J. Vis. Commun. Image R. 21 (2010) 453–461

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

# Multi-view video based multiple objects segmentation using graph cut and spatiotemporal projections

Qian Zhang *, King Ngi Ngan

*Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong*

ARTICLE INFO

*Article history:*
Received 1 July 2009
Accepted 14 September 2009
Available online 22 September 2009

*Keywords:*
Multi-view pre-processing
Depth reconstruction
Saliency model
Object extraction
Graph cut
Spatiotemporal projections
Multi-view segmentation
Multi-view video

ABSTRACT

In this paper, we present an automatic algorithm to segment multiple objects from multi-view video. The *Initial Interested Objects* (IIOs) are automatically extracted in the *key view* of the *initial frame* based on the saliency model. Multiple objects segmentation is decomposed into several sub-segmentation problems, and solved by minimizing the energy function using binary label graph cut. In the proposed novel energy function, the color and depth cues are integrated with the data term, which is then modified with *background penalty with occlusion reasoning*. In the smoothness term, *foreground contrast enhancement* is developed to strengthen the moving objects boundary, and at the same time attenuates the background contrast. To segment the multi-view video, the coarse predictions of the other views and the successive frame are projected by pixel-based disparity and motion compensation, respectively, which exploits the inherent spatiotemporal consistency. Uncertain band along the object boundary is shaped based on *activity* measure and refined with graph cut, resulting in a more accurate *Interested Objects* (IOs) layer across all views of the frames. The experiments are implemented on a couple of multi-view videos with real and complex scenes. Excellent subjective results have shown the robustness and efficiency of the proposed algorithm.



## 1. Introduction

In the recent decades, image/video segmentation has become an active research topic in image processing, computer vision and computer graphics, leading to significant breakthroughs on the development of its theories and technologies. Robust and accurate separation of foreground object from background has turned out to be a crucial prerequisite for many applications such as face segmentation in videotelephony [1], video object cut for pasting [2], and 3D modeling and reconstruction by joint segmentation [3]. Current segmentation methods can be categorized into two groups, region-based segmentation and boundary-based segmentation. Region-based segmentation methods aim to directly construct the region itself, while boundary-based segmentation methods tend to represent each region by its boundary. Some of the classical region-based segmentation methods are mean-shift [4], region growing [5], and graph partition (graph cut [6], grab cut [7]), as well as some popular image cutout tools such as Magic Wand in Photoshop. Active contour (snake) [8], level set [9] and GVF [10] are the representative approaches for boundary-based segmentation. Lazy snapping [11] designs a novel user interface for image cutout by inheriting the advantages of region-based and boundary-based methods.

Most of the interest has been focused on the research of single view segmentation, thus many advanced algorithms have emerged [12–15]. On the contrary, multiple view segmentation has not attracted much attention due to the limitation of image capturing technology and the difficulty to segment all the images simultaneously in real-time. However, multi-view images capturing the real-world environment from arbitrary viewpoints are capable of describing dynamic scene from different angles and can provide the observer more vivid and extensive viewing experience than the single-view image, resulting in more realistic and exciting visual effect. Additionally, depth information in the 3D scene can be reconstructed from multi-view images and assists in characterizing the visual objects more efficiently than the conventional 2D representation. Furthermore, efficient segmentation of IOs has played an important role in many multi-view applications, such as image-based rendering and 3D object model reconstruction. In image-based rendering, multi-view images are available for good visual rendering quality. The end-users may desire to render only the IOs instead of the whole scene, which makes the accurate segmentation of the objects desirable. For 3D object model reconstruction, integrating the 2D images captured from different views to reconstruct the 3D object model is a challenging problem. The first task is the efficient removal of background from these objects.

* Corresponding author.
*E-mail addresses:* qzhang@ee.cuhk.edu.hk (Q. Zhang), knngan@ee.cuhk.edu.hk (K.N. Ngan).

1047-3203/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2009.09.005

With the recent growing capability of the capturing devices, multi-view capturing system with dense or sparse camera array [16,17] can be built with ease, which motivates the development of multi-view techniques and its related applications. A multi-view image segmentation algorithm proposed in [18] aims to segment foreground object from a collection of 2D images taken from different viewpoints for 3D object reconstruction. It incorporates some useful and well-known algorithms including graph cut image segmentation, volumetric graph cut and learning shape priors. Quan et al. [19] investigated the issue of image-based plant modeling. They propose a plant modeling system for generating 3D models of natural-looking plant from a number of images captured by a hand-held camera with different views. Segmenting the leaves of a plant is a tough problem because of the occlusion and similarity of color between different overlapping leaves. In their approach, leaf segmentation problem is formulated as graph-based optimization aided by 3D and 2D information. To reconstruct the 3D geometry of static scene, an algorithm in [20] simultaneously deals with the depth map estimation and background separation in multi-view setting with several calibrated cameras. By exploiting the strong interdependency of two problems and minimizing a discrete energy functional using graph cut, this combined approach yields more correct depth estimate and better background separation on both real-world and synthetic scenes. The state-of-the-art work for bi-layer segmentation of the stereo video sequence is presented in [21]. By probabilistic fusion of stereo, color and contrast cues, it efficiently separates the foreground from background layer in real-time, and successfully applies to background substitution.

## 2. Overview of the proposed framework

In this paper, we propose an automatic and efficient algorithm to segment multiple objects from multi-view video. Fig. 1 shows the algorithm framework composed of three components: data pre-processing, offline-operations and online segmentation. We built a five-view camera system to capture the multi-view video data. Given the multi-view image sets $I_t^v$ captured at time instances $t$ from five different views $v \in \{0,1,2,3,4\}$, the objective is to obtain the labeling field $f_t^v$. After data acquisition, the raw sources undergo two pre-processing stages: color equalization and geometric calibration. Color equalization uniformizes the color responses across all views. Geometric calibration calculates the multiple camera parameters by the nonlinear algorithm in [22], used for correction of geometric distortion and for disparity estimation based on epipolar constraint.

