Unsupervised offline video object segmentation using object enhancement and region merging
Ryan, Ken
ProQuest Dissertations and Theses; 2007; ProQuest
pg. n/a

Unsupervised Offline Video Object Segmentation Using Object Enhancement and Region Merging

Ken Ryan

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Applied Science (Electrical and Computer Engineering) at

Concordia University

Montréal, Québec, Canada

November, 2006

© Ken Ryan, 2006

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-28927-3
Our file Notre référence
ISBN: 978-0-494-28927-3

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ABSTRACT

Unsupervised Offline Video Object Segmentation

Using Object Enhancement and Region Merging

Ken Ryan

Content-based representation of video sequences for applications such as MPEG-4 and MPEG-7 coding is an area of growing interest in video processing. One of the key steps to content-based representation is segmenting the video into a meaningful set of objects. Existing methods often accomplish this through the use of color, motion, or edge detection. Other approaches combine several features in an effort to improve on single-feature approaches. Recent work proposes the use of object trajectories to improve the segmentation of objects that have been tracked throughout a video clip.

This thesis proposes an unsupervised video object segmentation method that introduces a number of improvements to existing work in the area. The initial segmentation utilizes object color and motion variance to more accurately classify image pixels to their best fit region. Histogram-based merging is then employed to reduce over-segmentation of the first frame. During object tracking, segmentation quality measures based on object color and motion contrast are taken. These measures are then used to enhance video objects through selective pixel re-classification. After object enhancement, cumulative histogram-based merging, occlusion handling, and island detection are used to help group regions into meaningful objects. Objective and subjective tests were performed on a set of standard video test sequences which demonstrate improved accuracy and greater success in identifying the real objects in a video clip compared to two reference methods. Greater success and improved accuracy in identifying video objects is first demonstrated by subjectively examining selected frames from the test sequences. After this, objective results are obtained through the use of a set of measures that aim at evaluating the accuracy of object

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

boundaries and temporal stability through the use of color, motion and histograms.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Acknowledgments

I would like to express my sincere gratitude to my supervisors Dr. Aishy Amer and Dr. Langis Gagnon for all their support and valuable guidance during the pursuit of this research.

I would also like to thank my colleagues in the vidpro research group; Mohammed, Bin, Firas, Francois, El Helali, Hanif, Chang Su, Kumara, Julius, and Dr. Carlos Vazquez.

Thanks as well to Chris, Dave, Kostas, Ted and all my other colleagues and friends at Concordia.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Contents

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

List of Figures

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

List of Tables

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

List of Symbols

$P_s(\mathbf{X})$	MAP probability of spatial term in segmentation method [1].
$P(x_t^c(x,y)\|R_i)$	MAP probability of color term in segmentation method [1].
$P(x_t^f(x,y)\|R_i)$	MAP probability of motion term in segmentation method [1].
w_t^s	Weight for the spatial term in segmentation method [1].
w_t^c	Weight for the color term in segmentation method [1].
w_t^m	Weight for the motion term in segmentation method [1].
D_{max}	Maximum color and motion distance in maximin algorithm.
γ	Threshold used in maximin algorithm.
$\overline{\mathbf{C}}_{R_i}$	Color center of region R_i used in KMCC algorithm.
$\overline{\mathbf{M}}_{R_i}$	Motion center of region R_i used in KMCC algorithm.
$\overline{\mathbf{S}}_{R_i}$	Spatial center of region R_i used in KMCC algorithm.
$\mathbf{C}(\mathbf{p})$	Color vector value for image point $\mathbf{p}$.
$\mathbf{M}(\mathbf{p})$	Motion vector value for image point $\mathbf{p}$.
A_{R_i}	Area of region R_i in pixels.
$\overline{A}$	Average region area.
λ_1	Regularization parameter1.
λ_2	Regularization parameter2.
X	Image width.
Y	Image height.
μ	Region merging threshold used in KMCC algorithm.
N	Number of regions in current iteration of KMCC.
c_C	Color distance merging threshold for KMCC.
c_M	Motion distance merging threshold for KMCC.
c_S	Spatial distance merging threshold for KMCC.
I_t	Current frame.
I_{t-1}	Previous frame.
I_{INT}	Intermediate segmentation mask.
I_N	New region segmentation mask.
$\mathbf{P}_{R_i}$	Homogeneity measure for region R_i.
u_i	Horizontal component of estimated motion vector.
v_i	Vertical component of estimated motion vector.
E_{ave}	Average motion estimation error in bilinear motion calculation.
E_{all}	Total motion estimation error in bilinear motion calculation.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$\mathbf{U}^t(R_i)$	Region motion parameter vector used for region merging.
$\mathbf{U}(R_i)$	Trajectory matrix used for region merging.
$E^t_{m,n}(R_m)$	Motion compensation error for combined regions.
$E^t_m(R_m)$	Motion compensation error for individual region.
N^t_m	Size of region R_m.
$D^t_U(R_m, R_n)$	Motion difference measure.
$D_U(R_m, R_n)$	Trajectory difference measure.
$D_{U,min}$	Region merging termination threshold.
$\sigma^2_{R_i,C}$	Color variance of region R_i.
$\sigma^2_{R_i,M}$	Motion variance of region R_i.
α	Under segmentation threshold set experimentally to 0.02.
H_{R_i}	Histogram of region R_i.
H_{R_j}	Histogram of region R_j.
b	Histogram bin number.
S_{hist}	Histogram size.
β	Histogram merging threshold set experimentally set to 1.3.
$\mathbf{\Delta D_{R_i,max}}$	Maximum displacement of region R_i.
I	Video clip.
$\overline{A}_{R_i}$	Size of region R_i averaged over I.
$\chi^2(H_{R_i}, H_{R_j})$	histogram distance between regions R_i and R_j.

List of Abbreviations

EM	Expectation Maximization.
MAP	Maximum a Posteriori Probability.
MLE	Maximum likelihood estimate.
MLL	Maximum log likelihood.
KMCC	K-means with connectivity constraint algorithm.
pdf	Probability density function.
RGB	Color space with three components; red, green and blue.
YUV	Color space with an intensity component, and two color components.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 1

Introduction

Due to the increasing volume of video content available on the internet and in video archives, new methods of storing, transmitting and retrieving video are being developed [3–5]. For example, early MPEG-1 and MPEG-2 standards [6] focus on compressing video pixels for efficient transmission. More recently, video object-based compression standards like MPEG-4 overcome these traditional pixel-based approaches. In addition, standards that provide a framework for the description of the video content have emerged (e.g. MPEG-7). These video description standards aim at facilitating video processing applications such as the indexing of audio-visual content for content-based retrieval of video segments. The MPEG-7 standard, for instance, allows a high-level description of a video's content to be stored along with the video data [7–9]. The MPEG-7 standard does not aim at fixing the means of extracting the content of a video, it only provides a framework to describe and compress the video description [10]. Determining the video content and how to best describe it is the responsibility of the content provider. Developing software tools to automatically extract the video content is a important current research topic in video processing for which one of the key steps is segmenting the video into a meaningful set of objects.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1.1 Background and objective

The goal of video object segmentation is to classify the pixels of a video into groups that represent the objects in that video. For example, in a video shot of an intersection, we might classify each moving vehicle into its own group, and everything else into one group representing the background. However, this decision can vary depending on context. A different segmentation method might identify each of the cars tires as a separate object. For this reason, there can be several interpretations of what is a correct segmentation for a particular video sequence. Our objective is to segment video clips into semantically meaningful objects, focusing on the main objects of a sequence. A correct segmentation would consist of such video objects as person, automobile, and the background, as apposed to dividing the video into smaller objects such as, head, hands, tires, headlights, etc. While the human visual system is able to easily identify the objects in most videos, this remains a very challenging task for automated systems. Many approaches to video object segmentation have been proposed. These techniques rely on using several basic cues contained in video clips to determine the best segmentation. Two of the most commonly used cues are color and motion. The most basic pieces of information contained in each frame of a video are the color values of the pixels. In the case of the RGB color space, each pixel has three color values representing the red, green and blue components of the color. Other color spaces have also been developed, with the goal of producing color distributions that better represent the way the human visual system interprets color. For example, the YUV color space has one value, representing the intensity, two values representing the chrominance. Many segmentation methods rely on grouping pixels with similar color into the same object, often using algorithms developed for the segmentation of still images. While color is an important tool for segmentation, it is limited in its applicability. Clearly, real objects are not always homogeneous in

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

color, so segmentation techniques relying on color alone will not always yield satisfactory results. Another cue that can be used to segment video is motion. Motion in a video clip can be expressed as a set of motion vectors, one for each pixel, that describe the displacement of each pixel in the current frame with respect to the last frame. Several methods of calculating this displacement have been proposed. One commonly used method is block-based motion estimation, where the current frame is divided into blocks, and each block is matched with a block in the last frame by minimizing an error function. Once the block matching has been done, each point in the frame is assigned a motion vector that represents the displacement of the block it is contained in. Another way to represent motion in a video clip is through the use of parameterized models. For example, the block motion vectors can be used to estimate the 6 affine, or 8 bilinear parameters that model the camera motion in a video clip. The objective of this thesis is to develop a segmentation method that uses motion, color and other cues that are available in video sequences to meet the following requirements of the segmentation:

1. It should be unsupervised.
2. It should be applicable to a variety of video clips.
3. It should effectively segment videos with or without global motion.
4. It should segment objects which are moving in some frames, but stationary in others.
5. It should segment objects which are non-homogeneous in color or motion at the frame level.
6. It should segment the video into regions that correspond to the main objects being captured in the video.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1.2 Overview of proposed approach

The proposed segmentation method follows the common approach of segmenting the first frame (initial segmentation) and tracking segmented regions through the remainder of the clip. Since there is no prior knowledge of the video content, the first frame segmentation must be adaptable to numerous video characteristics. Since global motion may be present, techniques that require a stationary background, such as, frame differencing, are not good candidates to aid in the first frame segmentation. Techniques based solely on using object motion will not be able to segment objects which are not moving in the first frame. To increase the range of applicability of the algorithm, statistical clustering methods can be used. These techniques, such as the K-means, and other expectation maximization algorithms can be applied to a wide variety of videos, but only segment the image into regions which are homogeneous in color. Our first frame segmentation employs a variant of the K-means algorithm that includes terms for both color and motion. This way, both moving and non-moving objects can be segmented, as well as multi-colored objects. Due to noise, inaccuracies in motion estimation, and other factors, boundaries of segmented objects may not be accurate in all frames of a clip. To improve on the accuracy of object boundaries, a histogram-based object enhancement procedure is used. Since the first frame segmentation relies on using color and motion at the frame level, complex objects which are non-homogeneous with respect to either color or motion will not be segmented correctly. These objects will be over segmented. Since our segmentation method is off line, we are able to use information gathered through the tracking process to iteratively improve the segmentation. In this way, objects that could not be properly identified in the first frame can still be well segmented. To meet the requirements described in Sec. 1.1, our video object segmentation consists of an initial segmentation, followed by a tracking process, a histogram-based enhancement stage,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

and finally a region merging process. A high-level block diagram of the proposed method is shown in Fig. 1.1.

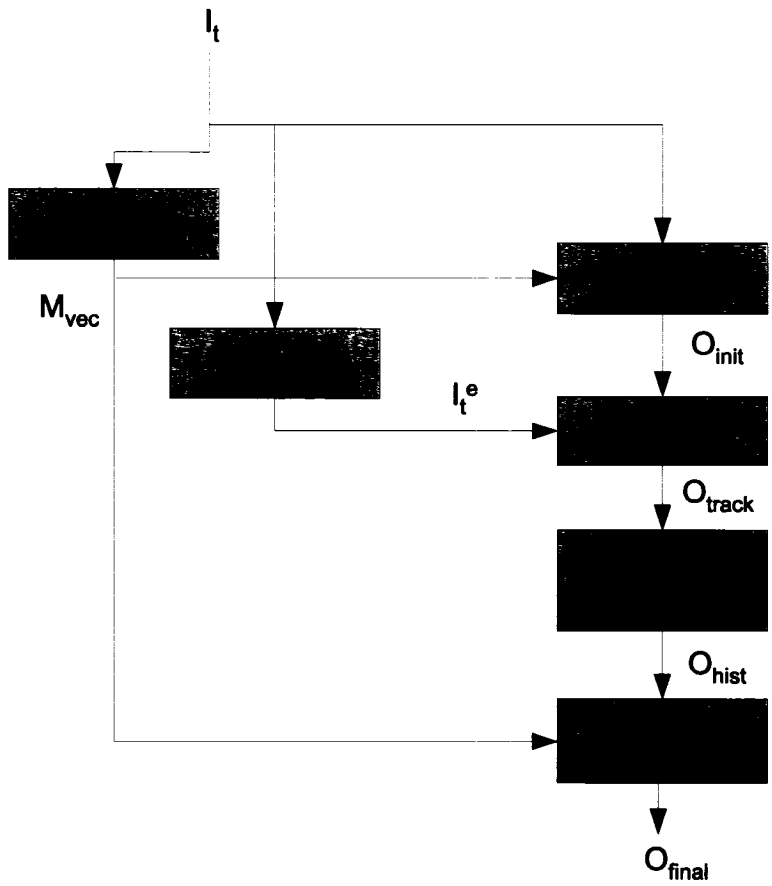


Figure 1.1: Block Diagram of the Proposed Segmentation Algorithm.

1.3 Contributions

This thesis introduces a number of innovations to video object segmentation. The following list describes the contributions of this thesis that are original to the best

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

knowledge of the author. [1]

- An improved initial segmentation where the following improvements have been made:

  1. Incorporating color and motion variance into an existing region clustering scheme.

  2. The addition of a histogram distance and motion variance-based merging stage to reduce over segmentation of the first frame.

- Histogram-based object enhancement, where a set of segmentation measures taken while tracking objects are used to improve the accuracy of object boundaries.

- Merging tracked objects based on cumulative histograms gathered throughout the video clip.

- Trajectory-based merging that has been extended to handle partial occlusion and isolated regions.

1.4 Thesis outline

The remainder of this thesis is organized as follows. Chapter 2 provides a review of related work in video object segmentation and describes two reference methods which have been studied and implemented. Chapter 3 describes the proposed segmentation method. Results are presented in Chapter 4 and Chapter 5 concludes the thesis.

[1] A paper based on this thesis was published in Proceedings of the 2006 IEEE International Conference on Multimedia and Expo [11]. A related journal paper is being prepared to be submitted to the IEEE Transactions on Multimedia.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 2

Literature Review

This chapter provides a review of the existing literature in video object segmentation. A general review of related work is presented in section 2.1. Sections 2.2 and 2.3 provide detailed descriptions of two reference methods that will be used as a basis of comparison for the method proposed in Chapter 3.

2.1 Related Work

The concept of segmenting video into layers was introduced by [12] and [13]. These papers describe how different regions of an image are segmented and stored as layers, which contain information, such as, an intensity map of the region and motion information. These layers correspond to the video object planes used in the MPEG verification model. The entire video clip can be represented by the segmented objects in each layer and the relative motion between layers. The authors of [13] use a robust estimation method to iteratively estimate the number of layers and the pixel assignments to each layer. In [12] an affine motion model is fitted to blocks of optical flow. Then, a K-means [14] approach is used to cluster the image points according to their affine parameters.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

As previously mentioned, many video segmentation approaches involve segmenting the first frame and tracking the segmented objects through the rest of the sequence. Normally, still image segmentation techniques combined with motion information are used to perform the initial pixel assignments.

2.1.1 Color-Based Approaches

Color characteristics are sometimes used to segment video objects. One recent example of this is [15], where color-based deformable models are used to segment and track objects. This method uses color constant gradients, and a model is proposed for estimation of sensor noise through these gradients. As a result, this method is robust when dealing with noisy data. As well, only color, and not intensity is used so that the method can deal with illumination changes. However, this method is only effective when dealing with homogeneous objects and does not handle occlusion.

2.1.2 Motion-Based Approaches

Motion-based approaches to video object segmentation are commonly employed as they often provide improved results on video clips for which color-based methods encounter problems. The authors of [16] present a number of region-based affine-parameter clustering methods using motion vector and intensity matching to align motion boundaries with real object boundaries. They then go on to use a specific combination of these methods to segment a number of video clips.

A different motion-based approach to segmentation is presented in [17]. Here, the motion estimation error along occluding boundaries of moving objects is studied. The authors show how the nature of this error can be used as a depth cue. Their segmentation approach involves segmenting the image based on color and motion independently. Then, by examining the motion estimation error at region boundaries,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

they are able to determine what are the occluding and occluded objects. In this way they are able to establish the relative depth of the image segments.

