A VOP Generation Tool: Automatic Segmentation of Moving Objects in Image Sequences Based on Spatio-Temporal Information

Munchurl Kim, Jae Gark Choi, *Member, IEEE*, Daehee Kim, Hyung Lee, Myoung Ho Lee, *Member, IEEE*, Chieteuk Ahn, and Yo-Sung Ho

***Abstract*— The new MPEG-4 video coding standard enables content-based functionalities. In order to support the philosophy of the MPEG-4 visual standard, each frame of video sequences should be represented in terms of video object planes (VOP's). In other words, video objects to be encoded in still pictures or video sequences should be prepared before the encoding process starts. Therefore, it requires a prior decomposition of sequences into VOP's so that each VOP represents a moving object. This paper addresses an image segmentation method for separating moving objects from the background in image sequences.**

The proposed method utilizes the following spatio-temporal information. 1) For localization of moving objects in the image sequence, two consecutive image frames in the temporal direction are examined and a hypothesis testing is performed by comparing two variance estimates from two consecutive difference images, which results in an F-test. 2) Spatial segmentation is performed to divide each image into semantic regions and to find precise object boundaries of the moving objects. The temporal segmentation yields a change detection mask that indicates moving areas (foreground) and nonmoving areas (background), and spatial segmentation produces spatial segmentation masks. A combination of the spatial and temporal segmentation masks produces VOP's faithfully. This paper presents various experimental results.

***Index Terms*— Object segmentation, change detection mask, temporal segmentation, spatial segmentation, and tracking.**

I. Introduction

OBJECT-BASED coding is one of the distinct features of the MPEG-4 standard, which is distinguishable from the previous standards, such as MPEG-1 and MPEG-2. In MPEG-4, the audiovisual contents are decomposed into objects that may represent a fixed background (still image), a talking person (video object), his/her voice (audio object), background music, etc. [1]. Object-based coding allows many useful functionalities: easy access to bitstreams of individual objects, object-based manipulation of bitstreams, various interaction with objects, and reuse of content information by scene composition, which are all suitable for multimedia applications.

In the MPEG-4 visual coding standard, each frame of the video sequence is represented by a set of snapshots in time of these objects, which are referred to as video object planes (VOP's) [1], and different encoding tools, such as shape, texture, and motion encoders, are applied to each VOP. However, MPEG-4 encoding assumes availability of segmented video objects in the form of VOP's.

Hence, video segmentation is an indispensable process for MPEG-4 coding, although only the syntax and semantics of video bitstreams and specification of the decoding process are normative parts of the MPEG-4 visual standard. Segmentation of image usually divides the image contents into semantic regions that can be dealt as separate objects. Here, a semantic object means a moving object through time evolution in image sequences.

During the first phase of the MPEG-4 standardization activity, the *multifunctional ad hoc group* of the MPEG-4 visual part had been working on development of automatic segmentation algorithms for moving objects in video sequences. The framework for video segmentation in MPEG-4 consists of three major components: temporal segmentation, spatial segmentation, and combination of spatio-temporal segmentation results [2]. Temporal segmentation localizes moving parts of objects in the image, and the spatial segmentation divides the image into semantic regions with precise object boundaries according to the given criteria. The combination of the spatial segmentation and the temporal segmentation produces robust segmentation results of moving objects. The spatio-temporal segmentation scheme adopted in the MPEG-4 visual part is depicted in Fig. 1.

Since moving objects in image sequences usually entail intensity changes, we take differences of intensity values between two successive image frames to find moving objects in space through time evolution [2]–[4]. A local window slides over the difference image that contains differences between two consecutive frames. The squared sum of differences or the sum of absolute differences in the local window is compared with a heuristically defined threshold [3], [4]. Aach *et al.* employed a test statistic that is the variance estimate divided by the true variance [7]. This test statistic has a χ^2 distribution, and a threshold can be theoretically determined for a given

Manuscript received October 1998; revised July 1999. This paper was recommended by Guest Editor A. Puri.

M. Kim, H. Lee, M. H. Lee, and C. Ahn are with the Broadcasting Technology Department, Radio and Broadcasting Technology Laboratory, Electronics and Telecommunications Research Institute, Taejon 305-350 Korea.

J. G. Choi is with Kyungil University, Kyungsan, Kyungsangbuk-Do 712-701 Korea.

D. Kim and Y.-S. Ho are with Kwangju Institute of Science and Technology, Kwangju 500-712 Korea.

Publisher Item Identifier S 1051-8215(99)09577-4.

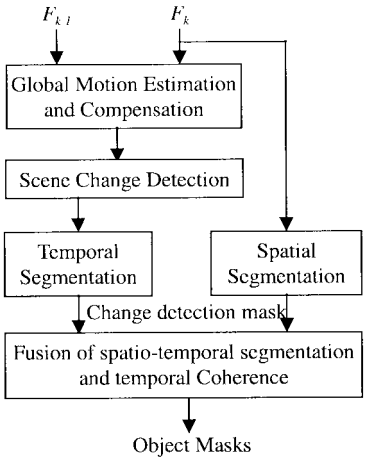


Fig. 1. Spatio-temporal segmentation scheme.

level of significance [7]. However, the true variance should be known *a priori* to compute the test statistic. In practice, the true value of variance is unknown and is dependent on the activity of the image.

In this paper, we propose a temporal segmentation method based on intensity differences, which includes a statistical hypothesis test based on variance comparison. The comparison of two variances leads to a test statistic with an F-distribution. Under a hypothesis, this test statistic does not require the true variance to be known *a priori*. We also introduce the spatial segmentation scheme adopted in the informative part of the MPEG-4 visual standard [2]. The spatial segmentation method is based on the watershed detection that is a region-growing algorithm. However, the watershed algorithm results in oversegmented regions. Therefore, it needs a region merging operation. Last, a combination of the spatial segmentation and the temporal segmentation is proposed, which considers both the spatially segmented regions and the temporally segmented regions so that moving object regions are separated from the background accurately.

This paper is organized as follows. Section II addresses preprocessing steps, such a global motion estimation/compensation and a scene-change detection method. Then, the temporal segmentation method proposed is explained. Spatial segmentation and the combination of spatio-temporal segmentation are addressed in Sections III and IV, respectively. Section V presents experimental results, and we conclude this paper in Section VI.

II. Localizing Moving Objects

A. Preprocessing

1) Global Motion Estimation and Compensation: For the global motion estimation, we utilize the affine six-parameter motion model. We first calculate local motion vectors for nonoverlapping 16×16 square blocks using the block matching algorithm [9]. Then we estimate six parameters using the least square method. In this process, we remove the outlier vectors and iterate the regression process until the estimates of six parameters converge. After the six parameters are obtained, an ON/OFF decision of global motion compensation (GMC) is made because the camera movement does not always exist in the video sequence. We adopt the *coefficient of multiple determination*, which is usually employed for the model adequacy test in statistical analysis. If the value of the coefficient of multiple determination is greater than the given threshold, the GMC is performed. Otherwise, the GMC is not performed.