In off-line operations, auxiliary information is calculated beforehand to support the online segmentation. Images with far views will lead to large search range of the disparity value, which makes the stereo matching error-prone and disparity estimation time-consuming. In order to reduce the projection error and avoid extensive computational load, we select view 2 as the *key view* to start the segmentation process. Motion field $M_{t,t-1}^v$ between successive frames, disparity field $D_t^{v_i,v_j}$ (target view $v_i$ with respect to reference view $v_j$) and occlusion map $O_t^{v_i,v_j}$ between two neighboring views are estimated offline. Based on the camera geometry and perspective projection model, depth can be reconstructed using the multiple disparity maps and the calibrated camera parameters. Depth maps $DE_t^v$ are reconstructed using two disparity maps between a particular view and its two neighboring views. The occluded pixels in either of the occlusion maps between $v$ and its two neighboring views are defined as occluded in the combined occlusion map $CO_t^v$.

The remainder of this paper focuses on the online segmentation. In Section 3, we introduce the multiple objects segmentation in the *key view* of the multi-view images. Section 4 is devoted to the multi-view video segmentation. Experimental results shown in the Section 5 validate the efficiency and robustness of the proposed algorithm. Finally, conclusions are drawn in Section 6.

## 3. Multiple objects segmentation for key view

In computer vision, image segmentation generally can be formulated as an energy minimization problem. Graph cut as a powerful energy minimization tool, has been widely used for solving many related vision and graphic problems with great success, such as stereo matching [23], multi-view reconstruction [24] and texture synthesis [25]. With its efficiency in segmentation as demonstrated by Boykov and Jolly [6], graph cut has generated extensive interest for image segmentation and spawned many related works [26–28].

### 3.1. Automatic IOs extraction based on saliency model

Most of the classical and start-of-the-art graph cut based segmentation algorithms require user's interventions to specify the initial foreground and background regions as hard constraints. Even though user's assistance is helpful to achieve good segmentation results, a major drawback is the dependence on such guidance. Initialization itself may be annoying to the user especially large quantities are needed. Furthermore, graph cut based segmentation

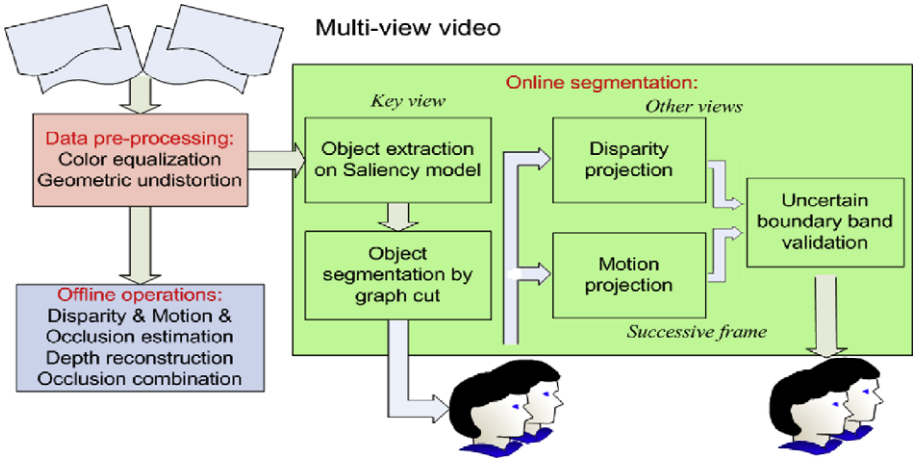


**Fig. 1.** The framework of the proposed algorithm.

suffers from incomplete initialization. These disadvantages motivate the object extraction in an automatic way.

Locating a semantic object in the visual environment is an effortless task for human observer but a challenging problem for computer, because far more perceptual information is presented than can be processed. The visual attention concept provides us with an intelligent mechanism to perceptually attract human's attention toward the location of IOs in a complicated scene. Itti et al. [29] proposed a saliency based visual attention model for scene analysis and attention location. Given a static image, this model employs color, intensity and orientation to compute a saliency map (SM) which encodes the conspicuity at each location in the visual input. The larger value in the SM indicates where more human attention is focused on. To improve the Itti's model, Han's SM [30] incorporates location cue based on the observation that human generally pays more attention to the object near to the center of the image. For salient object detection, SM proposed in [31] effectively combines a set of novel features including multi-scale contrast, center-surrounding histogram and color spatial distribution. However, the computation of these SMs is implicitly or explicitly based on the low-level features, which an IO may not always possess.

To achieve more efficient and robust saliency representation, higher-level visual features should be taken into consideration. Human attentions are generally more focused on the moving object than the static one in the video [32], which means IOs deserve a larger weight in the motion field. An IO appears to have similar depth values in the 3D scene, indicating that it has a uniform distribution in the depth field. Inspired by the work in [33], more sophisticated cues such as motion and depth are combined into our topographical SM. By thresholding, morphological operations and connected component analysis on the SM, IIOs can be automatically extracted as initialization to trigger the subsequent segmentation process. Fig. 2 shows the SMs of two images using higher-level features and the extracted IIOs, which are used to model the initial foreground regions.

### 3.2. Graph cut based multiple objects segmentation

Graph cut based method constructs a graph topology to minimize the specified energy function activated by the max-flow/min-cut algorithm, so that the min-cut on the graph is of minimum energy among all the cuts separating the terminals. The general formulation of energy function is given in (1):

$$E(f) = \sum_{(p\in P)} E_p(f_p) + \lambda \sum_{(p,q\in N)} E_{p,q}(f_p, f_q) \tag{1}$$

where $f$ is the labeling field, $P$ is the set of pixels and $N$ is the second-order neighborhood system. data term $E_p(f_p)$ is the likelihood energy and smoothness term $E_{p,q}(f_{p,q})$ is the prior energy. $\lambda$ is a parameter to weigh the importance of these two terms and is fixed as 15 in our experiments.

#### 3.2.1. Basic energy function

Traditional graph cut based segmentation using only color/contrast cues is error-prone especially on the regions with similar foreground/background features, leading to inaccurate results. It suggests a robust hybrid approach with more features. Stereo vision/depth information provided by multi-view data reveals a powerful representation of different layers in 3D scene and assists many multi-view applications [34–36].

