CONSTRAINED VIDEO OBJECT SEGMENTATION BY COLOR MASKS AND MPEG-7 DESCRIPTORS

Fatih Porikli

Mitsubishi Electric Research Labs
Murray Hill, NJ 07974, USA

Yao Wang

Polytechnic University
Brooklyn, NY 11201, USA

ABSTRACT

We present an automatic and computationally conservative boundary extraction method using available priori information in a framework consists of change detection mask, region growing, and trajectory motion. Instead of segmenting a entire video frame, only the regions belong to a target object specified by a set of rules are detected. One example of such rules is skin color features of a human body part. Processing domain is limited to the pixels that satisfy color or geometric rules. These rules are represented as a detection mask and implemented as a look-up table. As a result, significant computational reduction and real-time performance are achieved. The framework utilizes color consistency within a centroid-linkage growing technique to grow initial regions. Region seeds are selected among the pixels in the detection mask. The similarity thresholds are adapted from the MPEG-7 dominant color descriptors. The segmentation results of a frame diffused to the next frame and region statistics such as trajectory, percentage of the changed pixels, etc., are registered to determine the moving regions. A computational load comparison of the constrained region growing and regular region growing shows significant reduction in the complexity.

1. INTRODUCTION

Region-based segmentation that is a partitioning of an image into homogeneous regions has often been regarded as a first step in image and video analysis with applications in scene interpretation, object recognition and compression. The wide spectrum of approaches in the literature range from region growing [1] and split-merge in spatial domain, to histogram thresholding and clustering in color space [2], to physics-based modeling. Homogeneity criterion that defines a region is often established as color similarity, texture coherence, motion consistency, and edge properties. A fundamental concept used in region segmentation is the idea of region growing. In region growing, individual pixels that satisfy some neighborhood constraint are merged if their attributes are similar enough. Similarity of attributes is implemented in terms of a distance function and corresponding thresholds. These attributes can be assigned as color values or texture scores, yet color is more common due to its simplicity.

Existing region growing techniques intend to partition the entire image regardless of the application or the end goal. Whereas, processing the image as a whole may not be necessary for some cases. For instance, in gesture recognition, the accurate boundary is necessary only for the image regions that represent human body parts. In highway surveillance, the emphasis is given to the vehicle regions that obviously be within the side of the road. Such conditionals can be formulated in terms of a color range, i.e. skin colors, and geometrical limits for most applications. It has already been shown that color is a powerful descriptor that has practical use in the extraction of face location [3].

We designed a region growing framework to extract boundary of the target objects in video data by considering any available priori information. Instead of segmenting the entire image, only the parts that are fitting to the limits of the priori information are included in the processing domain. The priori information may indicate skin colors for face tracking, or orange color for tracking a basketball. The set of conditions is devised in terms of a detection mask that implemented as a look-up table to filter input images. We preferred to refer this mask as target color mask (TCM). Although the name does not manifest, any geometrical constrains on the pixel locations also shapes the mask. Application of the TCM generates a small list of pixels, which defines the domain of potential pixels. As a result, considerable reduction in the computational load of the region growing is achieved.

We also propose to utilize available MPEG-7 descriptors of the video sequence to adapt system parameters for computational redundancy [4]. MPEG-7 will be a standardized description of various types of multimedia information. By using the dominant color descriptor [5] to compute region growing thresholds, computation required for adaptation is eliminated.

Another computational reduction is accomplished by using the trajectory motion instead of optical flow or motion estimation. The trajectory motion approximates translational motion, which is adequate for most tracking scenarios.

The flow diagram of the framework is presented in Fig. 1. The automatic segmentation algorithm first decides similarity thresholds using descriptors. The seeds are then selected among the pixels of the current TCM. A region is expanded from a seed point using centroid linkage region growing if the homogeneity criterion is valid. The regions are refined by removing irregularities. Seeds of the next frame is determined from the previous regions and next frame's TCM. The CDM'are computed for the each region. The trajectory motion is computed. Object management controls the region statistics.