The focus of [18] is on extracting objects with similar motion. The two step process consists of generating 3D watershed volumes followed by a Bayesian merging of these volumes. In the first frame, markers are extracted which provide reliable seed regions for segmentation in subsequent frames. One weakness of this method is that it is assumed that the number of video objects is previously known.

In [19], deformable binary object models are used to segment and track objects. The models are updated from frame to frame and are therefore able to accommodate complex object motion as well as changes in shape. The models are updated using a modified watershed-based method. Like other methods, there is an initial detection/segmentation step followed by a tracking step. This method can handle moving backgrounds and partial occlusion. However, since the segmentation is based on motion and is done on the first frame, only objects that are present and moving in the first frame are detected. Newly appearing or stationary objects cannot be detected.

2.1.3 Edge-Based Approaches

There has also been significant research into using edge detection to segment video. The extracted edges are used to determine the boundaries of the segmented objects. One major difficulty with this approach is deciding which edges represent object boundaries and which are the result of other image properties, such as textured surfaces. The authors of [20] try to deal with this problem by using a multi-resolution approach to edge detection. A method of determining the optimal scale at each edge by examining edge strengths is presented. The edges at these optimal scales are then used to segment the image.

Boundary completion techniques are used in [21] to complete contours that are

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

smooth, but have low contrast. However, this method is vulnerable to problems when dealing with textures.

In [22], textures are dealt with explicitly by modeling them with textons. By combining texture cues with intervening boundary cues, this approach is able to deal with both textured and non-textured areas.

A different approach to improving edge-based segmentation is taken by [23]. This algorithm uses information from edges at multiple scales. Instead of trying to select the optimum scale for each edge, and then segmenting the image on the selected edges, this approach collects edge information at multiple scales and then does a simultaneous segmentation over all the edges. This method can capture both large and small scale image properties as well as deal with textured areas.

2.1.4 Segmentation Using Multiple Features

Much of the recent work focuses on using multiple video features to aid in segmentation. The authors of [24] use color and motion to segment objects in the first frame, which are then tracked by using their estimated motion to predict their location in the next frame. This method can also segment new objects that appear after the first frame.

In [1], a maximum a posteriori (MAP) framework is proposed. They assign weights to color and motion terms, which are adjusted at every pixel. They also model the spatial probability density function (pdf) of each region in order to impose temporal consistency.

A slightly different approach is employed by [2]. Instead of segmenting based on motion at the frame level only, regions which have been divided based on color, motion and position are tracked. The long-term trajectories of these regions are used to group them into an appropriate segmentation. Segmentation algorithms such as

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

this one, which perform multiple passes through a video clip are referred to as offline methods. Methods which only require knowledge of the current and previous frames being segmented are referred to as online methods.

2.2 Segmentation Using Multiple Features With Adaptive Weighting

The section introduces a segmentation method entitled "Object based segmentation of video using color, motion and spatial information", [1]. This method proposes a Maximum A Posteriori (MAP) framework to combine motion and color to segment the first frame, and uses the spatial pdf of the formed regions to track them through the remainder of the clip. This is an online segmentation method since only one pass is made through the video clip. However, it is not completely unsupervised, since the number of objects must be known prior to performing the segmentation. The algorithm consists of the following steps (Fig. 2.1):

1. Initial segmentation using expectation maximization (EM) algorithm.

2. Feature weight calculation and pdf modeling.

3. Region tracking using spatial pdfs and maximum likelihood estimate (MLE).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

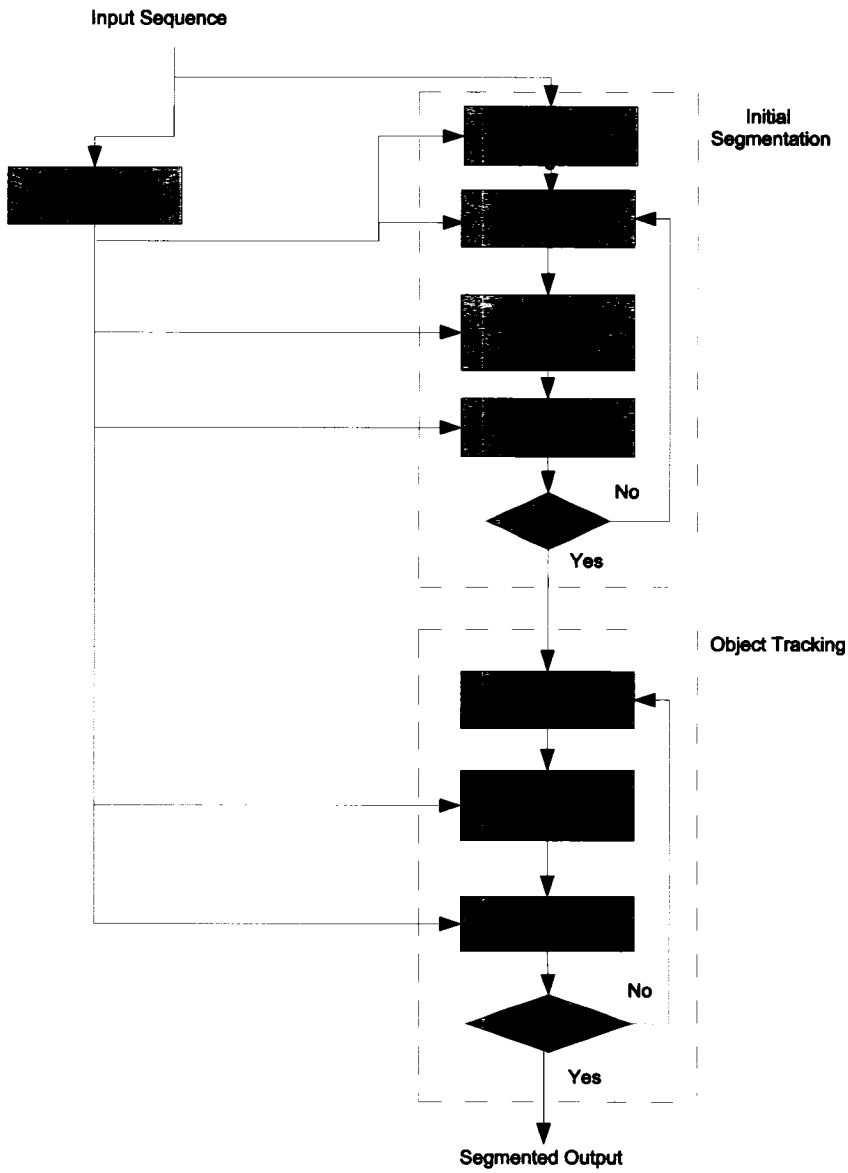


Figure 2.1: Block Diagram of Segmentation Algorithm Proposed in [1].

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2.2.1 Initial Segmentation

The initial segmentation uses an EM algorithm to break the image into seed regions roughly corresponding to the objects in the frame. First, motion vectors are estimated using the Lucas-Kanade [25] method, so that a pair of motion vectors are generated for each pixel. These motion vectors as well as the Y, U, and V color components of each pixel are fed into the EM algorithm which uses an iterative process to estimate a Gaussian mixture model for the data. This Gaussian model is then used to re-classify each pixel in the image to its region of maximum probability.

2.2.2 Feature Weight Calculation and PDF Modeling

The seed regions generated in the initial segmentation are used to calculate statistical models for each region in the first frame. The color and motion distributions are modeled by a 5-dimensional mixture of Gaussians over the [u,v,Y,U,V] space which represents the color and motion components for each pixel. The terms u and v represent the horizontal and vertical motion vectors. The image is represented in the YUV color space, where Y is the luminance component, and U,V are the chrominance components. Spatial pdfs are calculated by modeling each region as a sum of Gaussian distributions, with a distribution centered at each point in the region. This is accomplished by convolving the binary segmentation map of each region with a Gaussian kernel. Once the pdf of each region has been calculated, a maximum log likelihood (MLL) estimate is used to re-classify each pixel in the image. The MLL estimate calculates the probability of a pixel belonging to each region, based on a weighted sum of the MAP probabilities for each feature (Eq. 2.3).

The MLL estimate is derived from the MAP estimate (Eq. 2.1) as

$$P_t(x,y) = argmax_i\{P_s(\mathbf{X}) \times P(x_t^c(x,y)|R_i) \times P(x_t^m(x,y)|R_i)\}, \qquad (2.1)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

where $P_s(\mathbf{X})$, $P(x_t^c(x,y)|R_i)$, and $P(x_t^f(x,y)|R_i)$ are the MAP probabilities of the spatial, color and motion terms, respectively. By multiplying both sides of this equation by a log function, it can be re-written as a log likelihood

$$L_t(x,y) = argmax_i\{ln(P_s(\mathbf{X})) + ln(P(x_t^c(x,y)|R_i)) + ln(P(x_t^m(x,y)|R_i))\}. \tag{2.2}$$

Calculating the probability in this way allows the addition of weighting terms to each feature in the likelihood equation,

$$L_t(x,y) = argmax_i\{w_t^s L_t^s(x,y,i) + w_t^c L_t^c(x,y,i) + w_t^m L_t^m(x,y,i)\}, \tag{2.3}$$

where w_t^s, w_t^c, and w_t^m are the weights for the spatial, color and motion terms. The addition of feature weights allows the pixel re-classification to assign a heavier weight to features that are more reliable. The spatial weight for each pixel is fixed at 1, while the color and motion weights are adjusted at every pixel according to Eq. 2.4.

$$w_m = \rho_1 \rho_2 \tag{2.4}$$

$$w_m = 1 - \rho_1 \rho_2, \tag{2.5}$$

where ρ_1, and ρ_2 are variables that represent how well the motion of the pixel under examination matches with the highest MLL distribution, as well as the level of differentiation with the second highest MLL distribution. Both terms are regularized using modified sigmoid functions as shown in Eqs. 2.6 to 2.7.

$$\rho_1 = \{1 + exp(-a(max_i(L_t^m(x,y,i))))\}^{-1}, \tag{2.6}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

where

$$\rho_2 = \{1 + exp(-a(d-t))\}^{-1}, \tag{2.7}$$

$$d = |max(L_t^m(x,y)) - max2(\acute{L}_t^m(x,y))|. \tag{2.8}$$

2.2.3 Region Tracking

Regions are tracked using the region statistics from the previous frame to calculate the maximum likelihood estimate for each pixel in the current frame. The color and motion weights are re-calculated for each frame, and the spatial pdf term in the likelihood equation enforces temporal consistency. After pixels are re-assigned, region statistics are re-calculated and the process is repeated for every frame in the video sequence.

2.3 Segmentation Using Multiple Features With Trajectory-Based Region Merging

This section describes an unsupervised offline segmentation method aimed at providing segmentations suitable for content-based video applications. The method is entitled "Video object segmentation using Bayes-based temporal tracking and trajectory-based region merging", [2]. The algorithm is composed of three main steps (Fig. 2.2):

1. Initial segmentation using K-means with connectivity constraint (KMCC) algorithm over color, motion and spatial information.

2. Tracking algorithm using a Bayes classifier, and rule-based processing to re-assign changed pixels to existing regions and detect newly appearing objects.

3. Trajectory-based region merging procedure, to group objects based on long term motion.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

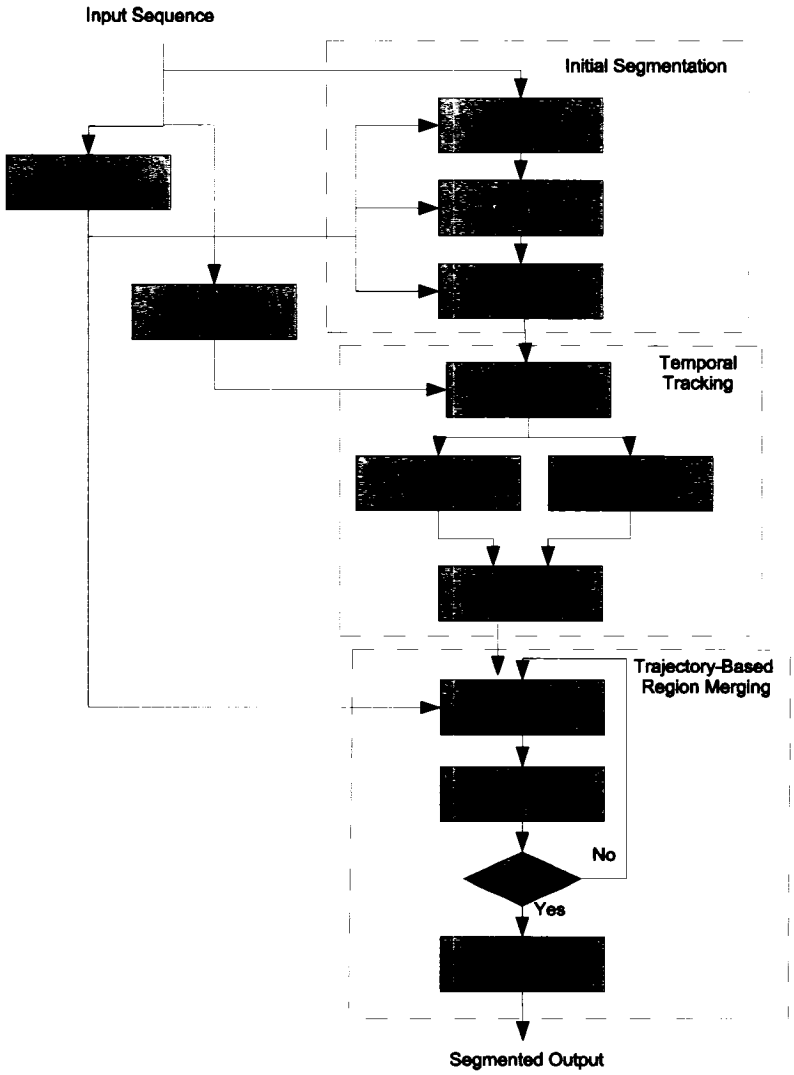


Figure 2.2: Block Diagram of Segmentation Algorithm Proposed in [2].

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2.3.1 Initial Segmentation

To segment the first frame, initial region centers are calculated, which are then used as input to the KMCC algorithm. After the KMCC algorithm converges, a first frame enhancement is applied to the segmented frame. These three steps are described in more detail below.

Initial Centers

To estimate initial object centers, color and motion features for each pixel are extracted. The CIE L*a*b* color space is chosen for its perceptual uniformity, and motion vectors are calculated using a full search block matching algorithm. The image is then broken down into blocks and a color feature vector and motion feature vector are assigned to each block [26]. Then, the initial number of centers is estimated using a variant of the maximin algorithm [27] consisting of the following steps:

1. Find the block with the maximum color and motion distance from first block D_{max}.

2. Examine every other block in the image, and accept any block that is at least $\gamma * D_{max}$ from every other block as a valid region. Where γ is set to 0.4.

Once the initial number of centers has been estimated, a K-means algorithm is applied to the blocks. The K-means algorithm clusters the blocks into K homogeneous regions, where K is the number of initial centers previously estimated.

Since the regions generated by the K-means algorithm are not necessarily continuous, (i.e., there can be many disconnected parts classified to the same region), a 4-connectivity component labeling procedure is used to divide the image into connected regions. The color, motion, and spatial centers of these connected components are then calculated and used as input to the KMCC algorithm.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

KMCC

The KMCC algorithm consists of the following steps:

1. Each pixel in the image is classified into one of the previously defined connected regions based on the distance function in Eq. 2.9.

$$D_{KMCC} = \left\|\mathbf{C}(p) - \overline{\mathbf{C}}_{R_i}\right\| + \lambda_1 \left\|\mathbf{M}(p) - \overline{\mathbf{M}}_{R_i}\right\| + \lambda_2 \frac{A_{R_i}}{\overline{A}} \left\|\mathbf{p} - \overline{\mathbf{S}}_{R_i}\right\|, \tag{2.9}$$

where $\overline{\mathbf{C}}_{R_i}$, $\overline{\mathbf{M}}_{R_i}$, and $\overline{\mathbf{S}}_{R_i}$ are the color, motion and spatial centers of region R_i, respectively. $\mathbf{C}(\mathbf{p})$ and $\mathbf{M}(\mathbf{p})$ are the color and motion vector values for image point $\mathbf{p}$. A_{R_i} is the area of region R_i in pixels, and $\overline{A}$ is the average region area. λ_1 and λ_2 are regularization parameters defined as

$$\lambda_1 = 2 \cdot \frac{D_{max}}{\sqrt{(2u_{max})^2 + (2v_{max})^2}}, \tag{2.10}$$

$$\lambda_2 = 0.1 \cdot \frac{D_{max}}{\sqrt{X^2 + Y^2}}, \tag{2.11}$$

where u_{max} and v_{max} are the maximum allowed block displacements used in the block-based motion estimation, and X and Y are the image width and height.

