Unsupervised video object segmentation and tracking based on new edge features

Byung-Gyu Kim *, Dong-Jo Park

Department of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology (KAIST),
373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea

Received 14 July 2003; received in revised form 28 November 2003
Available online 11 September 2004

Abstract

We present an efficient video segmentation and tracking strategy based on edge information to assist object-based video coding, motion estimation, and motion compensation for MPEG-4 and MPEG-7. The proposed algorithm utilizes the human visual perception to provide edge information. Three parameters are introduced and described based on edge information from the analysis of a local histogram. An edge function is defined to generate the edge information map, which can be thought as the gradient image. Then, an improved marker-based region growing and merging techniques are derived to separate the image regions. An efficient temporal segmentation and tracking algorithm is also developed in time domain when the initial segmentation is given. The proposed algorithm is tested on several standard sequences and demonstrates high reliability for video object segmentation and tracking.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Video object; Edge information; Segmentation; Tracking

1. Introduction

Object-based video coding is a distinct feature of the MPEG-4 or MPEG-7 standards. It was unavailable in the earlier MPEG-1 and MPEG-2 standards. These newer standards introduce the concept of the video object layer (VOL) to support content-based functionality at the user decoder (Koenen, 1999; MPEG, 1998). Under the concept of VOL, segmentation of video objects (VO) in the MPEG-4 and MPEG-7 coding schemes is inevitable in a frame.

In most algorithms for video object segmentation, moving objects in the time domain are usually considered to be meaningful objects. Thus, many developed algorithms for VO segmentation are based on the motion field and change detection between consecutive frames (Meier and Ngan, 1998; Meier and Ngan, 1999; Kim and Hwang, 1999;
Various algorithms that are based on edge and color information using color clustering, watershed, or edge masks have been suggested to detect the boundary of the video object (Salgado et al., 2000; Meier and Ngan, 1998; Salembier et al., 1998; Herrmann et al., 1999; Gu and Lee, 1997; Gambotto, 1992; Gu and Lee, 1998; Zhou et al., 2000; Deng and Manjunath, 2001; Gao et al., 2001; Kim and Kim, 2003). Approaches based on data clustering are iterative and require exhaustive computation. Edge detection based on edge masks requires spatial convolution for every pixel. In many cases of natural images, there are no perfect closed boundaries by using these edge masks. Thus, some additional post-processing such as edge linkage, edge clustering, and removal of isolated edges should be applied for the best result. Also, edges are often missed where the change of intensity or color is gradual.

We introduce an automatic video object segmentation and tracking algorithm based on newly defined edge information. Analysis of local histogram properties is performed to define three parameters in the analysis window for every pixel. Use of the modality and the defined parameters of the given histogram allows definition of a measure that can provide edge information as a gradient map (Gu and Lee, 1998; Gao et al., 2001; Kim and Kim, 2003).

This paper is organized as follows: Section 2 presents the unsupervised video object (VO) segmentation algorithm with new edge features. An efficient simultaneous segmentation and tracking algorithm is developed in detail in Section 3. The proposed algorithm is tested on standard sequences and results of VO segmentation are shown in Section 4. Conclusions are presented in Section 5.

2. The proposed video object segmentation in spatial domain

We define the edge features for the edge information based on an analysis of the local histogram rather than using spatial masks. An edge can be considered as adjacent regions having two or more distinct brightness or color values. If the intensity distribution of an image is multi-modal, the image is probably edge-contained. Otherwise, the image has no edge information.

2.1. Description of edge features

These parameters should specify the properties of the local histogram of the window image.

- Local contrast ($C$)
  Contrast has been exactly defined as a ratio of the absolute value of the difference between the intensity of the desired region and the intensity of the neighborhood to the luminance of the neighborhood (Graham and Clarence, 1965; Jain, 1997). The contrast gives important information regarding the edge of objects. To use this characteristic, the average contrast can be defined as the difference between the intensity of an object and the intensity of neighboring objects. We introduce the local contrast between two or more distinct regions based on the image histogram, as follows:

$$C = L_{right} - L_{left}, \quad (1)$$

where $v_i$ is an analyzed local minimum point, $L_{right}$ and $L_{left}$ are the intensity level at the maximum peak of the right-side to $v_i$ and that of the left-side to $v_i$, respectively. From Fig. 1, the local contrast $C$ is $|c_1 - c_2|$ by our definition. The local contrast of the image approximates the contrast between two distinct regions of the image. It has been reported that the distinctiveness between two regions is linearly proportional to this contrast in human visual perception.