2. DOMINANT COLOR DESCRIPTOR

The dominant color descriptor represents color attributes by depicting part or all of an image using a small number of colors. In the standard, the dominant colors are determined by successive divisions of color clusters using the generalized Lloyd algorithm in between and then merging of the color clusters. This algorithm measures the distances of color vectors to the color cluster centers,

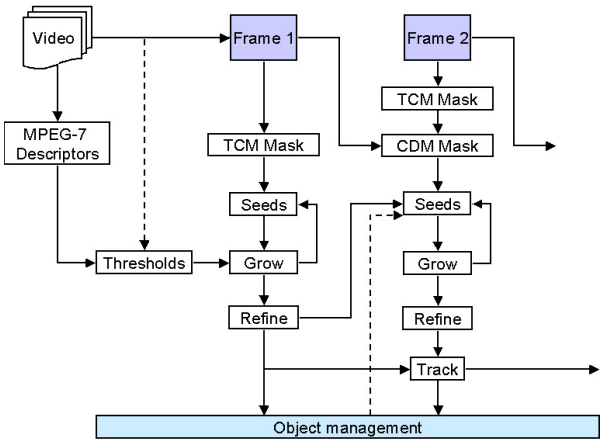

Fig. 1. The flow diagram of the presented algorithm.

and groups the color vectors in the cluster that has the smallest distance. First, all color vectors are assumed to be in the same cluster. For each cluster C_i, a color cluster center c_i is computed by means of averaging. After grouping the color vectors into the closest clusters, a distortion score D_i is computed for each cluster

$$D_i = \sum_p w_p \|x(p) - c_i\|^2 \qquad x(p) \in C_i \tag{1}$$

where c_i is the centroid of cluster C_i, and w_p is the perceptual weight for pixel p. This score is the sum of the distances of the color vectors to the cluster center, and it measures the number of color vectors that changed their clusters at each iteration. The perceptual weights are optional and calculated from the local pixel statistics to account for the fact that human vision perception is more sensitive to changes in smooth regions than in textured regions. Until the difference in the distortion scores becomes negligible the grouping is repeated. Then, each color cluster is divided into two new cluster centers by perturbing the center if the number of total clusters are less than a maximum number. As a final stage, the clusters that have close centers are grouped.

In case the MPEG-7 dominant colors are available, they are directly embedded into the region growing threshold formulation. Otherwise they are extracted from a sub-sampled initial frame. Since the extraction is done once for only a small number of color vectors; the computational load is negligible.

3. TARGET COLOR MASK

Depending on the application, the TCM filters video frames such that the remaining pixels are likely to constitute the object to be detected. It is denoted as $\mathcal{T}$, and formulated as a set of conditions

$$p \in \mathcal{T} \leftarrow \{\text{rule}_1(p); \text{rule}_2(p); ...\} \tag{2}$$

For applications such as gesture recognition, face tracking, etc., the TCM should detect such points that correspond to human skin. Nicely, these points can be identified by the presence of a certain set of chrominance values narrowly distributed in the color space. The limits of this distribution are obtained by fitting surfaces to

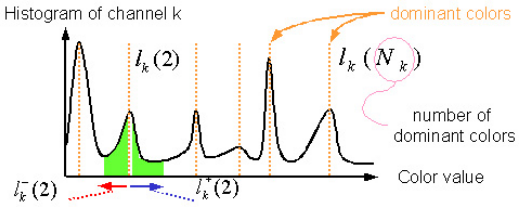

Fig. 2. Smoothed color histogram for a channel.

training data set that contains several human faces. The conditions of human skin color are determined as

$$\begin{array}{lll} \text{rule}_1 & : & |\tan^{-1}(\frac{b}{r}) - \frac{\pi}{4}| < \frac{\pi}{8} \\ \text{rule}_2 & : & |\tan^{-1}(\frac{g}{r}) - \frac{\pi}{6}| < \frac{\pi}{18}, \\ \text{rule}_3 & : & |\tan^{-1}(\frac{b}{g}) - \frac{\pi}{5}| < \frac{\pi}{15}, \\ \text{rule}_4 & : & \sqrt{r^2 + b^2 + g^2} > 0.15 \end{array} \tag{3}$$

where r, g, b are the color components of the pixel p.

By implementing the TCM as a look-up table as the memory restrictions allow, generation of the TCM require only matrix indexing which makes it suitable for real-time performance.

4. SEED ASSIGNMENT AND GROWING

Regions are grown from the seed pixels. By definition, a seed pixel should characterize its local neighborhood as relevant as possible. Such pixels having relatively smooth texture are good candidates to represent their local neighborhood. A color gradient magnitude $|\nabla I(p)|$ is computed for the pixels in the current TCM $\mathcal{T}$, and the seeds are chosen from the minimum gradient magnitude pixels iteratively. Selection and gradient computation are done for sup-sampled TCM for computational simplicity. A pixel that has the minimum gradient is selected as a seed pixel s_i, a region R_i is initiated, and region's centroid r^i is set to the seed pixel's color values. Then the adjoint pixels p are evaluated in 4-pixels neighborhood. Color distance $\Psi(s_i, p)$'s are computed. If the color distance is less than a threshold $\Psi(s_i, p) < \epsilon$, the pixel p is included in the region, assigned as an active pixel, and the centroid colors is updated by the averaged means. After a region R_i is grown, all the pixels of the region R_i is removed from the set $\mathcal{T}$. The seed selection can be formulated as

$$s_i = \arg\min_{\mathcal{T}} |\nabla I(p)| \quad ; \quad \mathcal{T} = S - \bigcup_{j=1}^{i} R_j \tag{4}$$

where $\mathcal{T}$ is the set of all pixels in the TCM initially. The next minimum in the remaining set is chosen, and selection process is iterated until no more pixel remains in $\mathcal{T}$.