2. The formed regions are broken down into their minimum number of connected components, and the color, motion, and spatial centers are re-calculated.

3. Regions considered too small to be meaningful (their area is less than 2% of image size) are dropped.

4. The connected components are examined to determine neighbour relationships of the regions, and neighbouring regions whose motion and color centers are

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

below a threshold, μ, are merged, where, μ, is defined as:

$$\begin{aligned} &\mu = 7.5 \quad if \quad C < 25, \\ &\mu = 15 \quad if \quad C > 75, \\ &\mu = 10 \quad otherwise, \end{aligned} \tag{2.12}$$

where C is a measure of the image contrast [2].

5. The number of regions and the region centers are re-calculated. If the remaining regions meet Eq. 2.13, the algorithm terminates.

$$\begin{aligned} &N = N_{old}, \\ &\overline{C}_{R_i} - \overline{C}_{R_i,old} < c_C, \\ &\overline{\mathbf{M}}_{R_i} - \overline{\mathbf{M}}_{R_i,old} < c_M, \\ &\overline{\mathbf{S}}_{R_i} - \overline{\mathbf{S}}_{R_i,old} < c_S, \end{aligned} \tag{2.13}$$

where N is the number of regions, and c_C, c_M, and c_S are color, motion and spatial distance thresholds.

6. If the algorithm has not converged by 20 iterations, then stop, otherwise return to step 1.

First frame enhancement

After convergence, the first frame segmentation is enhanced using a histogram-based Bayesian process. Pixels close to the edge of each region are marked as disputed, and a color histogram for each region is calculated using only the non-disputed points in that region. Then, the disputed pixels are re-classified into the region of highest probability based on the histogram value for that point in each neighbouring region.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2.3.2 Temporal Tracking

After the initial segmentation, a temporal tracking module is used to track the segmented regions through the remainder of the clip. The temporal tracking begins with a frame difference and thresholding, which marks regions as disputed or non-disputed. This is followed by tracking the existing regions. Next, new regions are detected and classified into their own segmentation mask. Finally, the segmentation mask for the new regions is merged with the mask for the existing regions.

Frame Difference and Thresholding

The first step in the tracking process is to determine which image points are unchanged from the previous frame, and which ones will need to be re-classified. To do this, a three by three Gaussian smoothing filter is applied to the current frame I_t, and the previous frame I_{t-1}, and an image color difference is calculated between the two frames. Pixels whose color difference are below an experimentally determined threshold are considered to belong to the same region as in the previous frame. All other pixels are marked as disputed, and are divided into connected disputed regions using a four-connectivity component labeling algorithm. This results in an intermediate segmentation mask consisting of non-disputed regions, and connected disputed regions, I_{INT}.

Tracking Existing Regions

Neighbour relationships for the intermediate segmentation mask I_{INT} are evaluated, and a Bayes classifier is used to assign the pixels in each disputed region to one of it's neighbouring non-disputed regions, using the histograms for each region from the previous frame as the a priori probability. The result of this process is a segmentation mask with all pixels in the current frame classified to one of the existing regions from

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the previous frame I_E.

Detection of New Regions

To detect new regions, a separate segmentation mask is created I_N. Here, two rules are used (labeled rules 1 and 2) to detect new regions. Using the intermediate mask I_{INT}, the homogeneity measure given in Eq. 2.14 is calculated for every region R_i.

$$\mathbf{P}_{R_i} = E\{C(p)|R_i\}, \tag{2.14}$$

where E is the expected value operator, and $C(p)$ is the MAP probability of each pixel p in R_i. Then, for each disputed region, the homogeneity measure is calculated as,

$$\mathbf{P}_{R_{merge}} = E\{max_{inondisp}(E\{C(p)|R_{i_{nondisp}}\})\}. \tag{2.15}$$

This measure gives an indication of how each disputed region would effect the homogeneity of its closest non-disputed neighbour, if the pair were to be merged. Any disputed region that dramatically lowers the homogeneity value of its nearest non-disputed region as in Eq. 2.16, is classified as a possible new region.

$$\frac{\mathbf{P}_{R_{merge}}}{min_i\{\mathbf{P}_{R_i}\}} < 0.05 \tag{2.16}$$

Next, histograms are calculated for each possible new region, and each pixel in each of these regions is re-classified to either the new region or to its nearest existing region based on its Bayesian probability. After this re-classification, a component labeling algorithm is again applied to the mask, and all new regions that exceed a predefined size threshold $t_{new} = 0.002 * X * Y$ are identified as valid new regions and marked in the new region segmentation mask I_N.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Segmentation Mask Fusion

The final step in tracking is to merge the new and existing segmentation masks I_E and I_N. Here, the mask for the existing regions I_E, is fused with the new region mask I_N. Any pixels that are not specified as valid new regions in the new region mask I_N, stay in their existing regions. Otherwise they are categorized as new regions and the following rules are then applied to the new regions.

- New regions are appended to their nearest existing neighbouring region if the color center distance is below a threshold.
- New regions are assigned to extinct regions, if their color centers are below a threshold and the spatial distance is below a threshold. This also allows for the tracking of fast moving objects whose regions of support do not overlap in consecutive frames.

2.3.3 Trajectory-Based Region Merging and Background Detection

The final stage of the segmentation algorithm is grouping the tracked regions into real objects. The regions that were segmented in the first frame and tracked through the clip consist of image areas that were homogeneous with respect to color and motion in the first frame. However, real objects often consist of more complex regions that are not necessarily homogeneous with respect to color or motion in any single frame. The result of this is an over segmentation of the clip. To determine which regions should be grouped together in order to represent the real objects in the clip, the long term trajectories of the regions are examined. The following three steps are employed to generate the final segmented output; Region Trajectory Calculation, Region Merging, and Background Region Detection.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Region Trajectory Calculation

For each frame, a block-based motion estimation is applied to generate motion vectors for every pixel. These motion vectors are then used to estimate the 8 parameters of the bilinear motion model (Eq. 2.17), for every region in the frame.

$$\begin{aligned} u &= a_0 + a_1 x + a_2 y + a_3 xy \\ v &= a_4 + a_5 x + a_6 y + a_7 xy \end{aligned} \tag{2.17}$$

The motion parameters are estimated using the least squares estimation method, where the estimation error is calculated according to Eq. 2.18.

$$E_{all} = \sum_{i=1}^{N} (u_i - \hat{u}_i)^2 + (v_i - \hat{v}_i)^2 \tag{2.18}$$

Where u_i, and v_i are the motion vectors generated from the estimated bilinear parameters, and $\hat{u}_i$ and $\hat{v}_i$ are the vectors obtained from the block-based motion estimation. To improve robustness, the bilinear motion estimation employs an iterative rejection scheme [28]. This is done by first calculating the estimation error of all blocks, according to Eq. 2.19

$$E_{ave} = \frac{1}{N} \sum_{i=1}^{N} (u_i - \hat{u}_i)^2 + (v_i - \hat{v}_i)^2 \tag{2.19}$$

Any blocks with an estimation error higher than the average E_{ave} are marked as rejected. The bilinear motion parameters are then recalculated using only the non-rejected blocks. This process repeats until the set of rejected blocks in the current iteration are the same as those from the previous iteration.

Motion parameters are estimated for every object so that a region motion param-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

eter vector, $\mathbf{U}^t(R_i)$, is calculated for each object in the frame as,

$$\mathbf{U}^t(R_i) = [a_0, a_1, ...a_7]. \tag{2.20}$$

The region motion parameter vector is calculated for every frame, so that a trajectory matrix $\mathbf{U}(R_i)$ is formed for each region, containing its bilinear parameters for every frame in the clip.

$$\mathbf{U}(R_i) = [\mathbf{U}^1(R_i), \mathbf{U}^2(R_i), ...\mathbf{U}^T(R_i)] \tag{2.21}$$

Region Merging

Region merging begins by defining spatiotemporal neighbours as all pairs of regions that are connected in every frame in which they co-exist. The following rules are employed in this stage.

- If a region is in the scene for less than 4 frames, then it is merged with it's nearest neighbour.
- If a region is too thin, it is merged with it's nearest neighbour.

Next, the trajectory matrix parameters are used to determine which regions will be merged into real objects. This is done by first merging each pair of spatiotemporal neighbours and calculating a new set of bilinear motion parameters for the merged pair. Then, these new parameters are used to motion compensate each region of the pair independently, and the compensation error is calculated. This compensation error is compared to the compensation error generated when using the original bilinear parameters of each region. Regions that belong to the same real object should be well represented by one set of motion parameters, so the motion error from the parameters of the merged pair should not be drastically higher than the original estimation error.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

A motion difference measure, $D_U^t(R_m, R_n)$, is defined for each frame as,

$$D_U^t(R_m, R_n) = \frac{E_{m,n}^t(R_m) - E_m^t(R_m)}{N_m^t} + \frac{E_{m,n}^t(R_n) - E_n^t(R_n)}{N_n^t}, \tag{2.22}$$

where $E_{m,n}^t(R_m)$ is the motion compensation error when the combined region parameters are used to compensate region R_m, $E_m^t(R_m)$ is the error using the original parameters to compensate R_m. Similarly, $E_{m,n}^t(R_n)$, and $E_n^t(R_n)$ are the combined and individual parameters for region R_n. N_m^t, and N_m^t are the sizes of regions R_m and R_n. A trajectory difference measure, $D_U(R_m, R_n)$, is formed by combining the motion similarity measures for all frames,

$$D_U(R_m, R_n) = \frac{\sum_{t=1}^{T-1} \Gamma_t(R_m)\Gamma_t(R_n) D_U^t(R_m, R_n)}{max\{1, \sum_{t=1}^{T-1} \Gamma_t(R_m)\Gamma_t(R_n)\}}, \tag{2.23}$$

where Γ_t is a function to express whether or not an object is present in a particular frame, defined as:

$$\begin{array}{l} \Gamma_t(R_i) = 1 \quad \text{if } R_i \text{ is present in frame t} \\ \qquad\qquad\quad 0 \quad \text{otherwise} \end{array} \tag{2.24}$$

The pair of regions with the lowest trajectory difference will be merged. Region trajectories are then re-calculated, and the merging process is repeated. The trajectory difference measure will increase after each merge. However, merging regions that belong to the same object should result in small increases in the difference measure, while merging regions belonging to different objects will produce a large increase. Therefore, the process terminates when the rate of error increase is at a maximum,

$$\frac{D_{U,k}}{D_{U,k}}, \forall k \in [K-2, 1], D_{U,k+1} > 0, D_{U,k} > D_{U,min}, \tag{2.25}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

where $D_{U,min}$ is set to 0.1.

Background Region Detection

After the image has been segmented into objects, the camera motion is estimated by the method proposed in [28], and objects whose trajectory is consistent with the camera motion are marked as background. This is done by comparing each object's trajectory to the background trajectory using the same similarity measurement as in the region merging stage.

2.4 Summary

The trend in multi feature-based object segmentation goes towards developing effective means to combine numerous video features at both the frame and sequence level. The authors of [1] propose an innovative technique for combining color, motion, and spatial features at the frame level with adaptive weighting. The authors of [2] propose a promising technique for using long term object trajectories taken over an entire video sequence to improve on an initial color, motion, and spatial feature-based segmentation. For these reasons, we have selected the papers in [1, 2] to implement and compare.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 3

Proposed Segmentation Method

3.1 Overview

The segmentation method we propose is based on [2]. The proposed method consists of an initial segmentation, object tracking, histogram-based object enhancement, and region merging. A detailed block diagram of the proposed approach is presented in Fig. 3.1.

We have introduced the following improvements:

1. Initial segmentation: We include motion and color variances in the distance function of the KMCC algorithm, and add histogram distance-based merging.

2. Histogram-based object enhancement: We take a set of segmentation measures while tracking objects to improve the accuracy of object boundaries.

3. Post-tracking merging: Regions are merged based on cumulative histograms gathered over the entire clip.

4. Trajectory-based merging: We handle partial occlusion and deal with isolated regions.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

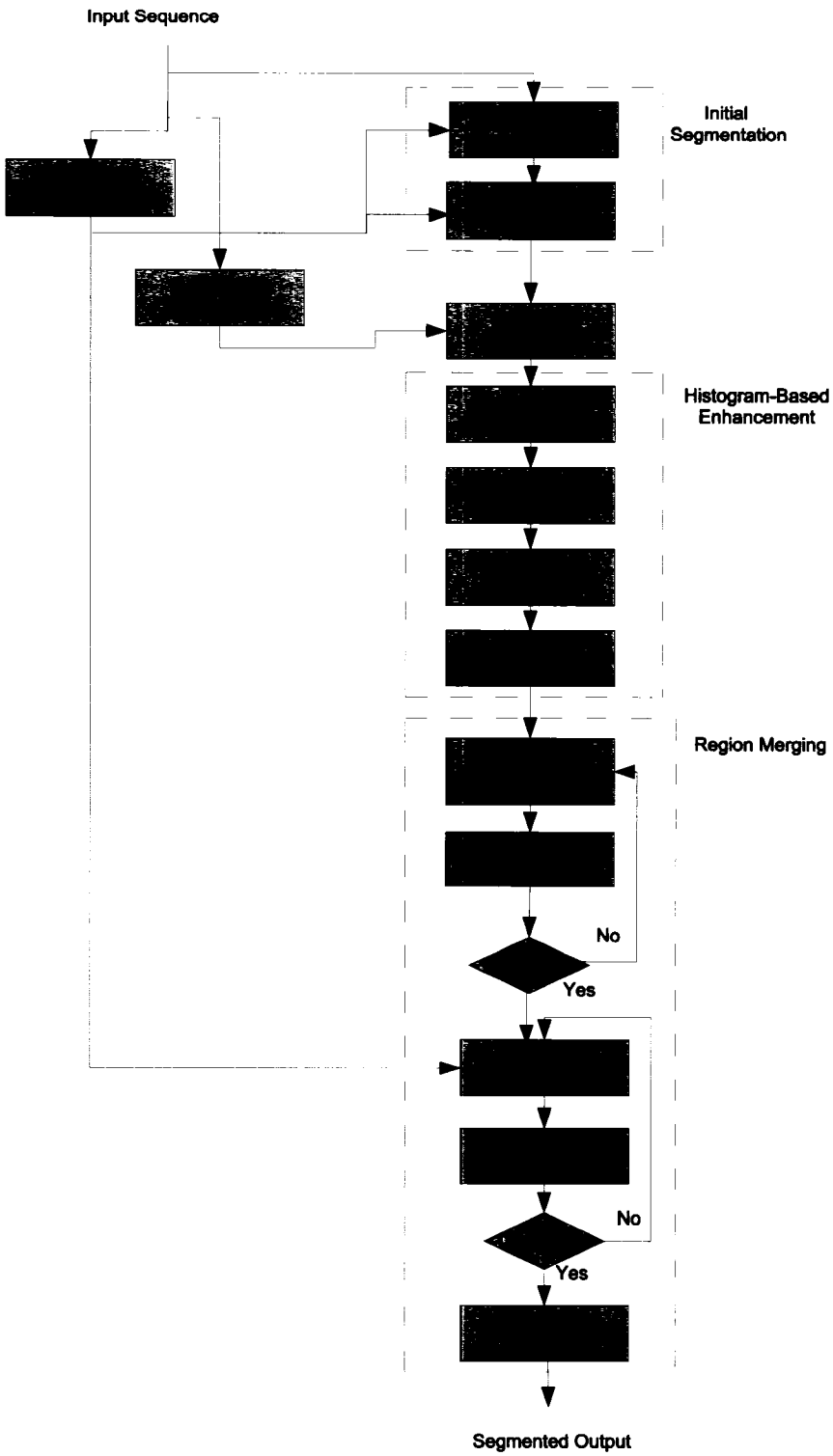


Figure 3.1: Detailed Block Diagram of the Proposed Segmentation Algorithm.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3.2 Initial Segmentation

The initial segmentation of [2] uses the euclidean distance of each pixel from each color and motion center to classify pixels. This is effective for regions that have relatively simple color and motion distributions, but can result in errors for more complex regions. In order to more accurately classify pixels higher order statistical information has to be taken into account. Fig. 3.2 shows a simple example PDF of two color regions. If color distance alone were to be used to classify pixels, all color values below 100 would be assigned to the distribution on the left. However, it is clear from looking at the pdfs that all color values above 60 should be assigned to the distribution on the right.