*3.2.1.1. Data term.* In the basic energy function, color (RGB) and depth information are combined to evaluate the likelihood of a certain pixel $p$ assigned to the label $f_p$:

$$\begin{aligned} &E_p(f_p) = E_{pc}(\theta_c; z_p; f_p) + E_{pd}(\theta_d; z_p; f_p) \\ &\begin{cases} E_{pc}(\theta_c; z_p; f_p) = -\log g(z_p|f_p, k_p) - \log w(f_p, k_p) \\ E_{pd}(\theta_d; z_p; f_p) = -\log h(z_p|f_p) \end{cases} \end{aligned} \tag{2}$$

$\theta_c$ and $\theta_d$ are the color and depth distributions modeled by the Gaussian Mixture Model (GMM) and the histogram model, respectively, based on the results in Fig. 2(c). $g(\cdot)$ denotes a Gaussian probability distribution and $w(\cdot)$ is the mixture weighting coefficient. $k_p$ is GMM component variable, set as 5 for foreground objects and 10 for the background. $z_p = \{d, r, g, b\}$ is a four-dimensional feature vector for pixel $p$, representing the depth and three color components.

*3.2.1.2. Smoothness term.* $E_{p,q}(f_p, f_q)$ measures the penalty of two neighboring pixels $p$ and $q$ with different labels and is defined as follow:

$$\begin{aligned} &E_{p,q}(f_p, f_q) = dist(p,q)^{-1} \exp(-diff(c_p, c_q)) \\ &diff(c_p, c_q) = \frac{1}{3}(\beta_r \cdot (r_p - r_q)^2 + \beta_g \cdot (g_p - g_q)^2 + \beta_b \cdot (b_p - b_q)^2) \end{aligned} \tag{3}$$

where $dist(p,q)$ and $diff(c_p, c_q)$ are the coordinate distance and average RGB color difference between $p$ and $q$, respectively. $\beta_r =$

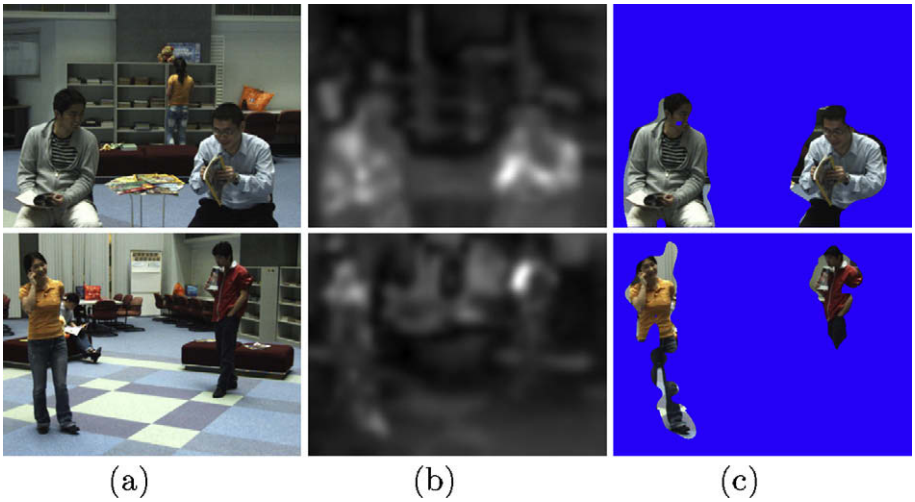


**Fig. 2.** Saliency map and object extraction: (a) input image, (b) saliency map using depth and motion and (c) extracted IIOs.

$(2\langle\|(r_p - r_q)^2\|\rangle)^{-1}$, where $\langle\cdot\rangle$ is the expectation operator for the red channel. $\beta_g$ and $\beta_b$ are defined similarly for the green and blue channels, respectively.

*3.2.2. Multiple objects segmentation using graph cut*

Comparing with the single object segmentation, multiple objects segmentation as a general case is investigated in our work. Based on the assumption that each object is not overlapped, we convert multiple objects segmentation into several sub-segmentation problems. For individual object, we construct a sub-graph for the pixels belonging to its "Object Rectangle", which is an enlarged rectangle to encompass the whole object and restricts the segmentation region. Bi-label graph cut is employed to minimize the energy function and segments each object from its background. Experimental results using basic energy function with different features are shown in Fig. 3(a)–(c).

*3.2.3. Modified energy function*

The segmentation quality using combined features in Fig. 3(c) has outperformed the ones using either single feature in Fig. 3(a) and (b). However, when the scenes contain complex background, notable segmentation inaccuracy around the objects still exists and leads to unsatisfactory results. These errors can be classified into two groups which are highlighted with rectangle and ellipse, respectively, as shown in Fig. 3(c). To tackle these two problems, we propose a modified energy function containing two novelties: *background penalty with occlusion reasoning* and *foreground contrast enhancement*.

*3.2.3.1. Background penalty with occlusion reasoning.* The segmentation errors in the rectangles occur because their color and depth information are very similar to the foreground data, thus using either of or combine these features fails to distinguish them from the object. In the multi-view images, focused object commonly appears in all the cameras. Since we capture the same scene at different view points, occluded background regions often occur around the object boundary. This important observation indicates that the occluded regions have a higher probability to be the background than the visible ones. Thus, we impose a background penalty factor $\alpha_{bp} = 3.5$ to enforce the background likelihood for the occluded pixels in $CO_t^v$, where $CO_t^v(p) = 128$ if $p$ is defined as occluded and *0* otherwise

$$E_p^*(f_p) = \alpha_{bp} \cdot E_p(f_p), \quad (f_p = 0, CO_t^v(p) = 128) \tag{4}$$

Fig. 4 shows the visualization of background probability map with and without occlusion penalty, where brighter pixels denote higher background probability. In Fig. 4(a), the original background probability map is generated using the basic data term, where the ambiguous regions around the object lead to the errors in Fig. 3(c). Fortunately, these ambiguous regions are defined as occluded in the combined occlusion map in Fig. 4(b), so that the background penalty factor can be imposed on to enforce their background probability and results in the improved map in Fig. 4(c).

*3.2.3.2. Foreground contrast enhancement.* The erroneous segmentations marked as ellipses in Fig. 3(c) are mainly caused by the strong color contrast in the background comparing to the weak contrast across the "true" object boundary as illustrated in Fig. 5(a), which is the same problem as discussed in [27]. The authors in [27] introduced a *background contrast attenuation* which can adaptively remove the background contrast while preserving the contrast across the foreground/background boundary. However, this scheme strongly depends on an additional background image, which increases its efficiency but weakens its flexibility.

