORK

A method for representing and structuring video shots has been presented. A robust 2D motion estimation is used to estimate and create a mosaic image representing the background information of the shot. This background information is then used to extract representative foreground regions, called key-regions. The major contribution of this paper deals with the foreground region extraction and analysis. The algorithm progressively builds a representation for key-regions taking into account the reliability of the shape information for each frame. Both key-regions and mosaic image create a compact and efficient representation of the content. Moreover, they allow the activity during the shot to be described. The most representative foreground regions of the scene are represented by templates allowing further indexing and analysis.

Possible extensions and improvements of the algorithm are been currently studied. New methods to improve the foreground region extraction in case of occlusions are being investigated. Also methods to classify the activity of key-regions (such as walking, running, seating if the case, for instance, of human key-regions) can be studied on the basis of the appearance and texture images of the extracted key-regions.

6. REFERENCES

[1] P. Aigrain, H.J. Zhang, and D. Petkovic, “Content-based representation and retrival of visual media: a state-of-the-art review,” *Multimedia Tools and Applications*, vol. 3, no. 3, pp. 179–202, November 1996.

[2] M. Yeung and B.-L. Yeo, “Time-constrained clustering for segmentation of video into story units,” in *International Conference on Pattern Recognition*, 1996, pp. 375–380.

[3] N.V. Patel and I.K. Sethi, “Video shot detection and characterization for videop databases,” *Pattern Recognition*, vol. 30, no. 4, pp. 607–626, April 1997.

[4] P. Salembier and F. Marqués, “Region-based representations of image and video: segmentation tools for multimedia services,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1147–1169, December 1999.

[5] H.J. Zhang, J. Wu, D. Zhong, S.W. Smoliar, “An integrated system for content-based video retrieval and browsing,” *Pattern Recognition*, vol. 30, no. 4, pp. 643–658, April 1997.

[6] H. Sawhney and S. Ayer, “Compact representations of videos through dominant and multiple motion estimation,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 814–830, August 1996.

[7] J.Y.A. Wang and E.H. Adelson, “Representing moving images with layers,” *IEEE Trans. on Image Processing*, vol. 3, no. 5, pp. 625–638, September 1994.

[8] M. Irani and P. Anandan, “Video indexing based on mosaic representations,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 86, no. 5, pp. 905–921, May 1998.

[9] I. Kompatsiaris and M. G. Strintzis, “Spatiotemporal segmentation and tracking of objects in image sequences,” in *International Conference on Image Preocessing, ICIP’99*, Kobe, Japan, 24-28 October 1999, pp. 155–158.

[10] P. Salembier, A. Oliveras, and L. Garrido, “Anti-extensive connected operators for image and sequence processing,” *IEEE Trans. on Image Processing*, vol. 7, no. 4, pp. 555–570, April 1998.

[11] F. Meyer and S. Beucher, “Morphological segmentation,” *Journal of Visual Communication and Image Representation*, vol. 1, no. 1, pp. 21–46, September 1990.

[12] L. Vincent and P. Soille, “Watersheds in digital spaces: an efficient algorithm based on immersion simulations,” *IEEE, Trans. on Pattern Analyis and Machine Intelligence*, vol. 39, no. 12, pp. 1845–1855, December 1991.