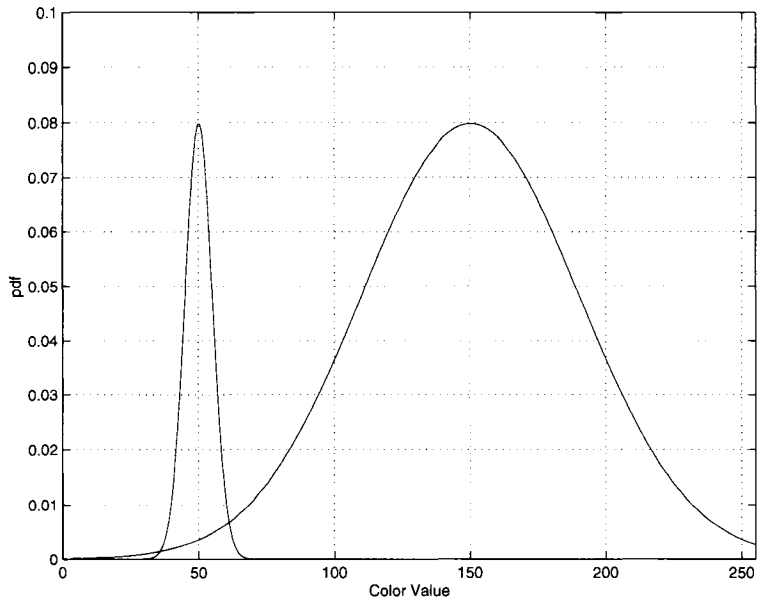


Figure 3.2: Simple Example Showing Color Distributions of Two Regions.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

We propose to include variance information about the color and motion distributions of each region in the KMCC distance function. After the initial centers are estimated, the feature variance of each region is calculated, and pixels are classified according to their distance from the center of each feature divided by the variance. So we propose the distance function in Eq. 3.1.

$$D_{KMCC} = \frac{\left\|\mathbf{C}(p)-\overline{\mathbf{C}}_{R_i}\right\|}{\sigma^2_{R_i,C}} + \lambda_1 \frac{\left\|\mathbf{M}(p)-\overline{\mathbf{M}}_{R_i}\right\|}{\sigma^2_{R_i,M}} + \lambda_2 \frac{A_{R_i}}{\overline{A}} \left\|\mathbf{p}-\overline{\mathbf{S}}_{R_i}\right\|, \tag{3.1}$$

where $\sigma^2_{R_i,C}$ and $\sigma^2_{R_i,M}$ are the color and motion variances of region R_i.

Classifying pixels in this way is more accurate than using only distances from region centers as in [2], since more information about the distribution of each region is being utilized. Also, this method divides the image into a smaller number of more complex regions, which reduces the over-segmentation normally associated with the KMCC algorithm. Reducing the over-segmentation of the first frame decreases the chances for error in later stages of the algorithm.

To improve the robustness of the initial segmentation, we examine the regions at the end of each iteration of the KMCC. If the algorithm converges to less than two regions, R_i, R_j, that meet Eq. 3.2, indicating under segmentation, the entire process resets and the original KMCC is used.

$$A_{R_i} > \alpha \times X \times Y \tag{3.2}$$

where A_{R_i} is the area of region R_i, and α set experimentally to 0.02.

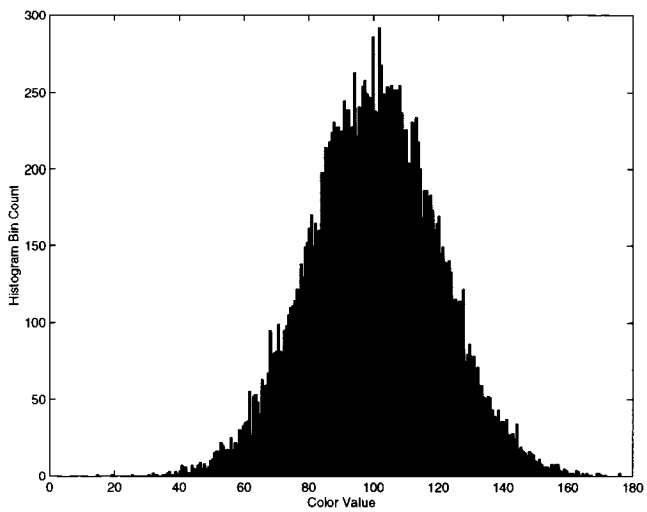
Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3.3 Histogram and Motion Variance-Based Region Merging

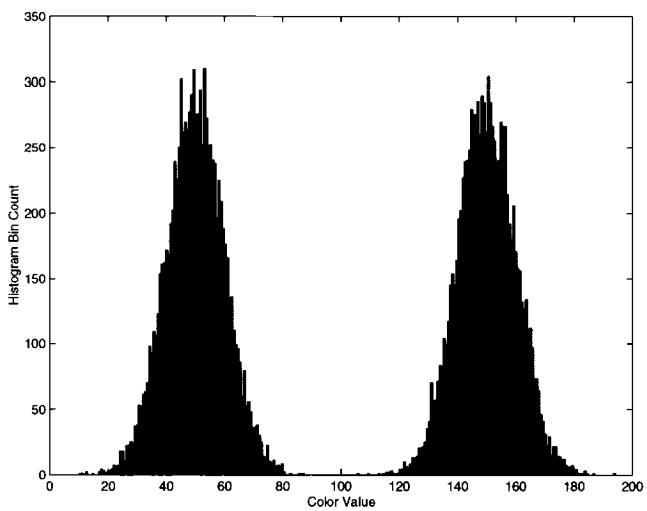
The next stage of the initial segmentation is a histogram and motion variance-based merging stage. As described in section 2.3, the reference KMCC algorithm incorporates merging of neighbouring regions whose color and motion centers are below a certain threshold. This merging process is another area that can be improved by using higher order statistical information about the regions being examined. We accomplish this through the use of color histograms and the motion variance of each region.

First, color histograms are calculated for each region. This is done by dividing the region into a three dimensional array of bins, where the value in each bin is the number of occurrences of that color in the region. This provides a more complete representation of a regions color distribution than using a motion center or a simple statistical representation, such as a Gaussian. Fig. 3.3 shows a hypothetical example of two regions with the same color centers, but different histograms.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.



(a) Histogram Centered at 100



(b) Histogram With a Center at 50 and a Center at 150

Figure 3.3: Example Showing Two Regions With the Same Color Center but Different Histograms.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Once color histograms have been calculated, the χ^2 histogram distance between each pair of neighboring regions is measured as

$$\forall R_i, R_j \in P_1, \quad \chi^2(H_{R_i}, H_{R_j}) = \sum_b \frac{(H_{R_i}(b) - H_{R_j}(b))^2}{(H_{R_i}(b) + H_{R_j}(b))}, \tag{3.3}$$

where P_1 is the set of all pairs of neighboring regions (R_i, R_j) in the first frame, H_{R_i} and H_{R_j} are the histograms of R_i and R_j, and b is the histogram bin. After the distances have been calculated, all neighboring regions satisfying Eqs. 3.4 and 3.5 are merged.

$$\chi^2(H_{R_i}, H_{R_j}) < \beta \cdot S_{hist} \tag{3.4}$$

$$\left\| \overline{\mathbf{M}}_{R_i} - \overline{\mathbf{M}}_{R_j} \right\| < \epsilon \cdot max(\sigma^2_{R_i,M}, \sigma^2_{R_j,M}) \tag{3.5}$$

where S_{hist} is the histogram size, β experimentally set to 1.3, and ϵ experimentally set to 2.

After merging, we re-evaluate the region motion centers and histograms and re-determine neighbor relationships. The merging continues until no more regions meet Eqs. 3.4 and 3.5.

By reducing over-segmentation compared with [2], we identify and merge regions in the first frame that better represent the true video objects.

3.4 Histogram-Based Object Enhancement

During object tracking, we measure the segmentation quality of each object in each frame. We use the following three measures to do this:

1. Color homogeneity of the region [2]. This is defined as the average of the MAP probabilities of every pixel in the region, and is determined from the color histograms for each region.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2. Color contrast across the object boundary [29]. This measure was shown in [29] to be an effective objective measure of segmentation quality. The object contour is traced, and all along the object boundary pairs of blocks are chosen, with each pair consisting of one block inside and one block outside the object. The mean color value for each block is calculated and the absolute difference between each pair of blocks is taken. The color contrast is the average of all these absolute differences along the object boundary.

3. Motion contrast across the object boundary [29]. This is calculated in a similar manner to the color contrast, except that motion vectors are used instead of color values.

After objects have been tracked through the entire clip, we examine these segmentation measures and each object's movements to determine which objects we will enhance, and for which frames we will perform the enhancement.

For a given object, most variation in object segmentation quality between frames is due to movement. Therefore, we are here mainly interested in moving objects. To this end, we examine the trajectories of all objects in the entire video clip and choose which ones to enhance as follows.

The (x, y) coordinates of each object's center in each frame are used to calculate the maximum displacement of every object in the clip. The displacement is taken with respect to the first frame. Objects whose maximum displacement is above a certain threshold are considered to have undergone significant motion and are candidates for enhancement (Eq. 3.6).

$$\begin{aligned} &\forall R_i \in I \quad \text{and} \quad t = \frac{\sqrt{A_{R_i}/\pi}}{2} \\ &\mathbf{\Delta D}_{R_i,max} > t : \text{enhance} \quad R_i \\ &\mathbf{\Delta D}_{R_i,max} \leq t : \text{keep} \quad R_i, \end{aligned} \tag{3.6}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

where $\mathbf{\Delta D_{R_i,max}}$ is the maximum displacement of region R_i over the entire clip I and $\overline{A}_{R_i}$ is the size of R_i averaged over I.

Once we have chosen which objects to enhance, we examine their segmentation quality measures for each frame and enhance objects according to the following rules:

1. If an object's color homogeneity in a given frame is below that same object's average color homogeneity for all frames, this indicates that pixels belonging outside the object have been classified inside the object in this frame. In this case, pixels within the object and close to the boundary will be marked as disputed and re-classified.

2. High color homogeneity with below average color contrast indicates that pixels belonging inside the object have been classified outside. In this case, pixels close to the boundary but outside the object will be re-classified.

3. High color homogeneity with high color contrast indicates a good segmentation. Nothing will be done.

We re-classify pixels through a Bayesian approach using histograms from key frames of the clip to determine the MAP probability of each disputed pixel. Out of every five frames, the frame with the highest homogeneity and contrast is a key frame. The disputed pixels in each frame are re-assigned based on each object's nearest key frame histogram.

After re-assigning pixels, we perform an error check based on the assumption that object enhancements should not result in drastic changes in object size. We measure the size of the object, and if it has increased in size by more than 200% or decreased by more than 70%, the test fails. If the object's motion contrast has decreased, the error check fails as well. Due to the use of block-based motion estimation, motion contrast is not effective for locating small inaccuracies in object boundaries, so it was not used

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

in selecting the frames needing improvement or the key frames. However, a decrease in motion contrast does indicate a significant reduction in boundary accuracy, making motion contrast an effective measure for error checking. If the enhanced object fails either of the error checks, the enhancement is rejected, otherwise it is accepted.

This enhancement stage improves the boundaries of tracked objects over that of [2]. This also allows more accurate motion parameters to be estimated for each object, improving the performance of the trajectory-based merging stage.

3.5 Post Tracking Region Merging

Post-tracking region merging simplifies the trajectory-based merging stage (Sec. 3.6). This is desirable, because trajectory-based merging can fail when an object's motion is too complicated (deformation or articulated motion), or when accurate motion vectors are not available (e.g., when objects are highly uniform in color).

Color histograms are used to merge regions which are spatio-temporal neighbors. Here we use cumulative histograms calculated from an object's pixels taken over all frames in the clip. Compared with histograms computed for an object in a single frame, cumulative histograms are less sensitive to noise, inaccurate object boundaries for particular frames, changing illumination, and occlusion. For example, an object with lighting that varies across it's surface in the first frame could be segmented into two regions, but as the object moves these illumination differences could even out, and the two halves of the object can be merged. As with the first frame histogram-based merging (Sec. 3.3), the χ^2 histogram distance (Eq. 3.3) is used to select regions to merge. This stage improves the segmentation of objects with complex motion that present problems for [2].

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3.6 Trajectory-Based Region Merging

We propose a trajectory-based merging that accounts for high occlusion of the background. The trajectory-based merging stage of [2] only examines regions which are spatio-temporal neighbors. However, since region connectivity is enforced during the initial segmentation with the KMCC algorithm, it is possible for the background to be initially segmented into multiple regions that are not spatio-temporal neighbors. One example is when there is a large object, extending from top to bottom in the middle of a frame. In these cases, the video cannot be segmented correctly without merging these non-neighboring background regions. To account for this, any region that contains a corner point, (0,0), (X-1,0), (0, Y-1), (X-1,Y-1), of a frame is considered to be a potential background region, and will be treated as a spatio-temporal neighbor of all other potential background regions in the clip for the purposes of trajectory-based merging. Note that the trajectories are still used to decide if to merge these potential background regions. Thus foreground objects with corner points can still be correctly identified (e.g., Fig. 4.11). With this change of the spatio-temporal neighbor criteria, we are able to correctly segment the disconnected pieces of the background, while still enforcing connectivity of all other objects. Furthermore, after the trajectory-based merging is finished, any island regions (those with only one spatio-temporal neighbor which is not a potential background region) are merged into their surrounding object.

3.7 Summary

In this chapter we have proposed an offline unsupervised video object segmentation method. The proposed method reduces first frame oversegmentation through the use of color and motion variance in the re-classification of pixels, as well as using histograms and motion variance to merge segmented regions. Tracked objects are

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

selectively enhanced through the use of segmentation quality measures after which two merging stages are employed to merge regions into meaningful objects.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 4

Results

4.1 Algorithm Parameters

The proposed segmentation algorithm utilizes the following fixed parameters which have been optimized through experimentation.

1. A_{R_i} First frame under segmentation threshold (Eq. 3.2), set to 0.02. This parameter is set so that if there is not at least one object in the first frame with an area greater than 2% of the image size, it is assumed that we have under segmented and the initial segmentation resets.

2. β Histogram merging threshold (Eq. 3.4), set to 1.3. The value of this parameter is chosen to provide an effective histogram-based merging stage, while preventing the merging of regions which do not belong to the same object.

3. ϵ Motion variance merging threshold (Eq. 3.5), set to 2. It is used along with the histogram merging threshold during the histogram and motion variance-based region merging stage.

4. t Threshold used to determine whether or not to enhance a given object (Eq.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3.6). The value is set to $\frac{\sqrt{\overline{A}_{R_i}/\pi}}{2}$. This provides a means of measuring an objects motion relative to its size. Larger objects would require a larger absolute displacement to be considered for enhancement.

4.2 Subjective Results

Comparison results are presented for a number of standard video test sequences (Table 4.1) in Figs. 4.1 to 4.14. Results for the method proposed in [2] are labeled reference method 1, and results for the method proposed in [1] are labeled reference method 2. In each case, our proposed method compares favorably with both reference methods.

Test Sequence	Dimensions	Number of Frames	Global Motion
Coastguard	352 x 288 (CIF)	300	Pan
Harbour	564 x 240	50	Pan
Mobile	352 x 288 (CIF)	100	Pan
Foreman	352 x 288 (CIF)	300	Pan
Basket Ball	352 x 288 (CIF)	20	Pan
Gameshow	352 x 288 (CIF)	600	Zoom
Tennis	352 x 288 (CIF)	60	Zoom
Miss	176 x 144 (QCIF)	150	None
Suzie	176 x 144 (QCIF)	150	None
Road1	352 x 288 (CIF)	30	None
Carphone	176 x 144 (QCIF)	95	None

Table 4.1: Test Sequences (total of 1855 frames).