In our work, we propose *foreground contrast enhancement* to enhance the contrast across foreground/background boundary and attenuate the background contrast. To make the color contrast representation more efficient, the average color difference is computed in the perceptually uniform $L^*a^*b$ color space. The global color smoothness term is defined in (5) and visualized in Fig. 5(b). Due to the proportionality of measured the color difference to the human perception, the superior performance of $L^*a^*b$ space over the non-uniform RGB color space in color difference evaluation has been demonstrated in [37], and it clearly benefits our segmentation task as demonstrated by the comparison of Fig. 5(f) and (g). The better result in Fig. 5(g) is achieved by using the smoothness term in Fig. 5(e) where the color contrast component is computed in the $L^*a^*b$ space.

$$E_{p,q}^{global}(f_p, f_q) = dist(p,q)^{-1} \exp(-diff(c_p, c_q))$$
$$diff(c_p, c_q) = \frac{1}{3}(\beta_L \cdot (L_p - L_q)^2 + \beta_a \cdot (a_p - a_q)^2 + \beta_b \cdot (b_p - b_q)^2) \tag{5}$$

By adopting the motion residual information, we attenuate the high color contrast in the background and enhance the object boundary contrast. When performing motion compensation, the moving object boundary is difficult to compensate which results in larger motion residual around it. This provides a useful cue to represent the

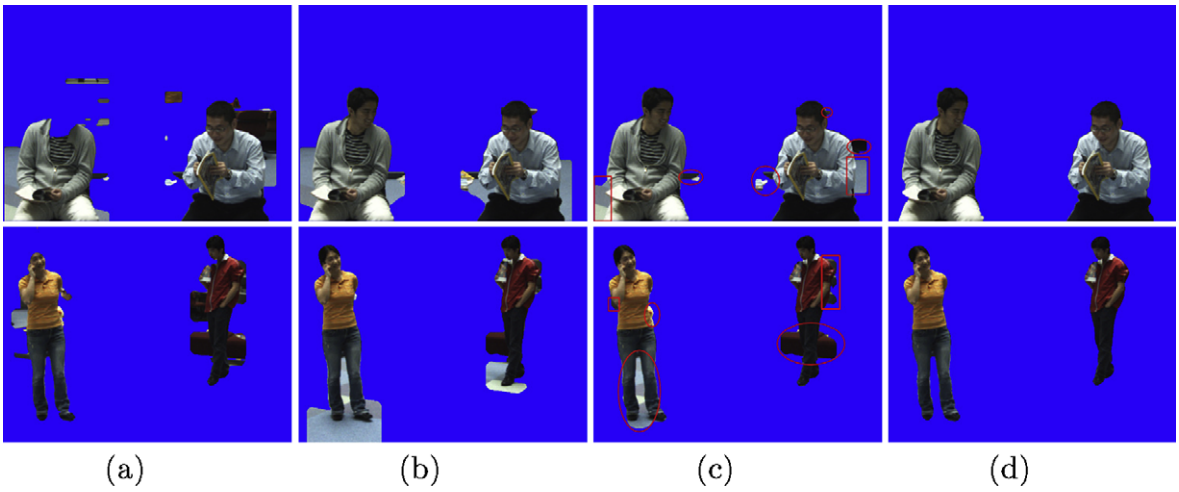


**Fig. 3.** Segmentation results using basic energy function and refinement using modified energy function: basic energy function using: (a) color, (b) depth, (c) combined color and depth and (d) results using modified energy function.

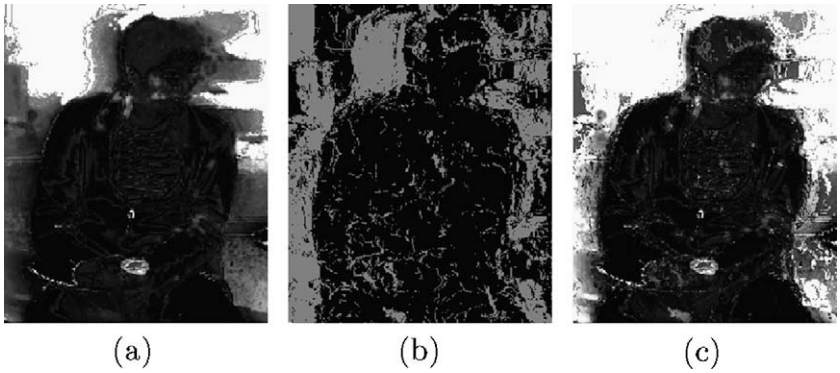


**Fig. 4.** Visualization of background penalty with occlusion reasoning: (a) background probability map without occlusion penalty, (b) combined occlusion map $CO_t^v$ and (c) background probability map with occlusion penalty.

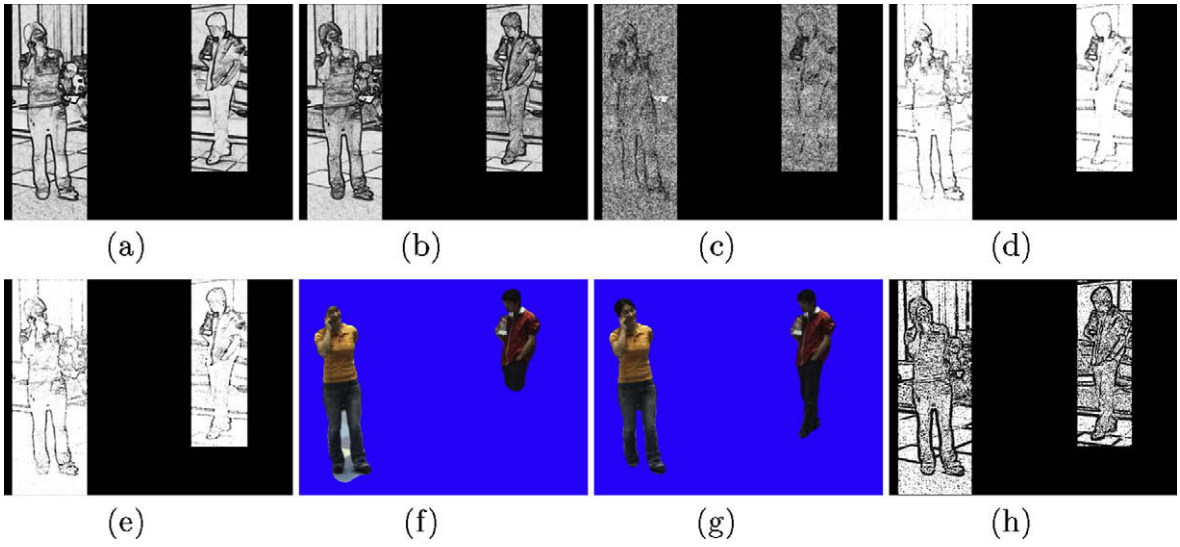


**Fig. 5.** Visualization of the smoothness term in "Object Rectangle" with intermediate results: (a) color contrast in RGB space, (b) color contrast in $L^*a^*b$ space, (c) motion residual contrast, (d) combined contrast of (a) and (c), (e) combined contrast of (b) and (c), (f) segmentation using the smoothness term of (d), (g) segmentation using the smoothness term of (e), (h) local color contrast of (b).

