Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Convex relaxation for a 3D spatiotemporal segmentation model using the primal-dual method*

Shi-yan WANG, Hui-min YU‡

(*Department of Information Science & Electronic Engineering, Zhejiang University, Hangzhou 310027, China*)

E-mail: {wangshiyan, yhm2005}@zju.edu.cn

Received Nov. 15, 2011; Revision accepted Mar. 23, 2012; Crosschecked May 9, 2012

Abstract: A method based on 3D videos is proposed for multi-target segmentation and tracking with a moving viewing system. A spatiotemporal energy functional is built up to perform motion segmentation and estimation simultaneously. To overcome the limitation of the local minimum problem with the level set method, a convex relaxation method is applied to the 3D spatiotemporal segmentation model. The relaxed convex model is independent of the initial condition. A primal-dual algorithm is used to improve computational efficiency. Several indoor experiments show the validity of the proposed method.

Key words: 3D spatiotemporal segmentation, Motion estimation, Total variation, Primal-dual
doi:10.1631/jzus.C1100331 **Document code:** A **CLC number:** TP391.7; TP317.4

1 Introduction

Motion analysis based on 3D videos is an important and challenging task in computer vision, especially when the viewing system is moving. Such a problem is of importance in applications such as 3DTV, video compression (MPEG-4), and autonomous navigation. The problem can be divided into two topics: motion segmentation and motion estimation, which can benefit from each other. While accurate segmentation releases motion estimation from the problem of ambiguity near motion boundaries, the motion information gives important hints on how to partition the image. Therefore, performing the two topics simultaneously is the key idea for the 3D spatiotemporal segmentation model in this paper.

In motion-based segmentation approaches, the optical flow or just the image difference is always handled as the pre-computed feature to feed into a standard segmentation method. Sequential methods (Dufaux *et al.*, 1994) infer that a field of the motion parameter is first computed and then segmented. Many recent approaches have performed optical flow estimation and segmentation simultaneously (Schnorr, 1991; Mémin and Pérez, 2002; Feghali and Mitiche, 2004; Paragios and Deriche, 2005). Feghali and Mitiche (2004) introduced an approach of tracking moving objects, in which the optical velocities on motion boundaries can be estimated from geometric properties of the 2D motion field in the spatiotemporal domain of image sequences. These approaches assume that the depth variations of the scenes are small or smooth compared to the distance from the camera to the observer. In real scenes, however, depth variations may be large. This means both the motion information and the structure of the scene contribute to the depth discontinuities, which are not easy to detect using 2D methods. To overcome these limitations of the existing methods, we consider using 3D models to solve this problem.

In recent decades, 3D motion estimation (Alvarez *et al.*, 2009) has attracted more and more attention. The classical approach (Heeger and Jepson, 1992) to estimating the 3D motion parameters is to refer its movement to the set of corresponding points found

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 60872069) and the National Basic Research Program (973) of China (No. 2012CB316400)

from consecutive frames using a single camera. A flow vector is used as the basic component of motion, aiming to recover 3D motion parameters from a 2D flow field. However, given monocular sequences only, the solution itself is not unique, which is a serious problem for motion estimation. The root cause is that the single camera cannot provide the depth information. Recently, a new kind of solid-state 3D camera based on the time-of-flight (TOF) principle has been developed and has an increasing use in image understanding, especially in robotics and autonomous navigation (Ohno *et al.*, 2006; Ye and Hegde, 2009). The 3D TOF camera can capture 3D scene data in real time. It provides us with the possibility of integrating the motion model and the evolution surface into a variational framework for precise segmentation and accurate estimation.

The standard active contour models (Kass *et al.*, 1988; Mumford and Shah, 1989; Caselles *et al.*, 1997) are among the most successful variational techniques for image segmentation. Much literature has been devoted to active contour models since the segmentation allows an easy combination of numerous low-level criteria such as edge consistency (Caselles *et al.*, 1997), intensity homogeneity (Chan and Vese, 2001), and shape knowledge (Leventon *et al.*, 2000). A well-known model is geodesic active contour (GAC) proposed by Caselles *et al.* (1997), which identifies objects by an edge detector. Another famous model is the Chan-Vese (CV) model based on regional information. In contrast to the above models that focus on a single image, in this paper we perform moving objects segmentation and tracking throughout a spatiotemporal domain.

Traditionally, active contour models are solved using the level set method (LSM) (Osher and Sethian, 1988). LSM has proved an important approach in computer vision and become a popular framework in image segmentation. It has the advantage of tracking objects with topology changes such as merging and breaking from the initial contour. However, LSM relies on a good initial condition because of the existence of local minima in energy. The problem is related to non-convexity of both the standard active contour model and level set formulation.

To find a robust method for 3D videos based spatiotemporal segmentation, we consider a 3D model to describe the relationship between segmentation and motion estimation. This model, besides developing a new derivation to include 3D transformation information into the optical flow constraint, utilizes the active contour model to construct an energy functional, based on the assumption that all background regions satisfy the 3D motion constraint well while the foreground regions do not. Unlike the popular way to perform contour propagation using LSM, the energy functional is relaxed to a convex one, avoiding the non-convexity problem. Energy minimization is solved with a dual formulation of the total variation (TV) norm (Chan *et al.*, 1999; Chambolle, 2004). Thus, 3D spatiotemporal segmentation is robust, fast, and independent of the initialization. The main contributions of our work are summarized as follows:

1. Propose a 3D model for spatiotemporal segmentation, allowing segmentation and motion estimation to benefit from each other.

2. Present a convex relaxation procedure of the proposed 3D spatiotemporal segmentation model, which overcomes the local minimum problem.

3. Use a fast primal-dual projection algorithm to minimize the functional, improving the computational efficiency.

2 Basic models

2.1 Active contour model

The general (two-phase) active contour segmentation model is as follows:

$$\min\left\{E_{\mathrm{AC}}(C)=\int_{C} g_{\mathrm{b}}(C,s)\mathrm{d}s+\lambda\int_{C^{\mathrm{in}}} g_{\mathrm{r}}^{\mathrm{in}}(C^{\mathrm{in}},x)\mathrm{d}x+\lambda\int_{C^{\mathrm{out}}} g_{\mathrm{r}}^{\mathrm{out}}(C^{\mathrm{out}},x)\mathrm{d}x\right\}, \tag{1}$$

where C stands for a closed curve in 2D images and a surface in 3D, C^{in} and C^{out} are the inside and outside regions of C in the image domain respectively, g_{b}: $\Omega\rightarrow\mathbb{R}$ is the boundary function (such as an edge detector function in the GAC model), $g_{\mathrm{r}}^{\mathrm{in}}$, $g_{\mathrm{r}}^{\mathrm{out}}$: $\Omega\rightarrow\mathbb{R}$ are arbitrary inside and outside region functions respectively, ds is the Euclidean element of length, dx is the region element, and λ is a free parameter.

In general, Eq. (1) can be used to encode most existing (boundary-, region-, shape-based) active contour models. When $g_b=1/(1+\beta|\nabla \boldsymbol{I}|^2)$ and $g_r^{in}=g_r^{out}=0$, Eq. (1) becomes the well-known GAC model, $E_{GAC}=\int_C 1/(1+\beta|\nabla \boldsymbol{I}|^2)\mathrm{d}s$. Here $\nabla \boldsymbol{I}$ is the image gradient and β is an arbitrary positive constant. This method uses an edge detector function to identify objects and has been widely used in many applications such as medical imaging (Malladi *et al.*, 1996; Jonasson *et al.*, 2005). Another famous model is the Chan-Vese (CV) model, which is a two-phase piecewise constant approximation of the Mumford-Shah model (Mumford and Shah, 1989), given by $g_b=1$, $g_r^{in}=(\mu_{in}-I)^2$, and $g_r^{out}=(\mu_{out}-I)^2$ such that $E_{CV}=\int_C \mathrm{d}s+\lambda\int_{C^{in}}(\mu_{in}-I)^2\mathrm{d}x+\lambda\int_{C^{out}}(\mu_{out}-I)^2\mathrm{d}x$.

