J. Vis. Commun. Image R. 22 (2011) 367–377

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Automatic body segmentation with graph cut and self-adaptive initialization level set (SAILS)

Qiang Liu [a,*], Hongliang Li [b], King Ngi Ngan [a]

[a] Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[b] School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu, China

ARTICLE INFO

Article history:
Received 26 July 2010
Accepted 2 February 2011
Available online 4 March 2011

Keywords:
Face detection
Motion estimation and compensation
Body segmentation
Object segmentation
Graph cut
Level set
Background contrast removal
Object detection

ABSTRACT

In this paper, we propose an automatic human body segmentation system which mainly consists of human body detection and object segmentation. Firstly, an automatic human body detector is designed to provide hard constraints on the object and background for segmentation. And a coarse-to-fine segmentation strategy is employed to deal with the situation of partly detected object. Secondly, background contrast removal (BCR) and self-adaptive initialization level set (SAILS) are proposed to solve the tough segmentation problems of the high contrast at object boundary and/or similar colors existing in the object and background. Finally, an object updating scheme is proposed to detect and segment new object when it appears in the scene. Experimental results demonstrate that our body segmentation system works very well in the live video and standard sequences with complex background.

1. Introduction

Segmentation is the problem of trying to extract objects from background in an image or a sequence of frames. Many segmentation algorithms aim to separate an image into meaningful objects using the low level features such as intensity, color, edge, and texture, etc. To satisfy the coming content-based multimedia services [1], segmentation of meaningful objects in unsupervised manner is urgently required in the real-world scenes. But for an arbitrary scene (e.g., dynamic background with varied objects), fully automatic object segmentation is still a monumental challenge to the state-of-the-art techniques [2–4] due to a wide variety of possible objects' combination.

Generally, automatic segmentation is supposed to consist of two stages, i.e., desired object detection and object extraction. The first stage is related to the pattern recognition techniques that rely on the prior information, whilst the second is concerned with classification to separate the closed object boundary. In order to achieve better segmentation results, many interactive methods have been proposed [5–10], which replace the first stage by the user defining the desired object to be extracted in advance. Since the complicated step of object detection is avoided at the cost of interactive effort made by users, these methods usually provide us with much better segmentation performance than automatic ways. However, those methods are time consuming and not suitable for the practical applications.

To avoid the expensive interactive operation, some specific interested objects (like human features: face or body) can be detected by designing appropriate detectors based on a physical model or a training scheme and some constraints are imposed on the real-world problems to achieve progress in this area. For example, given the case of focused objects and defocused background, Li et al. [11] extract the focused objects from low depth-of-field (DoF) video images. For a stationary background, Sun [12] used background subtraction to detect foreground objects and designed an adaptive contrast term to attenuate the contrast in the background which can improve segmentation result.

For the multi-view video scenario, object segmentation works very accurately and efficiently for a probabilistic fusion of multiple cues, i.e., depth, color, and contrast. Using the multi-camera, Zhang [13] employs a visual attention saliency map which is generated from depth, motion and wavelet features to initialize graph cut segmentation. In [14], a stereo-matching technology is proposed to enhance the segmented result and quantitatively evaluated by comparison with the ground-truth. Criminisi [15] replaced the stereo information with motion classification model on monocular sequence, which not only can avoid the complex calibration for stereo cameras but also can provide comparable segmentation accuracy with [14]. But the above two methods need initialization by learning from image frames manually labeled earlier.

* Corresponding author.
E-mail address: qliu@ee.cuhk.edu.hk (Q. Liu).

1047-3203/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2011.03.003

Li [16] proposed automatic face segmentation algorithm using face detection region to initialize graph cut to segment the human face out automatically. To speed up the detection process, a face rejection cascade is proposed to remove most of the negative samples while retaining all the face samples. Then a coarse-to-fine segmentation approach is proposed to solve the problem of the partly detected object. To refine the final segmented result, a matting method is designed to estimate the alpha-matte of human face.

Recent segmentation techniques [7–17] have shown the effectiveness of the color/contrast based model which is proposed by Boykov et al. [5]. This model considers both the color similarity to object/background color and the contrast strength along the object boundary. The final object layer is globally determined using the min-cut algorithm [25]. But, as demonstrated in [13–15], only using the color and contrast cues are insufficient. The segmentation result will be error-prone at the boundary with high contrast or an object will be misclassified when it interacts with the similar color background.

In our proposed automatic body segmentation system, a human body detector based on human features is designed to provide the hard constraints on the object and background for segmentation. And a coarse-to-fine segmentation strategy [16] is adopted to deal with the partly detected object in the first frame to be segmented. Background contrast removal (BCR) is proposed to remove the contrast in the clustered background and improve the visual quality of segmented results simultaneously. We also propose self-adaptive initialization level set (SAILS) to deal with similar color segmentation and speed up the evolution of level set. Finally, an object updating scheme for human body segmentation is proposed to detect and re-initialize the object. The proposed human body segmentation system can be applied to many real scenarios, such as video surveillance, videophone, video-conferencing, and web-chatting, etc. [18–20].

The paper is organized as follows: the proposed automatic body segmentation system is discussed in Section 2. In Section 3, experimental results and comparisons are provided to demonstrate the accuracy and efficiency of our proposed algorithm. Finally, conclusions are drawn in Section 4.

2. Proposed body segmentation

Human body segmentation is generally considered as the first step for object-based video coding or human-to-machine communication. Additionally, the segmented human body can be used in 3D reconstruction [21] and digital entertainment, such as tooning [22]. In this section, we introduce an automatic segmentation system which can find the interested object (human body) and extract it from the background accurately. The framework of our proposed human body segmentation system is demonstrated in Fig. 1.

2.1. Automatic human body detection

In the proposed system, we use the face detection algorithm in [23] and human body features detection in [24] as a pre-processing step to find the human body. The face detector can identify and locate the frontal-view human faces accurately and real-timely from the input image regardless of their positions, scales and illumination. The face detector consists of three parts, i.e., skin color filtering, rejector cascade, and cascades of boosted face classifier. The skin color filter and the rejector cascade can clean up the non-skin regions in the color image and remove most of the non-face candidates while retaining all the face samples. The potential face-like locations will be examined in the final boosted face classifier.

We implement the face detector in the live video and mark the detected face regions by red rectangles shown in Fig. 2.[1] Each face as a candidate of human body region will be recorded as a reference for the next frame. Only the face with the highest probability in the successive frames (for example, 20 frames) will be considered as a human body. If human face cannot be detected, the other learning-based human body feature detector will be activated to find the candidate in the current frame, more details referred to [24].

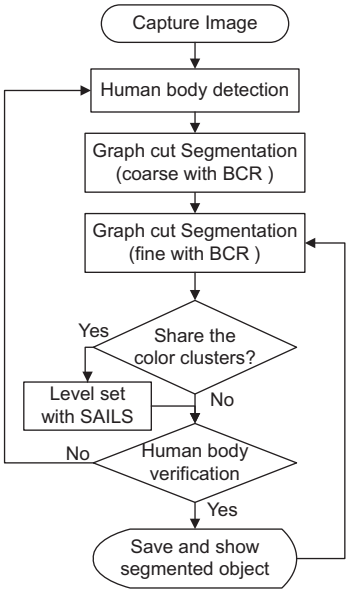


Fig. 1. The framework of the proposed automatic human body segmentation system.