Figs. 4.1 and 4.2 present results for the Gameshow and Miss America test sequences. The main objects in these clips consist of multiple colors, and motion that is difficult to model accurately (there is some movement of the neck and head, while the bodies mostly remain stationary). In both cases, significant improvement can be seen in our proposed method with respect to reference method 1. Reference method 2 performs poorly on the Gameshow sequence and has results comparable to ours for the Miss America sequence. In each case, improvement over reference method 1 is

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

due mainly to our improved first frame segmentation. The Miss America sequence is initially segmented into 2 regions, one corresponding to the actor, and one corresponding to the background, so that the merging stages consist of simply distinguishing a single object from the background. The Gameshow sequence is initially segmented into 6 regions, with 1 region corresponding to the actor, and the background divided into several regions, which are all correctly merged in the histogram and trajectory-based merging stages. In comparison, reference method 1 initially segments the actors of both these clips into several regions, corresponding to their head, shoulders, and torso. Due to the inconsistency of motion between the head and torso of the actors, the reference method's trajectory-based merging stage is unable to correctly merge all of the initially segmented regions.

Figs. 4.3 and 4.4 present results for the Foreman and Carphone sequences. These methods consist of moving faces with complex backgrounds. Significant improvement over both reference methods can be seen. Due to the complexity of the backgrounds, these clips are initially segmented into many regions (14 regions in the case of Foreman and 15 for Carphone). The histogram-based merging stages are important for these clips since they begin the merging process, reducing the chance for error in the final trajectory-based merging stage. To further illustrate this point, simulations were run using the proposed method, but with the histogram-based merging disabled. Fig. 4.5 shows results for selected frames demonstrating that parts of the background are misclassified when the histogram-based merging is not employed.

Figs. 4.6 and 4.7 present results for the Harbour and mobile test sequences. These sequences contain complex backgrounds with a moving camera, which present difficulties for both reference methods. In both cases, the proposed method exhibits significantly improved performance over both reference methods. The Harbour sequence also demonstrates the improved effectiveness of our trajectory-based merging

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

stage, obtained by accounting for background occlusion. Since the main object in the Harbour sequence is large enough to divide the background into 2 disconnected regions, it is important to account for this when performing region merging. Fig. 4.8 shows results for the proposed method when the trajectory-based merging does not account for background occlusion. It can clearly be seen that in this case the background is not correctly segmented, but is instead merged with the actor as part of the foreground.

Fig. 4.9 presents results for the Tennis sequence. The proposed method and both reference methods are able to correctly segment the main objects of this video sequence (the player and the tennis ball). This sequence also demonstrates the effectiveness of our proposed histogram-based object enhancement. Fig. 4.10 shows results for selected frames when the proposed method is run with and without the histogram-based object enhancement. In each case, the tennis ball was selected for enhancement and significant improvement of the object's boundary can be seen in the enhanced frames.

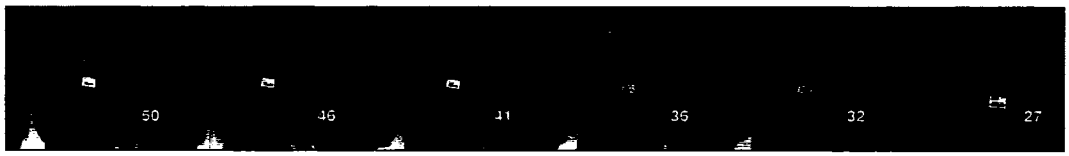
Fig. 4.11 presents results for the Suzie test Sequence. This is another head and shoulders clip with a simple background, but with some movement of the foreground objects. Due to the improved initial segmentation, the proposed method performs significantly better than reference method 1, and comparable to reference method 2.

Fig. 4.12 presents results for the Basketball test sequence. This sequence contains repaid object motion along with a fast moving camera, which presents difficulties for both reference methods. The proposed method is able deal with these characteristics and accurately segment the clip. Also, Figs. 4.12 and 4.13 show that the proposed method's histogram and trajectory-based merging stages can still be effective when histograms and trajectories are taken over relatively short periods.

Figs. 4.13 and 4.14 present results for the Road1 and Coastguard test sequences,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

for which reference method 1 performs well. In each case we achieve similarly strong results, correctly segmenting the car in the Road1 sequence and both boats in the coastguard sequence.



(a) Original Clip

(b) Proposed Method

(c) Reference Method 1

(d) Reference Method 2

Figure 4.1: Frames 1, 120, 240, 360, 480 and 600 of the Gameshow sequence (with some global motion).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(a) Original Clip

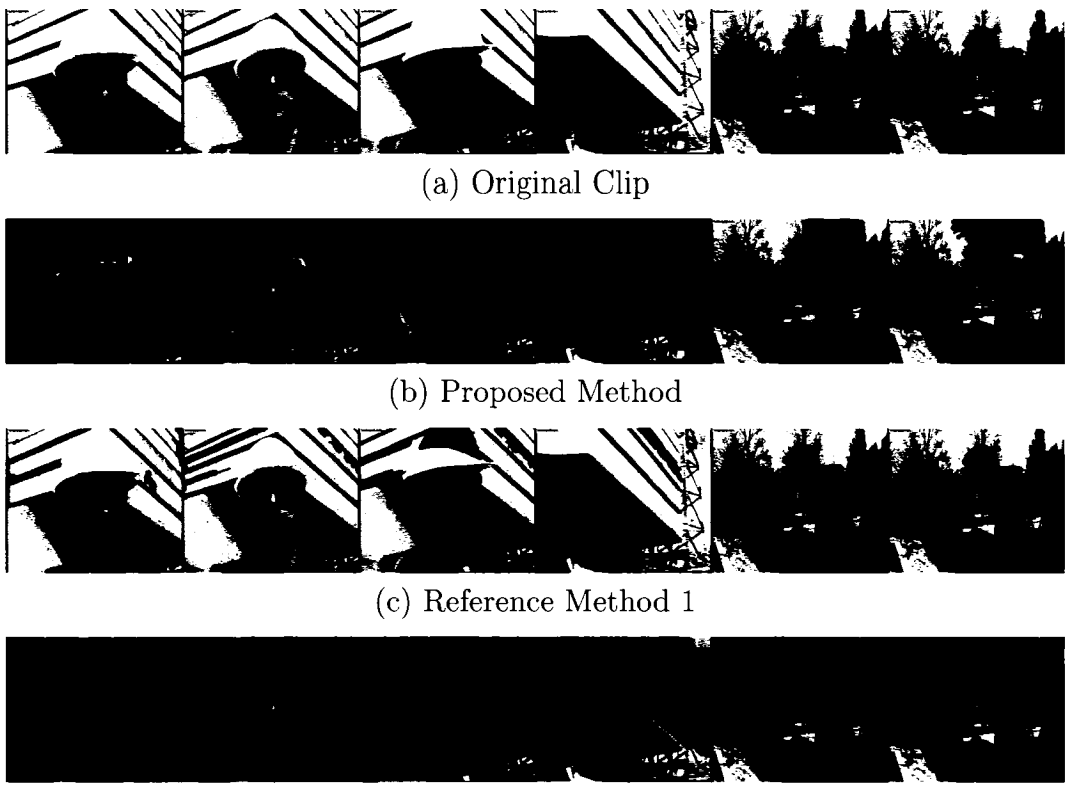
(b) Proposed Method

(c) Reference Method 1

(d) Reference Method 2

Figure 4.2: Frames 1, 30, 60, 90, 120 and 150 of the Miss America clip (without global motion).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.



(a) Original Clip

(b) Proposed Method

(c) Reference Method 1

(d) Reference Method 2

Figure 4.3: Frames 1, 60, 120, 180, 240, and 300 of the Foreman sequence (with global motion).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(a) Original Clip

(b) Proposed Method

(c) Reference Method 1

(d) Reference Method 2

Figure 4.4: Frames 1, 19, 38, 57, 76 and 95 of the Carphone sequence (without global motion).

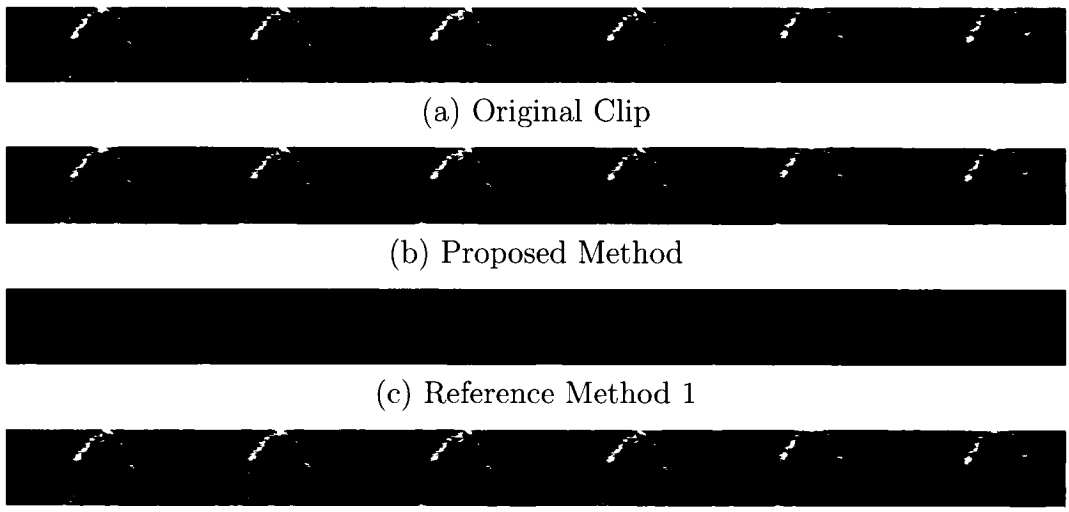
Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(a) With histogram-based merging enabled

(b) With histogram-based merging disabled

Figure 4.5: Proposed method results for frames 1 and 60 of the Foreman sequence.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.



(a) Original Clip

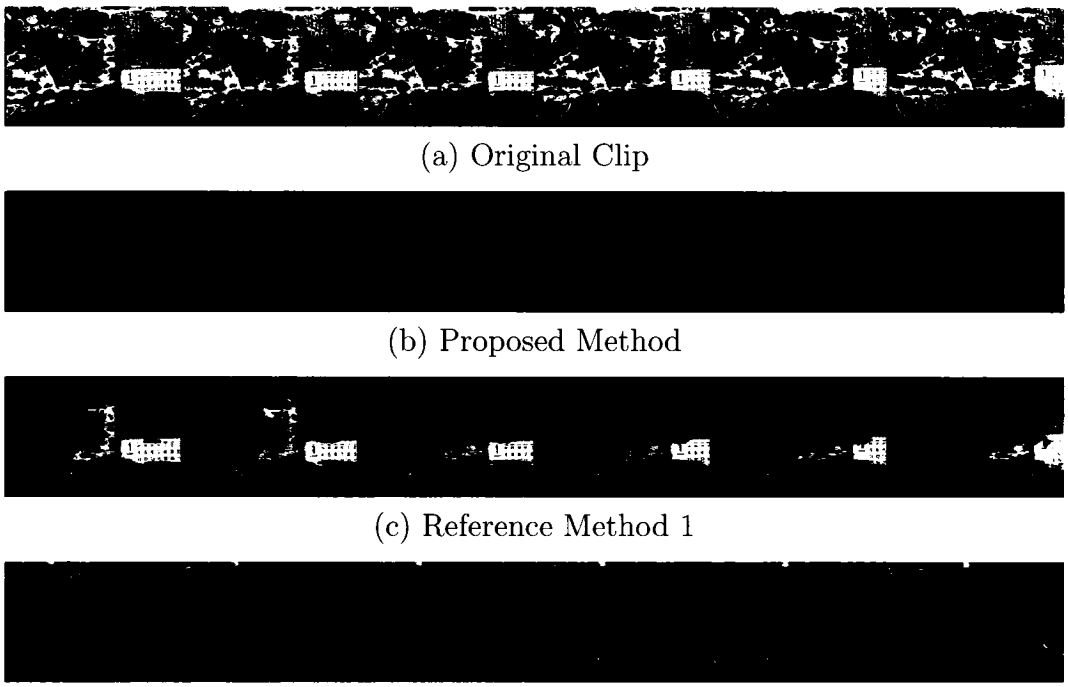
(b) Proposed Method

(c) Reference Method 1

(d) Reference Method 2

Figure 4.6: Frames 1, 10, 20, 30, 40 and 50 of the Harbour sequence (with global motion).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.



(a) Original Clip

(b) Proposed Method

(c) Reference Method 1

(d) Reference Method 2

Figure 4.7: Frames 1, 20, 40, 60, 80 and 100 of the Mobile sequence (with global motion).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(a) With accounting for background occlusion during trajectory-based region merging

(b) Without accounting for background occlusion during trajectory-based region merging

Figure 4.8: Results of the proposed method for frames 1 and 20 of the Harbour sequence.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(a) Original Clip

(b) Proposed Method

(c) Reference Method 1

(d) Reference Method 2

Figure 4.9: Frames 1, 12, 24, 36 and 60 of the Tennis sequence (with some global motion).

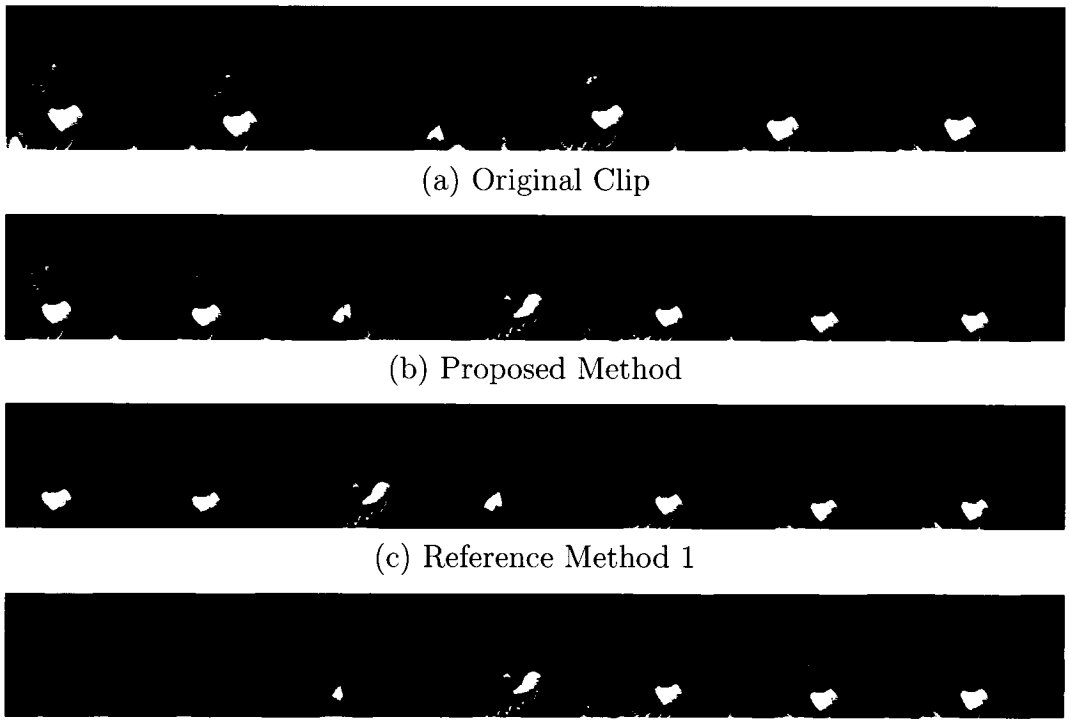
Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(a) After histogram-based enhancement of the ball

(b) Before histogram-based enhancement of the ball

Figure 4.10: Results of the proposed method for frames 14 and 15 of the Tennis sequence.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.



(a) Original Clip

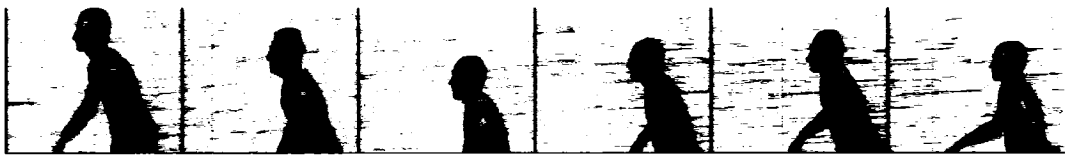
(b) Proposed Method

(c) Reference Method 1

(d) Reference Method 2

Figure 4.11: Frames 1, 30, 60, 90, 120 and 150 of the Suzie clip (without global motion).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.