Let (x, y) be a position in a frame before a camera motion occurs and (x', y') a location after the camera motion. Here, the affine motion model of six parameters is used to estimate the camera motion, and the resulting motion vector is expressed as follows:

$$\hat{v}_X(x, y) = x' - x \equiv a_1 x + a_2 y + a_3 \tag{1}$$

$$\hat{v}_Y(x, y) = y' - y \equiv a_4 x + a_5 y + a_6 \tag{2}$$

or

$$\begin{pmatrix} \hat{v}_X(x, y) \\ \hat{v}_Y(x, y) \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \\ a_4 & a_5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a_3 \\ a_6 \end{pmatrix}. \tag{3}$$

Suppose that the local displacement vectors of all 16×16 blocks are known. These vectors (v_x, v_y) can be obtained by the full search block-matching algorithm. In order to estimate the global motion vectors, an error function to be minimized is defined as

$$E(a) = \sum_{i=1}^{N} \{[v_X(x_i, y_i) - \hat{v}_X(x_i, y_i)]^2 + [v_Y(x_i, y_i) - \hat{v}_Y(x_i, y_i)]^2\} \tag{4}$$

where N is the number of motion vectors in the estimation of the six parameters.

By substituting (3) into (4), we have

$$E(a) = \sum_{i=1}^{N} \{[v_X(x_i, y_i) - (a_1 x + a_2 y + a_3)]^2 + [v_Y(x_i, y_i) - (a_4 x + a_5 y + a_6)]^2\}. \tag{5}$$

The optimal solution to the estimates of the six parameters is obtained by the least square method.

2) Scene-Change Detection: For scene-change detection, we can use two measures. One is the mean absolute difference (MAD) of luminance components between two consecutive images. If the MAD value is greater than a given threshold, it is considered that a scene change has occurred. The other measure is a signed difference of MAD (SDMAD), which is defined as follows:

$$\text{SDMAD}_k = \text{MAD}_k - \text{MAD}_{k-1}$$

where MAD_k is the MAD between the kth frame and the $k+1$st frame and MAD_{k-1} is the MAD between the $k-1$st frame and the kth frame.

The SDMAD is the second derivative of a frame F_k and exhibits large positive and negative peaks at image frames that have scene changes. Due to large positive and negative peaks at the preceding and the following scene-changed frames, the detection performance of scene change is insensitive to the selected threshold value. The SDMAD may fluctuate in small

positive and negative values around zeros when the scene change does not occur. When there is a scene change, we do not perform the temporal segmentation.

B. Temporal Segmentation

In order to develop an image segmentation method that separates moving objects as foreground from stationary background, statistical hypothesis tests are worth using. In general, the motion of moving object entails intensity changes in magnitude so that intensity changes are one of important cues for locating moving object in time and space. These intensity changes can be represented with the difference image between two successive images, both of which embed moving object.

Intensity difference in stationary background between two consecutive frames is from thermal noise of the image capturing device such as charge-coupled device and quantization noise etc., which are often modeled as normal distribution [7], [8]. The model of normal distribution is reasonable because of the central limit theorem. For temporal segmentation, a test for the equality of two variances from each background and a location under consideration of equality is performed. The idea of this approach is that the intensity variation of a moving region is different from that of stationary background because the motion of the moving object changes in intensity of the corresponding region. Therefore, a statistical hypothesis testing is incorporated based on a variance test to detect intensity change.

Let D_k be the frame difference between F_{k-1} and F_k. The difference image is modeled with a normal distribution. A set of random variables in the observation window W is compared with that of the background. Therefore the null hypothesis made on variance is assumed that the two random samples observed inside and outside W were drawn from the same distribution. In other words, the two variances σ_1^2 and σ_2^2 are the same. The hypothesis made on variance is written such that

$$\begin{aligned} H_0&: \sigma_1^2 = \sigma_2^2 \\ H_1&: \sigma_1^2 < \sigma_2^2. \end{aligned} \tag{6}$$

The hypothesis test in (6) is a left-tailed test. The null hypothesis H_0 implies that the true variance σ_1^2 from background and the true variance σ_2^2 from an observation are the same. The alternative hypothesis is that if the observation window W passes over a region of a moving object, the variance estimate $\hat{\sigma}_1^2$ of pixel differences in W should be larger than the variance estimate $\hat{\sigma}_2^2$ of pixel differences in background. This is because the moving object in the region usually entails large changes in intensity, resulting in a large value of the variance estimate for the pixel difference in the region between two consecutive frames.

In order to perform a hypothesis testing on a parameter θ, that is, the variance σ^2, a test statistic needs to include the parameter assumed. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. The random variable

$$(n-1)S^2/\sigma^2 = \sum_{i=1}^{n}(X_i - \bar{X})/\sigma^2$$

has a chi-squared distribution with $n-1$ degree of freedom. Let V_1 and V_2 be independent random variables governed by chi-squire distributions with n_1 and n_2 degrees of freedom, respectively. Furthermore, for simplicity, the pixel differences of the frame difference D_k are assumed to be identically independently distributed random variables. Any F random variable can be written as the ratio of two independent chi-squared random variables, each divided by their respective degrees of freedom. Then the random variable V

$$V = \frac{V_2/n_2}{V_1/n_1} \tag{7}$$

has an F-distribution with n_1 and n_2 degrees of freedom. Here, the random variables S_1^2/σ_1^2 and S_2^2/σ_2^2 have χ^2 distributions with n_2-1 and n_1-1 degrees of freedom, respectively, where S_1^2 and S_2^2 are estimators of variances σ_1^2 and σ_2^2, and n_1 and n_2 are the numbers of samples, respectively. Therefore, a test statistic

$$V = \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \tag{8}$$

will have an F-distribution with n_1-1 and n_2-1 degrees of freedom. Under the null hypothesis, that is, $\sigma_1^2 = \sigma_2^2$, the test statistic V is reduced to the ratio S_2^2/S_1^2. Looking at the ratio performs the hypothesis test established on variance comparison. Note that the true variance is not required to be *a priori* known for computing V if the null hypothesis $H_0: \sigma_1^2 = \sigma_2^2$ is true. When the null hypothesis is true, we can then assert S_1^2 and S_2^2 to be close in value, forcing S_2^2/S_1^2 to be close to 1. If the ratio is significantly larger than 1, then we conclude that the two variances from each population are different. This happens when S_2^2 is estimated in the observation window passing over a region that was affected by moving objects. The intensity difference of pixels in the region with a moving object shows a large variation in intensity so that the resulting variance estimates in the region are usually large.

The test statistic V computed from the window and background is used in computing a change detection mask by testing the hypothesis established in (6). The change detection mask (CDM) is a binary image that indicates foreground (moving object regions) and background (nonmoving regions).

It is desirable to reject H_0 if $V > V_{\text{thr}}$, where the threshold V_{thr} is chosen so that when H_0 is true, $\Pr(V < V_{\text{thr}}) = \alpha$ regardless of μ_1 and μ_2. V_{thr} is the α percentage point of the F-distribution with n_2-1 and n_1-1 degrees of freedom. This leads to the conclusion that the pixel under consideration in the window is declared as being changed in intensity because its variance estimated at the center of the W is different from that of stationary background. If $V < V_{\text{thr}}$, then the null hypothesis is accepted and the alternative hypothesis is rejected. In this case, the pixel intensity at the location under consideration in the current frame is declared unchanged compared to that at the same location in the previous frame. Note that our hypothesis test does not require the information of the true variances σ_1^2 and σ_2^2 *a priori* in computing the test statistic under the null hypothesis. There is a close relationship between the test of a hypothesis about the parameter θ and the confidence interval for θ. If the confidence interval $[0, V_{\text{thr}}]$ is a $100(1-\alpha)$

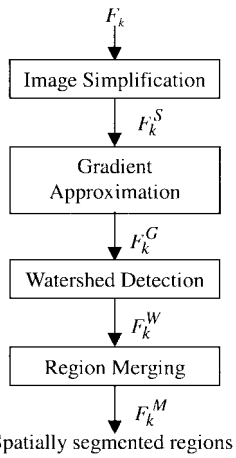


Fig. 2. Spatial segmentation process.

confidence interval for the parameter θ, then the hypothesis (6) will lead to rejection of H_0 if and only if V is not in the interval $[0, V_{\text{thr}}]$.