- Region ratio ($R$)
  For the given image with two regions, the distinctiveness between regions also depends on the areas of two regions. This is called as the areal effects (Graham and Clarence, 1965). An inverse relationship between the size of an object and its perceptible luminance has been reported (Graham and Clarence, 1965). Objects
with small retinal images have high luminance thresholds and objects with larger retinal images have lower perceptible thresholds. For bi-modal distribution of the image in Fig. 1, the distribution of the gray level shows that the image has two distinct regions. Under this situation, the region ratio ($R$) is defined as the ratio of the minimum area ($B$ in Fig. 1) to the maximum area ($A$ in Fig. 1). We can also see that the distinctiveness between image regions increases as the region ratio increases.

- **Edge potential ($P$)**

  In an image with several regions, the homogeneity of a region becomes larger as the variation of the intensity decreases. In human visual response, it has been reported that the distinctiveness between regions gets larger as the homogeneities of the adjacent regions increase (Graham and Clarence, 1965; Jain, 1997). Thus, we introduce another parameter to take into account the variation of the intensity in the region. For any local minimum $v_i$, the edge potential can be expressed as follows:

  $P = \text{Min}\{p_{\text{left}}^{v_i}, p_{\text{right}}^{v_i}\} - p(v_i), \tag{2}$

  where $p_{\text{left}}^{v_i}$ is the probability of the maximum peak on the left-side for $v_i$, $p_{\text{right}}^{v_i}$ is the probability of the maximum peak on the right-side for $v_i$ and $p(v_i)$ is the probability at the local minimum $v_i$.

The image contains more edge information as the edge potential increases. The concept of edge potential is illustrated in Fig. 1 where the edge potential $P$ is $|P_2 - P_3|$ for the local minimum $v_i$.

### 2.2. Spatial segmentation algorithm

We define an edge function for the analyzed parameters $R$, $P$, and $C$ to obtain edge information using the defined parameters, as follows:

$$D(R_i, P_i, C_i) = w_C \frac{C_i}{C_{\text{max}}} + w_R \frac{R_i}{R_{\text{max}}} + w_P \frac{P_i}{P_{\text{max}}} \text{ for each } v_i, \tag{3}$$

where $w_C$, $w_R$, and $w_P$ are weighting factors which are set to be 1. Also, $C_{\text{max}} = 255$, $R_{\text{max}} = 1.0$, and $P_{\text{max}} = 0.5$.

For case of multiple valleys in the histogram, three parameters can be determined as $\{C_i, R_i, P_i\}$ for each local minimum point $v_i$. If there are two local minimum points $v_1$ and $v_2$, two sets of parameters $\{C_1, R_1, P_1\}$ and $\{C_2, R_2, P_2\}$ are computed for each minimum $v_1$ and $v_2$.

After obtaining these feature parameters, the edge information $D(\cdot)$ can be computed by Eq. (3) for each local minimum $v_i$. Thus, the edge information $D(\cdot)$ can exist as many as the number of local minimum $v_i$ dose. By the definition of the edge function, this measure becomes larger as parameters go to their maximum values. So, we take the maximum value among the computed the edge information $D(\cdot)$’s as edge response at the given image.

Based on $D(R_i, P_i, C_i)$, the amount of the edge information with the windowed image is computed for the entire pixels. Since this can be thought as a gradient image (edge map) of the original image, this edge map is a good indicator of whether an area is in the interior of the region or near the boundary. For color image segmentation in RGB space, the edge information image for a pixel $(i,j)$ is generated by taking the maximum value among the analyzed edge measures $\{D_R(\cdot), D_G(\cdot), D_B(\cdot)\}$ for $R$, $G$ and $B$ components.