To locate the human body, we define a frontal human body template shown in Fig. 3(a). The template is designed according to the structure of human body consisting of the center of human face (region 1), which is located by the face detector, and the hair part (region 2) is defined in the upper part of the detected face. The body parts (regions 3 and 4) are extracted from the area below the detected face with a size double that of the face, which also can be detected by [24]. Human body (with positive values) and neighboring regions (with negative values) are located as the hard constraints for the segmentation algorithm as described in Fig. 3(b) and the blue region (region 0) in Fig. 3(a), out of the computation region, will not be considered in order to save computation. The characteristics of those regions will be classified by *k*-means and modeled as Gaussian distributions.

2.2. Object segmentation

We suppose that the background image $\mathcal{I}^{\mathcal{B}}$ is known and stationary, and the current image $\mathcal{I}$ is to be processed. $\mathcal{I}_i^{\mathcal{B}}$ and $\mathcal{I}_i$ denote the color pixel value of the *ith* pixel in $\mathcal{I}^{\mathcal{B}}$ and $\mathcal{I}$, respectively. Let Ω be the set of all pixels in $\mathcal{I}$ and Λ be the set of all the adjacent pixel pairs (4- or 8-neighbor). So the object and background separation can be considered as a binary labeling problem, namely to assign a unique label x_i to each pixel $i \in \Omega$, i.e., $x_i \in \{\mathcal{O}, \mathcal{B}\}$.

After the object (human body) is detected, the characteristics of the object and background, like color distributions $\mathcal{P}(\mathcal{I}|\mathcal{O})$ and $\mathcal{P}(\mathcal{I}|\mathcal{B})$, are modeled from the detected regions and its surrounding regions, respectively. Background color can be modeled as a mixture of a global color model $P(\mathcal{I}|x = \mathcal{B})$ and a more accurate per-pixel color model $p(\mathcal{I}_i|x_i = \mathcal{B})$ as done in [12,14] which can produce a more accurate segmentation.

2.2.1. The basic model for segmentation

Generally, two important properties of the image to be considered for the segmentation task are homogeneity within the object

[1] Please note that Figures 2–17 will appear in B/W in print and colour in the web version. Based on this, please approve the footnote 1 which explains this.

Fig. 2. Face detection in the real-time video (only one of the detected faces will be selected for segmentation).

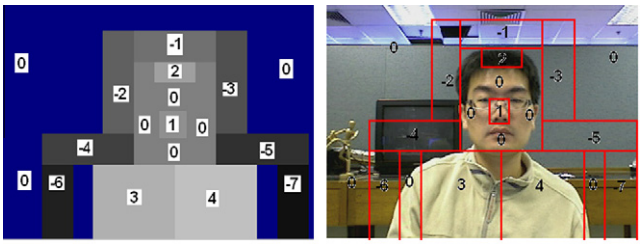


Fig. 3. The template of the frontal face and shoulder; the regions with positive value denote the human body and minus value regions present background.

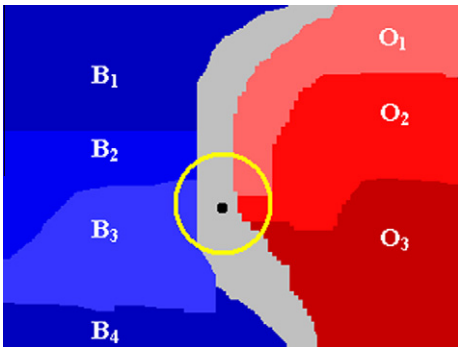


Fig. 4. The spatial weighting coefficient computation, O_i and B_i denote the object and background clusters.

and contrast at the object boundary. So the general formulation of energy function for image segmentation can be described as follows:

$$E(N) = \sum_{i\in\Omega} E_1(n_i) + \lambda \sum_{(i,j)\in\Lambda} E_2(n_i, n_j). \tag{1}$$

Color Term: The first term $E_1(n_i)$ called *color term* in the right-hand side of (1) to measure the smoothness in the region Ω. This term defines the likelihood of a given node n_i belonging to the object and background. Gaussian mixture models (GMMs) are used to model the detected human body region and its surroundings. Hence, for a given node n_i (pixel or region) with the color value $\mathcal{I}_i$ and label x_i, the likelihood represented by $p(n_i|x_i = \mathcal{O})$ and $p(n_i|x_i = \mathcal{B})$ to the object $\mathcal{O}$ and background $\mathcal{B}$ are defined as follows:

$$p(n_i|x_i = \mathcal{O}) = \frac{\log(d_i^{\mathcal{O}})}{\log(d_i^{\mathcal{O}}) + \log(d_i^{\mathcal{B}})},$$
$$p(n_i|x_i = \mathcal{B}) = \frac{\log(d_i^{\mathcal{B}})}{\log(d_i^{\mathcal{O}}) + \log(d_i^{\mathcal{B}})} \tag{2}$$

with

$$d_i^{\mathcal{O}} = \sum_{k=1}^{K^{\mathcal{O}}} \omega_k^{\mathcal{O}} \cdot N(\mathcal{I}_i|\mu_k^{\mathcal{O}}, \Sigma_k^{\mathcal{O}}) \quad \text{and} \quad d_i^{\mathcal{B}} = \sum_{k=1}^{K^{\mathcal{B}}} \omega_k^{\mathcal{B}} \cdot N(\mathcal{I}_i|\mu_k^{\mathcal{B}}, \Sigma_k^{\mathcal{B}}),$$

where ω_k denotes the weight corresponding to the percentage of the spatial samples for the *kth* component of the GMMs. μ and Σ represent the mean and the covariance matrices, respectively. $K^{\mathcal{O}}$ and $K^{\mathcal{B}}$ denote the number of the components of the GMMs used in the object and background, respectively. An example for how to calculate the weight ω_k is illustrated in Fig. 4. The red and blue colors represent the clusters of the object and background. The gray region is computation region. The weight of the pixel (labeled as black color) can be calculated according to the percentage of the clustered pixels within the yellow circle region.

For a known and stationary background, a pixel-wise single isotopic Gaussian distribution $p_{\mathcal{B}}(\mathcal{I}_i)$ is also added to model background:

$$p_{\mathcal{B}}(\mathcal{I}_i) = N(\mathcal{I}_i|\mu_i^{\mathcal{B}}, \Sigma_i^{\mathcal{B}}), \tag{3}$$

where $\mu_i^{\mathcal{B}} = \frac{1}{N}\sum_{n=1}^{N} \mathcal{I}_{n,i}^{\mathcal{B}}$, and $\Sigma_i^{\mathcal{B}} = \sigma_{N,i}^2$ (N is the number of frames to be learnt). The means $\mu_i^{\mathcal{B}}$ and variances σ_i^2 were learnt at the initialization stage. The pixel-wise color model is more precise than the global color model but it is sensitive to noise, like luminance change and small movement in the background. The global background color model is more robust to those disturbances but less precise. Therefore, the mixture of the two models can achieve a better performance than one of the two models separately.

$$E_1(n_i) = \alpha p(n_i|x_i = \mathcal{B}) + (1 - \alpha)p_{\mathcal{B}}(n_i), \tag{4}$$

where α is a weighing factor between the global and the pixel-wise background color model.

Contrast Term: The second term $E_2(n_i, n_j)$ in the right-side of (1) is called *contrast term* which sets a penalty for discontinuity between the two adjacent nodes n_i and n_j (two pixels or two regions). It becomes larger when the smaller change is between two nodes, which means a smaller probability of an edge appearing between the two. The general format is defined in terms of the local intensity gradient as the existing work [5,6,8,16]. We employ the exponential function based on the gradient intensity as follows:

$$E_2(n_i, n_j) = |x_i - x_j| \exp(-\beta d_{ij}), \tag{5}$$

with

$$d_{ij} = \|\mathcal{I}_i - \mathcal{I}_j\|^2,$$

where β is a robust parameter to weigh the color contrast, and can be set to $\beta = (2\langle\|\mathcal{I}_i - \mathcal{I}_j\|^2\rangle)^{-1}$ [6], where $\langle\rangle$ is the expectation operator. The parameter λ in (1) weighs the influence of the contrast term.