moving object boundary. The motion residual $MR_t^v$ is defined as follow:

$$MR_t^v = abs(I_t^v - R_t^v), R_t^v = (I_{t-1}^v + M_{t,t-1}^v) \tag{6}$$

$R_t^v$ is the reconstructed image from $I_{t-1}^v$ and the motion field $M_{t,t-1}^v$. In the motion residual contrast shown in Fig. 5(c), the static background is quite smooth when compared with the high contrast along moving object boundary. Thus, the smoothness term in (7) achieves *foreground contrast enhancement* by combining the color and motion residual contrasts to enhance the contrast across the moving object boundary, thereby attenuates the background contrast effectively, as illustrated in Fig. 5(e).

$$\begin{aligned} E_{p,q}^{foreEh}(f_p,f_q) &= E_{p,q}^{global}(f_p,f_q) + E_{p,q}^{motionRs}(f_p,f_q) \\ E_{p,q}^{motionRs}(f_p,f_q) &= dist(p,q)^{-1}\exp(-\beta_{mr}\cdot(mr_p - mr_q)^2) \end{aligned} \tag{7}$$

where $\beta_{mr} = (2\langle\|(mr_p - mr_q)^2\|\rangle)^{-1}$, and $mr_p, mr_q$ are the motion residual of $p$ and $q$.

Because of the dynamic gestures, certain parts of the object will keep static for a period of time in the video sequence, resulting in no motion information and hence motion residual; for example, the leg and foot of the moving human. Directly combining the color contrast and motion residual contrasts will not only attenuate the background contrast but also weaken the "true" foreground contrast in these static regions as in Fig. 5(e). As a result, they are considered as the background and eliminated in Fig. 5(g). To have a tradeoff between the attenuation of high background contrast and preservation of "true" object contrast, we define a local color contrast to enhance the discontinuity distribution in its neighborhood, which was the similar idea as in [26]. We calculate the local mean $\mu$ and the local variance $\delta$ of contrast in the each local pattern, to keep the contrast which has higher value than the mean using the following equation:

$$E_{p,q}^{local}(f_p,f_q) = \begin{cases} \exp((E_{p,q}^{global}(f_p,f_q) - \mu_{p,q})^2/2*\delta_{p,q}^2), & \text{if } E_{p,q}^{global}(f_p,f_q) > \mu_{p,q} \\ 1, & \text{otherwise} \end{cases} \tag{8}$$

The final smoothness term is the combination of the local color contrast and the motion residual contrast. The comparison between the segmentation results using the modified energy function and the basic energy function are provided in Fig. 3(d) and (c).

## 4. Multi-view video segmentation

In the above work, we have dealt with the segmentation in a single *key view* of *initial frame*. In many applications, accurate ob-

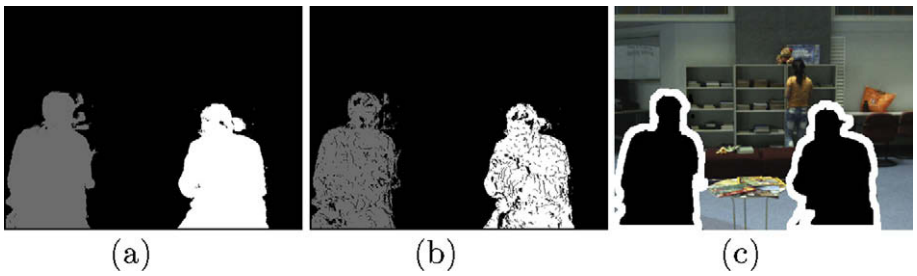


**Fig. 6.** Disparity projection and uncertain band: (a) prediction mask of view 3 without visibility constraint, (b) prediction mask of view 3 with visibility constraint, (c) uncertain band based on the post-processing of (b).

ject segmentation for all views of the frames of a video is required. In this section, we extend the segmentation algorithm to multi-view video.

### 4.1. Disparity projection under visibility constraint

Based on the segmentation result of the *key view*, the coarse predictions of the other views can be projected by pixel-based disparity compensation, which exploits the spatial consistency among inter-view images. However, disparity vectors cannot be estimated correctly for the occluded areas, introducing serious prediction errors as in Fig. 6(a) and the undesired effect for the subsequent process. Since only the IOs should be projected in the target view, which are defined as visible (not occluded) in $CO_t^{v_i}$, thus the projection is performed under visibility constraint in (9):

$$P_t^{v_i}(p) = f_t^{v_j}(p + D_t^{v_i,v_j}(p)), \quad (CO_t^{v_i} = 0) \tag{9}$$

### 4.2. Motion projection for video tracking

Segmenting the consecutive frame is achievable as the motion information is known. Motion prediction is a form of tracking, which