The size of the analysis window determines the size of the image region that can be detected. Thus,
the size of the window should be chosen based on features of the image.

The characteristics of the generated edge map allow us to use a region growing method to segment the image. Region growing consists of determining marker points and expanding from marker locations. Region growing is followed by region merging and filtering to yield the final segmentation result.

### 2.2.1. Marker extraction

A set of initial marker areas are selected as the basis for region growing. In general, segmentation results of various region growing algorithms strongly depend on whether markers are good or not. If the extracted markers are very noisy (i.e., useless regions where their sizes are very small), this makes the algorithm slow to partition the image into the final segmentation. Also, more procedures like some regions filtering based on the color and size of the region should be required for the final result. This causes an increase of processing time of the algorithm.

In this work, all pixels are quantified by the pre-defined quantification level with the acquired edge information map. The distribution of the quantified value is then estimated. With the estimated distribution, the proper threshold value $T$ is selected for marker areas by entropic thresholding (Pun, 1981). Pixels with values of the edge map less than $T$ are considered as marker points. Remaining pixels are the uncertainty region for the growing procedure.

### 2.2.2. Marker growing

For marker extraction, the edge map is divided into the two categories of marker areas and uncertainty regions. A marker area is assigned a label, then the uncertainty regions are grouped into the marker area. A modified approach is used for implementation of this procedure, as follows:

1. The average values of colors $R$, $G$ and $B$ for each labeled marker region are computed. For an unclassified pixel with only one labeled region in its neighborhood, the pixel is assigned to the region. If the pixel has more than one labeled region in its neighborhood, it is assigned to the label of several regions if the shortest color distance to this pixel for each assigned region is less than the proper threshold value. The proper threshold value is determined as the minimum color distance between neighboring regions.

2. The remaining uncertainty pixels are assigned to the label regions with the shortest color distance to each pixel. This process is repeated until there are no remaining uncertainty pixels.

Region growing is followed by region merging and filtering to yield the final segmentation result. We use a minimum size of 50 pixels for a region. Small areas that are less than the defined size are merged into the neighborhood region with the shortest color distance to the small area.

### 3. Temporal region segmentation and tracking

Various methods have been proposed to track objects in a temporal domain (Gambotto, 1992; Gu and Lee, 1998; Zhou et al., 2000; Deng and Manjunath, 2001), most of which require motion information for the segmented regions to achieve successful tracking in the regions of interest. Others use the marker propagation technique to the next frame and region growing (Gambotto, 1992; Deng and Manjunath, 2001).

We assume that the video scene is continuous and there are no abrupt changes within the video data of interest. For a given initial segmentation by the proposed spatial algorithm, the main steps of the proposed tracking scheme are (1) region projection for the region of interest (ROI) and (2) probabilistic pixel classification based on color/intensity information of the neighboring regions. In marker propagation and growing techniques, they require to edge information in advance. For motion-based tracking methods, motion estimation procedure is very time-exhaustive process. However, our method is more efficient since the present scheme does not require these edge information and exhaustive motion estimation.
3.1. Regional projection

For the segmented regions as the initialization, a region \( R_i(t-1) \) is projected onto the current frame \( F(t) \). Let the projected region on the current frame be \( P_i(t) \). Fig. 2(a) illustrates the proposed regional projection of a segmented region of the previous frame.

We introduce a region correlation (RC) to estimate the coherence between the segmented region \( R_i(t-1) \) and projected region \( P_i(t) \). The region correlation is composed of two terms of the area correlation (AC) and color correlation (CC).