2.2.2. Course-to-fine segmentation

The coarse-to-fine segmentation strategy we proposed in [16] is very efficient to segment the partly detected object. For body segmentation, the body region is derived from the detected human features and the body template, which is different from the interactive ways in [5–7]. For example, to specify an object in [7], a user will mark several different color lines on the image to indicate whether the marked regions are object or background. If the segmented results are not satisfactory, the user can add or remove the marked regions and try segmenting it again. Unlike the interactive way, we build the GMMs model from the partly detected body regions to segment the object out without supervision.

Coarse-to-fine segmentation can provide a good solution which consists of two steps, i.e., coarse segmentation and fine segmentation. In coarse segmentation, a mixture of GMMs is employed to model object and background from the four regions with positive values, i.e., 1, 2, 3 and 4 and the surrounding regions labeled with minus values i.e. −1, −2 in Fig. 3, respectively. The corresponding means and variances are considered as the initial clusters for the object and background. Then, for each pixel in the computation region, we calculate the weighted likelihood to the object and background clusters according to (2) and (4) and the contrast with neighboring pixels according to (5). Note that in the color term E_1, the weight of each component is estimated based on the spatial samples within a defined circle ($r = 20$ pixels) centered in the current node. And 4 and 7 components are used to model object and background by GMMs, respectively. Finally, we use the min-cut [25] to perform the global optimization.

Finer segmentation aims to refine the result of the coarse segmentation. The object regions are defined as the set of pixels belonging to the body terminal obtained from the coarse segmentation, while the background region consisting of those surroundings are classified to the background terminal. We use 10 and 20 components of GMMs to describe the human body and background, respectively. The means and variances of GMMs are estimated using the *k*-means algorithm. Then, the similar process in the coarse segmentation is adopted to compute the term E_1 and E_2 and get the global optimum by the min-cut algorithm.

Generally, it has a good performance using above basic model to segment images with smooth background and sharp contrast at the object boundary. However, when the object boundary meet strong edges or the similar colors in the background, notable segmentation errors may occur around it as shown in the second column of Fig. 5(b). This would generate annoying visual artifacts in the video sequence. Why does this happen? The reason is that the basic model of graph cut contains two terms for the color and contrast. Inevitably, when object boundaries pass strong edges or suffer from similar colors, the two energy terms will be influenced heavily and bias the final segmented result in respect that the strong edge results in the minimized contrast term E_2 and the similar colors make for a minimum color term E_1. To solve the problems, we propose background contrast removal (BCR) and self-adaptive initialization level set (SAILS) to achieve a good performance in the complex background.

2.2.3. Background contrast removal (BCR)

Since the basic model considers both color term and contrast term simultaneously, a potential problem is that the final segmentation boundary will be attracted to high contrast edges in a cluttered background. Although this kind of error may be small around the boundary or only occur in partial frames, the flickering artifact in the video due to this error can be very distracting and unpleasant in the final composited video.

Given a known background, background subtraction is a straightforward way to remove the contrast existing in the background $\mathcal{I}^{\mathcal{B}}$ from the current image $\mathcal{I}$. Nevertheless, background subtraction is very sensitive to noise and illuminance changes. And similar color existing in object and background results in holes in the detected object. Although more sophisticated techniques [26–28] have been proposed to overcome these problems, the results are still error-prone. In the following section, we propose a new effective way to remove the contrast in background and get an improved segmented result.

In the basic model, every pixel in background has built a Gaussian distribution model as in (3). In the current image, the likelihood of $\mathcal{I}_i$ belonging to background $p(\mathcal{I}_i|x_i = \mathcal{B})$ has already been computed in the color term E_1 of (4). A normalized background probability distribution $\overline{p_{\mathcal{B}}}(\mathcal{I}_i)$ is employed into the contrast term of the basic model to remove the contrast in background. The modified contrast term can be expressed as:

$$E_2(n_i, n_j) = |x_i - x_j| \exp(-\beta d_{ij}) \tag{6}$$

with

$$d_{ij} = \frac{\|\mathcal{I}_i - \mathcal{I}_j\|^2}{1 + \kappa \overline{p_{\mathcal{B}}}(\mathcal{I}_i)},$$

where κ is a scaling factor and sets $\kappa = \max\{\|\mathcal{I}_i - \mathcal{I}_j\|^2\}$. The contrast d_{ij} should be decreased when the pixel has a higher probability belonging to background and be preserved when the pixel belongs to the object. However, using above model, the edge of object and background is decreased too. To solve this problem, an adaptive factor is introduced to d_{ij}:

$$d_{ij} = \frac{\|\mathcal{I}_i - \mathcal{I}_j\|^2}{1 + \kappa \overline{p_{\mathcal{B}}}(\mathcal{I}_i) \exp\left(-\frac{\mathcal{D}_{ij}}{\sigma^2}\right)}, \tag{7}$$

where $\mathcal{D}_{ij} = \max\left\{\|\mathcal{I}_i - \mathcal{I}_i^{\mathcal{B}}\|^2, \|\mathcal{I}_j - \mathcal{I}_j^{\mathcal{B}}\|^2\right\}$. When the pixel pair (I_i, I_j) has a high probability of belonging to the background, the strength of decreasing contrast should be large $\left(\overline{p_{\mathcal{B}}} \to 1, \exp\left(-\frac{\mathcal{D}_{ij}}{\sigma^2}\right) \to 1\right)$. Otherwise, the contrast probably caused by the object and background should be preserved $\left(\exp\left(-\frac{\mathcal{D}_{ij}}{\sigma^2}\right) \to 0\right)$. The experimental result is demonstrated in the Fig. 5(c).

We compare the proposed method with Sun's [12] on the test sequences. The contrast maps and the corresponding segmented

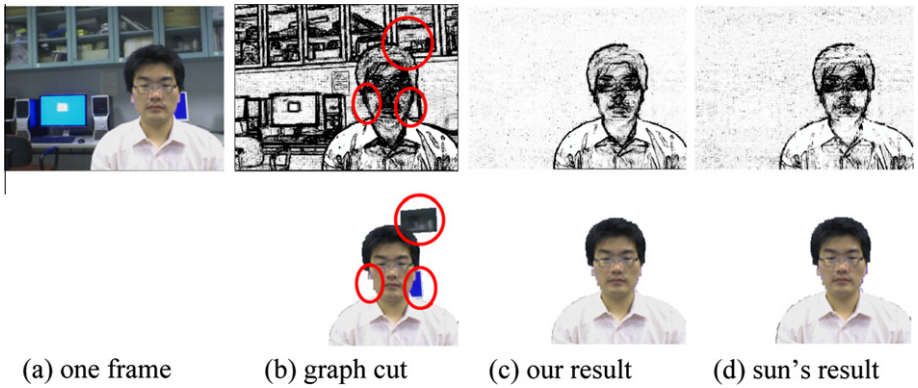


Fig. 5. Comparisons of the basic model, graph cut with BCR and [12]: ((a) is the input frame; (b) are the contrast map (upper) and the segmented result (down) by the basic model of graph cut; (c) are BCR result and the corresponding segmented result; (d) are contrast map and segmented result by [15]. (For display, the contrast map for each pixel n is computed as $\sqrt{d_{n,n_x}^2 + d_{n,n_y}^2}$, where n_x and n_y are two adjacent pixels on the left and below of pixel n).

results are shown in Fig. 5(c) and (d), respectively. Our method can achieve the comparable performance to Sun's but has 20% saving in computation. Comparing (b) with (c) in Fig. 5, the body segmentation errors caused by background contrast can be substantially reduced by BCR.