2.2 3D spatiotemporal segmentation model

Note that optical flow is the fundamental feature in the motion-based segmentation approaches, and we adopt the hypothesis that the flow field can be segmented into connected sets of flow vectors, while each set is consistent with a rigid 3D motion. Before undertaking the variational framework, the relationship between 3D motion parameters and optical flow will be discussed in this subsection.

Consider a coordinate system *OXYZ* at the optical center of the TOF camera (Fig. 1). Under perspective projection, the equation relating $p(x, y)$ on the image to $P(X, Y, Z)$ in the environment is

$$x=f\cdot X/Z,\quad y=f\cdot Y/Z, \tag{2}$$

where f is the focal length.

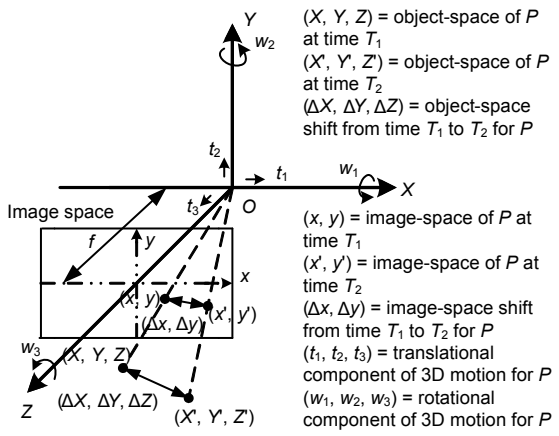

Fig. 1 The camera coordinate system

Let $\boldsymbol{T}=(t_1, t_2, t_3)$ and $\boldsymbol{W}=(w_1, w_2, w_3)$ be the translational and rotational components of the 3D motion of the background, respectively. Suppose (u_{op}, v_{op}) is the 2D velocity of point (x, y) in the image space related to the 3D motion of the projected point (X, Y, Z) on the background. Then the relationship between optical flow and the 3D motion parameters is given as follows (Longuet-Higgins and Prazdny, 1980):

$$u_{op}=\frac{1}{Z}(ft_1-xt_3)-\frac{xy}{f}\omega_1+\frac{f^2+x^2}{f}\omega_2-y\omega_3, \tag{3a}$$

$$v_{op}=\frac{1}{Z}(ft_2-yt_3)+\frac{xy}{f}\omega_2-\frac{f^2+y^2}{f}\omega_1+x\omega_3, \tag{3b}$$

where Z is the depth information in the object space. Eqs. (3a) and (3b) describe the image motion vector field caused by 3D motion, which is also named 'flow field'. Also, the following condition must be satisfied for Eq. (3): the motion of both the camera and the targets is small between two frames, or the time interval between two frames is short enough. Eq. (3), under this condition, is employed as the basic equation of our analysis throughout the paper.

According to the optical flow constraint (OFC) equation (Horn and Schunck, 1981), we have

$$I_x u_{op}+I_y v_{op}+I_t=0. \tag{4}$$

Here I_x, I_y, and I_t are the image spatiotemporal derivatives.

Substituting Eq. (3) into Eq. (4), for each point in the background region, we have

$$I_{op}(\boldsymbol{\theta}, I_t)=\boldsymbol{\theta}\cdot\boldsymbol{\rho}+I_t=0, \tag{5}$$

where

$$\begin{aligned}\boldsymbol{\theta}&=(t_1,\ t_2,\ t_3,\ w_1,\ w_2,\ w_3),\\ \boldsymbol{\rho}&=(fI_x/Z,\ fI_y/Z,\ (-xI_x-yI_y)/Z,\\ &\quad -fI_y-y(xI_x+yI_y)/f,\\ &\quad fI_x+x(xI_x+yI_y)/f,\ xI_y-yI_x)^{\mathrm{T}}.\end{aligned}$$

Here $\boldsymbol{\rho}$ is a constant vector for each pixel. Eq. (5) describes the relationship among the image gradient, depth information, and 3D motion parameters (called the '3D optical flow constraint' in the following). This constraint also meets the condition that the surface of

the real-world environment is approximately piecewise smooth. If $\boldsymbol{\theta}$ is estimated correctly, I_{op} of the background region approaches zero; in the meantime, I_{op} of the foreground may be large, which makes I_{op} a distinct feature to distinguish between the background and foreground. Therefore, we choose the following observation model to design the region functions:

$$\begin{cases} g_{\mathrm{r}}^{\mathrm{in}}(x)=(\boldsymbol{\theta}\cdot\boldsymbol{\rho}+I_t)^2, & x\in C^{\mathrm{in}}, \\ g_{\mathrm{r}}^{\mathrm{out}}(x)=\alpha \mathrm{e}^{-(\boldsymbol{\theta}\cdot\boldsymbol{\rho}+I_t)^2}, & x\in C^{\mathrm{out}}, \end{cases} \tag{6}$$

where α is a scale parameter to control the competition between the inside (background) and outside (foreground) regions. Besides, g_{b}=1 is defined as a penalty on the 3D spatiotemporal surface. Thus, the segmentation model based on 3D videos can be written as follows:

$$E=\int_C \mathrm{d}s+\lambda\int_{C^{\mathrm{in}}}(\boldsymbol{\theta}\cdot\boldsymbol{\rho}+I_t)^2\mathrm{d}x+\lambda\int_{C^{\mathrm{out}}}\alpha \mathrm{e}^{-(\boldsymbol{\theta}\cdot\boldsymbol{\rho}+I_t)^2}\mathrm{d}x. \tag{7}$$

Here C represents a closed surface because the integral is throughout the spatiotemporal domain. This segmentation model has the following characteristics: (1) The approach estimates the parameters while evolving the surface; (2) It allows segmentation with a moving viewing system and does not need prior information of camera motion; (3) It has no limitation with respect to the number of targets, which means any object that has non-homogeneous motion with the background can be partitioned and tracked.

3 Convex relaxation for energy functional

Energy minimization like Eq. (7) is usually solved by the level set method (LSM) (Osher and Sethian, 1988). The key idea of LSM is to implicitly represent a curve/surface C in an image plane $\Omega\rightarrow\mathbb{R}^n$ as the zero level of an embedding function in such a way that it can handle topological changes such as breaking and merging, and is of great significance for stable numerical computation. However, LSM is highly sensitive to initial conditions. Actually, a good segmentation result depends much on a 'good' initial position of the contour. Conversely, 'bad' initial conditions usually lead to unsatisfactory results. The problem is related to the existence of local minima in the energy functional. Take the CV model as an example. The active contour with a 'bad' initial position of a brain image in Fig. 2a (see p.435) cannot segment all interested regions in Fig. 2b, as it gets 'stuck' in a local minimum. In contrast, another initial condition in Fig. 2c results in satisfactory segmentation, as shown in Fig. 2d. This example shows that LSM computes only a local minimum since the level set formulation of energy is not convex.