Following is a summary of the CDM computation procedure.

1) Select a background region and compute the variance estimate S_1^2 of pixel differences inside the region in D_k.
2) Set the center of an observation window of size N at a location (x, y) in D_k.
3) Compute variance estimate S_2^2 of the pixel differences inside the observation windows $W_N(x, y)$.
4) Compute the test statistic

$$V = S_2^2 / S_1^2.$$

5) For a given significance level α, read the corresponding threshold V_{thr} from a table.
6) If $V > V_{\text{thr}}$, the location under consider is declared as foreground. Else it is assigned background.
7) Repeat steps 1) through 5) for all pixel locations in D_k.

III. Spatial Segmentation

Spatial segmentation splits the entire image into homogeneous regions in terms of intensity. The different homogeneous regions are distinguished by their encompassing boundaries that can be obtained from the spatial segmentation. Salembier *et al.* [13] proposed a method for spatial segmentation based on a morphological segmentation that utilizes morphological filters and a watershed algorithm. However, the watershed usually introduces oversegmentation problems, which cause many partitioned regions even in a single object region [16]. A region merging process should be devised to avoid the oversegmentation after watershed in the proposed spatial segmentation. Fig. 2 depicts the proposed spatial segmentation process. The spatial segmentation algorithm consists of the following steps: simplification, gradient approximation, watershed detection, and region merging. The details of each step are explained in the following subsections.

A. Simplification of Image

For spatial segmentation, images are first simplified for ease of spatial segmentation. Many morphological filters are often designed to smooth noisy gray-level images by a determined composition of opening and closing with a given structuring element [12]. Here, erosion and dilation with a flat structuring element are used in constructing a morphological filter for image simplification

$$\varepsilon_n(F_k(x,y)) = \min(F_k(x+l, y+m)) \quad \forall (l,m) \in R_n \tag{9}$$

$$\delta_n(F_k(x,y)) = \max(F_k(x-l, y-m)) \quad \forall (l,m) \in R_n. \tag{10}$$

Morphological opening and closing operations are given by

$$\gamma_n(F_k) = \delta_n(\varepsilon_n(F)) \tag{11}$$

$$\varphi_n(F_k) = \varepsilon_n(\delta_n(F)). \tag{12}$$

It can be noticed that in order to simplify the original image, the morphological opening operation $\gamma_n(F_k)$ removes the bright components that do not fit within the structuring elements and the morphological closing operator $\varphi_n(F_k)$ eliminates the dark components compared to the structuring components. For the case of the images containing fluctuating scale values, the bright and dark components should be considered for the simplification. With the boundary preservation, morphological opening/closing by reconstruction filters are used for the purpose of image simplification [13]. That is, the usage of these filter pairs removes regions that are smaller than a given size but preserves the contours of the remaining objects in the image [13]. The opening/closing by reconstruction filters is given by partial reconstruction.

Morphological opening by reconstruction of opening

$$\gamma^{(\text{rec})}(\gamma_n(F_k), F_k). \tag{13}$$

Morphological closing by reconstruction of closing

$$\varphi^{(\text{rec})}(\varphi_n(F_k), F_k). \tag{14}$$

By removing unwanted details in texture of the regions, the image is homogenized in terms of region textures. So this morphological operation helps spatial segmentation obtain more semantic region partitioning in the later stages. The image simplification is not necessary applied only to the gray-level image. For the regions that show low contrasts, the simplification may results in merging of the two regions that should be separated. Fig. 3(b) shows the simplification results of gray level by *opening/closing by reconstruction* process. The size of structuring elements was 5×5. The simplified image exhibits that detailed textures of hair and clothes of the original image are smoothed out, but the object boundaries are preserved. The size of the structuring elements is dependent on image sizes and types.

B. Gradient Approximation

The spatial gradient of the simplified image is approximated by the use of a morphological gradient operator [12]–[14]. Through the image simplification, the inside of each homogeneous region is small gradient, but large gradients are induced

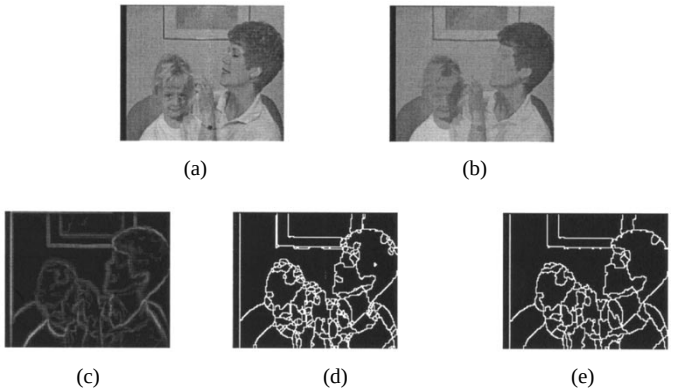


Fig. 3. Spatial segmentation. (a) Original image, (b) simplified image, (c) gradient image, (d) region split after watershed, and (e) region split after merging.

along the region boundaries among different homogeneous regions in the image. So the spatial gradient can be used as an input of watershed algorithm to partition an image into homogeneous intensity regions. The morphological gradient operator used is expressed such as

$$G(f) = \delta_1(f) - \varepsilon_1(f). \tag{15}$$

Note that the morphological gradient is positive. The positive gradients usually indicate borders between neighbor regions and allow the watershed to split an image into regions that are divided by the borders.

In the regions where intensity variations are small, adjacent regions are ambiguous so that watershed algorithm may yields false contours due to the resulting small gradients. To overcome this problem, we incorporate color information into gradient computation in a way that the largest values among the gradients obtained in $w_1 G_{\mathrm{Y}}$, $w_2 G_{\mathrm{Cr}}$, and $w_3 G_{\mathrm{Cb}}$ are chosen where w_1, w_2, and w_3 are weighting factors applied to the gradients of G_{Y}, G_{Cr}, and G_{Cb}, respectively

$$\max\{w_1 G_{\mathrm{Y}}, w_2 G_{\mathrm{Cr}}, w_3 G_{\mathrm{Cb}}\}. \tag{16}$$

In many practical cases, the resulting gradient images exhibits many noisy gradients, which usually cause the watershed algorithm to oversegment the image. In order to prevent this, the gradient images are clipped by a certain threshold value. In other words, gradient values less than the threshold value are set to zero; otherwise they remain same.

The gradient image obtained from the simplified image is shown in Fig. 3(c) and exhibits large values along the region boundaries. Note that in this experiment, small gradients less than ten were set to zero for removing noisy small gradients. The weighting factors (w_1, w_2, and w_3) were set to one, two, and two, respectively, that is, the gradients for Cr and Cb were emphasized twice. This gradient image is used as input image for the watershed detection.

C. Watershed Detection

The image is often interpreted as geographical surface in mathematical morphology, and its gray level is regarded as altitude. As for geographical surfaces, image structure exhibits inclines, hills, and plateaus over the image plane.

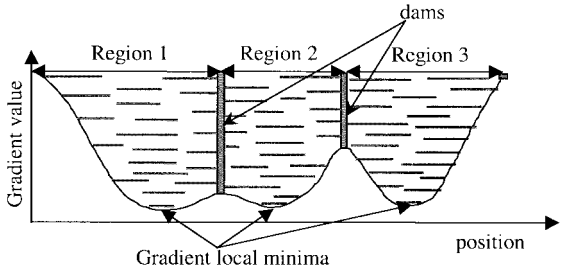


Fig. 4. Watershed flooding.