2.3. Self-adaptive initialization level set (SAILS)

Segmentation can be interpreted as a region-based or contour-based process. In a region-base approach [12,29,30], the algorithm, likes graph cut we have discussed, aims at learning the statistics of the object and background so that it is able to find a boundary to distinguish between the two. But region-based algorithms cannot provide directly control over the boundary location. Instead, the user updates the initialized pixels to change the boundary in the hope of getting a better segmented result. The problem may be aggravated when the statistics of the object and background are similar. Contour-based methods aim to find the contour which is usually presumed to be the most salient edge in the image. Contour-based methods allow the user more directly control of the boundary by defining different types of energy functions. With reference to the above discussion, we propose an efficient initialization scheme known as self-adaptive initialization level set (SAILS) to be integrated into the body segmentation system.

2.3.1. Level set without re-initialization in evolution

The Gateaux derivative (or first variation) [31] of the functional $\varepsilon(\varphi)$ is denoted by $\frac{\partial\varepsilon(\varphi)}{\partial\varphi}$ and the following evolution equation

$$\frac{\partial\varphi}{\partial t} = -\frac{\partial\varepsilon(\varphi)}{\partial\varphi} \tag{8}$$

is the *gradient flow* that minimizes the functional $\varepsilon(\varphi)$. Now, the total energy functional to control the evoluted curve can be defined as:

$$\varepsilon(\varphi) = \gamma\mathcal{P}(\varphi) + \varepsilon_{g,\eta,v}(\varphi), \tag{9}$$

where $\mathcal{P}(\varphi)$ is *internal energy* term to penalize the deviation of the level set function from a signed distance function and $\varepsilon_{g,\eta,v}(\varphi)$ is *external energy* that drives the motion of the zero level curve of φ. $\gamma(>0)$ is a weighting parameter that controls the effect of internal energy.

Internal energy: $\mathcal{P}(\varphi)$ is defined as (10) to characterize how close a function φ is to a signed distance function [32] in $\Omega \subset \Re^2$ and as the penalty for the deviation of φ from a signed distance function in (9).

$$\mathcal{P}(\varphi) = \int_\Omega \frac{1}{2}(|\nabla\varphi| - 1)^2\, dxdy. \tag{10}$$

Due to the penalizing effect of the internal energy, the evolving function φ will be automatically maintained as an approximate signed distance function. Therefore the re-initialization procedure is completely eliminated in the evolution process.

External energy: $\varepsilon_{g,\eta,v}(\varphi)$ drives the zero level set towards the object boundaries. We define the external energy for a function $\varphi(x,y)$ as below:

$$\varepsilon_{g,\eta,v}(\varphi) = \eta\mathcal{L}_g(\varphi) + v\mathcal{A}_g(\varphi), \tag{11}$$

where $\eta(>0)$ and v are constants, and the term $\mathcal{L}_g(\varphi)$ and $\mathcal{A}_g(\varphi)$ are defined by:

$$\mathcal{L}_g(\varphi) = \int_\Omega g\delta(\varphi)|\nabla\varphi| dxdy \tag{12}$$

and

$$\mathcal{A}_g(\varphi) = \int_\Omega g\mathcal{H}(-\varphi) dxdy \tag{13}$$

respectively, where $\delta(\varphi)$ is the Dirac function, $\mathcal{H}$ is the Heaviside function and g is the edge indicator function which is defined as:

$$g = \frac{1}{1 + |\nabla\mathcal{G}_\sigma * \mathcal{I}|^2},$$

where $\mathcal{G}_\sigma$ is the Gaussian kernel with standard deviation σ. For physical meanings of the energy terms in (11), the energy term $\mathcal{L}_g(\varphi)$ penalizes the length of the zero level curve of φ and $\mathcal{A}_g(\varphi)$ is the area of $\Omega_\varphi^- = \{(x,y)|\varphi(x,y) < 0\}$ which can be considered as the weighted area of Ω_φ^-. By calculus of variations [31], the Gateaux derivative of the functional $\varepsilon(\varphi)$ in (9) can be written as

$$\frac{\partial\varepsilon(\varphi)}{\partial\varphi} = -\gamma\left[\Delta\varphi - div\left(\frac{\nabla\varphi}{|\nabla\varphi|}\right)\right] - \eta\delta(\varphi)div\left(g\frac{\nabla\varphi}{|\nabla\varphi|}\right) - vg\delta(\varphi),$$

where Δ is the Laplacian operator. Therefore, the function φ that minimizes this functional satisfies the Euler–Lagrange equation $\frac{\partial\varepsilon(\varphi)}{\partial\varphi} = 0$. The steepest descent process for minimization of the functional $\varepsilon(\varphi)$ is the following gradient flow:

$$\frac{\partial\varphi}{\partial t} = \gamma\left[\Delta\varphi - div\left(\frac{\nabla\varphi}{|\nabla\varphi|}\right)\right] + \eta\delta(\varphi)div\left(g\frac{\nabla\varphi}{|\nabla\varphi|}\right) + vg\delta(\varphi). \tag{14}$$

This gradient flow is the evolution equation of the level set function.

In the experiment, we approximate the Dirac function $\delta(\varphi)$ in (14) to be slightly smoothed as given by the following function $\delta_\alpha(\varphi)$:

$$\delta_\alpha(\varphi) = \begin{cases} 0, & |\varphi| > \alpha, \\ \frac{1}{2\alpha}\left[1 + cos(\frac{\pi\varphi}{\alpha})\right], & |\varphi| \leqslant \alpha, \end{cases} \tag{15}$$

we use the regularized Dirac $\delta_\alpha(\varphi)$ with $\alpha = 1.5$ for all the experiments in this paper.

On the other hand, all the spatial partial derivatives $\frac{\partial\varphi}{\partial x}$ and $\frac{\partial\varphi}{\partial y}$ are approximated by the central difference, and the temporal partial derivative $\frac{\partial\varphi}{\partial t}$ is approximated by the forward difference. The approximation of (14) by the above difference scheme can be simply written as

$$\frac{\varphi_{i,j}^{k+1} - \varphi_{i,j}^k}{\tau} = \mathcal{L}(\varphi_{i,j}^k), \tag{16}$$

where $\mathcal{L}(\varphi_{i,j}^k)$ is the approximation of the right hand side in (14) by the above spatial difference scheme. The time step τ can be chosen significantly larger than that used in the traditional level set methods. We use $\tau = 50$ in our experiments. The difference Eq. (16) can be expressed as the following evolutive iteration:

$$\varphi_{i,j}^{k+1} = \varphi_{i,j}^k + \tau\mathcal{L}(\varphi_{i,j}^k). \tag{17}$$

2.3.2. Self-adaptive initialization

Comparing with traditional level set methods [33–36], the above method completely eliminates the procedure to re-initialize the level set function φ as a signed distance function. In the video segmentation, another initialization is need between each frame. It is common practice to use the segmented result of the previous frame as the initial state φ_0 and impose the dilation operation on it to predict disturbance of the object in the current frame. Let Ω_0 be a subset (an object) in the image domain Ω, and $\partial\Omega_0$ be all the points on the boundaries of Ω_0, which can be efficiently identified by some simple morphological operations. Then, the initial function φ_0 is defined as

$$\varphi_0(x,y) = \begin{cases} -\rho, & (x,y) \in \Omega_0 - \partial\Omega_0, \\ 0, & (x,y) \in \partial\Omega_0, \\ \rho, & (x,y) \in \Omega - \Omega_0, \end{cases} \tag{18}$$

where ρ is a constant. We set $\rho = 4$ and suggest choosing ρ larger than 2α, where α is the width of the window in (15) the definition of the regularized Dirac function δ_α.