To solve the problems associated with non-convexity, many works (Chan *et al.*, 2004; Bresson *et al.*, 2007; Goldstein *et al.*, 2009) have been undertaken on the globally continuous convex models, based on total variation regularized energies. Chan *et al.* (2004) proposed a global minimization approach for two well-known models. One is for binary image denoising, and the other is the model of active contours without edges of CV. Bresson *et al.* (2007) developed a global segmentation approach for the snake model, the Rudin–Osher–Fatemi denoising model, and the Mumford–Shah segmentation model. They also established theorems with proofs to determine the existence of a global minimum of these models. A globally continuous minimization approach is important because the global minimum can be found using any optimization algorithm and its solution is independent of the initial condition, avoiding the drawbacks of LSM, including initialization and re-initialization of the distance function.

Here we emphasize 'continuous' to avoid confusion with discrete global minimization approaches such as graph cut (Boykov and Kolmogorov, 2004; Boykov and Funka-Lea, 2006), which can also provide a global minimum. Boykov and Funka-Lea (2006) proposed a fast graph cut algorithm based on max-flow/min-cut (Boykov and Kolmogorov, 2004) for N-dimensional image segmentation. Unlike continuous global minimization approaches, however, the discrete ones do not have sub-pixel accuracy. Also, graph cuts highly depend on the grids. In fact, 'bad' grids can lead to systematic metrification errors. In addition, a discrete graph cut algorithm induces anisotropy and needs special schemes for memory allocation if we process 3D images. Continuous models do not have these limitations. Inspired by their work, we extend the continuous global minimization approach to the 3D spatiotemporal segmentation model.

The level set formulation of energy functional (7) is

$$\min_{\boldsymbol{\Phi},\boldsymbol{\theta}}\left\{E_{\mathrm{LSM}}=\int_{\Omega}|\nabla\boldsymbol{\Phi}|\delta(\boldsymbol{\Phi})\mathrm{d}x+\lambda\int_{\Omega}g_{\mathrm{r}}^{\mathrm{in}}(\boldsymbol{\theta})H(\boldsymbol{\Phi})\mathrm{d}x+\lambda\int_{\Omega}g_{\mathrm{r}}^{\mathrm{out}}(\boldsymbol{\theta})(1-H(\boldsymbol{\Phi}))\mathrm{d}x\right\},\tag{8}$$

where $\boldsymbol{\Phi}$ is the level set function implemented by the signed distance function, δ is the Dirac operator which identifies the zero level set, H is the Heaviside operator (Merriman *et al.*, 1994) which provides the inside and outside regions, and $g_{\mathrm{r}}^{\mathrm{in}}$, $g_{\mathrm{r}}^{\mathrm{out}}$ are the region functions defined in Section 2.2.

An alternate iterative procedure is adopted to solve multi-variable optimization problem (8). After updating $g_{\mathrm{r}}^{\mathrm{in}}$, $g_{\mathrm{r}}^{\mathrm{out}}$, the Euler-Lagrange equation for the level set function is as follows:

$$\frac{\partial\boldsymbol{\Phi}}{\partial t}=\left(\operatorname{div}\frac{\nabla\boldsymbol{\Phi}}{|\nabla\boldsymbol{\Phi}|}+\lambda(g_{\mathrm{r}}^{\mathrm{in}}(\boldsymbol{\theta})-g_{\mathrm{r}}^{\mathrm{out}}(\boldsymbol{\theta}))\right)\delta(\boldsymbol{\Phi}).\tag{9}$$

If δ is regularized not to vanish over the whole image domain Ω, we can remove this function in Eq. (9) without changing the optimality condition:

$$\frac{\partial\boldsymbol{\Phi}}{\partial t}=\operatorname{div}\frac{\nabla\boldsymbol{\Phi}}{|\nabla\boldsymbol{\Phi}|}+\lambda(g_{\mathrm{r}}^{\mathrm{in}}(\boldsymbol{\theta})-g_{\mathrm{r}}^{\mathrm{out}}(\boldsymbol{\theta})).\tag{10}$$

Eq. (10) is the gradient descent equation of the following energy functional:

$$\min_{\boldsymbol{\Phi}\in\{0,1\}}\left\{E=\int_{\Omega}|\nabla\boldsymbol{\Phi}|\mathrm{d}x+\lambda\int_{\Omega}g_{\mathrm{r}}^{\mathrm{in}}\boldsymbol{\Phi}\mathrm{d}x+\lambda\int_{\Omega}g_{\mathrm{r}}^{\mathrm{out}}(1-\boldsymbol{\Phi})\mathrm{d}x\right\}.\tag{11}$$

Function $\boldsymbol{\Phi}$ of LSM is defined in {0, 1}, which is a non-convex set. Thus, Eq. (11) is still not a global optimization problem. If we relax $\boldsymbol{\Phi}\in\{0, 1\}$ to a continuous interval [0, 1] and change the notation $\boldsymbol{\Phi}$ to $\boldsymbol{u}$ to avoid confusion with LSM, energy functional (11) is equal to

$$\min_{\boldsymbol{u}\in[0,1]}\left\{E_{\mathrm{G}}=\int_{\Omega}|\nabla\boldsymbol{u}|\mathrm{d}x+\lambda\int_{\Omega}r(x,\boldsymbol{\theta})\boldsymbol{u}\mathrm{d}x\right\},\tag{12}$$

where $r(x,\boldsymbol{\theta})=g_{\mathrm{r}}^{\mathrm{in}}-g_{\mathrm{r}}^{\mathrm{out}}=(\boldsymbol{\theta}\cdot\boldsymbol{\rho}+I_t)^2-\alpha\mathrm{e}^{-(\boldsymbol{\theta}\cdot\boldsymbol{\rho}+I_t)^2}$.

Eq. (12) is the convex spatiotemporal segmentation model based on 3D videos. We have the following remarks on the energy functional: (1) $\int_{\Omega}|\nabla\boldsymbol{u}|\mathrm{d}x$ is the total variation of $\boldsymbol{u}$, equivalent to the curvature in LSM. (2) Energy Eq. (12) is convex in function $\boldsymbol{u}$, which makes a global minimization solution to the segmentation problem possible (only if $g_{\mathrm{r}}^{\mathrm{in}}$ and $g_{\mathrm{r}}^{\mathrm{out}}$ are fixed). (3) Since the standard active contour is homogeneous of degree 1 in function $\boldsymbol{u}$, $\boldsymbol{u}$ is restricted to [0, 1] to have a stationary state. In addition, $\boldsymbol{u}$ is a 3D matrix because the energy Eq. (12) is defined on the spatiotemporal domain. (4) The inside and outside regions are determined by thresholding the function $\Omega_{\mathrm{in}}=\{x|\boldsymbol{u}(x)>\sigma\}$, $\Omega_{\mathrm{out}}=\Omega_{\mathrm{in}}^{\mathrm{c}}$ for certain $\sigma\in(0, 1)$, where $\Omega_{\mathrm{in}}^{\mathrm{c}}$ is the complement of Ω_{in}.

4 Energy minimization

There are two groups of variables in the energy functional, namely the spatiotemporal segmentation surface and the background motion parameters. Thus, the optimization problem is to minimize the energy simultaneously with respect to $\boldsymbol{u}$ and $\boldsymbol{\theta}$. Fig. 3 gives an alternate iterative approach to Eq. (12).