Usually, a semantic region is distinguished by its inclines from the surrounding contiguous regions in the image. The gradient exhibits large values at the inclines. So a watershed algorithm is applied to the gradient image to partition the image into homogeneous intensity regions. A number of watershed algorithms have been proposed in the literature, but here we use an immersion-based watershed algorithm that is simple and computationally efficient.

The watershed algorithm is a region-growing algorithm, and it assigns pixels in the uncertainty area to the most similar neighbor region with some segmentation criterion such as difference of intensity values. The final result of the algorithm is a tessellation of the input image according to its different catchment basins.

Fig. 4 depicts a one-dimensional graphical illustration of the watershed algorithm in an intuitive way. The local gradient minima are detected and used as seeds for region growing. In the image, these local minima are derived from homogeneous intensity regions. As the tessellation of the gradient image is immersed, each catchment basin is filled up from its lowest altitude selected as a seed. When the surface of the water in each catchment basin reaches the top of the crest, a dam is made for the water not to overflow into the nearby catchment basins. Last, the water surface in each catchment basin reaches the same level but is confined within the region encompassed by the crest lines or dams built on the crest in the gradient image. Then each region is assigned a unique label.

The watershed algorithm is highly sensitive to gradient noise, which yields many catchment basins, resulting in oversegmentation. The resulting image of watershed detection is shown in Fig. 3(d). Since the watershed detection could result in oversegmentation, the region-merging algorithm based on intensity homogeneity should be incorporated to solve the oversegmentation problem.

D. Region Merging

As mentioned in the previous section, the watershed algorithm was applied to partition the images into small regions that are homogeneous in terms of luminance and color. Therefore, the image partition into many homogeneous regions results in oversegmentation. As shown in Fig. 3(d), the watershed detection led the image partition to oversegmentation so that a region-merging step is necessitated to avoid the oversegmentation problem. As the images are simplified for region partition, the oversegmented partition can also be relaxed using a spatio-temporal similarity measure. In the region-

merging step of the algorithm using a graph-based clustering, small regions are recursively merged into their neighbored regions to which the small regions are most similar based on the spatio-temporal similarity. This region clustering yields larger and meaningful regions, which are homogeneous and different from their neighbors. An adjunct scheme of region representation can be developed using a graph theory. This scheme represents both regions and their boundaries explicitly, and makes the storing and indexing of their semantic properties possible. In the following, a spatio-temporal similarity measure is proposed, and the region merging strategy is introduced using the spatio-temporal similarity measure.

1) Spatio-Temporal Similarity Measure: For region merging, it must be determined which regions are merged to their neighbors in terms of certain predefined criteria. A spatio-temporal similarity is measured from a region under consideration and neighbored regions for region merging. The spatio-temporal similarity $\mathrm{SM_{st}}$ consists of a temporal similarity SM_t and a spatial similarity SM_s and is given by

$$\mathrm{SM_{st}} = \alpha \mathrm{SM}_t + \beta \mathrm{SM}_s \tag{17}$$

where α and β are weighting factors. SM_s is the average intensity of a region and measures the similarities against its neighbor regions in the current frame. On the other hand, SM_t indicates the average sum of absolute differences between two successive frames, which measures the similarity of a region under consideration in temporal dimension. For rigid motion, the regions of a moving object show temporal coherence to another so the regions should be represented within a single object region and could be easily distinguished from their neighbored regions that are stationary. Since object motion entails luminance change in intensity, the absolute difference corresponding to the moved regions shows generally larger values. Therefore, the temporal similarity SM_t of a moving region will exhibit larger values compared to its stationary neighbored regions. So it could be a good discriminator distinguishing moving regions from stationary regions.

2) Region Merging The region merging process is depicted in Fig. 5. As mentioned, the region merging is done using a graph theory. As shown in Fig. 5, the small regions that are smaller than n pixels are selected from the spatial segmentation mask after watershed detection. Each small region is then merged to its most similar region in terms of the spatio-temporal similarity measure $\mathrm{SM_{st}}$ that represents how a region is closer to a neighbored region. In other words, we compare the similarity measure $\mathrm{SM_{st}}$ of the region under consideration with its neighbors. Then the region under consideration is merged to the region where the difference between two similarity measures is smallest. This region merging process is repeated until no more small regions smaller than n pixels do exist.

Fig. 3(e) illustrates a spatial segmentation mask after region merging. For the image, α and β were set to 1.8 and 0.1, respectively. The maximum region size *MaxRnSz* was set to 50, which yields a good performance for the region merging. The value n controls the amount of meaningful regions in a segmented image. We start the merging process with a region size *RnSz* $= 10$, and repeat merging iteratively with increasing

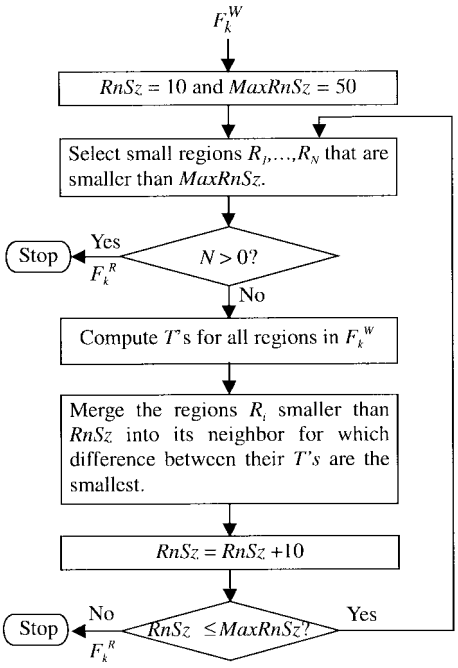


Fig. 5. Region merging process.

RnSz by 10 until the value *RnSz* reaches *MaxRnSz*. In Fig. 3(e), the merged spatial segmentation mask shows larger and more semantic regions. The small regions on top of the heads and around the mother's forehead are shown to be merged into the hair regions and the forehead region.

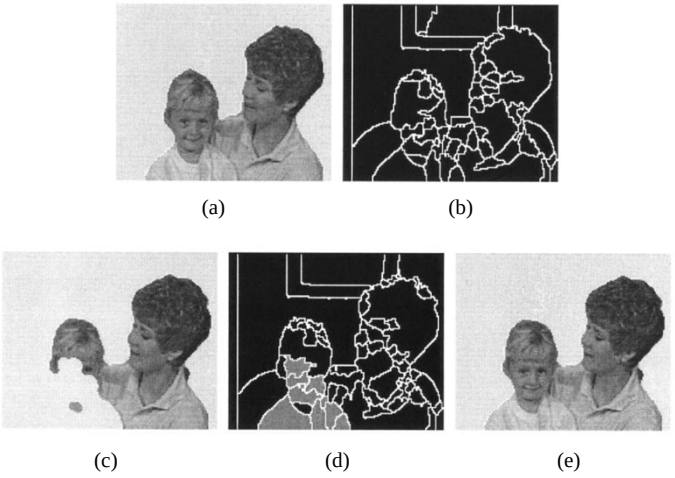
IV. Fusion of Spatio-Temporal Segmentation and Temporal Coherence

In a spatial segmentation mask F_k^M obtained from the spatial segmentation, each region represents coherence in intensity and motion. With the fusion module that combines the spatial segmentation mask and the temporal segmentation mask, foreground regions are discriminated from background regions, so yielding VOP's for foreground and background, respectively. The partition of the image into foreground and background is illustrated in Fig. 6.

The decision of foreground and background (FG/BG) is made in two steps. First, FG/BG decision is made based on combining the change detection masks with the spatial segmentation in the current frame; second, the FG/BG decision process is performed, based on region projection and tracking into the current frame. The details of the FG/BG decision are explained in the following two subsections.