Area correlation (AC): The area correlation is generally defined as the ratio of the coherent area between two segmented regions to the area of the segmented region \( R_i(t-1) \). The segmented result of the current frame \( F(t) \) is needed for this computation. The area correlation is estimated as follows to avoid this procedure:

\[
AC_i(t) = \sum_{(x,y) \in P_i(t)} \text{Bin}(x,y)
\]

\( \text{if} \)

\[
\left| \bar{R}_R(t-1) - C_R(x,y) \right| < \kappa \sigma_R(t-1) \}
\]

\[
\left| \bar{R}_G(t-1) - C_G(x,y) \right| < \kappa \sigma_G(t-1) \}
\]

\[
\left| \bar{R}_B(t-1) - C_B(x,y) \right| < \kappa \sigma_B(t-1) \}
\]

where \( \text{Bin}(x,y) \) is a binary function which has a value 0 or 1 according to the given condition, \( \{ \bar{R}_R(t-1), \bar{R}_G(t-1), \bar{R}_B(t-1) \} \) denote mean color components of RGB color space of the segmented region \( R_i(t-1) \). \( C_R(x,y), C_G(x,y) \) and \( C_B(x,y) \) are color components at pixel \( (x,y) \) in the current frame (at time \( t \)). Also, \( \sigma_R(t-1), \sigma_G(t-1) \) and \( \sigma_B(t-1) \) denote variances of RGB color components of the segmented region \( R_i(t-1) \) and \( \kappa \) denotes the predefined weighting factor. Using this equation, a pixel is used to estimate the area correlation if the color similarity is sufficiently reliable.

Color correlation (CC): The intersection histogram is used to measure the color similarity between the regions (Chen and Chen, 2002). The color correlation is defined from the intersection histogram between the segmented region \( R_i(t-1) \) and the projected region \( P_i(t) \) (Chen and Chen, 2002).

With the defined two correlation terms, the RC is defined as

\[
RC_i(t) = \beta AC_i(t) + (1 - \beta) CC_i(t).
\]

Here, \( \beta \) is a weight factor which determines the degree of contribution to the RC.

The region correlation (RC) can provide movement information for the object region. For the estimated RC, we can make sense that the motion of the interested region becomes larger as it is smaller.

Region of interest (ROI): To set the ROI in the current frame \( F(t) \), the motion and bandwidth of the region are first acquired. We estimate the motion of the object region using a simple translational model because exact motion information is not required.
If two-dimensional rotations are assumed to be small enough, the inter-frame translation \( v(t) \) can be computed as the variation of the center of gravities as shown in Fig. 2. The motion of the object region can be given as \( 2v(t) \) as long as the area correlation (AC) exists between the segmented region \( R_i(t-1) \) and the projected region \( P_i(t) \).

The bandwidth \( (B_i(t) \) in Fig. 2) for ROI is introduced to adapt for the change of object shape and motion. The adaptive bandwidth is written by using \( RC_i(t) \) as

\[
B_i(t) = (B_{\text{min}} - B_{\text{max}}) \cdot RC_i(t) + B_{\text{max}},
\]

where \( B_{\text{min}} \) and \( B_{\text{max}} \) are the constants which can be selected based on the motion and size of the tracked region.

The values \( B_{\text{min}} \) and \( B_{\text{max}} \) describe the minimum and maximum bandwidths as the tracked object’s motion which highly depends on the given sequence. For a stationary object in time domain, the \( B_{\text{min}} \) must be zero, ideally. But there is small variation of the object’s shape actually owing to some noise component. So, we choose \( B_{\text{min}} \) as 2 for considering very small variation of object’s shape experimentally.

Similarly, \( B_{\text{max}} \) should be properly selected according to the given sequence and tracked object’s speed in time. This value was set as 10 through experiments in our study.

By using Eq. (7), the ROI that will be processed in the stage of pixel classification is efficiently selected with consideration of the motion and change of the shape of the object region.

### 3.2. Pixel classification

The goal of this stage is to partition pixels of the \( R_i(t) \) into the desired object or into the background. To do this, we use information from the neighboring regions around the desired object. In Fig. 2(b), the given regions including the desired region \( R_i \) and its neighboring regions \( N_j \) \( (j = 1, \ldots, J) \) are shown. For the \( k \)th pixel in the ROI, we define the probability which it belongs to any \( i \)th region as

\[
P_{ki} = \frac{s_{ki}}{\sum_{l=1}^{i} s_{kl} + s_{ki}},
\]

where \( s_{kj} \) denotes a similarity measure based on Euclidean color distance between the color vector of a pixel \( k \) and the region \( j \). Based on the defined probabilities of pixels in the region of interest (ROI), the \( k \)th pixel is classified into a label or region which has the largest probability value. This pixel classification stage can be easily implemented as a minimum distance classifier based on the defined color similarity measure.