While $\|\boldsymbol{u}^k-\boldsymbol{u}^{k-1}\|>\mathrm{tol}$ **do**

Estimate motion parameters $\boldsymbol{\theta}^k=(\boldsymbol{T}^k,\boldsymbol{W}^k)$;

Update region functions $g_{\mathrm{in}}^k=(\boldsymbol{\theta}\cdot\boldsymbol{\rho}+I_t)^2$,

$g_{\mathrm{out}}^k=\alpha\mathrm{e}^{-(\boldsymbol{\theta}\cdot\boldsymbol{\rho}+I_t)^2}$;

Solve $\boldsymbol{u}^k=\arg\min_{0\le\boldsymbol{u}\le1}\left(|\nabla\boldsymbol{u}|+\lambda r(x,\boldsymbol{\theta})\boldsymbol{u}\right)$;

Find $\Omega_{\mathrm{in}}^k=\{x\,|\,\boldsymbol{u}^k(x)>\sigma\}$, and $\Omega_{\mathrm{out}}^k=(\Omega_{\mathrm{in}}^k)^{\mathrm{c}}$;

Until convergence

Fig. 3 Standard minimization based on the partial differential equation

4.1 Motion estimation

For fixed function $\boldsymbol{u}$, motion estimation can derive from the Euler-Lagrange equations which minimize energy functional (12):

$$\frac{\partial t_i}{\partial\tau}=-\frac{\partial E_{\mathrm{G}}(\boldsymbol{u},\boldsymbol{\theta})}{\partial t_i},\ \frac{\partial\omega_i}{\partial\tau}=-\frac{\partial E_{\mathrm{G}}(\boldsymbol{u},\boldsymbol{\theta})}{\partial\omega_i},\ i=1,2,3.\tag{13}$$

Here we use τ as the variable of time to distinguish from the motion parameter $\boldsymbol{t}$. In practice, we omit the

integral term introduced by g_r^{out}, which cannot provide new information to update $\boldsymbol{T}$ or $\boldsymbol{W}$. This not only reduces computation complexity, but also consists with the assumption that each point in the background region should meet the 3D optical flow constraint. Thus, we have the following iteration equations for the translational and rotational components of 3D motion:

$$\begin{cases} t_1^{k+1}=t_1^k+\Delta\tau\left(-\int_{\Omega_{\text{in}}^k}\frac{2fI_x}{Z}(\boldsymbol{\rho}\cdot\boldsymbol{\theta}+I_t)\mathrm{d}x\right), \\ t_2^{k+1}=t_2^k+\Delta\tau\left(-\int_{\Omega_{\text{in}}^k}\frac{2fI_y}{Z}(\boldsymbol{\rho}\cdot\boldsymbol{\theta}+I_t)\mathrm{d}x\right), \\ t_3^{k+1}=t_3^k+\Delta\tau\left(-\int_{\Omega_{\text{in}}^k}\frac{2(xI_x+yI_y)}{Z}(\boldsymbol{\rho}\cdot\boldsymbol{\theta}+I_t)\mathrm{d}x\right), \end{cases} \tag{14a}$$

$$\begin{cases} \omega_1^{k+1}=\omega_1^k+\Delta\tau\left(-\int_{\Omega_{\text{in}}^k}2(-fI_y-\frac{y}{f}(xI_x+yI_y))(\boldsymbol{\rho}\cdot\boldsymbol{\theta}+I_t)\mathrm{d}x\right), \\ \omega_2^{k+1}=\omega_2^k+\Delta\tau\left(-\int_{\Omega_{\text{in}}^k}2(fI_x+\frac{x}{f}(xI_x+yI_y))(\boldsymbol{\rho}\cdot\boldsymbol{\theta}+I_t)\mathrm{d}x\right), \\ \omega_3^{k+1}=\omega_3^k+\Delta\tau\left(-\int_{\Omega_{\text{in}}^k}2(xI_y-yI_x)(\boldsymbol{\rho}\cdot\boldsymbol{\theta}+I_t)\mathrm{d}x\right), \end{cases} \tag{14b}$$

where $\Delta\tau$ is the time step for motion estimation.

4.2 Surface optimization

4.2.1 Steepest descent minimization scheme

Energy functional (12) with respect to the spatio-temporal surface can be minimized using any optimization algorithm such as the standard gradient descent method since it is convex in $\boldsymbol{u}$, only if g_r^{in} and g_r^{out} are fixed. The standard minimization flow of functional E_G with respect to $\boldsymbol{u}$ is

$$\boldsymbol{u}_t=\nabla\cdot\left(\frac{\nabla\boldsymbol{u}}{|\nabla\boldsymbol{u}|}\right)+\lambda r(x,\boldsymbol{\theta}). \tag{15}$$

The numerical bottleneck of Eq. (15) comes from the total variation part in energy functional (12) that suffers from serious nonlinearity and non-differentiability. An artificial time marching was introduced by Rudin *et al.* (1992) to solve such gradient flow as in the form of Eq. (15). This method is slow due to the regularization process of the TV-norm and the strict constraints on the time step size. Also, it computes the solutions of not the exact energy E_G but its approximation $\int_\Omega\sqrt{|\nabla\boldsymbol{u}|^2+\varepsilon}\mathrm{d}x$, where ε is a small positive constant to avoid numerical instability. Many techniques have been proposed to overcome this difficulty. Chan *et al.* (2004) enforced the problem on the convex space of $\boldsymbol{u}$: $\Omega\to[0, 1]$ via an exact penalizer $\upsilon(\boldsymbol{u})=\max\{0, 2|\boldsymbol{u}-0.5|-1\}$. The drawback of this method is that the penalty function is non-differentiable either.

4.2.2 Primal-dual method

The primal-dual method (Chan *et al.*, 1999; Chambolle, 2004) is an efficient method for total variation minimization. Chan *et al.* (1999) introduced an additional variable $\boldsymbol{p}=\nabla\boldsymbol{u}/|\nabla\boldsymbol{u}|=(p^1, p^2, p^3)$ to remove some of the singularities caused by the non-differentiability of TV-norm before applying a linearity technique such as Newton's method. Thus, the total variation can be rewritten as follows:

$$\mathrm{TV}(\boldsymbol{u})=\int_\Omega|\nabla\boldsymbol{u}|\mathrm{d}x=\sup_{|\boldsymbol{p}|\le1}\left\{\int_\Omega\boldsymbol{u}(x)\mathrm{div}\,\boldsymbol{p}\,\mathrm{d}x\right\}. \tag{16}$$

Here $\boldsymbol{u}$ and $\boldsymbol{p}$ are the so-called primal and dual variables, respectively. Similar dual formulation was proposed by Chambolle (2004), whose algorithm is faster even if the convergence of Chambolle (2004)'s scheme is linear and Chan *et al.* (1999)'s scheme is quadratic.

We use a convex regularization of the variational model (12):

$$E_G(\boldsymbol{u},\boldsymbol{v},\boldsymbol{\theta})=\int_\Omega\left(|\nabla\boldsymbol{u}|+\lambda r(x,\boldsymbol{\theta})\boldsymbol{v}+\frac{1}{2\varepsilon}\|\boldsymbol{u}-\boldsymbol{v}\|_{\text{L2}}^2\right)\mathrm{d}x, \tag{17}$$

where ε is a positive constant and $\boldsymbol{v}$ is an auxiliary variable to enforce $\boldsymbol{u}\approx\boldsymbol{v}$ for sufficiently small ε. This convex minimization problem can be optimized by alternating steps as follows:

1. When $\boldsymbol{v}$ is fixed, we search for $\boldsymbol{u}$ as a solution of

$$\boldsymbol{u}^k=\arg\min_{\boldsymbol{u}}\int_\Omega\left(|\nabla\boldsymbol{u}|+\frac{1}{2\varepsilon}\|\boldsymbol{u}-\boldsymbol{v}\|_{\text{L2}}^2\right)\mathrm{d}x. \tag{18}$$

Based on dual formulation of the TV-norm, the solution of Eq. (18) is given by

$$\boldsymbol{u} = \boldsymbol{v} - \varepsilon \cdot \operatorname{div} \boldsymbol{p}. \quad (19)$$

Substituting Eq. (19) for minimal $\boldsymbol{u}$ into Eq. (18) gives

$$\min_{|\boldsymbol{p}|\leq 1} \int_{\Omega} \left[(\boldsymbol{v} - \varepsilon \operatorname{div} \boldsymbol{p}) \operatorname{div} \boldsymbol{p} + \frac{\varepsilon}{2} || \operatorname{div} \boldsymbol{p} ||^2 \right] \mathrm{d}x. \quad (20)$$

As Chambolle (2004) showed, we can use a semi-implicit gradient algorithm for the solution of $\boldsymbol{p}$:

$$\boldsymbol{p}^{k+1} = \frac{\boldsymbol{p}^k + \Delta t \cdot \nabla(\operatorname{div} \boldsymbol{p}^k - \boldsymbol{v}/\varepsilon)}{1 + \Delta t \, |\nabla(\operatorname{div} \boldsymbol{p}^k - \boldsymbol{v}/\varepsilon)|}, \quad \boldsymbol{p}^0 = 0, \quad (21)$$

where the time step Δt=1/8.