(a) Original Clip

(b) Proposed Method

(c) Reference Method 1

(d) Reference Method 2

Figure 4.12: Frames 1, 4, 8, 12, 16, and 20 of the Basketball sequence (with global motion).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(a) Original Clip

(b) Proposed Method

(c) Reference Method 1

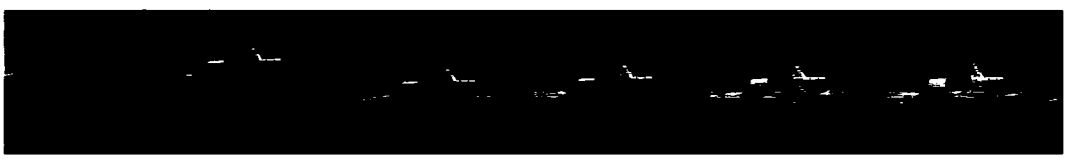
(d) Reference Method 2

Figure 4.13: Frames 1, 5, 10, 15, 20, 25 and 30 of the Road Sequence (without global motion).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(a) Original Clip

(b) Proposed Method



(c) Reference Method 1

(d) Reference Method 2

Figure 4.14: Frames 1, 60, 120, 180, 240, and 300 of the Coastguard sequence (with global motion).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

4.3 Objective Results

This section presents a set of objective results [29] for our test sequences. There are three measures presented for each clip; color contrast, histogram distance, and motion contrast. The color and motion contrast measures trace the object boundary and compare color values and motion vectors inside and outside the object. The histogram distance measures calculate the stability of object histograms throughout the clip.

The color and motion contrast measures are calculated by first tracing the object contour and then drawing a set of normal lines at equally spaced locations across the object boundary. The points on either side of these normal lines are selected to be the centers of blocks inside and outside of the object. In this way, a set of sample blocks containing pixels on either side of the object boundary are constructed. The pixels inside these blocks are used to calculate the object's color contrast as in Eq. 4.1,

$$0 \leq d_{color}(t) = 1 - \frac{1}{K_t} \cdot \sum_{i=1}^{K_t} \delta_{color}(t|i) \leq 1 \tag{4.1}$$

where

$$\delta_{color}(t|i) = \frac{\left\| C_O^i(t) - C_I^i(t) \right\|}{\sqrt{3 \cdot 255^2}}. \tag{4.2}$$

K_t is the total number of normal lines used to calculate the blocks inside and outside the object in frame t. $C_o^i(t)$ and $C_I^i(t)$ are the average color values for the 3 by 3 blocks on the outside and inside of each normal line.

The motion contrast across the object boundary is calculated using the same set of sample blocks inside and outside of the object, according to Eq. 4.3

$$0 \leq d_{motion}(t) = 1 - \frac{\sum_{i=1}^{K_t} \delta_{motion}(t|i)}{\sum_{i=1}^{K_t} w_i} \leq 1 \tag{4.3}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

where

$$\delta_{motion}(t|i) = (1 - exp\left(-\frac{|\mathbf{v}_O^i(t) - \mathbf{v}_I^i(t)))|}{\sigma^2}\right)) \cdot w_i. \tag{4.4}$$

$\mathbf{v}_O^i(t)$ and $\mathbf{v}_I^i(t)$ are the average values of the motion vectors in the sample blocks outside and inside the object boundary. The weighting term w_i is calculated as

$$0 \leq w_i = R(\mathbf{v}_O^i(t)) \cdot R(\mathbf{v}_I^i(t)) \leq 1 \tag{4.5}$$

where

$$R(\mathbf{v}^i(t)) = exp\left(-\frac{|\mathbf{v}^i(t) - \mathbf{b}^i(t+1)|^2}{2\sigma_m^2}\right) \cdot exp\left(-\frac{|c(p^i|t) - c(p^i + \mathbf{v}^i(t|t+1)|^2}{\sigma_c^2}\right). \tag{4.6}$$

The term $\mathbf{b}^i(t+1)$ is the backwards motion vector in frame $t+1$ at the location $c(p^i + \mathbf{v}^i(t))$. In this way, the motion reliability term $R(\mathbf{v}^i(t))$ estimates the reliability of the motion vectors at each point by measuring the similarity of the backward and forward motion vectors, and the difference in pixel intensities at the estimated displacements.

The histogram distance measure for each object is determined by calculating the object's χ^2 histogram distance between each frame and the first frame, as in Eq. 4.7

$$0 \leq w_i = \chi^2(H_t, H_{ref}) = \frac{1}{(N_{H_t} + N_{H_{ref}})} \cdot \sum_b \frac{(r_1 \cdot H_{R_i}(b) - r_1 \cdot H_{R_j}(b))^2}{(H_{R_i}(b) + H_{R_j}(b))} \leq 1 \tag{4.7}$$

where H_t is the histogram for the current frame, and H_{ref} is the histogram for the first frame. The normalization factors r_1, r_2, N_{H_t}, and $N_{H_{ref}}$ are defined as

$$r_1 = \sqrt{\frac{N_{H_{ref}}}{N_{H_t}}}, \tag{4.8}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$r_2 = \frac{1}{r_1}, \tag{4.9}$$

$$N_{H_t} = \sum_b H_t(j) \tag{4.10}$$

and

$$N_{H_{ref}} = \sum_b H_{ref}(j) \tag{4.11}$$

where N_{H_t} and $N_{H_{ref}}$ are the sizes of the current and reference histograms.

The objective measures for each of our test sequences are presented in Figs. 4.15 to 4.25. In each graph, lower normalized values of the color and motion contrast measures indicate more accurate segmentation and lower values of the histogram distance measure means the object histogram is more stable over the clip, indicating better object tracking.

Figs. 4.15, 4.16 and 4.17 present objective measures for the Gameshow, Basketball and Harbour test sequences. In these graphs it can clearly be seen that the objective measures confirm the improved performance of the proposed method over both reference methods. Improvement (lower values) can be seen in each of the color, histogram and motion measures, indicating improved accuracy and stability of the proposed segmentation method.

Figs. 4.18, 4.19, and 4.20 present objective results for the Foreman, Mobile and Carphone sequences. In these clips, the backgrounds contain several colors and textures so that the sharpest color contrast does not always occur on the boundary of the main object of the clip. For this reason, the color contrast and color histogram measures do not produce a reliable indication of segmentation quality. For these clips, the motion contrast is the most reliable quality measure. From the graphs it can be seen that the motion contrast of the proposed method shows improvement

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

with respect to the reference methods.

Fig. 4.21 shows the objective results for the Miss America sequence. In this clip, it can be seen that there is high color contrast within the object combined with non-rigid motion. The result of this is that reference method 1 can achieve low color and motion contrast measures without entirely segmenting the main object of the video sequence. The proposed method achieves objective results that are similar to those of reference method 2, as would be expected from viewing the subjective results.

Fig. 4.22 presents objective results for the Suzie test sequence. As in the case of the Miss America sequence, reference method 1 is able to achieve good segmentation measures by segmenting a smaller object with high color and motion contrast. Also, in agreement with the subjective results, the proposed method achieves similar objective measures as reference method 2.

Figs. 4.23, 4.24 and 4.25 present objective results for the Tennis, Coastguard and Road1 test sequences. For these clips, the proposed method achieves objective results that are similar to reference method 1, and improved over reference method 2, as would be expected from observing the subjective results.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

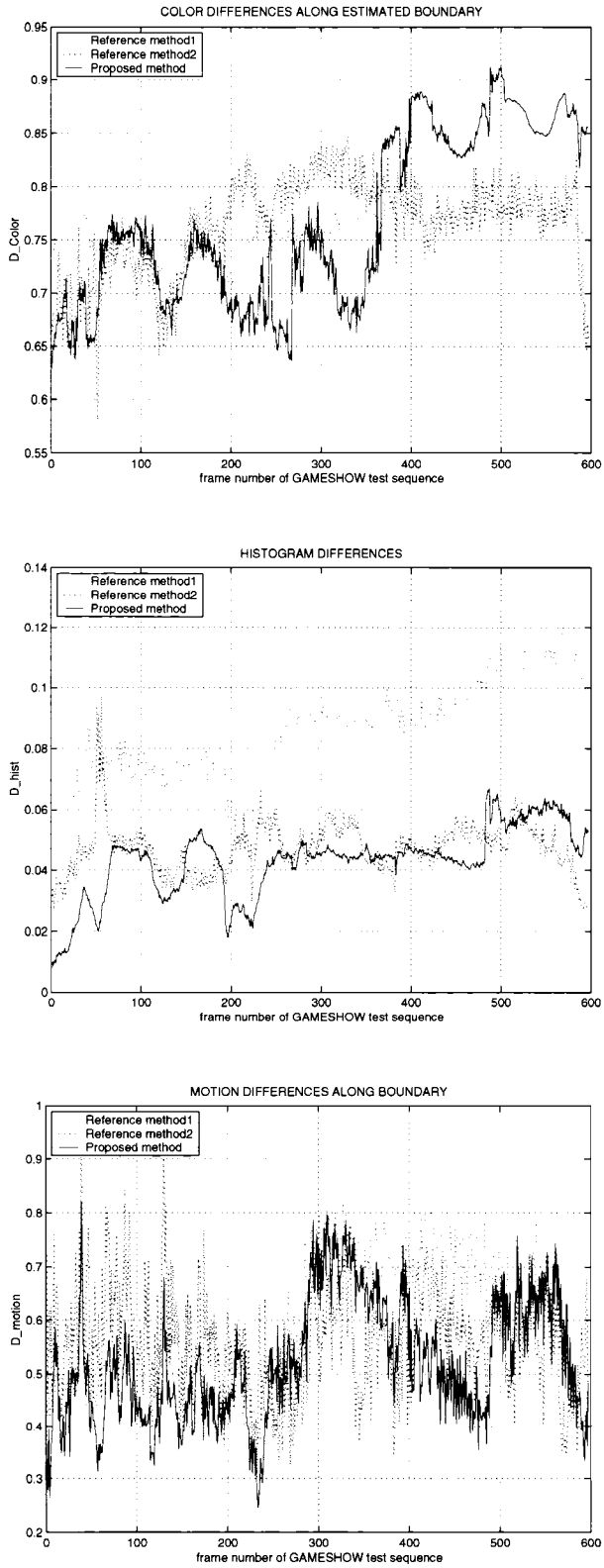


Figure 4.15: Objective Results for the Gameshow Sequence.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

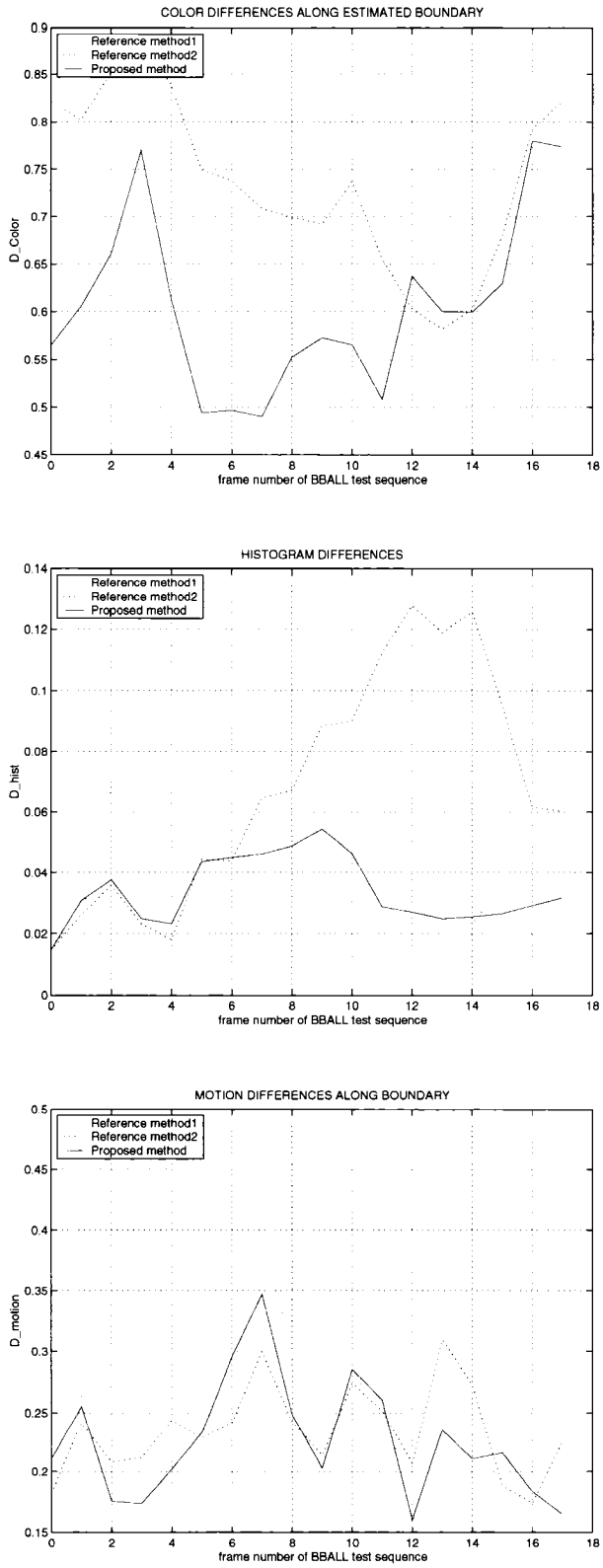


Figure 4.16: Objective Results for the Basketball Sequence.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

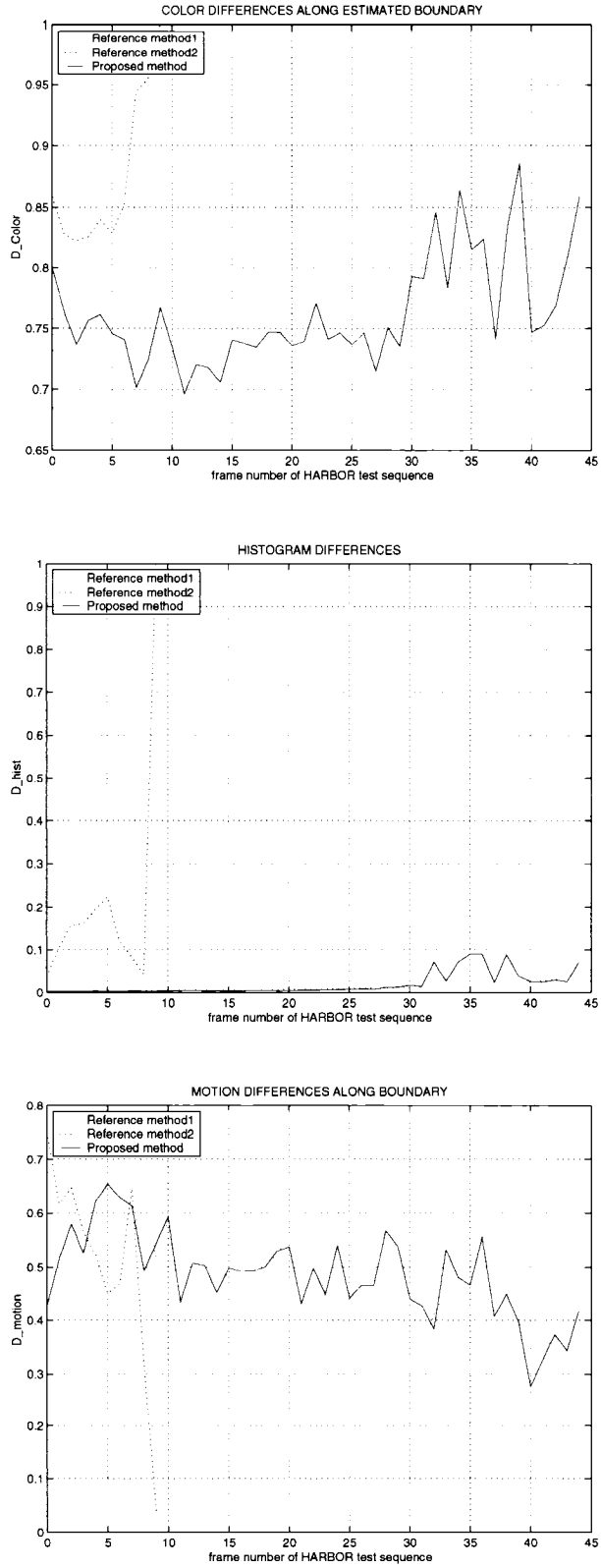


Figure 4.17: Objective Results for the Harbour Sequence.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