A. Foreground/Background Decision Based on the Change Detection Mask

First, the spatial segmentation mask F_k^M is superposed on top of the change detection mask CDM_i obtained from temporal segmentation between the preceding frame F_{k-1} and the following frame F_k. When the majority part of a spatially segmented region $R_{i,k}$ is covered by the foreground parts of CDM_k, the whole region of $R_{i,k}$ is declared as the foreground; otherwise the whole region of $R_{i,k}$ is declared



(a) (b) (c) (d) (e)

Fig. 6. Region tracking and FG/BG decision. *Mother & Daughter* sequence: (a) final segmentation mask in the thirtieth frame; (b) spatial segmentation mask in the thirtieth frame; (c) FG/BG decision results with the change detection mask only in the thirty-third frame; (d) spatial segmentation mask in the thirty-third frame; and (e) final segmentation mask with region tracking in the thirty-third frame.

as the background. This will be effective when object masks do not cover the whole area of an object consistently through time evolution. The resulting image is a labeled image indicating foreground and background. In this case, the parts of the change detection mask exceeding the region boundary (dashed line) are cut by the region boundary in the spatial segmentation mask. Furthermore, parts of the region unfilled by the foreground of the change detection mask in the spatial segmentation mask are all declared as foreground within the region.

B. FG/BG Decision Based on Region Tracking

Second, historical attributes can also be taken into account in FG/BG decisions. In generic video sequences, a moving object often moves continuously through temporal evolution. That is, a region that is a part of foreground in a frame is likely to be a foreground in the next frame. Also, when an object moves slowly at a certain time instance, the temporal segmentation may fail to localize the object region because the intensity change due to motion is not significant. This insignificant intensity change introduces the missing probability problem of a hypothesis test. In other words, temporal segmentation methods based on a hypothesis test have the missing probability (Type II error) problem inherently. Therefore, a region tracking method is incorporated to overcome the problem of missing probability and to enhance the temporal coherence of the moving object segmentation.

Let a region $R_{i,k-1}$ be a foreground region in the previous frame F_{k-1}. A polygon matching is used to track $R_{i,k-1}$ in the current frame F_k. The matching is done by computing an intensity similarity as a matching criterion. This criterion compares an average of absolute intensity differences in $R_{i,k-1}$ with those in the regions of the current frame. When the best matched region is found at a certain location for $R_{j,k}$ in F_k, $R_{i,k-1}$ is projected on top of $R_{j,k}$ at the location to be centered, which is called $R_{i,k}^{\mathrm{proj}}$.

If the projected region $R_{i,k}^{\mathrm{proj}}$ covers most of the spatially partitioned region $R_{j,k}$ in the current frame, the region $R_{j,k}$ is also declared as foreground. Let $N_{R_{i,k}^{\mathrm{proj}} \cap R_{j,k}}$ be the number of pixels within the union $(R_{i,k-1}^{\mathrm{proj}} \cap R_{j,k})$ of the two regions $R_{i,k-1}^{\mathrm{proj}}$ and $R_{j,k}$. A decision rule is defined as

$$P = \frac{N_{R_{i,k}^{\mathrm{proj}} \cap R_{j,k}}}{N_{R_{j,k}}} \begin{cases} \geq \tau : \mathrm{FG} \\ < \tau : \mathrm{BG} \end{cases} \tag{18}$$

where $N_{R_{j,k}}$ is the number of pixels in $R_{j,k}$. If the value of P is greater than or equal to a given threshold τ, the whole region $R_{j,k}$ is considered as foreground; otherwise background. The FG/BG decision results based on the region tracking method are shown in Figs. 6 and 7. The value τ was set to 0.9. That is, if more than 90% of $R_{j,k}$ is covered by the region union, $R_{j,k}$ is declared as foreground. The region tracking process and FG/BG decision are performed for all foreground regions in F_{k-1}.

Fig. 6(a) indicates the previous final segmentation masks in which the foreground parts are filled with the original images of the thirtieth frame for the *Mother & Daughter* sequence. The corresponding spatial segmentation mask is shown in Fig. 6(b). The FG/BG decision result based on the CDM obtained between frame 30 and 33 is shown in Fig. 6(c). It can be noticed that most of the daughter's chest in frame 33 is not detected as foreground with the CDM. This is because the corresponding regions failed to be detected as foreground in the CDM. Fig. 6(d) shows the spatial segmentation mask in which the regions that are foreground but declared as background are gray-colored. In the region tracking, the regions belonging to the foreground in Fig. 6(b) are projected onto frame 33 and then matched with the regions in Fig. 6(d). Since the background regions in frame 33 are mostly covered by their corresponding foreground regions in frame 30, their regions labeled as background are changed to foreground. The final segmentation mask is obtained as shown in Fig. 6(e). For the *Claire* sequence, as shown in Fig. 7(d), almost all regions in the chest part of *Claire* remain as background after FG/BG decision with the CDM. However, they are declared as foreground by incorporating the region tracking, which is shown in Fig. 7(e).

Notice that when the scene change is detected, the region tracking is not incorporated in the foreground/background decision at the scene change. That is, the foreground/background decision is only made based on the change detection mask.

V. Experimental Results

The spatio-temporal segmentation described has been experimentally investigated by means of computer simulations. Two test sequences, *Mother & Daughter* and *Claire* with the QCIF format (176 × 144 pixel elements), have been used with a reduced frame frequency of 10 Hz.

A. Temporal Segmentation

The temporal segmentation has been performed by testing the hypothesis established on variances between background

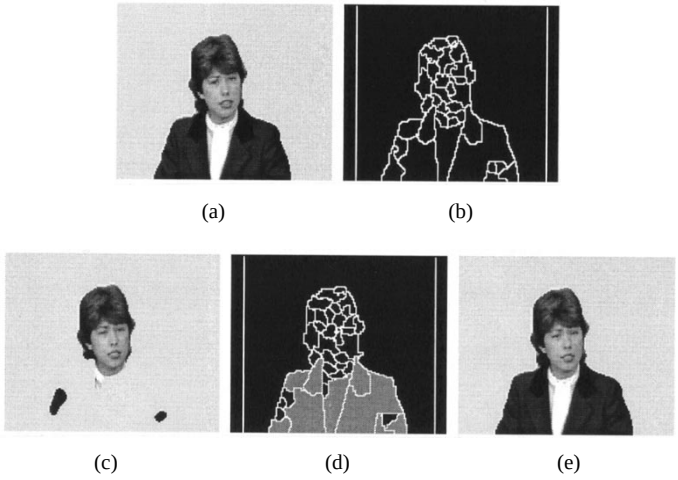


Fig. 7. Region tracking and FG/BG decision. *Claire* sequence: (a) final segmentation mask in the 147th frame; (b) spatial segmentation mask in the 147th frame; (c) FG/BG decision results with the change detection mask only in the 150th frame; (d) spatial segmentation mask in the 150th frame; (e) final segmentation mask with region tracking in the 150th frame.