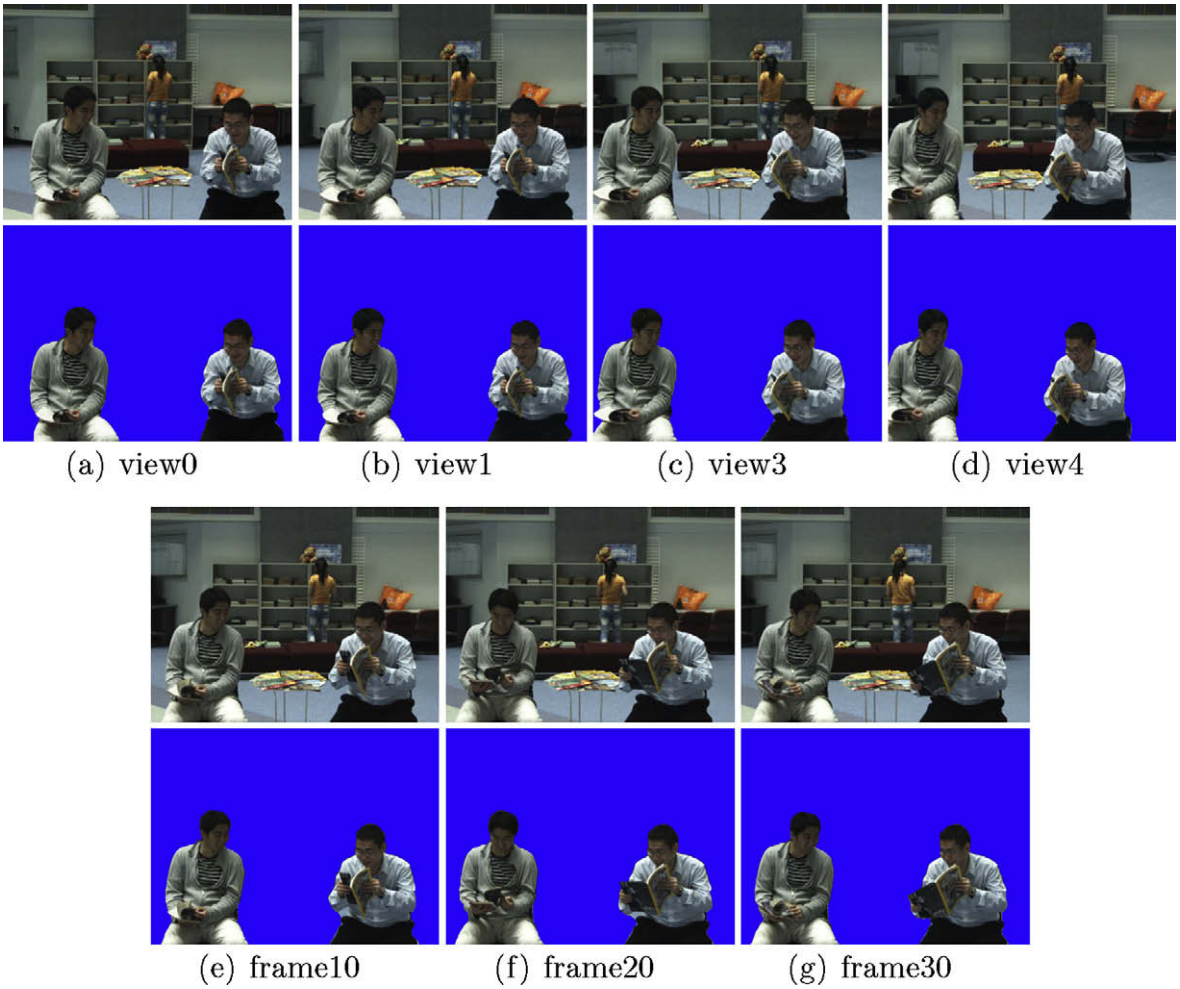


**Fig. 7.** Multi-view video segmentation of *Reading Sequence*.

enforces the temporal consistency between adjacent frames in video. The coarse prediction of the current frame is projected by pixel-based motion compensation from the mask of its previous frame:

$$P_t^v(p) = f_{t-1}^v(p + M_{t,t-1}^v(p)) \quad (10)$$

### 4.3. Uncertain boundary band validation

Because of the existence of noise and non-homogeneity in the estimated field, and despite performing post-processing after the predictions, inaccuracy still exist along the object boundary. To improve the segmentation results, we contrust an uncertain band along the object boundary as in Fig. 6(c) based on an *activity* measure. We define the *activity* of a pixel as the motion variance within its second-order neighborhood. The pixel with the highest *activity*is searched within the neighborhood of each contour pixel, and a band centered at the most active pixel is defined as uncertain region. The pixels lying in the inner band are labeled as foreground ($f_p = n$), and outer band pixels are background ($f_p = 0$). The indices of pixels in the uncertain band are set to be $255 - n$. Labeling field for the uncertain band is validated using the algorithm in Section 3.2 to yield more accurate segmentation layers.

## 5. Experimental results

The efficiency and robustness of the proposed algorithm are demonstrated on two types of multi-view videos simulating different scenarios, which were captured by our five-view camera system in indoor scenes, with resolution of 640 * 480 at frame rate of 30 frames per second (fps).

### 5.1. Segmentation of IOs with similar and low depth

In the *Reading Sequence* (in Fig. 7), both IOs are located in the low depth of field and share the similar depth value, which simulates the video conferencing application. Even though there is a moving object in the background, it is in different depth layer from two IOs which are close to the camera. By incorporating depth and motion information, the saliency values of this background moving object as known in the top row of Fig. 2(b) are much lower than those of the two IOs due to the great depth disparity between them. The input image with the segmentation results of other views and successive every 10 frames are provided in Fig. 7. From these results, it is clearly that the two IOs are segmented precisely and the accuracy is preserved in the multi-view video using the projection technique, with the effective removal of the complex background including still and moving objects. The excellent segmentation performance validates the algorithm efficiency and shows the promising application for video conferencing scenario.

### 5.2. Segmentation of IOs with different depths

Different from the *Reading Sequence*, the two IOs in the *Calling Sequence* (Fig. 8) appear at different depth levels in the 3D scene.