2. When $\boldsymbol{u}$ is fixed, we search for $\boldsymbol{v}$ as a solution of

$$\boldsymbol{v}^k = \arg\min_{v} \int_{\Omega} \left(\lambda r(x, \boldsymbol{\theta}) \boldsymbol{v} + \frac{1}{2\varepsilon} || \boldsymbol{u} - \boldsymbol{v} ||_{\mathrm{L2}}^2 \right) \mathrm{d}x. \quad (22)$$

Thus, the solution of Eq. (22) is given by

$$\boldsymbol{v} = \min\{\max\{\boldsymbol{u}(x) - \varepsilon\lambda r(x, \boldsymbol{\theta}),\ 0\},\ 1\}. \quad (23)$$

Taking the update of motion estimation into consideration, the primal-dual method for the proposed 3D spatiotemporal segmentation model can be formulated as the algorithm given in Fig. 4.

While $|| \boldsymbol{u}^{k+1} - \boldsymbol{u}^k ||$> tol **do**

Estimate motion parameters

$\boldsymbol{\theta}^{k+1} = \arg\min_{\boldsymbol{\theta}} \sum_{\Omega_{\mathrm{in}}^k} (\boldsymbol{\theta} \cdot \boldsymbol{\rho} + I_t)^2$;

Update region functions g_{in}^{k+1}, g_{out}^{k+1};

For $\boldsymbol{v}$ being fixed, $\boldsymbol{u}^{k+1} = \boldsymbol{v}^k - \varepsilon \cdot \operatorname{div} \boldsymbol{p}^k$;

For $\boldsymbol{u}$ being fixed,

$\boldsymbol{v}^{k+1} = \min\{\max\{\boldsymbol{u}^{k+1}(x) - \varepsilon\lambda r(x, \boldsymbol{\theta}^{k+1}),\ 0\},\ 1\}$;

Find $\Omega_{\mathrm{in}}^{k+1} = \{x \,|\, \boldsymbol{u}^{k+1}(x) > \sigma\}$, and $\Omega_{\mathrm{out}}^{k+1} = (\Omega_{\mathrm{in}}^{k+1})^{\mathrm{c}}$;

Until convergence

Fig. 4 Primal-dual method for the proposed 3D spatiotemporal segmentation model

5 Experimental results

All the experimental sequences are captured by a 3D TOF camera, SwissRanger SR-3000 from Mesa, which is capable of producing 3D images in real-time. Both intensity and range images have 176×144 pixels in each frame, and the range image encodes a pixel's depth information linearly. All numerical experiments were performed under Windows Version. We used MATLAB 7.5.0 (R2007b) to load and save images, in addition to calling the codes.

5.1 Single target

In the 'car' example, both the car (leftward moving) and the SR-3000 camera move at different speeds. We use frames 101–105 to validate our algorithm. Motion parameters $\boldsymbol{T}$ and $\boldsymbol{W}$ are initialized to zero. Since the energy functional (12) is convex with respect to $\boldsymbol{u}$, the initialization of function $\boldsymbol{u}$ can be an arbitrary value in the interval [0, 1]. We fix the initialization of all variables, determine the spatiotemporal derivative for the given pairs of consecutive frames, and iterate the minimization of energy functional (Eq. (12)), alternating the motion estimation (Eq. (14)) and surface propagation (Eq. (19)). Fig. 5a shows the value map of $\boldsymbol{u}$ for frame 103, in which the value approaches 1 if the pixel is in the background region. In contrast to other implicit representations of segmentation such as the signed distance function, the value of $\boldsymbol{u}$ has a clear probabilistic interpretation. Actually the function $\boldsymbol{u}(\boldsymbol{x})$: $\mathbb{R}^d \rightarrow [0, 1]$ specifies the probability that a pixel $\boldsymbol{x} \in \mathbb{R}^d$ belongs to the background.

Fig. 5b gives the segmentation curve of frame 103 with the threshold σ=0.5. Fig. 5e is its corresponding depth map captured by the TOF camera. A comparison of our method to that of Feghali and Mitiche (2004) is shown in Fig. 5c. Obviously, the 3D model based method yields a significant improvement in comparison to the approach where the motion is estimated for each region by a 2D flow field (u_{op}, v_{op}) computed from the spatiotemporal domain of image sequences. The optic flow field is given in Fig. 5d by mapping the estimated parameters $\boldsymbol{T}$ and $\boldsymbol{W}$ into an image field according to Eqs. (3a) and (3b).

5.2 Multiple targets

In this example, we validate the proposed method with multiple targets. The cars and the TOF camera are moving at their own directions and speeds. We choose frames 256–262 of the 'car' sequence as the spatiotemporal domain. Fig. 6a gives the 3D spatiotemporal surface after the primal-dual method

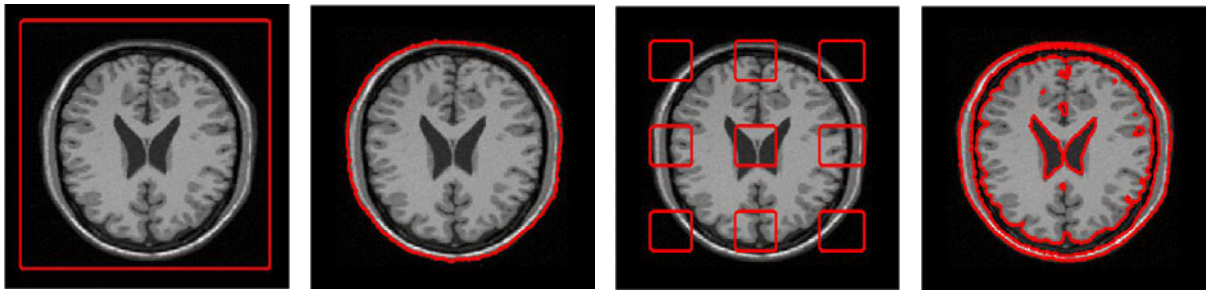

Fig. 2 Comparison of different initializations for the Chan-Vese model
(a) Initial curves I; (b) Segmentation result of (a); (c) Initial curves II; (d) Segmentation result of (c)

is performed, in which two targets are successfully tracked, and each surface demonstrates the trajectory of its corresponding target. As can be seen from Fig. 6a, each time-slice of the surface is the segmentation curve of its corresponding frame. Take frame 261 as an example. Fig. 6b depicts the optical flow field, and the corresponding depth map encoded in HSV is given in Fig. 6c.