To demonstrate the effectiveness of the proposed initialization scheme, we apply it to the given sequence in Fig. 6. Generally, the number of iterations for level set should be set as the maximum iterations of all frames in the sequence to ensure good segmentation results in every frame. But this would induce unnecessary iterations and might result in over-convergence.

To avoid above potential problems, an adaptive iterations for level set in each frame is needed. Simultaneously, a measurement should be proposed to judge whether the evolution is convergent or not. When it is convergent, the iterations is stopped right away. At each evolution stage, the area of object is increased or decreased depending on v in (11). But the variation of the converging area should be smaller when the evolution approaches the end. Some frames are sampled to exploit the relationship between the areas changing and the number of iterations shown in Fig. 7(a). As expected, we can define a convergence threshold Thd_c to approximatively measure whether the evolution is convergent or not. Using different Thd_c, we can get different grades of convergent results described as Fig. 7(b)–(d). Generally, good result can be obtained with $Thdc = 20$.

In order to further assess the proposed convergence measurement scheme for level set, we implement it on the test sequence using different dilations and thresholds. Fig. 8 gives the number of iterations in each frame at different preset values of (dilation, Thd_c) pairs.

From the segmented results in Fig. 8(a) and (b), we can see that with the same Thd_c, different initialization for level set can obtain the same segmentation results. This means our proposed convergence measurement scheme can control the convergence state of level set. In all our experiments, we set the default values of Thd_c to 20 to obtain good segmentation results on average. The range of Thd_c is quite narrow and stable. There is no notable change in segmentation results when we change the threshold from 15 to 25.

Additionally, from Fig. 8(d) (the green line with triangle and the brown line with square, the blue line with circle and the black line

Fig. 6. The initialization for level set between frames ($\varepsilon = 1.5, \tau = 50; \mu = 0.24/\tau; \lambda = 5; \mu = 1.5$ the same settings for the following experiments).

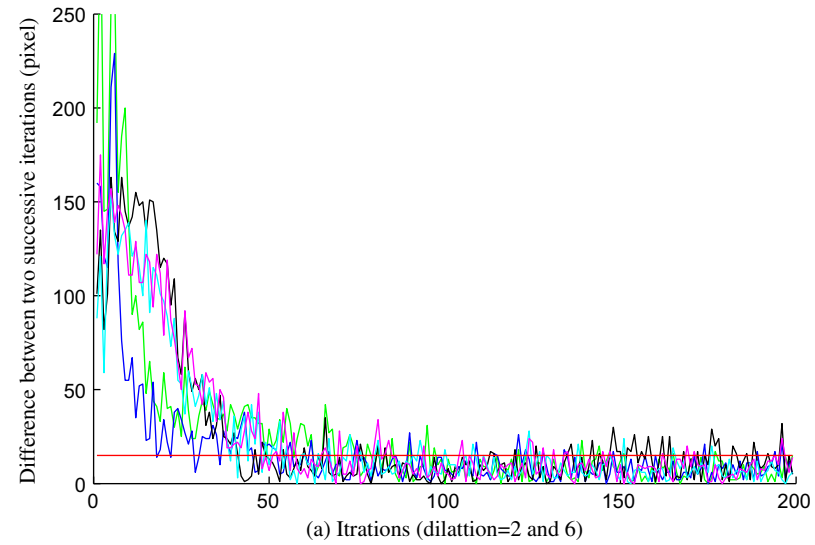
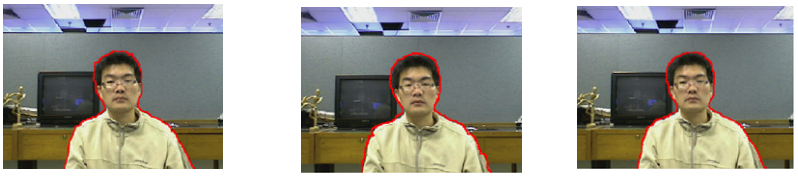
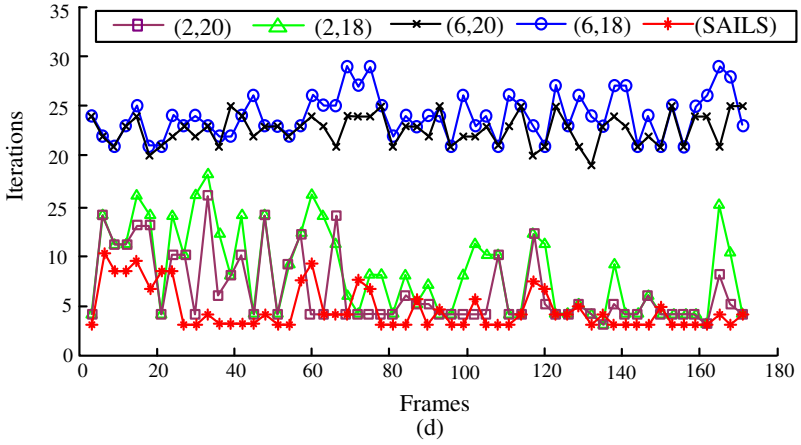


Fig. 7. The relationship between area difference and the number of iterations. The red line is the Thd_c line in (a). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) (dilation=2 and 6, Thd_c =20) (b) (dilation=2 and 6, Thd_c =18) (C) SAILS' result



(d)

Fig. 8. The relationship of iterations, dilation and Thd_c.

with cross) we find that the number of iterations is proportional to the degree of dilation which directly affects the speed of evolution. Hence, an adaptive dilation is needed to replace the fixed one.

Intuitively, the optimal initial mask is that just covers the object in the current frame. In the proposed frame initialization process, it includes two steps: (a) position estimation and (b) adaptive dilation. Here, we propose to use motion estimation and compensation (MEC) [37] to predict the initial mask replacing using the result of the previous frame. MEC is an extensively used method in video coding to estimate the movement of an object in two successive frames and predict the location of the object in the current frame, which can provide us with a much better result than the previous one when the object has a big movement. In the experiment, block-based MEC is employed and the block size is 8×8. Additionally, dilation is performed to cover the error induced by blocky MEC. We set the dilation strength $d = \sigma_V + 1$, where σ_V is the standard deviation of the motion vectors and 1 is the residual dilation. The segmentation result and corresponding iterations used in evolution are shown in Fig. 8(c) and (d) (the red line with star). Comparing with the fixed dilation, the number of iterations in each evolution process is decreased greatly, less than 20 with the same final segmented result.

Our proposed SAILS not only controls the segmentation result adaptively but also employs the optimum number of iterations to each frame to save computation. As an initialization schemer for graph cut segmentation, SAILS is only activated when the GMMs of object and background share the similar color clusters in the computation region.

2.4. Object updating

When a human body is extracted from the background, its properties such as motion, size and shape similarity are verified to be those of the human body in the successive frames. In our system, for simplification only the human body size is employed to confirm the final segmentation. Human body detector will be activated to search and re-initialize new object for segmentation if the condition is not satisfied. Fig. 9 demonstrates the efficiency of our updating scheme, when a man disappears and reappears in the scene.