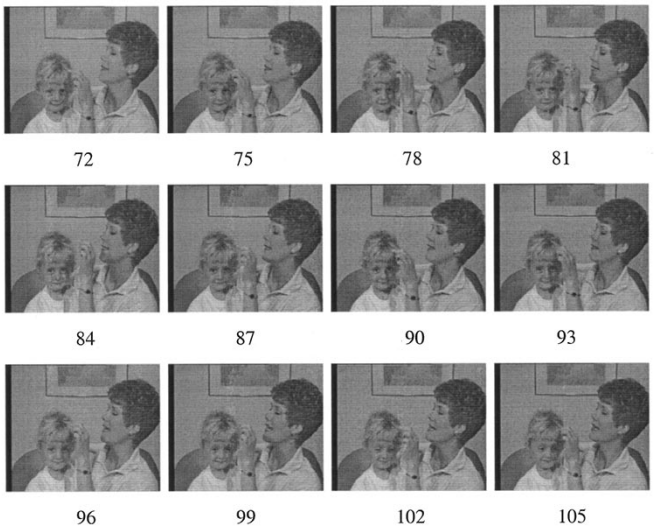


Fig. 8. Original images: *Mother & Daughter* with QCIF.

Fig. 9. Original images: *Claire* with QCIF.

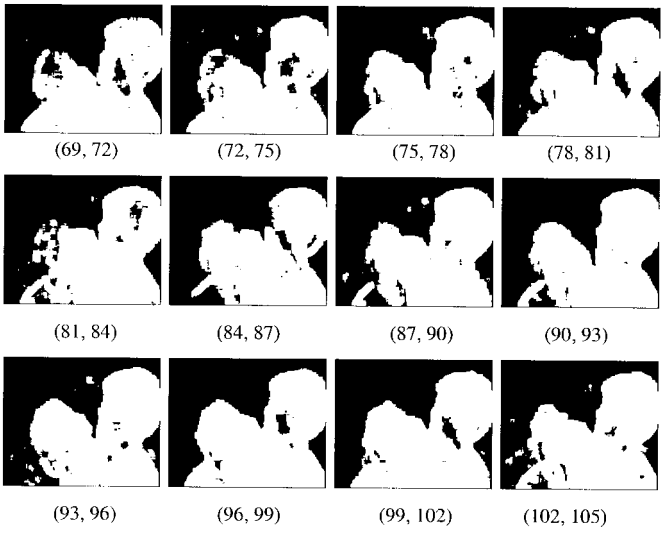


Fig. 10. Change detection masks: *Mother & Daughter* sequence.

and locations under consideration. The test statistic (6) was used to test the hypothesis established on variance comparison.

Figs. 8 and 9 show the original frames of *Mother & Daughter* and *Claire* sequences from frame 72 to frame 105 and from 144 to frame 177, respectively. Figs. 10 and 11 exhibit the corresponding change detection masks obtained between the pairs of two frames (one frame in a three-frame distance with the other) over the frame range. The bright areas indicate foreground, that is, parts of moving objects and the dark areas are background. The level of significance α was set to 1×10^{-3} for threshold selection of testing the hypothesis. It can be noticed that foreground (mother and daughter in Fig. 8 and *Claire* in Fig. 9) and background are well distinguished in all the change detection masks. The characteristics of the *Mother & Daughter* sequence are that the head of the mother has relatively larger motion (mostly small rotation) but the chest part exhibits small motion. This is the same in the *Claire* sequence. The whole chest of the daughter exhibits a little motion through the entire sequence. The hairs of the mother and daughter and Claire have sufficient textures while their closes have relatively simple and homogeneous textures. For the detection of change in intensity, sufficient textures in moving objects of image sequences make it easier for the intensity change detection. However, it is difficult to detect intensity change in the moving objects having simple and homogeneous textures.

In Fig. 10, it is observed that background parts (dark regions) of the change detection masks are not significantly disturbed by salt and pepper noises (bright spots and small regions in the change detection masks). On the other hand, the moving objects (mother and daughter) are well covered by foreground parts (white regions) for the given significance level $\alpha = 1 \times 10^{-3}$. In Fig. 11, due to small motion of the chest part of Claire as in the mother and daughter, a large part of the chest region is declared as background.

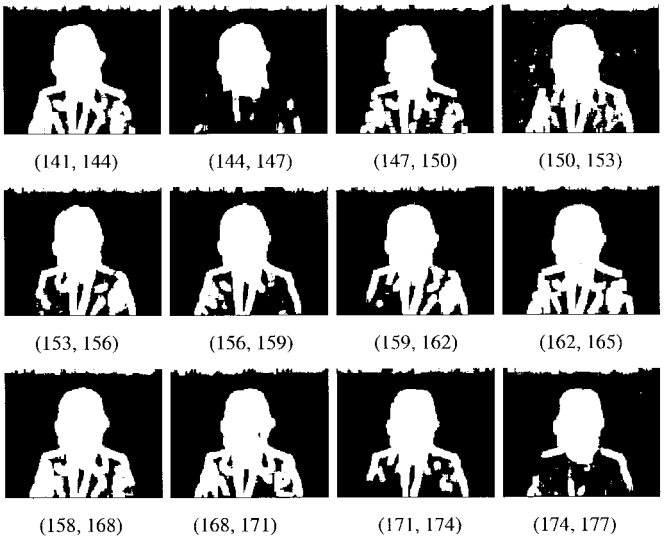


Fig. 11. Change detection masks: *Claire* sequence.

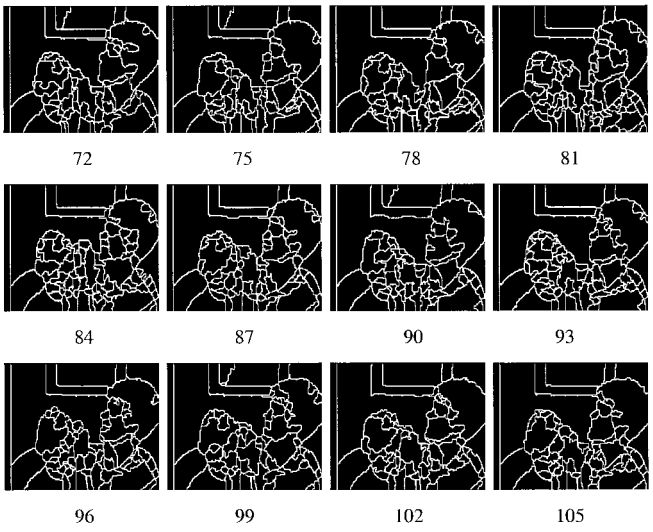


Fig. 12. Spatial segmentation masks: *Mother & Daughter* with QCIF.

Since the test statistic is computed as the ratio of a variance estimate of the pixel values in an observation window to the variance estimate of background, the values of the test statistic are large when the center of the window passes through the insides of moving object regions in the frame differences. For foreground in Fig. 10, Some parts in the *Daughter*'s chest and some parts of the *Daughter*'s head are not declared as foreground due to her small motion. Also, in Fig. 11, large foreground parts of *Claire*'s chest are not detected between frame 144 and 147, between frame 144 and 147, between frame 144 and 147, and between frame 144 and 147. Relatively, the head part showing large motion is well identified as foreground. The small regions in the background were removed by median filtering with a size of 5×5 while the holes inside the foreground were filled up in the change detection masks in a further processing. The test static used for computing the change detection mask shows a good discriminating power between locations having the intensities of pixels changed and locations having pixel intensities unchanged between two consecutive frames.