Once the object region is updated, the statistics of color component for a newly segmented region is modified for propagation to the next frame.

### 4. Results and discussion

A variety of images were tested to validate the performance of the proposed VO segmentation and tracking algorithm. These images were selected frames of standard MPEG image sequences with the size of QCIF.

For test of the spatial segmentation, several images from standard MPEG sequences were selected to validate the performance of the proposed algorithm. The images selected were from the Claire, Tabletennis, Flower garden, Lena, and Foreman sequences. The conventional marker-based watershed algorithm and probabilistic relaxation method based on multi-resolution analysis (MRA) were used to compare results with the proposed algorithm (Salgado et al., 2000; Gao et al., 2001; Kim and Kim, 2003). In these methods, the gradient image is need to extract the marker areas. Most of them utilize the Canny operator or morphological gradient. In this paper, the Canny operator with the size of \( 5 \times 5 \) and \( \sigma = 2 \) was used for the gradient image of the watershed algorithm (Gao et al., 2001; Kim and Kim, 2003) and probabilistic relaxation method based on MRA (Salgado et al., 2000) with transform level \( m = 1 \).

By thresholding the acquired gradient image, the markers were extracted for region growth. To do this, we controlled threshold values for images. These values were differently selected 30 and 50 as type of the given image. If threshold value becomes larger, then we get less the controlled markers. Otherwise, the number of the controlled markers increases, significantly. The former may
cause a loss of the details for region segmentation. The later guarantees the details of the segmented image, but includes many noisy clutters. The size of the analysis window was set as $4 \times 4$ or $2 \times 2$ in the proposed algorithm. 1-level wavelets transform was implemented in the probabilistic relaxation method based on MRA (Salgado et al., 2000).

Fig. 3 illustrates results for gradient information map of the Claire. For the result of the Canny operator, collars of women’s cloth have weak contrast to the other of that. So, the spatial gradient based on the Canny operator was relatively low value in the boundary between the collar and other of women’s cloth. This is due to the consideration of the gradient response by the spatial intensity difference, simply. But it can be seen that the proposed method gave a stable response for edge information as shown in Fig. 3(c) owing to the local histogram analysis.

We investigated the computational amount and the consumed time for extracting the edge information in the proposed algorithm. As described in advance, most algorithms for vision applications employed the spatial mask such as the Sobel, Prewitt and Canny operators to examine the edge possibility of a pixel. Thus, we compare the part of the local histogram analysis in our algorithm with the other’s spatial mask for extraction of the edge information.

The main procedures for examining the edge information are listed in Table 1. In the proposed method, three steps for the edge information are executed. In these steps, the histogram modality analysis takes most portion of the exhausted time in the overall steps to generate the edge information map. Since this procedure is based on the distribution of intensity levels, we employ the fast scheme based on the multi-resolution analysis for analysis of the local histogram is previously introduced (Kim et al., 2003). In this study, we set $m = 2$ for wavelets transform. Also, $\theta_{\text{var}} = \pm 4$ for refinement. The above three steps are included to