A comparison of the proposed method to the approach of Wang and Yu (2011) using LSM is shown in Fig. 7. Figs. 7a and 7b give the segmentation results with respect to the non-convex energy functional (7) by LSM with two different initial curves. As shown in Fig. 7b, the small car cannot be tracked under the second condition. The reason is that LSM highly depends on the initialization; thus, it is easily trapped into local minima. In contrast, the proposed method is globally convex with any initialization of function $\boldsymbol{u}$. We design four different initializations to validate Algorithm 2. The first two initializations of $\boldsymbol{u}$ are shown in Figs. 7c and 7d, in which the white region represents 1 and the black region represents 0. The third is a random uniform distribution of $\boldsymbol{u}$ in [0, 1], and the last is a gradual change of $\boldsymbol{u}$ along the x axis in [0, 1]. As shown in Figs. 7c–7f, both cars are successfully partitioned under all these initial conditions.

5.3 Non-rigid target

Fig. 8 shows the segmentation results of a sequence including rigid and non-rigid targets. Again, we choose four consecutive frames 147–150 from the 'person and car' sequence, in which the person is circumvolving, and the camera and car have their own motion. Fig. 8a shows the final curve of frame 147 with σ=0.5. Fig. 8b gives the map of function $\boldsymbol{u}$, and Fig. 8c is its corresponding depth map captured by SR-3000. Figs. 8d–8f display the segmentation results of frames 148–150.

5.4 Parameter selection

Parameter selection is a key step in energy-based methods. Without loss of generality, we set the parameters α=λ=1, ε=1/5, Δt=1/8, σ=0.5 for the primal-dual method in all the experiments.

As discussed in Section 2, there are two scale parameters α and λ in the proposed energy functional. α balances the forces between the background and foreground, and λ measures the forces between the regions and surfaces. Therefore, both α and λ are selected according to the characteristics of the experimental 3D videos. In addition, the empirical values of α and λ are between 0 and 10, to satisfy the requirements of most 3D videos.

In the primal-dual algorithm, the time step Δt should satisfy Δt<1 to guarantee stability. Here we choose Δt=1/8. Moreover, ε is a positive parameter to enforce $\boldsymbol{u}\approx\boldsymbol{v}$ for sufficiently small ε. Theoretically, the smaller the ε (i.e., as $\varepsilon\to 0$), the more accurate the approximation of $\boldsymbol{v}$. However, the scheme slows down as the accuracy increases. Therefore, attention should be given to both factors when selecting ε.

Another important parameter is the threshold function σ, which is a key factor in determining the segmentation result. For the 'person and car' sequence, Fig. 9a displays the map of function $\boldsymbol{u}$ after convergence for frame 148 shown in Fig. 9b, where the contour lines $C=\{\boldsymbol{x}|\boldsymbol{u}(\boldsymbol{x})=\sigma\}$ are given with respect to σ=0.5, 0.7, 0.9. The segmentation curves in Figs. 9c–9e correspond to the three values of σ.

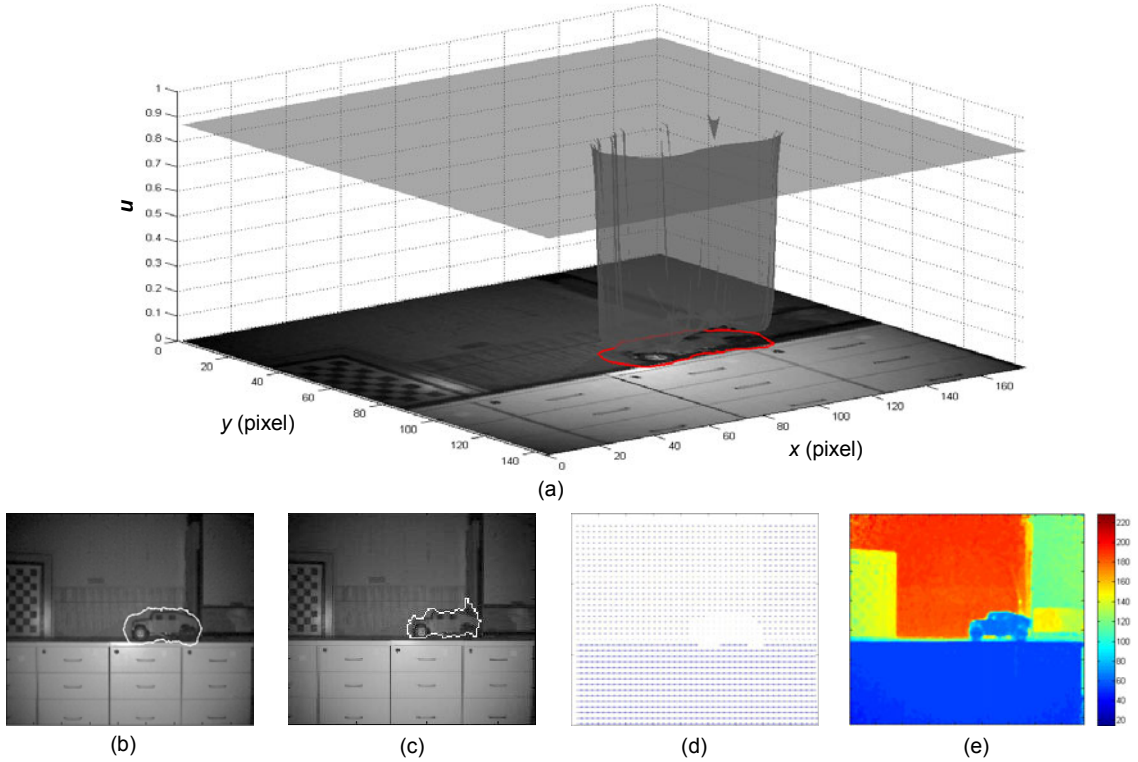

Fig. 5 Segmentation results of the 'car' sequence

(a) Function ***u*** of frame 103; (b) Segmentation result of frame 103 with our method; (c) Segmentation result with the 2D method (Feghali and Mitiche, 2004); (d) The corresponding optical flow; (e) Color encoded depth map of frame 103