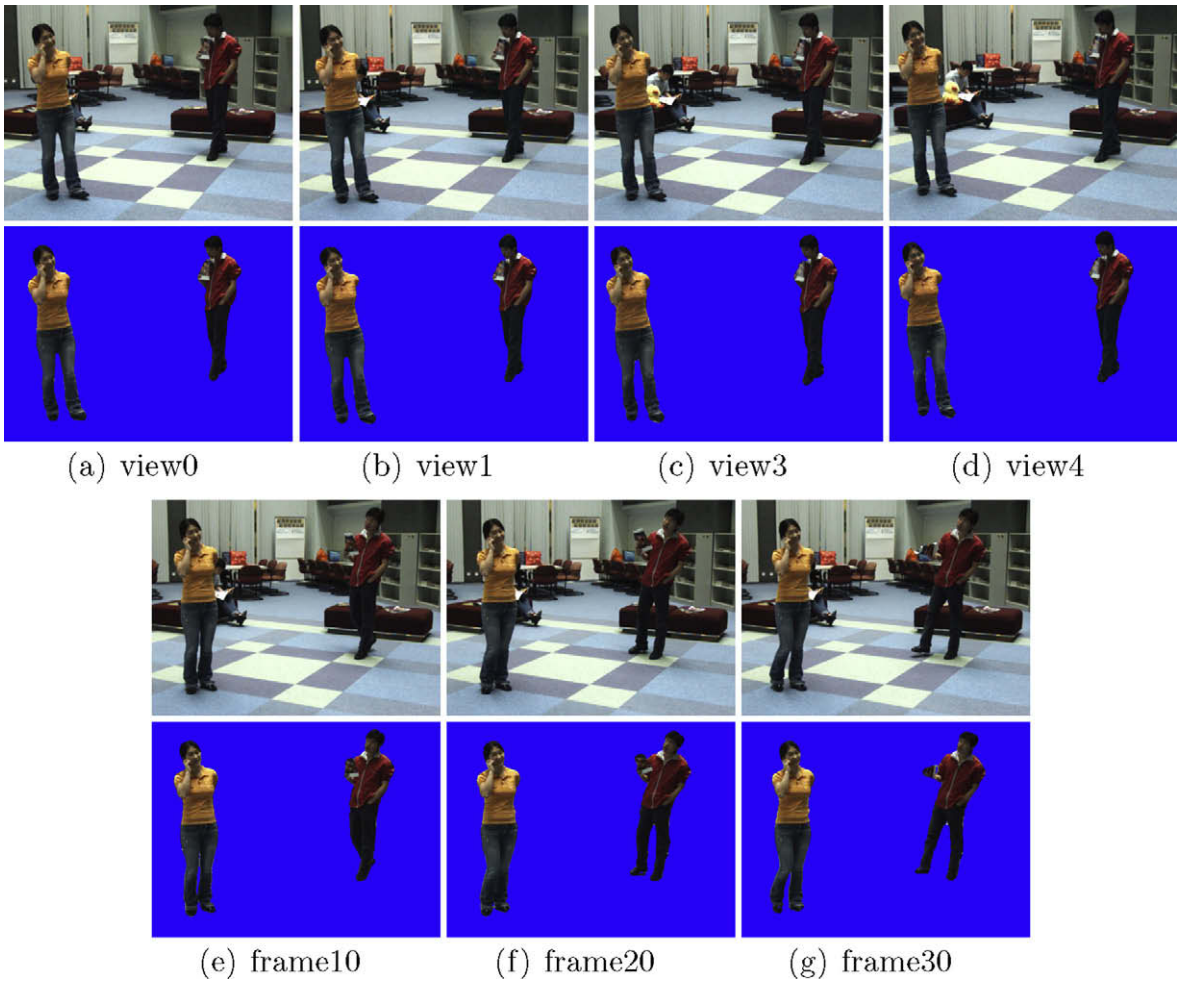


**Fig. 8.** Multi-view video segmentation of *Calling Sequence*.


**Fig. 9.** Segmentation comparison using *IU-JW Sequence* frame 30: (a) left view, (b) right view, (c) result by our proposed algorithm and (d) result by Kolmogorov's algorithm.

Complete objects (human bodies) are captured because of the further distance from the cameras. An object in the background is considered not an IOs due to its stationary in the video sequence. The input image with the segmentation results of other views and successive every 10 frames are shown in Fig. 8. The cluttered background, color mixing around object boundary and the stationary of part of object in this video increases the segmentation complexity, which cannot be handled with the basic energy function as illustrated in the second row of Fig. 3(c). However, by adopting the modified energy function, the algorithm can successful segment the IOs, and keep good separation of the foreground/background layer across the video. The satisfactory segmentation quality on this challenging video further supports the efficiency of our proposed algorithm and demonstrates its robustness.

### 5.3. Comparison with others' methods

We performed a comparison with Kolmogorov's bilayer segmentation algorithm [21] using their test images in the *IU-JW Sequence*. The input stereo image pair after color equalization, the segmentation results of the left view using our proposed algorithm and the bilayer segmentation are shown in Fig. 9. Judging from the results of the two algorithms, our algorithm achieved more accurate segmentation, which has benefited from the improvement due to the *foreground contrast enhancement*, and the efficient background removal because of the background penalty.

To further validate the superiority of our algorithm over other's, we compared our proposed algorithm with an existing method employing multi-way cut with $\alpha$-expansion [33] using our test images. The experimental results using the multi-way cut in the *key view* of *initial frame* in *Reading Sequence* and *Calling Sequence* are presented in Fig. 10, both of which are based on the same initialization results as our algorithm (shown in Fig. 2(c)) for fair comparison. From the results using multi-way cut in Fig. 10 and our results in Fig. 3(d), it is clear that our algorithm offers noticeable improvement in segmentation quality. The major drawback of the multi-way cut is its dependence on the initialization process, and the segmentation results suffer if the initialization results are poor. Also, segmentation errors emerge in the areas with high background contrast. Moreover, our algorithm outperforms the multi-way cut method in the computational efficiency. To produce the results in Fig. 10, 10 iterations are required to perform the $\alpha$-expansion for each label $f \in \{0, 1, 2\}$ in the whole image, which is extremely time-consuming. However, our proposed algorithm applies the bi-label graph cut for the foreground label in the restricted segmentation regions with only one iteration, thus greatly reduces the computational time.


**Fig. 10.** Segmentation results using multi-way cut on the *key view* of *initial frame* in: (a) *Reading Sequence* and (b) *Calling Sequence*.

## 6. Conclusions

In this paper, we propose an automatic segmentation algorithm for multiple objects from multi-view video. After data pre-processing, offline operations are carried out to yield motion and disparity information facilitating the online segmentation. IIOs are extracted in an unsupervised manner in the *key view* of *initial frame* based on the saliency model, where a single topological saliency map is calculated by combining motion and depth information. Multiple objects segmentation is decomposed into several sub-segmentation problems and solved using bi-label graph cut individually. In the proposed novel energy function, foreground/background likelihood is evaluated by fusing color, depth and occlusion cues. *Foreground contrast enhancement* by efficiently combining the color contrast with the motion residual contrast is employed to measure the smoothness penalty. To enforce the spatiotemporal consistency in the multi-view video, the coarse predictions of the other views and the next frame are projected by disparity compensation and motion compensation, respectively. Uncertain band around the object boundary is constructed and refined to obtain more accurate results. The experiment was implemented on two representative multi-view videos. Accurate segmentation results with good visual quality and subjective comparison with others' methods attest to the efficiency and robustness of our proposed algorithm.