<table>
<thead>
<tr>
<th>Table 1</th>
<th>The main steps for the extraction of the edge information in two approaches (size of $N \times N$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial mask-based method</td>
<td>Local histogram analysis-based method</td>
</tr>
<tr>
<td>1. Two-dimensional convolution (x- and y-axes)</td>
<td>1. Histogram generation</td>
</tr>
<tr>
<td>2. Computation of magnitude</td>
<td>2. Histogram modality analysis (MRA)</td>
</tr>
<tr>
<td>3. Computation of the edge information</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>CPU-time of the conventional convolution-based method (unit: ms, for 10,000 pixels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of mask</td>
<td>Total consumed time</td>
</tr>
<tr>
<td>$3 \times 3$</td>
<td>10</td>
</tr>
<tr>
<td>$5 \times 5$</td>
<td>20</td>
</tr>
<tr>
<td>$7 \times 7$</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>CPU-time of the proposed method (unit: ms, for 10,000 pixels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of window</td>
<td>Histogram processings</td>
</tr>
<tr>
<td></td>
<td>Generation</td>
</tr>
<tr>
<td>$2 \times 2$</td>
<td>6.6</td>
</tr>
<tr>
<td>$4 \times 4$</td>
<td>2.7</td>
</tr>
<tr>
<td>$6 \times 6$</td>
<td>6.75</td>
</tr>
<tr>
<td>$8 \times 8$</td>
<td>6.75</td>
</tr>
</tbody>
</table>
Fig. 4. Segmentation results: the first column—original, the second—the conventional marker-based watershed, the third—the probabilistic relaxation based on multi-resolution analysis and the last—the proposed algorithm.
compute the consumed CPU-time of the proposed approach.

Tables 2 and 3 show results of the consumed CPU-time for generating the edge information map. From these tables, it can be known that the proposed local histogram analysis-based approach is faster than the convolution-based method for extracting the edge information map.

Fig. 4 shows the final results of segmentation. Visually, the watershed algorithm based on the Canny operator and probabilistic relaxation method seem to perform better than the proposed method. This is due to the larger number of segments for the watershed method and probabilistic relaxation method than for our algorithm. However, the proposed algorithm gives a reliable and detailed segmentation result with a smaller number of segments (or classes) for all tested images. Especially, the results of the probabilistic relaxation method based on MRA are very rough and bad in visual.

We employ three measurements to evaluate the objective segmentation performance of the presented algorithm. These are the number of segments \((M)\), goodness \((F)\), and signal-to-noise ratio \((\text{SNR})\). Generally, there exists a trade-off between the number of segments and the details of segmentation. Too many segments can give satisfactory results based upon visual observation while causing an oversegmentation problem in many marker-based segmentation algorithms. Too few segments can result in loss of information from many detailed regions of the original image.

The goodness function \(F\) was defined as follows (Liu and Yang, 1998):

\[
F(I) = \sqrt{M} \times \sum_{i=1}^{M} \frac{e_i^2}{\sqrt{A_i}},
\]

where \(M\) is the number of regions in the segmented image, \(A_i\) is the number of pixels in the \(i\)th region, and \(e_i\) is the sum of the Euclidean distance of the color vectors between the original image and the segmented image in the \(i\)th region. Eq. (9) is composed of two terms: The first term, \(\sqrt{M}\), penalizes segmentation that forms too many regions. The second term, \(\frac{e_i^2}{\sqrt{A_i}}\), is a local measure that penalizes small regions or regions with a large color error since \(e_i\) indicates whether or not a region is

<table>
<thead>
<tr>
<th>Contents</th>
<th>(N)</th>
<th>(F)</th>
<th>(\text{SNR} , (\text{db}))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Claire</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watershed-based</td>
<td>41</td>
<td>37.06</td>
<td>31.79</td>
</tr>
<tr>
<td>Prob. relaxation (MRA)</td>
<td>84</td>
<td>45.00</td>
<td>31.27</td>
</tr>
<tr>
<td><strong>Proposed</strong></td>
<td>36</td>
<td>32.73</td>
<td>31.31</td>
</tr>
<tr>
<td><strong>Foreman</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watershed-based</td>
<td>105</td>
<td>174.13</td>
<td>30.88</td>
</tr>
<tr>
<td>Prob. relaxation (MRA)</td>
<td>158</td>
<td>189.65</td>
<td>29.07</td>
</tr>
<tr>
<td><strong>Proposed</strong></td>
<td>85</td>
<td>140.08</td>
<td>30.48</td>
</tr>
<tr>
<td><strong>Garden</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watershed-based</td>
<td>200</td>
<td>535.09</td>
<td>25.94</td>
</tr>
<tr>
<td>Prob. relaxation (MRA)</td>
<td>152</td>
<td>229.92</td>
<td>26.51</td>
</tr>
<tr>
<td><strong>Proposed</strong></td>
<td>130</td>
<td>34.40</td>
<td>40.38</td>
</tr>
<tr>
<td><strong>Lena</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watershed-based</td>
<td>150</td>
<td>322.52</td>
<td>28.50</td>
</tr>
<tr>
<td>Prob. relaxation (MRA)</td>
<td>204</td>
<td>263.91</td>
<td>26.93</td>
</tr>
<tr>
<td><strong>Proposed</strong></td>
<td>109</td>
<td>210</td>
<td>27.37</td>
</tr>
<tr>
<td><strong>Susi</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watershed-based</td>
<td>69</td>
<td>67.16</td>
<td>33.79</td>
</tr>
<tr>
<td>Prob. relaxation (MRA)</td>
<td>85</td>
<td>48.41</td>
<td>35.34</td>
</tr>
<tr>
<td><strong>Proposed</strong></td>
<td>67</td>
<td>66.16</td>
<td>33.13</td>
</tr>
</tbody>
</table>
Fig. 5. The temporal tracking results: the Coast guard ((a)–(f)) and Foreman ((g)–(l)) sequences.
assigned an appropriate color. We scale down the variable $F$ by a factor $1/1000$. Smaller values of $F$ result in better segmentation.