The dominant colors are utilized to determine the color distance metric. For each color channel, the dominant colors are ordered and denoted by $l_k(i)$ where $k = r, g, b$ as shown in Fig.2. Let N_r, N_g, and N_b be the number of the dominant colors at the corresponding channels, i.e., $l_k(i) < l_k(i+1)$, $i = 1..N_k$ for color channel k, and initially $N_r = N_g = N_b$. If any consecutive pair $l_k(i)$ and $l_k(i+1)$ are close to each other in a channel they are merged, therefore N_k's are not necessarily same after the merging. For instance, let three dominant colors be $(1.0, 1.0, 1.0)$, $(1.0, 0.5, 0.0)$, and $(1.0, 0.0, 1.0)$. After merging we obtain $l_r : \{1.0\}$, $l_g : \{0.0, 0.5, 1.0\}$, $l_b : \{0.0, 1.0\}$ where $N_r = 1, N_g =$

$3, N_b = 2$. For each $l_k(i)$, two values $l_k^-(i)$ and $l_k^+(i)$ that correspond to the distances from the nearest colors $l_k(i{+}1)$ and $l_k(i{-}1)$ on the both sides are determined such that

$$\begin{aligned} l_k^-(i) &= \frac{1}{2}(l_k(i) - l_k(i-1), \\ l_k^+(i) &= \frac{1}{2}(l_k(i+1) - l_k(i)) \end{aligned} \quad (5)$$

and, the colors between $[l_k^-, l_k^+]$ form a color bin. Using the current seed pixel s, three lengths l_r, l_g, l_b are computed

$$l_k = \begin{cases} l_k^-(i) & l_k(i) - l_k^-(i) < I_k(s) \le l_k(i) \\ l_k^+(i) & l_k(i) < I_k(s) \le l_k(i) + l_k^+(i) \end{cases} \quad (6)$$

where $k = r, g, b$. These lengths represent the distances between the dominant colors at the corresponding color channel, and are utilized to determine a Lorentzian-based distance measure between the centroid r_i and the candidate pixel p

$$\Psi(r_i, p) = \sum_k N_k \log(1 + \frac{|I_k(r_i) - I_k(p)|}{l_k}) \quad (7)$$

where $k = r, g, b$. We scaled the channel differences $|I_k(r_i) - I_k(p)|$ with the corresponding lengths l_k's for normalization. The addition term keeps the logarithmic distance measure positive. The Lorentzian term is sensitive enough towards the small color differences while it prevents the computed distance from inflating for a slightly large color difference in a single channel although the color differences in the other channels are very small. Considering a channel that has more distinctive dominant colors provides more information for segmentation, the channel distances are weighted by the corresponding N_k's. In the implementation, the divider lengths l_k's are integrated with the weight terms for computational simplicity. Then, the distance threshold is set as

$$\epsilon = (N_r + N_g + N_b) \quad (8)$$

which means that the $I_r(p)$, $I_g(p)$, $I_b(p)$ are all within the same color bins with the centroid. Thresholds are calculated for each seed pixels after they are selected. This method is computationally insignificant since it involves a small number of dominant colors. Also, only the thresholds for the dominant colors those fall in the color range of the priori conditions are required.

While growing a region not only the pixels included in the TCM but any neighboring pixel adjoint to the current region boundary is evaluated. In this sense, the seed points are selected from the TCM pixels, but the TCM does not restrict the candidate pixels in the growing stage. Any extra condition on the pixel positions in the growing stage directly increases the computational load.

After regions are extracted for the current frame, the seed points of the next frame are selected from the corresponding TCM. If a seed point in the next frame is adjoint to a region in the current frame then it is marked with the same region index. Such frame-wise diffusion prevents from having disconnected regions parts.