## Acknowledgment

This work was supported in part by the Research Grants Council of the Hong Kong SAR (Project CUHK415707).

## References

[1] D. Chai, K.N. Ngan, Face segmentation using skin color map in videophone applications, IEEE Transactions on Circuits and Systems for Video Technology 9 (4) (1999) 551–564.
[2] Y. Li, J. Sun, H.Y. Shum, Video object cut and paste, ACM Transactions on Graphics 24 (2005) 595–600.
[3] L. Quan, J.D. Wang, P. Tan, L. Yuan, Image-based modeling by joint segmentation, International Journal of Computer Vision 75 (1) (2007) 135–150.
[4] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (5) (2002) 603–619.
[5] L.G. Shapiro, G.C. Stockman, Computer Vision, Prentice-Hall, New Jersey, 2001, ISBN 0-13-030796-3. pp. 279–325.

[6] Y. Boykov, M.P. Jolly, Interactive graph cuts for optimal boundary and region segmentation of objects in N–D images, in: Proceedings of the IEEE International Conference on Computer Vision, 2001, pp. 105–112.
[7] C. Rother, V. Kolmogorov, A. Blake, Grabcut: interactive foreground extraction using iterated graph cuts, ACM Transactions on Graphics 23 (2004) 309–314.
[8] M. Kass, A. Witkin, D. Terzopoulos, Snakes: active contour models, International Journal of Computer Vision (1987) 259–268.
[9] S. Osher, J.A. Sethian, Fronts propagating with curvature-dependent speed: algorithms based on Hamilton–Jacobi formulations, Journal of Computational Physics 79 (1988) 12–49.
[10] C. Xu, J.L. Prince, Snakes, shapes, and gradient vector flow, IEEE Transactions on Image Processing 7 (3) (1998) 359–369.
[11] Y. Li, J. Sun, C.K. Tang, H.Y. Shum, Lazy snapping, ACM Transactions on Graphics 23 (2004) 303–308.
[12] G. Sfikas, C. Nikou, N. Galatsanos, Edge preserving spatially varying mixtures for image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–7.
[13] V. Lempitsky, C. Rother, A. Blake, LogCut – efficient graph cut optimization for Markov random fields, in: Proceedings of the IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[14] Y.W. Tai, J.Y. Jia, C.K. Tang, Soft color segmentation and its applications, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (9) (2007) 1520–1537.
[15] Y.C. Huang, Q.S. Liu, D. Metaxas, Video object segmentation by hypergraph cut, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[16] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, C. Zhang, Multi-view imaging and 3DTV, IEEE Signal Processing Magazine 24 (2007) 10–21.
[17] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, R. Szeliski, High-quality video view interpolation using a layered representation, ACM Transactions on Graphics 23 (2004) 600–608.
[18] Y.P. Tasi, C.H. Ko, Y.P. Hung, Z.C. Shih, Background removal of multiview images by learning shape priors, IEEE Transactions on Image Processing 16 (2007) 2607–2616.
[19] L. Quan, P. Tan, G. Zeng, L. Yuan, J.D. Wang, S.B. Kang, Image-based plant modeling, ACM Transactions on Graphics 25 (3) (2006) 599–604.
[20] B. Goldlcke, M.A. Magnor, Joint 3D-reconstruction and background removal separation in multiple views using graph cuts, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2003, pp. 683–688.
[21] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, C. Rother, Probabilistic fusion of stereo with color and contrast for bi-layer segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (9) (2006) 1480–1492.
[22] C.H. Cui, W.X. Yang, K.N. Ngan, External calibration of multi-camera system based on pair-wise estimation, Advances in Image and Video Technology-PSIVT, Lecture Notes in Computer Science, vol. 4872, Springer, Berlin/Heidelberg, 2007, ISBN 978-3-540-77128-9. pp. 497–509.
[23] Y. Boykov, O. Veksler, R. Zabih, Markov random fields with efficient approximation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1998, pp. 648–655.
[24] V. Kolmogorov, R. Zabih, Multi-camera scene reconstruction via graph cuts, in: European Conference on Computer Vision and Pattern Recognition, 2002, pp. 82–96.
[25] V. Kwatra, A. Sch?dl, I. Essa, G. Turk, A. Bobick, Graphcut texture: image and video synthesis using graph cuts, ACM Transactions on Graphics 22 (3) (2003) 277–286.
[26] J. Wang, P. Bhat, R.A. Colburn, M. Agrawala, M.F. Cohen, Interactive video cutout, ACM Transactions on Graphics 24 (2005) 585–594.
[27] J. Sun, W.W. Zhang, X.O. Tang, H. Y Shum, Background cut, ECCV (2006).
[28] S. Vicente, V. Kolmogorov, C. Rother, Graph cut based image segmentation with connectivity prior, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[29] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 1254–1259.
[30] J. Han, K.N. Ngan, M. Li, H. Zhang, Unsupervised extraction of visual attention objects in color images, IEEE Transactions on Circuits and Systems for Video Technology 16 (1) (2006) 141–145.
[31] T. Liu, J. Sun, N.N. Zheng, X.O. Tang, H.Y. Shum, Learning to detect A salient object, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[32] L. Itti, P. Baldi, A principled approach to detecting surprising events in video, IEEE Int. Conf. Computer Vision (2005) 631–637.
[33] W.X. Yang, K.N. Ngan, Unsupervised multiple object segmentation of multiview images, Advanced Concepts for Intelligent Vision Systems Conference (2007) 178–189.
[34] A.D. Doulamis, N.D. Doulamis, K.S. Ntalianis, S.D. Kollias, Unsupervised semantic object segmentation of stereoscopic video sequence, in: Proceedings of the International Conference on Information Intelligence and Systems, 1999.
[35] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, C. Rother, Bi-layer segmentation of binocular stereo video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 407–414.
[36] Dongbo Min, Donghyun Kim, SangUn Yun, Kwanghoon Sohn, 2D/3D freeview video generation for 3DTV system, Signal Processing: Image Communication 24 (2009) 31–48.
[37] G. Paschos, Perceptually uniform colour space for colour texture analysis: an empirical evaluation, IEEE Transactions on Image Processing 10 (6) (2001) 932–937.