3. Experimental results

3.1. Evaluation using real-time videos and standard sequences

We implement the proposed body segmentation on the real-time videos, which were captured by a Philips web camera. The frame rate of the camera is set to 30 fps with the frame size of QVGA (320×240) and "full automatic control" is turned off to keep the background stationary. And standard test sequences [40] with human body are also employed to test the performance of the proposed system in Fig. 16.

The body segmentation system is evaluated in the cases of indoor and outdoor with different poses, sizes, movements and the different lighting with object moving. Some representative frames are shown in the odd rows of Fig. 17. The corresponding initial frame and segmentation results are given in the following even

Fig. 9. Automatic updating scheme for new human body.

rows. From the segmented results, we can see that the human body is segmented accurately in above cases.

3.2. Comparison with other methods

We firstly compare our method with the state-of-the-art segmentation methods, like the region-based graph cut [12] and the contour-based level set [32]. Some experimental results are listed in Fig. 10 and Fig. 11 and the improvements are enclosed by white circles. In Fig. 10, the first row is the original frames and the corresponding segmented results by [12] and our method are listed in the second row and the last row, respectively. Because of the attribution of level set, our method can deal with the similar color (in the white circles) at the object boundary much better than the region-based graph cut. In Fig. 11 the red curve is the final segmented results by [32] and our method. Comparing with [32], our method can avoid suffering the effect of strong edge at the object boundary due to use BCR.

In order to present the differences from [16] and verify the importance of our proposed body segmentation system, we test the two methods on the following two aspects:

1. test the two methods using the same sequence;
2. test the two methods with same initialization for object segmentation using same sequence;

In test 1, we run the two methods on the sequence which was presented in Fig. 10, and the segmented results are shown in Fig. 12 in which the 1st row and last row are the output results of [16] and our method, respectively. From the two sets of images, we can see that [16] only can provide human face segmentation, but our methods can deal with human body segmentation. In order to further assess the core algorithm (segmentation part) of the two approaches, we modified the initialization part of [16] to make them have the same initialization input for segmentation in test 2. We run [16] on the above sequence and segmented results are listed in the 2nd row of Fig. 12. Compare the 2nd row of Fig. 12 with the 2nd row of Fig. 10, we can see that the similar results were produced by [12] and [16], because they used the same segmentation algorithm originally proposed in [5] which suffers the bias to high contrast at boundary and similar color. But in our system, a much better results are produced because we improve graph cut by the BCR which can remove the effect of the contrast at the object boundary and the SAILS which can deal with similar colors segmentation. One More example is given in Fig. 13.

To be more convincing, we quantitatively evaluate the accuracy of our approach with the two most popular body segmentation methods, i.e., the "background cut" [12] and the "Bi-layer segmentation" [14] on "AC" video (The left view contains complex background and similar color with the human body in the abnormal lighting condition.) which is a stereo video sequence provided by [14]. The ground truth foreground/background segmentation is provided in every 5 frames. The segmentation error measurement we adopted is the same as in [14] which is measured as the percentage of misclassified pixels with respect to the ground truth. For the known background, we use the image mosaic from the

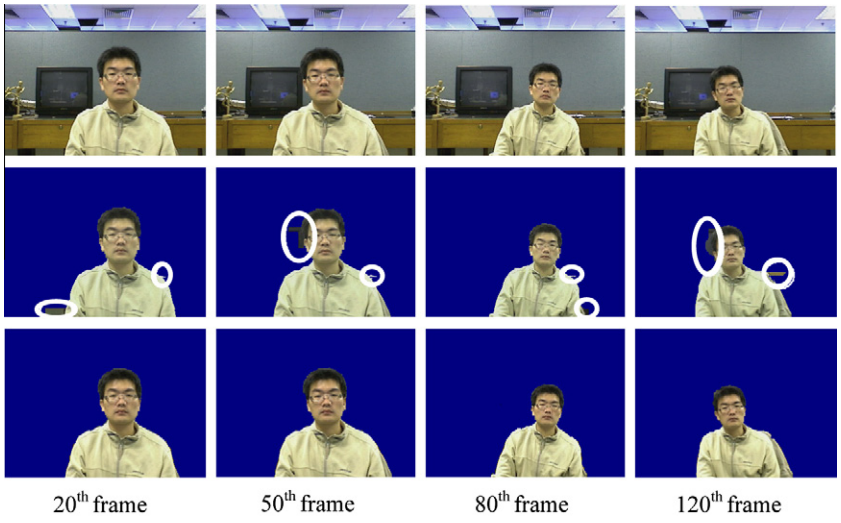


Fig. 10. Comparisons our method (3rd row) with graph cut [12] (2nd row).

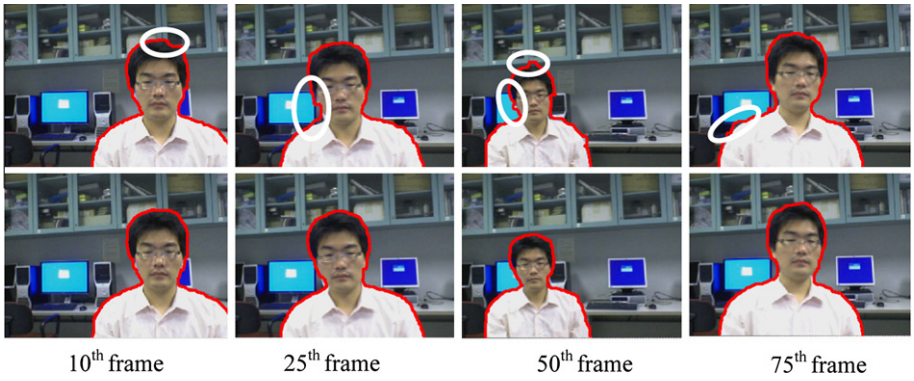


Fig. 11. Comparisons our method (2nd row) with level set [32] (1st row).

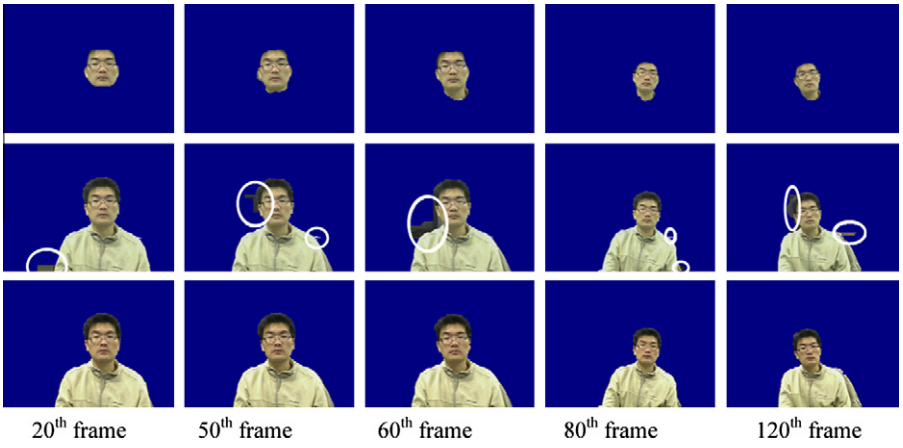


Fig. 12. Compare our approach with [16] using the same sequence in Fig. 10 (the 1st row is the segmented results of [16]; the 2nd row is the segmented results of [16] with the same initialization as our method; and the last row is our segmented results.