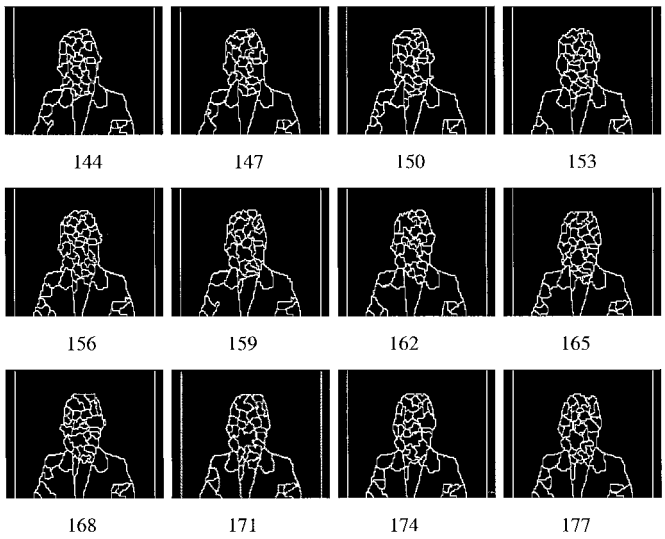


Fig. 13. Spatial segmentation masks: *Claire* sequence.

B. Spatial Segmentation

While the change detection masks were computed from two consecutive frames (the preceding and following frames), the spatial segmentation is performed on the current frame. First, the following images are simplified with an open/close by reconstruction filter with a 5×5 structuring element. The resulting simplified images were used as inputs to the watershed algorithm. After watershed detection, oversegmented regions smaller than 50 pixels were merged to neighbored large regions.

For the *Mother & Daughter* sequence, the spatial image partitions (spatial segmentation masks) are shown in Fig. 12. Since background exhibits a simpler texture structure than those of the objects (mother and daughter) in Fig. 8, spatially segmented regions of background in the merged spatial segmentation masks are shown to be semantically well partitioned after the region merging process. That is, the wall and couch at the daughter's side in background is represented almost as a single region through the frames, and a picture in frame behind the mother and daughter and the frame were also partitioned, respectively. On the other hand, the whole objects are partitioned into many regions due to their complex structures of textures. In addition to the complex structures, the spatially segmented regions of the objects are slightly or significantly different frame by frame. This is because object motion changes the structures of texture, and the luminance of some parts of texture may change due to change in illumination angle. The structural change in regions of the daughter is smaller than in that of the mother because the daughter's motion is relatively small while the mother somewhat moves. The regions of the mother's face and neck show somewhat consistent structures, while some parts of the mother's shoulder are shown to be changed due to some

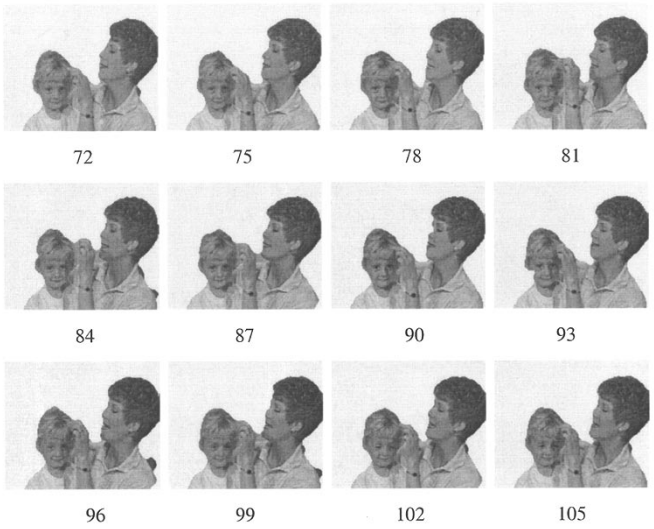


Fig. 14. Final segmentation masks with original images filled in the foreground: *Mother & Daughter* with QCIF.

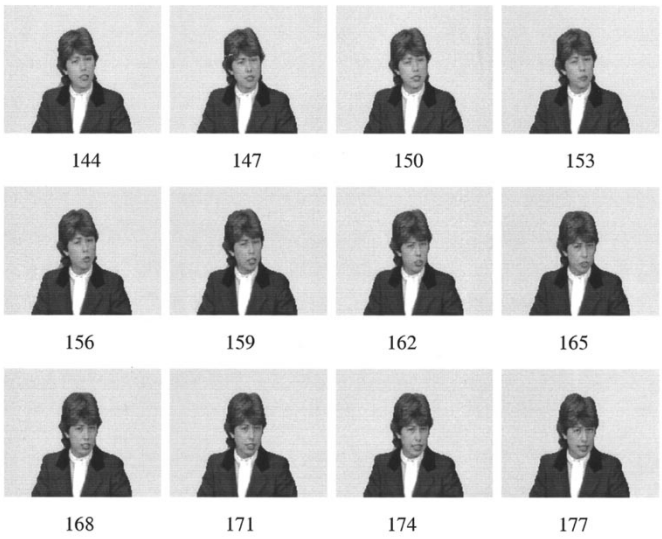


Fig. 15. Final segmentation masks with original images in the foreground: *Claire* with QCIF.

warped parts of the cloth that cause changes in intensity frame by frame.

For the *Claire* sequence, the spatial image partitions (spatial segmentation masks) are shown in Fig. 13. The background and the parts of Claire's jacket are homogeneous so that they are well partitioned into semantically meaningful regions. The boundary shapes of the region partition vary in the parts of the left and right arms in different frames. This is because small chest motion of Claire changes the wrinkle's shape of the arm parts.

C. Fusion of Spatio-Temporal Segmentation

After temporal segmentation and spatial segmentation are performed, both segmentation results are here combined to yield the segmentation masks that first break out images into foreground and background. Next, the region tracking is applied to the background regions. The background regions that correspond to foreground in the previous frame are corrected to be foreground. Fig. 14 shows the final segmentation results for *Mother & Daughter*. It can be noticed that the moving objects are well captured by the change detection masks and the spatial segmentation masks precisely represent the object boundaries. Some unfilled portions of the changed regions in the change detection masks are also fully covered by the region tracking. It is also observed that the unfilled foreground regions (daughter's right shoulder) in the change detection masks from frames 84 to 91 are fully declared as foreground via the region tracking. This improves the segmentation performance of temporal coherence of moving objects, especially when the objects move slowly for a certain period of time frames. Fig. 15 shows the final segmentation results for the *Claire* sequences. In spite of unstable results of temporal segmentation, the final segmentation masks exhibit nearly perfect separation between foreground and background by incorporating the region tracking.

VI. Conclusion and Discussion

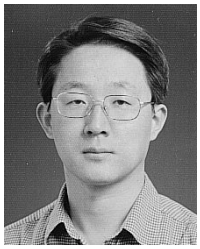
An automatic VOP generation method for the support of an object-based coding standard MPEG-4 has been presented that continuously separates moving objects in image frames through time evolution. The proposed method utilizes temporal information for localizing moving objects and spatial information for the acquisition of precise object boundaries and semantic region partition. For the temporal, a statistical hypothesis testing is proposed that allows for automatic localization of moving object. The temporal segmentation method does not suffer from required information (true mean and/or variance values) *a priori* and predefined threshold values depending on image types. The combination of the spatio-temporal segmentation improves segmentation accuracy and temporal coherence of moving object boundaries.

Unfortunately, image segmentation is recognized as an ill-posed problem and still remains unsolved. In addition, MPEG-4 assumes that VOP's in video are available prior to the encoding process. Therefore, real-time generation of the VOP's is a very attractive problem, and the accurate segmentation of high quality is required for high-level image analysis such as object recognition, image understanding, and scene interpretation and is necessary for authoring multimedia content, etc. When good segmentation performance becomes achievable in generic images and video, it will enrich the current MPEG-4 and MPEG-7 that is related to content-based indexing and retrieval of multimedia data bases.