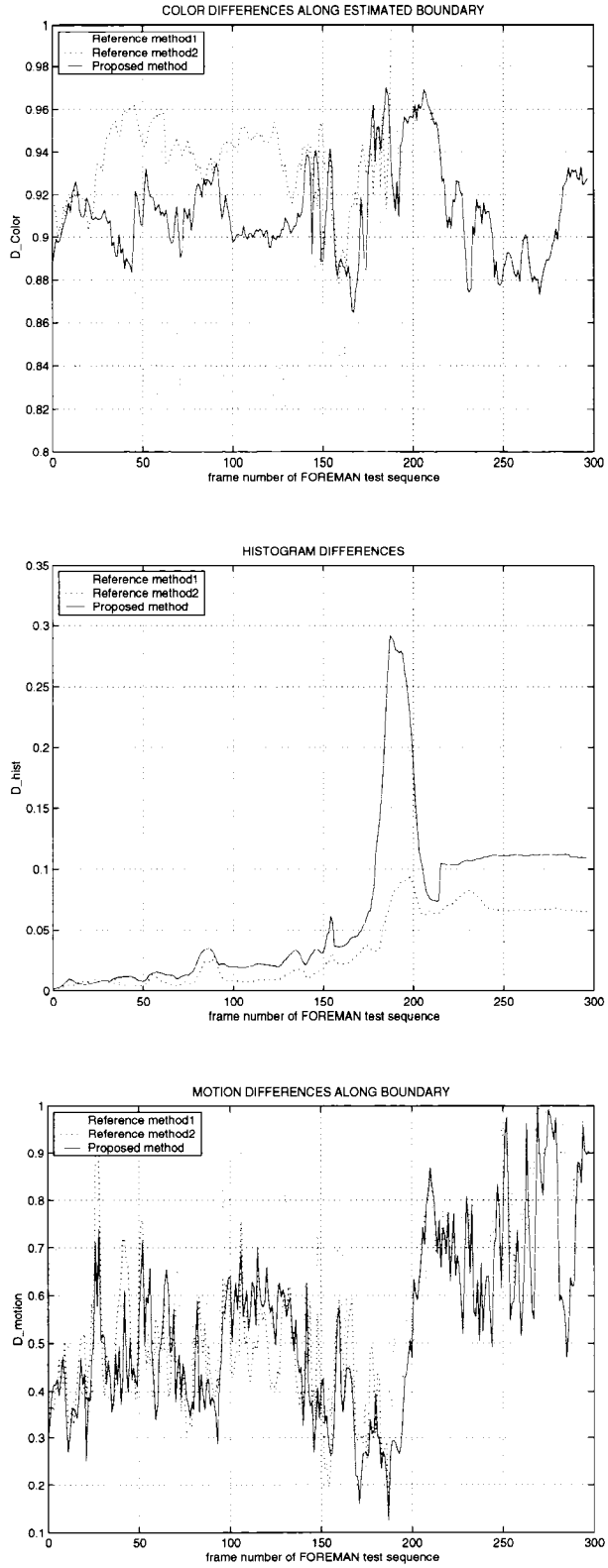


Figure 4.18: Objective Results for the Foreman Sequence.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

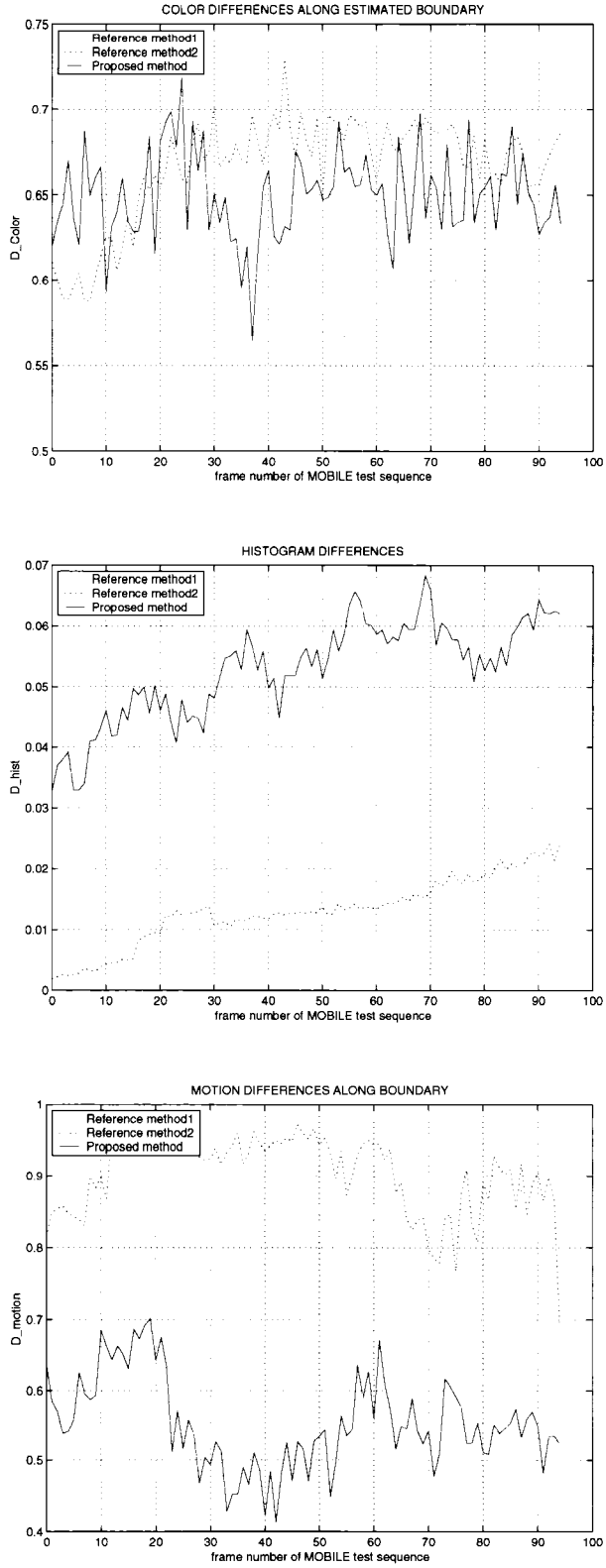


Figure 4.19: Objective Results for the Mobile Sequence.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

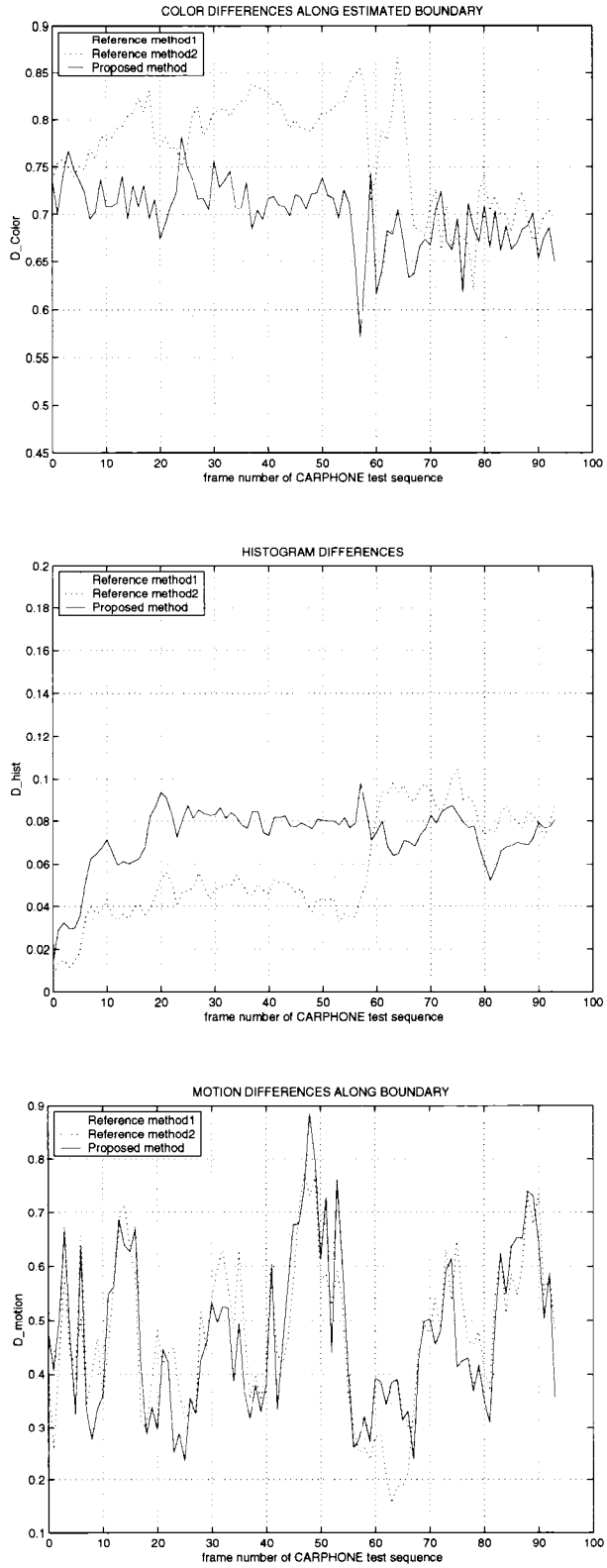


Figure 4.20: Objective Results for the Carphone Sequence.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

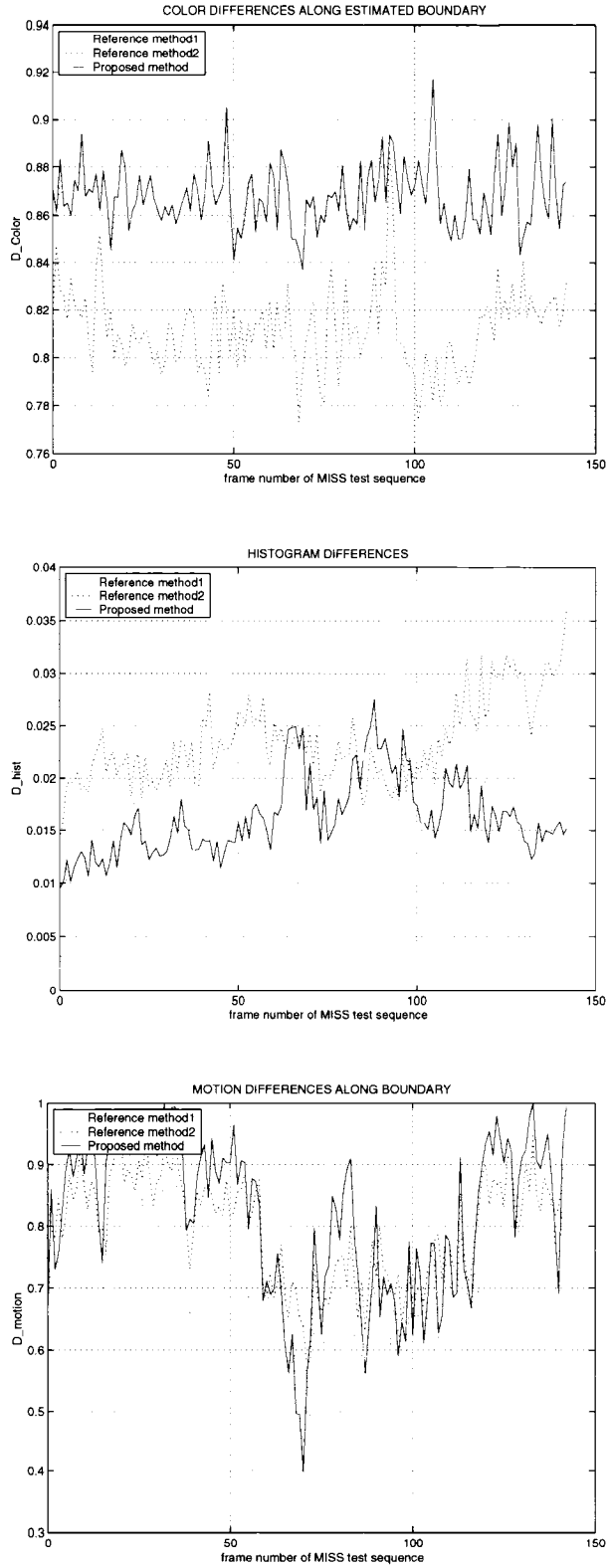


Figure 4.21: Objective Results for the Miss America Sequence.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

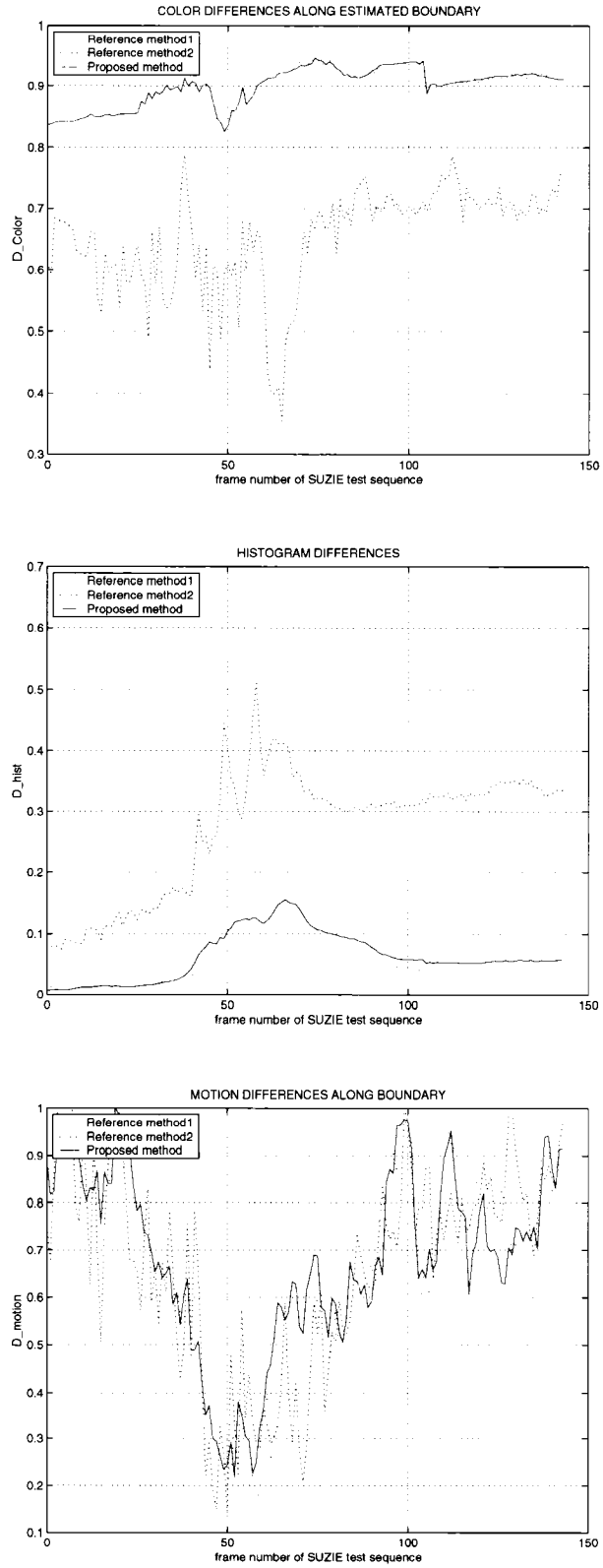


Figure 4.22: Objective Results for the Suzie Sequence.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