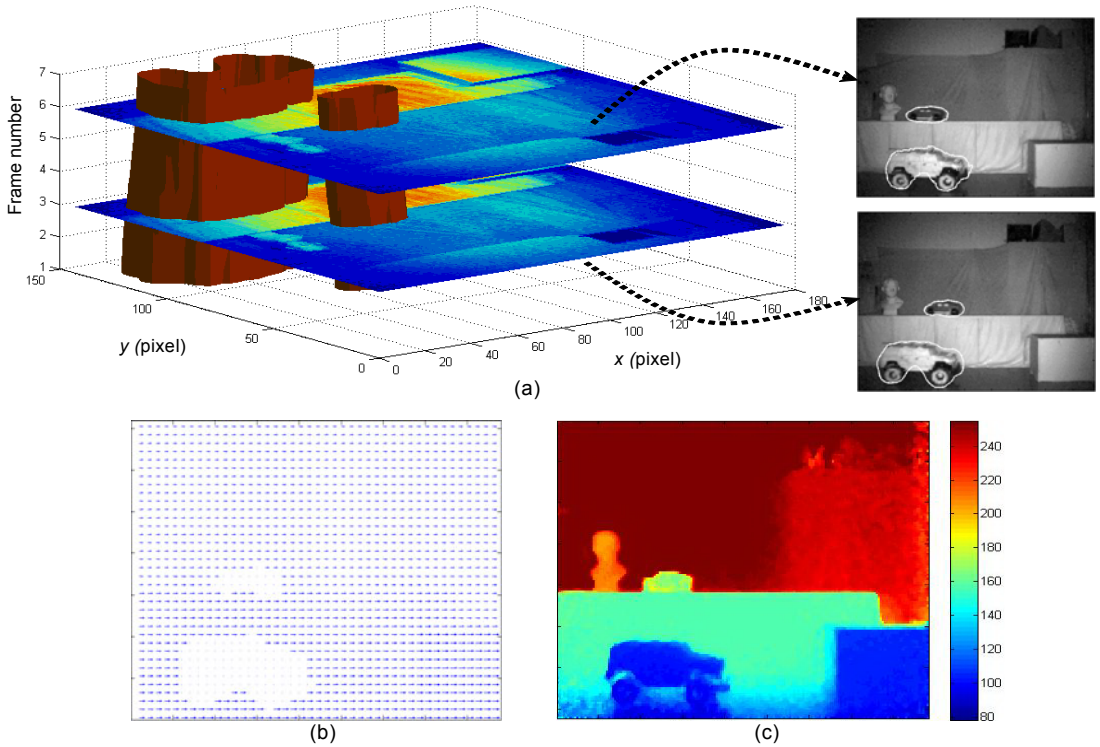

Fig. 6 Segmentation results of the 'car' sequence

(a) 3D spatiotemporal surface and its corresponding segmentation curves; (b) Optical flow of frame 261; (c) Color encoded depth map of frame 261

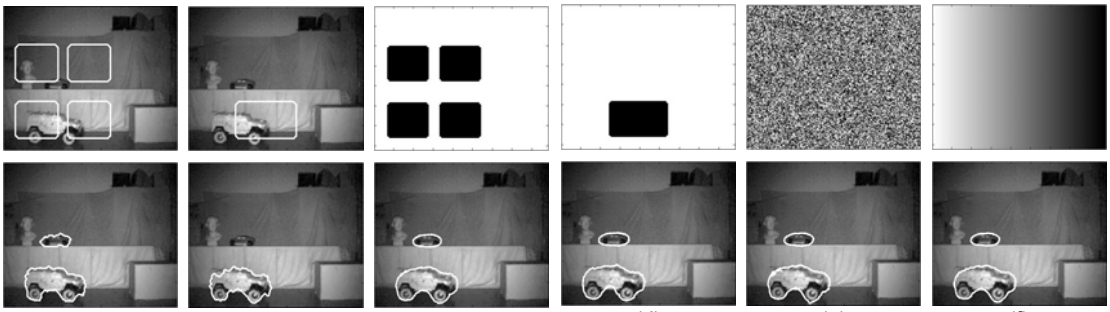

Fig. 7 Comparison of the proposed method with the level set method (LSM)

(a) Segmentation by LSM with initial curves I; (b) Segmentation by LSM with initial curves II; (c) Segmentation by the proposed method with initialization I of $\boldsymbol{u}$ (white region represents 1 and black region represents 0); (d) Segmentation with initialization II of $\boldsymbol{u}$; (e) Segmentation with a random uniform distribution of $\boldsymbol{u}$ in [0, 1]; (f) Segmentation with a gradual change of $\boldsymbol{u}$ in [0, 1]

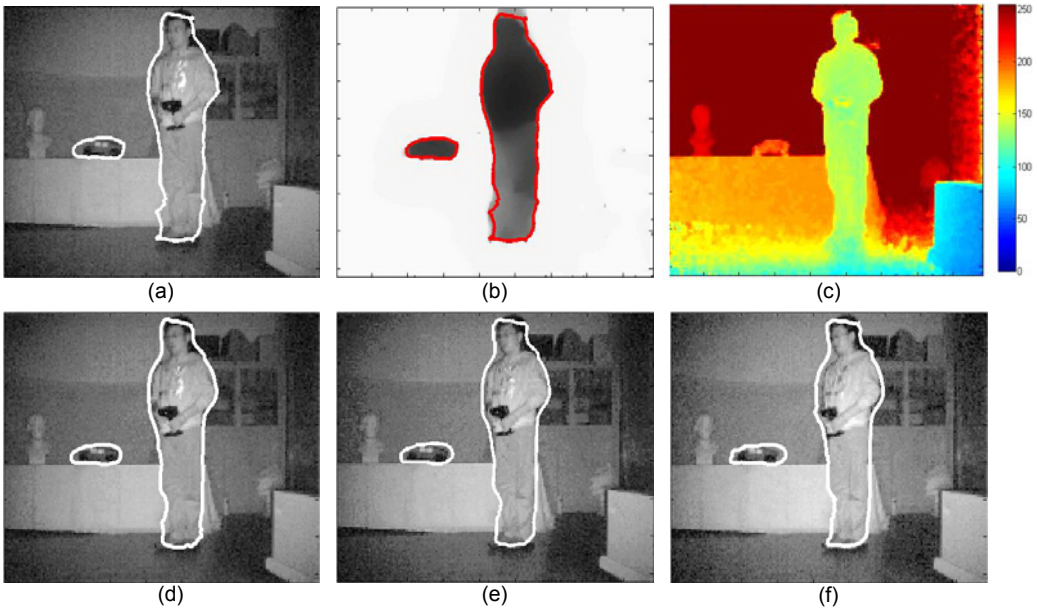

Fig. 8 Segmentation results of the 'person and car' sequence

(a) Final curve of frame 147; (b) Function $\boldsymbol{u}$ of frame 147; (c) Color encoded depth map of (a); (d–f) Segmentation results of frames 148–150

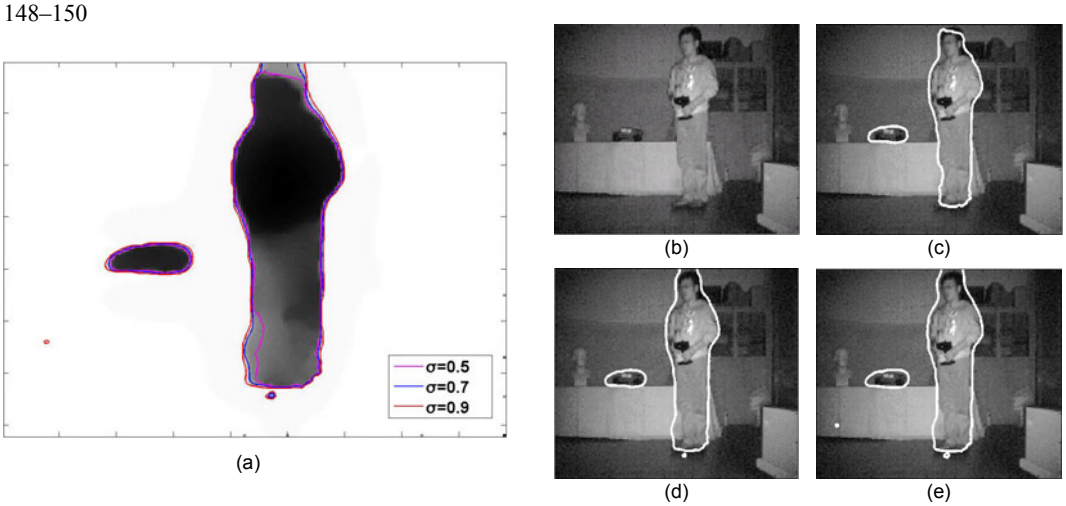

Fig. 9 Segmentation results under different thresholds σ

(a) Contour lines with respect to σ=0.5, 0.7, 0.9 on the map of function $\boldsymbol{u}$; (b) Frame 148; (c) Segmentation result with σ=0.5; (d) Segmentation result with σ=0.7; (e) Segmentation result with σ=0.9

6 Conclusions

A novel method is presented for 3D spatio-temporal segmentation and motion estimation with a moving viewing system. Different from existing motion-based segmentation algorithms that use level sets, our proposed method is a 3D videos based method, and the main benefit is that it is performed by embedding 3D motion parameters and the evolution surface into an energy functional and applying joint segmentation and motion estimation. To overcome the limitation of the local minimum problem with the level set method, we give the convex formulation of the energy functional for the 3D spatiotemporal segmentation model. A primal-dual method is applied to solve the global convex minimization problem, improving the computational efficiency. All experimental sequences are provided by a 3D TOF camera. Various indoor experiments demonstrate the high efficiency of the proposed model in comparison with traditional methods.