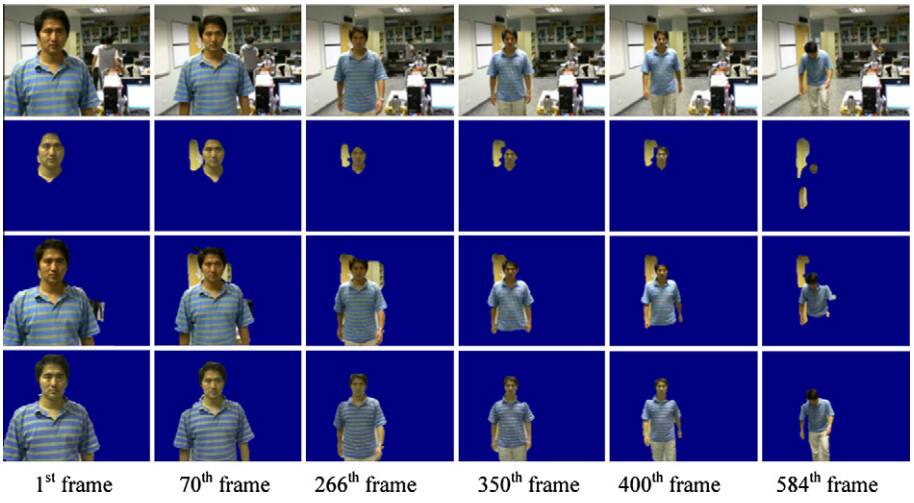


Fig. 13. Experimental results of comparison our approach with [16](the 1st row is the input frame images; the 2nd row is the results of [16]; the 3rd row is the segmentation results of [16] with the same initialization as our method; and the last row is our results).

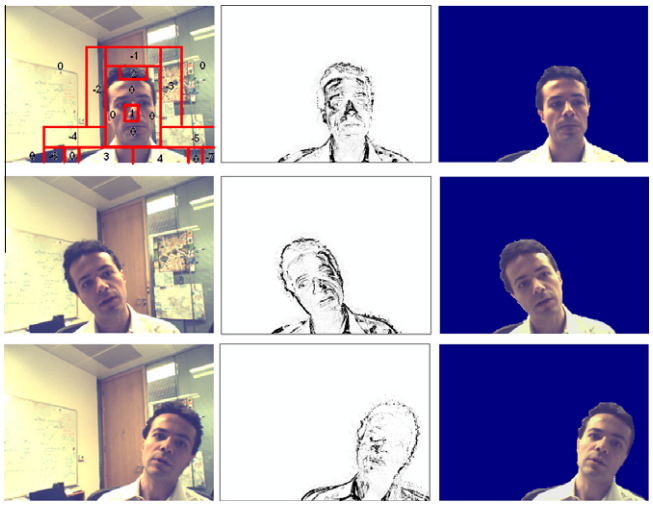


Fig. 14. Segmentation results of "AC" (the first column is the original frames, the second column is the removed contrast maps and the last column is the segmented results).

frames extracted from the left view of the video. The results of BCR and the final segmentation in "AC" are shown in Fig. 14. The original video and the ground truth segmentation are obtained from: http://research.microsoft.com/vision/cambridge/i2i/DSWeb.htm.

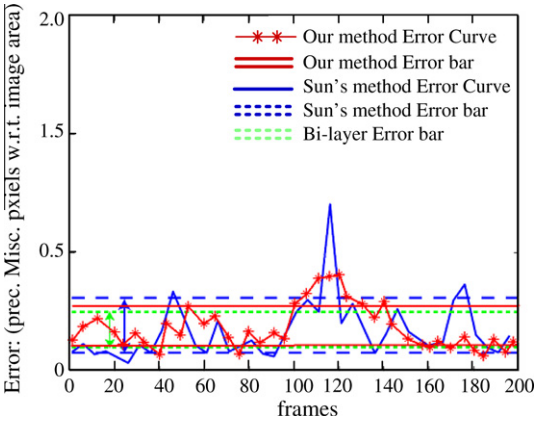


Fig. 15. Error statistics of comparison with "background cut" and "Bi-layer segmentation" on "AC".

We evaluate the segmented error statistically by comparing with the results of "background cut" and "Bi-layer segmentation" in Fig. 15. The red solid line with star is the error curve with respect to the ground truth in every 5 frames and two horizontal red lines represents the standard deviation error bar [14] of our method. The blue and green lines show the same for "background cut" and "Bi-layer segmentation", respectively.

From the error statistics figure, we can see that the proposed automatic body segmentation method is better than the "background cut" method due to the use of SAILS and comparable with the well-known "Bi-layer segmentation" method because of a lack of the depth information. More experimental results are given in Figs. 16 and 17 with the live videos and standard sequence in the different conditions, like different lighting in the indoor case and outdoor case.

3.3. Computational complexity analysis

In the proposed automatic body segmentation system, the body detection employs the color filtering and rejector cascades to speed up the object detection. So the computation load for the final boosting stage will be reduced significantly. The detection window scans the video image with a zooming factor of 1.5 and a shifting step of 2-pixel. The detector runs at 13*ms* in the initial frame with QVGA size on a CPU of 1.85 GHz.

The coarse-to-fine segmentation strategy will be performed first when a human body is located. In the successive frames, the tri-map tracking [16] based segmentation is employed, which speeds up the segmentation because only the pixels in the tri-map region need to be calculated. Adaptive dilation using MEC is employed to speed up the evolution of level set. On the above

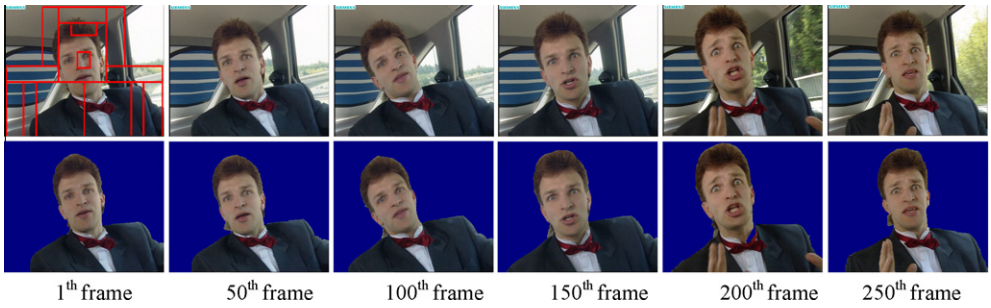


Fig. 16. Segmentation results for the standard video sequences (Carphone (352 × 288, 250 frames) and the background are extracted by the image mosaic from the selected frames).

(a) segmentation results for the real-time video sequences with different poses and clutter background

(b) segmentation results for the real-time video sequences in outdoor with moving objects in the background

Fig. 17. Segmentation results for real-time video sequences.

platform, it takes 40*ms* per-frame for QVGA on average. For the level set segmentation, the process is convergent with a few iterations (<20*ms*) due to self-adaptive initialization.

3.4. Extensions

Matting [16,38,39] is an important operation to separate a foreground object with an arbitrary shape from an image by estimating the opacity of the foreground element at every pixel. In our system, we only use a simple matting method to matte the human body by computing the transparency of the boundary pixels according to the distance to the object. A new matting algorithm should be designed to estimate the alpha-matte based on the binary object mask.

At present, our system can detect multiply human bodies but only can segment one object at a time with stationary background. In the future work, Multi-object segmentation in the dynamic background, more challenging cases, will be carried on. Because of the limited length of this paper, more details on further development can be found on the website [41].