5. CHANGE DETECTION MASK

For the grown regions, the CDM scores are computed to evaluate which regions are moving. Unlike the TCM, the CDM keeps a difference score for the pixels of the grown regions. Since taking pixel-wise frame difference is sensitive to the image noise, we employed a local window matching approach. There is a trade-off between the size of the window and the sensitivity of the CDM; larger the size of the matching window, less sensitive the CDM becomes to the image noise, whereas, more resistant to region movements. We observed that a window size between 3×3 is a good compromise. First, window-wise differences $\delta(p)$ is computed for a point p in the current and points q_n in the following frame $t{+}1$ within a local window $\wp_1$ as

$$\delta(p, q_k) = \sum_{i,j \in \wp_1} \sum_n |I_k(p(i,j,t)) - I_k(q_n(i,j,t{+}1))| \quad (9)$$

where i, j are coordinates in the window, k is the color components. To compensate small motion, points q_n are selected in another window $\wp_2$ around p. The minimum of the computed distances $\delta(p, q_n)$'s is assigned as the difference score $I_c(p)$ of the point p

$$I_{cdm}(p) = \min_n \delta(p, q_n), \qquad q_n \in \wp_2. \quad (10)$$

6. OBJECT MANAGEMENT

While growing the regions, the center of masses are also computed. Object management keeps a record of these centers to compute motion trajectories. Motion trajectory is a feature associated to a moving region, defined as the spatiotemporal localization of one of its representative points such as its centroid. For each region R_i, a trajectory $\mathbf{t}_i(t) = [X_i(t), Y_i(t)]^T$ is extracted by computing the averaged region coordinates

$$\mathbf{t}_i(t) = \begin{bmatrix} X_i(t) \\ Y_i(t) \end{bmatrix} = \begin{bmatrix} \frac{1}{N_t(i)} \sum x \\ \frac{1}{N_t(i)} \sum y \end{bmatrix} ; \quad (x, y) \in R_i^t. \quad (11)$$

Above, R_i^t is the corresponding region at frame t. $N_t(i)$ the area of the R_i^t. A concentration score for the changed points in each region is computed after compensating pixels with respect to the region trajectories

$$R_i^{cdm} = \sum_t \sum_{x,y \in R_i^t} I_{cdm}(x - X_i(t), y - Y_i(t), t). \quad (12)$$

The change detection scores of the regions, R_i^{cdm}, are uniformly normalized to $[0, 1]$ range. It is observed that regions having values higher than $0.7 - 0.8$ generally correspond to moving regions.

7. DISCUSSION

This algorithm is tested with several color sequences with size 352×288. Figure 4 shows samples from test sequences (a,c,e,g), and their corresponding TCM's (b,d,f,h). In the head-and-shoulder sequences skin color features are used to construct the TCM. For the football sequence, a white color feature is used as the TCM (Fig.4-f) to detect the ball, and an aspect ratio constraint for refinement. For the road sequence, road boundaries are utilized as the TCM (Fig.4-h), and the CDM for refinement. For evaluation, the proposed is compared with two algorithms. The first one is a widely accepted region growing algorithm that partitions all of the input frames by a similar centroid-linkage method with the same distance threshold adaptation technique presented in this paper. The second algorithm detects the target colors and then applies morphological opening and closing with a $5x5$ block shaped structuring element. As an output, the first algorithm generated similar object boundaries, although it violated target object boundaries at some frames. However, the first standard region growing

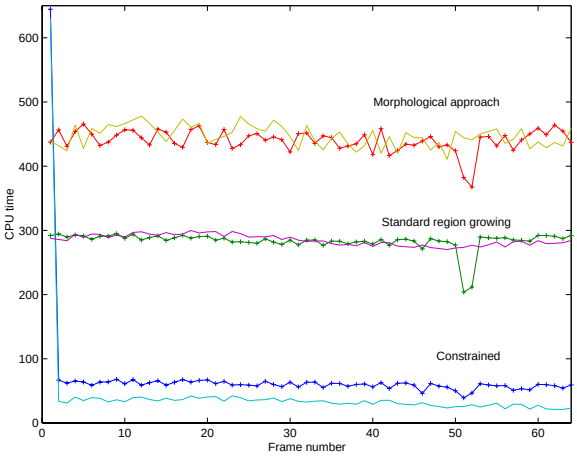

Fig. 3. Frame-wise processing times of *Akiyo* (straight lines) and *Girl* (marked lines) with no CDM. For the constrained method, the dominant colors are extracted only at the first frame in which the processing time is relatively higher. Since all pixels are involved, the CDM computation amplifies the CPU times of the standard and morphological methods much more than the constrained method.

algorithm spent 300ms on average to segment each frame. The second algorithm produced noisy and inaccurate boundaries with spatially disconnected regions. Besides, it consumed 450ms on average for each frame. In Fig. 3, frame-wise processing time results are shown for the proposed method, standard region growing, and morphological approach. On the other hand, although formation of the TCM look-up table required more computational power and memory at the beginning of the segmentation process, the presented method correctly outlined the target object boundaries in most of the frames while achieving the real-time performance by using only 40-60ms of the CPU time.As a result, the proposed method automatically segmented the target objects using available priori information. Significant computational reduction is accomplished, and object boundaries are accurately extracted. A novel threshold adaptation method is utilized. The proposed method requires no background indexing, and it can determine the translational motions without any exhaustive motion computation.