For further work we will investigate the extension of this two-phase global minimization approach to multi-phase variational models.

References

Alvarez, L., Castano, C.A., Garcia, M., Krissian, K., Mazorra, L., Salgado, A., Sanchez, J., 2009. A new energy-based method for 3D motion estimation of incompressible PIV flows. *Comput. Vis. Image Understand.*, **113**(7):802-810. [doi:10.1016/j.cviu.2009.01.005]

Boykov, Y., Funka-Lea, G., 2006. Graph cuts and efficient N-D image segmentation. *Int. J. Comput. Vis.*, **70**(2):109-131. [doi:10.1007/s11263-006-7934-5]

Boykov, Y., Kolmogorov, V., 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, **26**(9):1124-1137. [doi:10.1109/TPAMI.2004.60]

Bresson, X., Esedoglu, S., Vandergheynst, P., Thiran, J., Osher, S., 2007. Fast global minimization of the active contour/snake model. *J. Math. Imag. Vis.*, **28**(2):151-167. [doi:10.1007/s10851-007-0002-0]

Caselles, V., Kimmell, R., Sapiro, G., 1997. Geodesic active contours. *Int. J. Comput. Vis.*, **22**(1):61-79. [doi:10.1023/A:1007979827043]

Chambolle, A., 2004. An algorithm for total variation minimization and applications. *J. Math. Imag. Vis.*, **20**(1-2):89-97. [doi:10.1023/B:JMIV.0000011325.36760.1e]

Chan, T.F., Vese, L.A., 2001. Active contours without edges. *IEEE Trans. Image Process.*, **10**(2):266-277. [doi:10.1109/83.902291]

Chan, T.F., Golub, G.H., Mulet, P., 1999. A nonlinear primal-dual method for total variation-based image restoration. *SIAM J. Sci. Comput.*, **20**(6):1964-1977. [doi:10.1137/S1064827596299767]

Chan, T.F., Esedoglu, S., Nikolova, M., 2004. Algorithms for finding global minimizers of image segmentation and denoising models. *SIAM J. Appl. Math.*, **66**:1632-1648. [doi:10.1137/040615286]

Dufaux, F., Moccagatta, I., Moscheni, F., Nicolas, H., 1994. Vector quantization-based motion field segmentation under the entropy criterion. *J. Vis. Commun. Image Represent.*, **5**:356-369. [doi:10.1006/jvci.1994.1034]

Feghali, R., Mitiche, A., 2004. Spatiotemporal motion boundary detection and motion boundary velocity estimation for tracking moving objects with a moving camera: a level sets PDEs approach with concurrent camera motion compensation. *IEEE Trans. Image Process.*, **13**(11):1473-1490. [doi:10.1109/TIP.2004.836158]

Goldstein, T., Bresson, X., Osher, S., 2009. Geometric applications of the split Bregman method: segmentation and surface reconstruction. *J. Sci. Comput.*, **45**(1-3):272-293. [doi:10.1007/s10915-009-9331-z]

Heeger, D.J., Jepson, A.D., 1992. Subspace methods for recovering rigid motion I: algorithm and implementation. *Int. J. Comput. Vis.*, **7**(2):95-117. [doi:10.1007/BF00128130]

Horn, B.K.P., Schunck, B.G., 1981. Determining optical flow. *Artif. Intell.*, **17**(1-3):185-203. [doi:10.1016/0004-3702(81)90024-2]

Jonasson, L., Bresson, X., Hagmann, P., Cuisenaire, O., Meuli, R., Thiran, J.P., 2005. White matter fiber tract segmentation in DT-MRI using geometric flows. *Med. Image Anal.*, **9**(3):223-236. [doi:10.1016/j.media.2004.07.004]

Kass, M., Witkin, A., Terzopoulos, D., 1988. Snakes: active contour models. *Int. J. Comput. Vis.*, **1**(4):321-331. [doi:10.1007/BF00133570]

Leventon, M.E., Grimson, W.E.L., Faugeras, O., 2000. Statistical Shape Influence in Geodesic Active Contours. IEEE Conf. on Computer Vision and Pattern Recognition, **1**:316-323. [doi:10.1109/CVPR.2000.855835]

Longuet-Higgins, H.C., Prazdny, K., 1980. The interpretation of a moving retinal image. *Proc. R. Soc. Lond. B*, **208**(1173):385-397. [doi:10.1098/rspb.1980.0057]

Malladi, R., Kimmel, R., Adalsteinsson, D., Sapiro, G., Caselles, V., Sethian, J.A., 1996. A Geometric Approach to Segmentation and Analysis of 3D Medical Images. Proc. Workshop on Mathematical Methods in Biomedical Image Analysis, p.244-252. [doi:10.1109/MMBIA.1996.534076]

Merriman, B., Bence, J.K., Osher, S.J., 1994. Motion of multiple junctions: a level set approach. *J. Comput. Phys.*, **112**(2):334-363. [doi:10.1006/jcph.1994.1105]

Mémin, E., Pérez, P., 2002. Hierarchical estimation and segmentation of dense motion fields. *Int. J. Comput. Vis.*, **46**(2):129-155. [doi:10.1023/A:1013539930159]

Mumford, D., Shah, J., 1989. Optimal approximations of piecewise smooth functions and associated variational

problems. *Commun. Pure Appl. Math.*, **42**(5):577-685. [doi:10.1002/cpa.3160420503]

Ohno, K., Nomura, T., Tadokoro, S., 2006. Real-Time Robot Trajectory Estimation and 3D Map Construction Using 3D Camera. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, p.5279-5285. [doi:10.1109/IROS.2006.282027]

Osher, S., Sethian, J., 1988. Fronts propagating with curvature dependent speed: algorithms based on the Hamilton-Jacobi formulation. *J. Comput. Phys.*, **79**(1):12-49. [doi:10.1016/0021-9991(88)90002-2]

Paragios, N., Deriche, R., 2005. Geodesic active regions and level set methods for motion estimation and tracking. *Comput. Vis. Image Understand.*, **97**(3):259-282. [doi:10.1016/j.cviu.2003.04.001]

Rudin, L.I., Osher, S., Fatemi, E., 1992. Nonlinear total variation based noise removal algorithms. *Phys. D*, **60**(1-4):259-268. [doi:10.1016/0167-2789(92)90242-F]

Schnorr, C., 1991. Determining optical flow for irregular domains by minimizing quadratic functionals of a certain class. *Int. J. Comput. Vis.*, **6**(1):25-38. [doi:10.1007/BF00127124]

Wang, S., Yu, H., 2011. A Variational Approach for Egomotion Estimation and Segmentation Based on 3D TOF Camera. 4th Int. Congress on Image and Signal Processing, p.1160-1164. [doi:10.1109/CISP.2011.6100402]

Ye, C., Hegde, G.P.M., 2009. Robust Edge Extraction for SwissRange SR-3000 Range Image. Proc. IEEE Int. Conf. on Robotics and Automation, p.2437-2442. [doi:10.1109/ROBOT.2009.5152559]