References

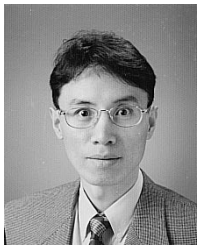
[1] ISO/IEC JTC1/SC29/WG11, "Overview of the MPEG-4 standard," MPEG98/N2323, Dublin, Ireland, July 1998.
[2] ISO/IEC JTC1/SC29/WG11, MPEG-4 Visual Final Committee Draft, Dublin, Ireland, July 1998.
[3] M. Hotter and R. Thoma, "Image segmentation based on object oriented mapping parameter estimation," *Signal Processing*, vol. 15, pp. 315–334, 1988.
[4] H. G. Musmann, M. Hotter, and J. Ostermann, "Object-oriented analysis-synthesis coding of moving images," *Signal Processing: Image Commun.*, vol. 1, pp. 117–138, 1989.

[5] J. G. Choi, M. Kim, M. Ho Lee, and C. Ahn, "Automatic segmentation based on spatio-temporal information," ISO/IEC JTC1/SC29/WG11 MPEG97/m2091, Bristol, U.K., Apr. 1997.
[6] J. G. Choi, S.-W. Lee, and S.-D. Kim, "Spatio-temporal video segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 279–286, Apr. 1997.
[7] T. Aach and A. Kaup, "Statistical model-based change detection in moving video," *Signal Processing*, vol. 31, pp. 165–180, 1993.
[8] Y. Z. Hsu, H. H. Nagel, and G. Rekers, "New likelihood test methods for change detection in image sequences," *Comput. Vision, Graphics, Image Processing*, vol. 26, pp. 73–106, 1984.
[9] M. Biering, "Displacement by hierarchical block matching," in *Proc. SPIE Visual Communications and Image Processing (VCIP'88)*, Cambridge, MA, Nov. 1988, vol. 1001, pp. 942–951.
[10] J. S. Milton and J. C. Arnold, *Introduction.* New York: McGraw-Hill, 1995.
[11] F. Meyer and S. Beucher, "Morphological segmentation," *J. Visual Commun. Image Represent.*, vol. 1, no. 1, pp. 21–46, Sept. 1990.
[12] P. Salembire, "Morphological multiscale segmentation for image coding," *Signal Processing*, vol. 38, pp. 359–386, 1994.
[13] P. Salembier and M. Pardas, "Hierarchical morphological segmentation for image sequence coding," *IEEE Trans. Image Processing*, vol. 3, no. 5, pp. 639–651, 1994.
[14] P. Salembire, L. Torres, F. Meyer, and C. Gu, "Region-based video coding using mathematical morphology," *Proc. IEEE*, vol. 83, pp. 359–386, June 1995.
[15] L. Vincent and P. Soille, "Watershed in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 583–598, June 1991.
[16] L. Najman and M. Schmitt, "Geodesic saliency of watershed contours and hierarchical segmentation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, no. 12, pp. 1163–1210, 1994.



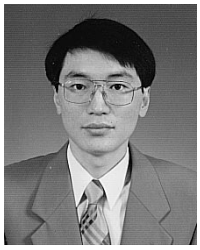
Munchurl Kim received the B.E. degree in electronics from Kyungpook National University, Korea in 1989 and the M.E. and Ph.D. degrees in electrical and computer engineering from the University of Florida, Gainesville, in 1992 and 1996, respectively.

He joined the Electronics and Telecommunications Research Institute (ETRI), Korea, where he worked on MPEG-4 standardization-related research. Since 1998, he has been involved in a new standardization research activity of multimedia content description interfaces (MPEG-7). In the course of MPEG standardization, he has been contributing more than 20 proposals, served as the Team Leader on evaluation of Video Description Scheme proposals in MPEG-7 in Lancaster U.K., in 1999, and chaired the breakout group on the development of generic visual description in MPEG-7 at the MPEG Seoul meeting, March 1999. In 1998, he chaired the Multimedia Information Retrieval III and IV sections at the International Conference on Computational Intelligence and Multimedia Applications, Churchill, Australia, and was a Panelist in the Feature Extraction and Segmentation Panel at the International Workshop on Very Low Bit Rate Video Coding (VLBV'98), Urbana, IL. His current areas of interest are video analysis, visual information description and retrieval, and multimedia interactive services.



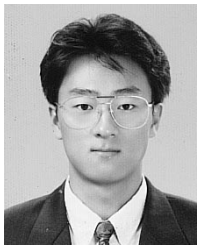
Jae Gark Choi (S'95–M'97) received the B.S. degree in electronic engineering from Kyoungpook National University, Taegu, Korea, in 1984 and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology, Seoul, Korea, in 1987 and 1997, respectively.

He was with the Electronics and Telecommunications Research Institute, Taejon, Korea, from 1987 to 1998, working on problems related to visual communications. Since 1998, he has been a full-time Lecturer in the Department of Control and Instrumentation Engineering, at the Kyungil University, Kyungsan, Korea. His research interests include multimedia signal processing, image coding, motion estimation, and image segmentation.



Daehee Kim received the B.S. degree in control and instrumentation engineering from the University of Seoul, Seoul, Korea, in 1995 and the M.S. degree in information and communications department from Kwanju Institute Science and Technology (K-JIST), Kwangju, Korea, in 1997, where he is currently pursuing the Ph.D. degree.

His research interests include digital image processing, digital video compression, and image segmentation.



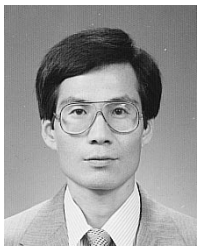
Hyung Lee received the B.S. degree in computer science and the M.S. degree in computer engineering from Chungnam National University, Taejon, Korea, in 1995 and 1997, respectively, where he is currently pursuing the Ph.D. degree in the Computer Engineering Department.

His research interests include parallel processing systems for image processing and object tracking.

Myoung Ho Lee (S'94–M'96) received the B.S. and M.S. degrees in electronics engineering from Soongsil University, Korea, in 1983 and 1985, respectively, and the Ph.D. degree in communication engineering from Osaka University, Osaka, Japan in 1996.

Since 1985, he has been with the Electronics and Telecommunications Research Institute (ETRI), Korea, where he has been involved in developing videophone, MPEG-4 coding technology, and MPEG-4 codec. His current research interests lie in image and video coding, video communication, image processing, and computer graphics.

Dr. Lee is a member of IEICE.



Chieteuk Ahn received the B.E. and M.E. degrees from Seoul National University, Seoul, Korea, in 1980 and 1982, respectively, and the Ph.D. degree from the University of Florida, Gainesville, in 1991.

He is a Principal Member of Technical Staff in the Broadcasting Department of the Electronics and Telecommunications Research Institute (ETRI), Korea. Since joining ETRI in 1982, he has been involved in developing digital switching system, MPEG-2 video encoding chipsets, and SD/HDTV encoders. His recent work has been focused on MPEG-4 standardization as well as developing interactive broadcasting technology. He has been an HOD of MPEG-Korea and SC29-Korea since 1996. His main interests are in areas of visual signal processing and communications.

Yo-Sung Ho received the B.S. and M.S. degrees in electronic engineering from Seoul National University, Seoul, Korea, in 1981 and 1983, respectively, and the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, in 1990.

He joined the Electronics and Telecommunications Research Institute (ETRI), Daejon, Korea, in 1983. From 1990 to 1993, he was with Philips Laboratories, Briarcliff Manor, NY, where he was involved in development of the Advanced Digital High-Definition Television (AD-HDTV) system. In 1993, he rejoined the Technical Staff of ETRI and was involved in development of the Korean DBS digital television and high-definition television systems. Since 1995, he has been with the Kwangju Institute of Science and Technology, where he is currently Associate Professor of Information and Communications Department. His research interests include digital video and audio coding, image restoration, content-based signal representation, three-dimensional model coding, and video indexing and retrieval.