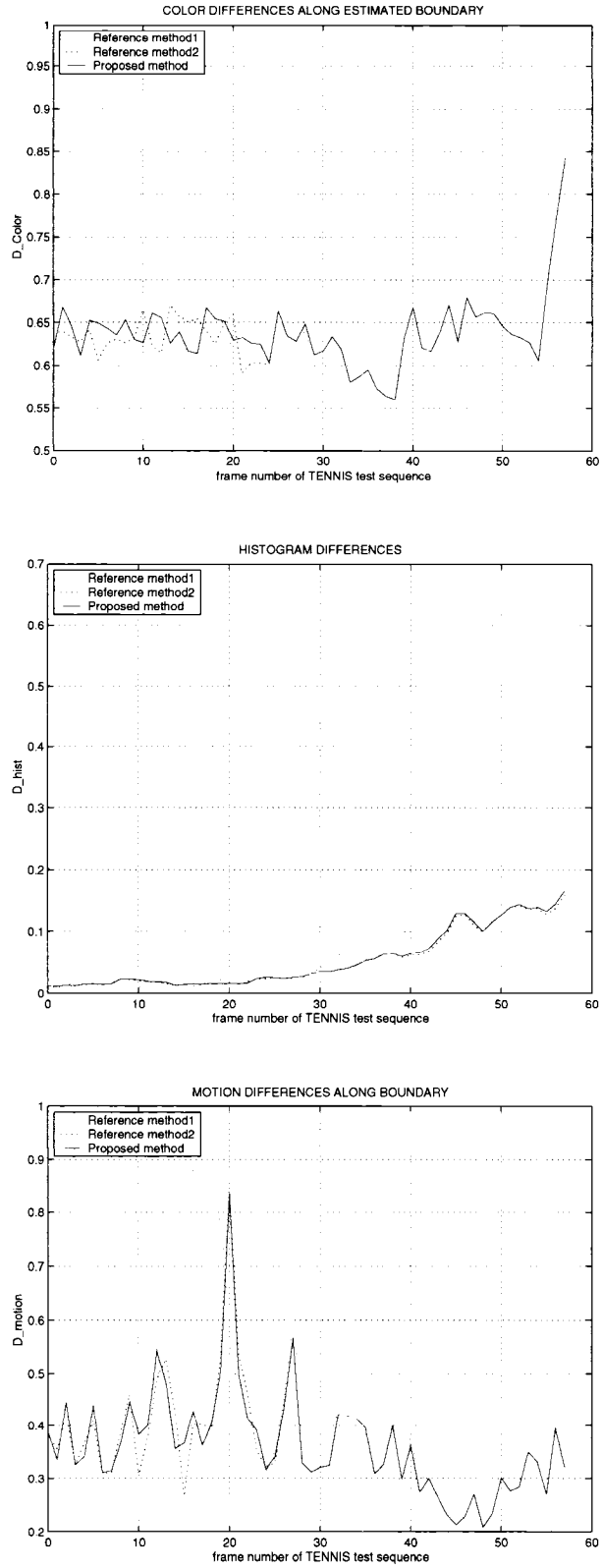


Figure 4.23: Objective Results for the Tennis Sequence.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

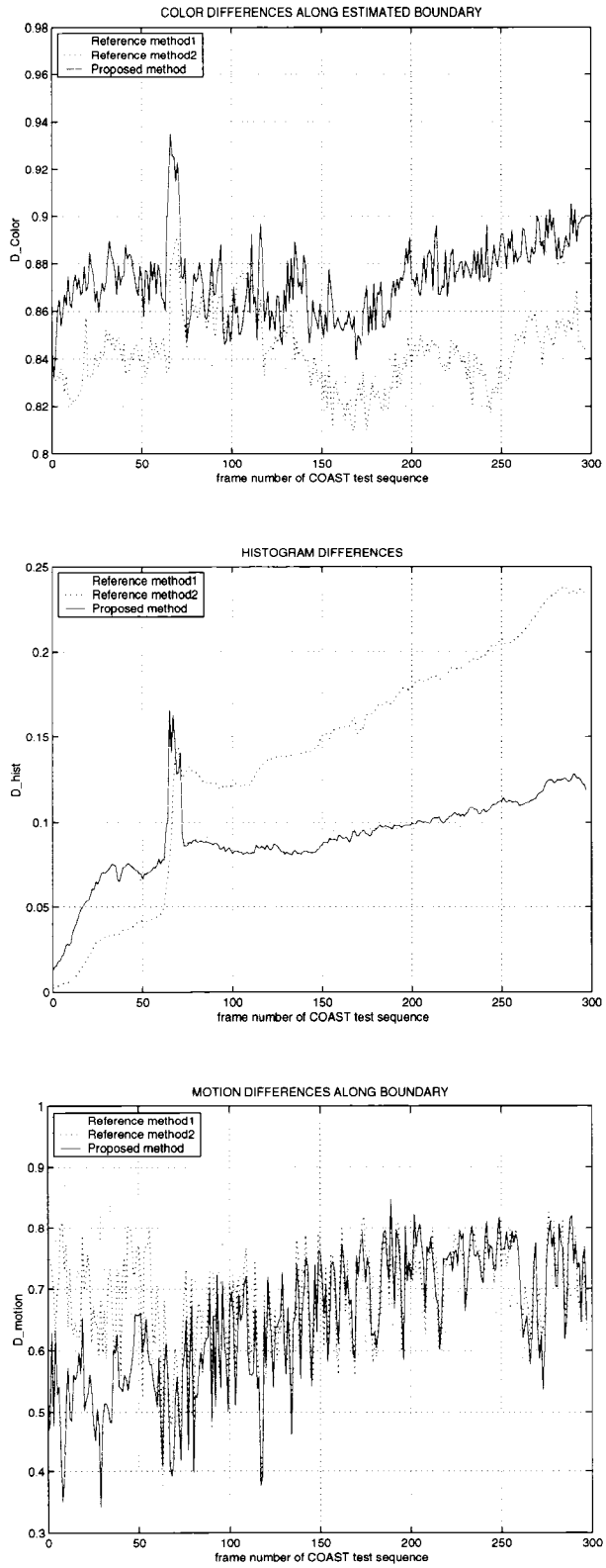


Figure 4.24: Objective Results for the Coastguard Sequence.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

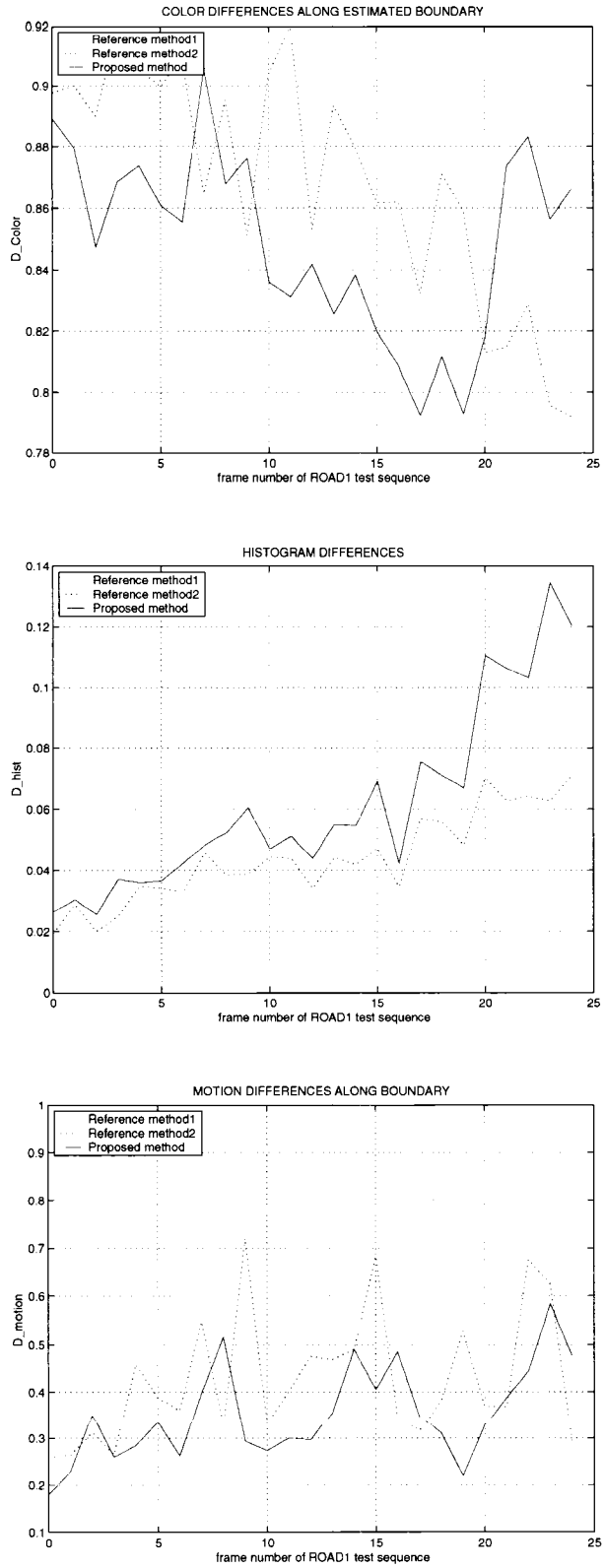


Figure 4.25: Objective Results for the Road1 Sequence.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

4.4 Computation Time

Simulations were performed on an Intel Pentium processor running at 1.5 Ghz. The simulation run times for all test sequences are given in table 4.2. Simulation times can vary depending on the complexity of the sequence under test. The main bottleneck of run time performance is the trajectory-based merging stage. This is because the merging involves iteratively running through the entire clip and re-estimating bilinear motion parameters for each region identified in the initial segmentation. Scenes that are more complex will normally have more regions identified in the initial segmentation, resulting in more iterations of the trajectory-based merging being necessary to identify the real objects.

Although the proposed segmentation method is not real-time, this does limit its practical applicability. In MPEG-7 applications for archive mining, content-based extraction is done off-line on the database. Therefore, CPU time is not a critical factor. The main requirement is that the segmentation be unsupervised, so that content extraction on large databases can be done with minimal user intervention.

Test Sequence	First Frame CPU Time	Average CPU Time Per Frame
Miss	20s	3.2s
Suzie	33s	10.15s
Harbour	121s	121.5s
Mobile	100s	37.85s
Basket Ball	68s	69.85
Gameshow	63s	25.28s
Tennis	65s	62.33s
Coastguard	62s	46.37s
Road1	75s	25.57s
Foreman	59s	32.33s
Carphone	27s	7.08s

Table 4.2: CPU Run Time for Test Sequences.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

4.5 Summary

This chapter has presented both subjective and objective results for a number of video test sequences. The proposed method's computational complexity was discussed and a description of the algorithms parameters provided. From the results, we can conclude that the proposed method shows improved performance with respect to both reference methods.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 5

Conclusion

Video object segmentation remains a challenging topic in video processing. In this thesis we have proposed an offline unsupervised video object segmentation algorithm that is applicable to a variety of video sequences. The proposed method consists of an initial segmentation, object tracking, histogram-based object enhancement, and region merging. The proposed method introduces a number of innovations for video object segmentation. These include reducing over-segmentation of the first frame, using segmentation quality measures to enhance object accuracy, merging tracked regions based on histograms, and accounting for background occlusion. Experimental results have been presented which demonstrate that our algorithm meets all of the objectives outline in Sec. 1.1, as well as demonstrating improved performance over two reference methods.

This proposed system meets its objectives through the following contributions:

- An improved initial segmentation where the following improvements have been made:

    1. Incorporating color and motion variance into an existing region clustering scheme.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2. The addition of a histogram distance and motion variance-based merging stage to reduce over segmentation of the first frame.

- Histogram-based object enhancement, where a set of segmentation measures taken while tracking objects are used to improve the accuracy of object boundaries.

- Merging tracked objects based on cumulative histograms gathered throughout the video clip.

- Trajectory-based merging that has been extended to handle partial occlusion and isolated regions.

Possible extensions to the proposed methodology include improved handling of heavy object occlusion, as well as mechanisms to deal with object splitting and merging. These two issues present problems for many segmentation methods, and designing methods to deal with them is an active area of research in video processing.

Another area where improvement is possible is in the execution speed of the algorithm. For example, by reducing the number of frames used in the trajectory-based merging stage, significant improvement in execution speed may be attainable. As can be seen in Figs. 4.12 and 4.13, an accurate segmentation can sometimes be achieved by examining object trajectories over a relatively small number of frames. By selectively applying the trajectory-based merging to sub-segments of longer clips, accurate segmentations may be attainable at reduced computational complexity. The challenge to this approach is determining when to apply it and which sub-segments to choose.

Finally, improvement may also be obtained by making algorithm parameters more adaptable to video content. For example, clips with global motion might have different optimal settings for certain thresholds than clips without.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Bibliography

[1] S. Khan and M. Shah, "Object based segmentation of video using color, motion and spatial information," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, December 2001, vol. 2, pp. II–746 – II–751.

[2] V. Mezaris, I. Kompatsiaris, and M.G. Strintzis, "Video object segmentation using bayes-based temporal tracking and trajectory-based region merging," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, pp. 782 – 795, June 2004.

[3] R. Schettini, "Open issues and research trends in content based image retrieval," in *Workshop on Image Mining*, Centre National de la Recherche Scientifique, Paris, 2003.

[4] C. Cotsaces, N. Nikolaidis, and I. Pitas, "Video shot boundary detection and condensed representation: A review," *IEEE Signal Proc. Mag.*, vol. 23, no. 2, pp. 28–37, 2006.

[5] Y. Rui, Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "A unified framework for video summarization, browsing and retrieval," in *MERL Technical Report*, 2004, pp. TR2004–115.

[6] L. Chiariglione, "The development of an integrated audiovisual coding standard: Mpeg," in *Proceedings of the IEEE*, February 1995, vol. 83, pp. 151–157.

[7] B. Furht and O. Marques, *Handbook of Video Databases: Design and Applications*, CRC Press, 2004.

[8] L. Gagnon, "R&D status of ERIC-7 and MADIS: two systems for MPEG-7 indexing/search of audio-visual content," in *Proc. SPIE Conference on Multimedia Systems and Applications VIII (SPIE #6015)*, October 2005, pp. 341–352.

[9] A. Amer and C. Regazzoni, "Introduction to the special issue on video object processing for surveillance applications," *Real-Time Imaging*, vol. 11, pp. 1–5, 2005.

[10] J. Hunter, "An overview of the mpeg-7 description definition language (ddl)," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 765 – 772, June 2001.

[11] K. Ryan, A. Amer, and L. Gagnon, "Video object segmentation based on object enhancement and region merging," in *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo*, July 2006, pp. 273 – 276.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[12] J. Wang and E. Adelson, "Representing moving images with layers," *IEEE Transactions on Image Processing*, vol. 3, pp. 625 – 638, September 1994.

[13] T. Darrell and A.P. Pentland, "Cooperative robust estimation using layers of support," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 474 – 487, May 1995.

[14] J. McQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematics Statistics and Probability*, 1967, vol. 1, p. 281296.

[15] T. Gevers, "Robust segmentation and tracking of colored objects in video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, pp. 776 – 781, June 2004.

[16] P.E. Eren, Y. Altunbasak, and Tekalp, "Region-based affine motion segmentation using color information," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1997, vol. 4, pp. 3005 – 3008.

[17] L. Bergen and F. Meyer, "A novel approach to depth ordering in monocular image sequences," in *IEEE International Conference on Computer Vision and Pattern Recognition*, June 2000, vol. 2, pp. 536 – 541.

[18] Y. P. Tsai, C.-C. Lai, Y.-P. Hung, and Z.-C. Shih, "A bayesian approach to video object segmentation via merging 3-d watershed volumes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 175 – 180, January 2005.

[19] H. Xu, A.A. Younis, and M.R. Kabuka, "Automatic moving object extraction for content-based applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, pp. 796 – 812, June 2004.

[20] T. Lindeberg, "Scale-space for discrete signals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 234 – 254, March 1990.

[21] L.R. Williams and D.W. Jacobs, "Local parallel computation of stochastic completion fields," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 1996, vol. 1, pp. 161 – 168.

[22] J. Malik, S. Belongie, J. Shi, and T. Leung, "Textons, contours and regions: cue integration in image segmentation," in *Seventh IEEE International Conference on Computer Vision*, September 1999, vol. 2, pp. 918 – 925.

[23] S.X. Yu, "Segmentation using multiscale cues," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2004, vol. 1, pp. 247 – 254.

[24] D. Wang, "Unsupervised video segmentation based on watersheds and temporal tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, pp. 539 – 546, September 1998.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[25] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Internation Joint Conference on Artificial Intelligence*, 1981, pp. 674–679.

[26] N. V. Boulgouris, I. Kompatsiaris, V. Mezaris, D. Simitopoulos, and M. G. Strintzis, "Segmentation and content-based watermarking for color image and image region indexing and retrieval," *EURASIP Journal Applied Signal Processing*, pp. 418–231, April 2002.

[27] D. Torrieri and K. Bakhru, "The maximin algorithm for adaptive arrays and frequency-hopping communications," *IEEE Transactions on Antennas and Propagation*, vol. 32, no. 9, pp. 919– 928, September 1984.

[28] T. Yu and Y. Zhang, "Retrieval of video clips using global motion information," *Electronics Letters*, vol. 37, pp. 893 – 895, July 2001.

[29] C. Erdem, B. Sankur, and A. Tekalp, "Performance measures for video object segmentation and tracking," *IEEE Transactions on Image Processing*, vol. 13, no. 7, July 2004.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.