4. Conclusion

In this paper, we proposed an automatic human body segmentation system for real-time videos. Firstly, a human body detector based on the structure of the human body is designed and learning-based human body features detection is proposed to find the object in the scene. Secondly, a coarse-to-fine segmentation approach is adopted to solve the partly detection problem and background contrast removal is proposed to improve the performance of graph cut algorithm in the clustered background. Thirdly, due to the shortcoming of region-based segmentation, ambiguous boundary, we proposed the SAILS to find the boundary with similar color. Lastly, in our system an object updating scheme for segmentation is proposed to detect and re-initialize new object.

The experimental results demonstrated that our system can provide effective and robust segmentation for the human body, and can achieve similar or superior results compared to those of the well-known techniques.

References

[1] Ç.E. Erdem, F. Ernst, A. Redert, E. Hendriks, Temporal stabilization of video object segmentation for 3D-TV applications, Signal Processing: Image Communication 20 (2) (2005) 151–167.
[2] J.Y.A. Wang, E.H. Adelson, Layered representation for motion analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, New York, 1993, pp. 361–366.
[3] J. Wills, S. Agarwal, S. Belongie, What went where, in: IEEE Conference on Computer Vision and Pattern Recognition, 2003, pp. 37–44.
[4] H. Li, K.N. Ngan, Automatic video segmentation and tracking for content-based multimedia services, IEEE Communications Magazine, USA 45 (1) (2007) 27–33.
[5] Y. Boykov, M.P. Jolly, Interactive graph cuts for optimal boundary & region segmentation of objects in *n-d* images, in: Proceedings of the International Conference on Computer Vision, vol. 1, 2001, pp. 105–112.
[6] C. Rother, V. Kolmogorov, A. Blake, GrabCut - interactive foreground extraction using iterated graph cuts, in: Proceedings of ACM SIGGRAPH, 2004.
[7] Y. Li, J. Sun, C.K. Tang, H.Y. Shum, Lazy snapping, in: Proceedings of ACM SIGGRAPH, 2004.
[8] Y. Li, J. Sun, H.Y. Shum, Video object cut and paste, in: Proceedings of ACM SIGGRAPH, 2005, pp. 595–600.
[9] J. Wang, P. Bhat, R.A. Colburn, M. Agrawala, M.F. Cohen, Interactive video cutout, in: Proceedings of ACM SIGGRAPH, 2005, pp. 585–594.
[10] Huitao Luo, Alexandros Eleftheriadis, An interactive authoring system for video object segmentation and annotation, Signal Processing: Image Communication 17 (7) (2002) 559–572.
[11] H. Li, K.N. Ngan, Unsupervised video segmentation with low depth of field, IEEE Transactions on Circuits and Systems for Video Technology, USA 17 (12) (2007) 1742–1751.
[12] J. Sun, W. Zhang, X. Tang, H.-Y. Shum, Background cut, in: Proceedings of European Conference on Computer Vision, vol.2, 2006, pp. 628–641.
[13] Q. Zhang, K.N. Ngan, Multi-view video based multiple objects segmentation using graph cut and spatiotemporal projections, Journal of Visual Communications and Image Representation, USA 21 (5–6) (2010) 453–461.
[14] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, C. Rother, Bi-layer segmentation of binocular stereo video, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 1186–1193.
[15] A. Criminisi, G. Cross, A. Blake, V. Kolmogorov, Bilayer segmentation of live video, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 53–60.
[16] H. Li, K.N. Ngan, Q. Liu, FaceSeg: Automatic face segmentation for real-time video, IEEE Transactions on Multimedia, USA 11 (1) (2009) 77–88.
[17] H. Li, K.N. Ngan, Saliency model based face segmentation and tracking in head-and-shoulder video sequences, Journal of Visual Communications and Image Representation, Europe 19 (5) (2008) 320–333.
[18] X. Ji, Z. Wei, Y. Feng, Effective vehicle detection technique for traffic surveillance systems, Journal of Visual Communication and Image Representation 17 (3) (2006) 647–658.
[19] Changick Kim, Jenq-Neng Hwang, Fast and automatic video object segmentation and tracking for content-based applications, IEEE Transactions on Circuits and Systems for Video Technology 12 (2) (2002).
[20] N. Atzpadin, P. Kauff, O. Schreer, Stereo analysis by hybrid recursive matching for real-time immersive video conferencing, IEEE Transactions on Circuits and Systems for Video Technology 14 (3) (2004) 321–334.
[21] Wen-Chao Chen, Hong-Long Chou, Zen Chen, A quality controllable multi-view object reconstruction method for 3D imaging systems, Journal of Visual Communication and Image Representation 21 (5–6) (2010) 427–441.
[22] J. Wang, Yingqing Xu, Heung-Yeung Shum, Michael F. Cohen, Video tooning, in: ACM SIGGRAPH 2004 Papers, Los Angeles, California, August 08–12, 2004.
[23] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2001, pp. 511–518.
[24] L. Bourdev, J. Malik, Poselets: Body part detectors trained using 3D human pose annotations, in: ICCV, 2009. p. 5.
[25] Y. Boykov, V. Kolmogorov, An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (9) (2004) 1124–1137.
[26] Y. Sheikh, M. Shah, Bayesian object detection in dynamic scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 1778–1792.
[27] O. Tuzel, F. Porikli, Peter Meer, A bayesian approach to background modeling, in: IEEE Workshop on Machine Vision for Intelligent Vehicles, ISSN: 1063-6919, vol. 3, 2005, pp.58–63.
[28] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland, Pfinder: Real-time tracking of the human body, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1997) 780–785.
[29] Ç.E. Erdem, Video object segmentation and tracking using region-based statistics, Signal Processing: Image Communication 22 (10) (2007) 891–905.
[30] Ioannis Patras, Emile A. Hendriks, Reginald L. Lagendijk, Semi-automatic object-based video segmentation with labeling of color segments, Signal Processing: Image Communication 18 (1) (2003) 51–65.
[31] L. Evans, Partial Differential Equations, Providence: American Mathematical Society, 1998.
[32] C. Li, C.Y. Xu, C.F. Gui, M.D. Fox, Level set evolution without re-initialization: a new variational formulation, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 20–25 June 2005 pp. 430–436.
[33] J.A. Sethian, Level Set Methods and Fast Marching Methods, Cambridge University Press, Cambridge, 1999.
[34] S. Osher, R. Fedkiw, Level Set Methods and Dynamic Implicit Surfaces, Springer-Verlag, New York, 2002.
[35] T. Chan, L. Vese, Active contours without edges, IEEE Transactions on Image Processing 10 (2001) 266–277.
[36] B. Vemuri, Y. Chen, Joint image registration and segmentation, Geometric Level Set Methods in Imaging Vision and Graphics, Springer, New York, 2003. 251–269.
[37] I.E. Richardson, H.264 and MPEG-4 Video Compression: Video Coding for Next-generation Multimedia, John Wiley, Sons Ltd., Sussex, England, 2003.
[38] Y.Y. Chuang, B. Curless, D.H. Salesin, R. Szeliski, A Bayesian approach to digital matting, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2001, pp. 264–271.
[39] J. Sun, J. Jia, C.K. Tang, H.Y. Shum, Poisson matting, in: Proceedings of ACM SIGGRAPH, 2004.
[40] H. Li, K.N. Ngan, Saliency model based face segmentation and tracking in head-and-shoulder video sequences, Journal of Visual Communications and Image Representation, Europe 19 (5) (2008) 320–333.
[41] Available: <http://ivp.ee.cuhk.edu.hk/projects/itf.shtml>.