The averaged SNR ($\text{SNR}$) is widely used in image enhancement, image restoration and image compression (Jain, 1997). Similar to the Goodness function $F$, the segmentation result is compared with the original image. Usually, the mean square SNR of the segmented image to the original image is written as

$$\text{SNR} = \frac{\sum \sum \hat{f}(i,j)^2}{\sum \sum [f(i,j) - \hat{f}(i,j)]^2},$$

(10)

where $\hat{f}(i,j)$ is the segmented image and $f(i,j)$ is the original image.

For applying the image segmentation method with $M$ segments, the averaged SNR can be written as

$$\text{SNR} = \frac{1}{M} \sum_{i=1}^{M} \text{SNR}_i.$$  (11)

As the value of the SNR increases, the segmentation algorithm yields better results for any given number of image regions.

Segmentation results are shown in Table 4. The initial number of markers was chosen as the number of segments in Table 4 for comparison with the watershed method based on the Canny operator (Gao et al., 2001; Kim and Kim, 2003) and the probabilistic histogram relaxation based on multi-resolution analysis (Salgado et al., 2000).

The proposed algorithm based on local histogram analysis can reduce the number of noisy segments significantly compared with two existing methods. Thus, the proposed algorithm is able to remove small noise clusters during marker extraction. All images have lower Goodness values than watershed algorithm and probabilistic histogram relaxation using multi-resolution analysis images using fewer segments.

The SNR value indicates that there is little loss of information from the results of the watershed segmentation and the proposed algorithm while the number of segments is less than for watershed segmentation and probabilistic histogram relaxation based on multi-resolution analysis.

Two standard sequences were used to demonstrate the feasibility of the proposed tracking algorithm. We select as $\beta = 0.7$ and $\kappa = 3$ for estimation of the RC.

In Fig. 5(a)–(f), several regions, including the ship, are tracked in the Coast guard sequence. The tracked regions demonstrate reliable visual segmentation results. The tracking results from the Foreman sequence are displayed in Fig. 5(g)–(l). The color of the man’s hat is similar to the background wall color. However, we can see that the hat is segmented and tracked well.

The proposed scheme in the tracking mode provides fast tracking since edge information and motion estimation are not required.

5. Conclusions

We have developed a novel video segmentation and tracking algorithm based on the newly defined concept of edge information. The introduced edge information is determined from analysis of the local histogram. A discriminant function based on local histogram analysis is introduced to construct an edge information map that describes the existence of two or more distinct brightness regions in a window image. According to the characteristics of the gradient image, a modified region-growing method is applied for region partitioning. A fast and efficient segmentation and tracking algorithm is also devised to pursue the object regions in the temporal domain.

Unlike other methods that use spatial gradient operators, the proposed edge feature can extract the edge region where a change of intensity or color is weak. This can give better results in the stage of region growing. Experimental results are provided to demonstrate the feasibility of the proposed algorithm.

References