8. REFERENCES

[1] R. Taylor and P. Lewis, Color image segmentation using boundary relaxation, ICPR, Vol.3, 1992, 721–724

[2] Y. Ohta and T. Kanade and T. Sakai, Color information for region segmentation, CGIP, No.13, 1980, 22–241

[3] D. Chai and K. N. Ngan, Face segmentation using skin-color map in videophone applications. IEEE Trans. Circuits and Systems for Video Technology, **9**, no. 4, 1999, 551-561

[4] ISO/IEC JTC1/SC29/WG11 N4031, Coding of Moving Pictures and Audio, Singapore, March 2001

[5] B.S. Manjunath, J.R. Ohm, V. Vasudevan, and A. Yamada, "Color and Texture Descriptors", IEEE Trans. Circuits and Systems for Video Technology, Vol. 11, No. 6, June 2001

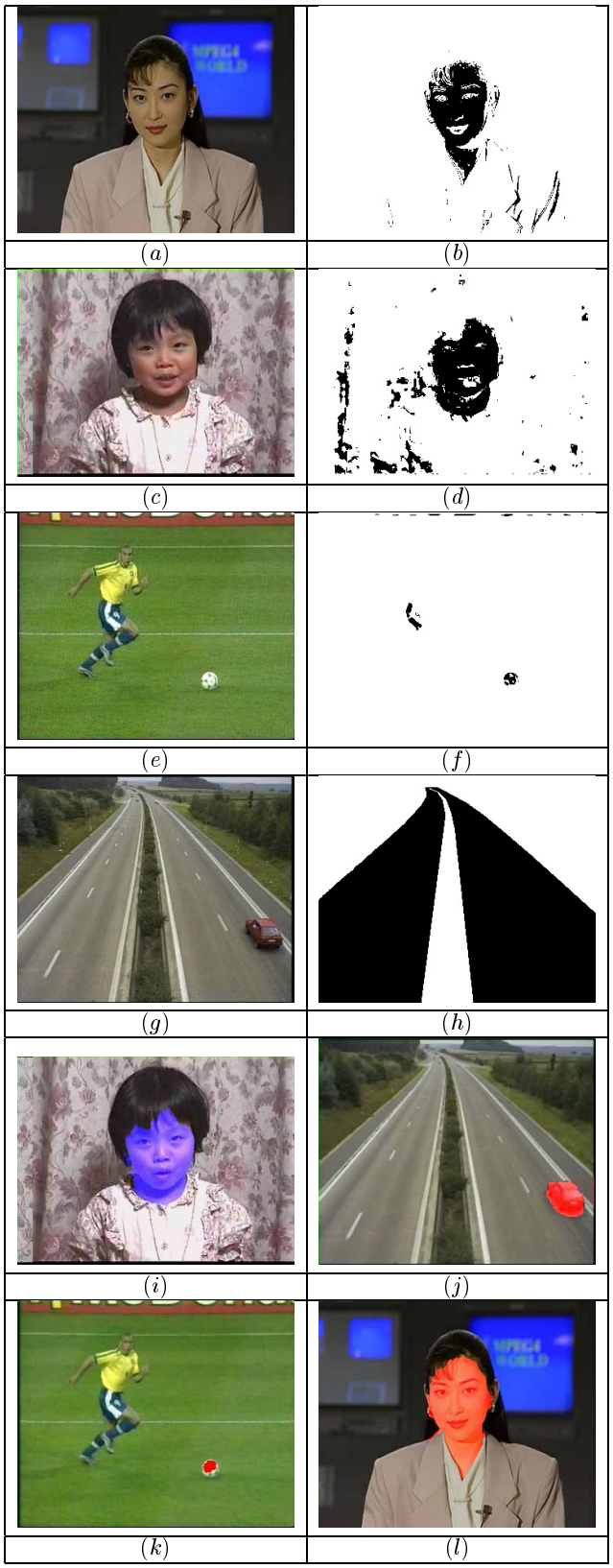

Fig. 4. Original frames (a,c,e,g), the TCM's (b,d,f,h), results of the constrained method using the skin color as the priori information for *Girl*-(i) and *Akiyo*-(l), using white ball color for *Football*-(k), and road boundaries for *Road*-(j).