JOINT SPACE-TIME IMAGE SEQUENCE SEGMENTATION: OBJECT TUNNELS AND OCCLUSION VOLUMES

Mirko Ristivojević and Janusz Konrad

Department of Electrical and Computer Engineering, Boston University
8 Saint Mary's Street, Boston, MA 02215 [mirko,jkonrad]@bu.edu

ABSTRACT

Spatial segmentation of image sequences is usually computed based on motion between two frames. Some recent approaches extend this to joint segmentation in space-time; the resulting 3-D segmentation (in $x-y-t$ space) can be interpreted as a volume "carved out" by a moving object in the image sequence domain, or the so-called "object tunnel". In this paper, we extend this concept to explicit modeling of occlusion events in the $x-y-t$ space. In addition to the modeling of object evolution, we also model occluded and newly-exposed areas in the background and in the object by means of "occlusion volume", a new space-time concept. We propose a variational formulation of the problem that we solve using the multiphase level set method. We show experimental results for synthetic and natural image sequences.

1. INTRODUCTION

In most studies to date, image sequences have been processed and analyzed in groups of two frames. Therefore, they cannot take longer-term dynamics into account. In order to speed up convergence, subsequent segmentations can be initialized by previous-frame results. An alternative approach, although closely related, is tracking of segments between frames. Some early work using multiple frames includes motion detection using 3-D MRF models and "video-cube" segmentation based on marker selection and volume growing. More recently, a novel concept of object tracking as spatio-temporal boundary detection has been proposed by El-Feghali *et al.* [1]. This new, multiple-image framework has lead to interesting space-time image sequence segmentation methods developed by Mansouri *et al.* [2] and, independently, by the authors [3]. The resulting 3-D, or volumetric, segmentation in $x-y-t$ space can be interpreted as a volume "carved out" by a moving object in the image sequence domain. We call such volumes "object tunnels".

In terms of the detection of occlusions, methods proposed to date are primarily based on the analysis of 2-3 frames [4, 5, 6], although an exception is the approach developed by Chahine and Konrad [7], where 5 and 7 frames were used. None of the above methods treats occlusions as a continuing event across many frames. In this paper, we extend our video segmentation work by explicitly modeling the evolution of objects and background using motion trajectories. We include explicit models of the occluded and newly-exposed areas for both the object and background. This is in contrast to our recent work [8] where we applied such model only to the background. Similarly to that work, we measure object and background intensity variations along motion trajectories spanning the whole temporal support of the image sequence. Clearly, parts of the background are occluded or exposed within this support and cannot be accurately modeled either by object or background motion trajectories. As the result, they will be randomly included in either object or background segmentation volumes, thus creating errors. The more frames of the sequence are processed, the more serious these errors become. Similar errors occur when an object gets occluded or exposed. In this paper, we explicitly model these regions as occluded and newly-exposed volumes for objects and for background. We use variational framework for the formulation and level-set methodology for the solution. As for motion trajectories, we use a parametric model associated either with objects or background.

This work was supported by the National Science Foundation (NSF) under grant CCR-0209055.

2. MULTIPHASE MOTION-BASED SEGMENTATION

We want to partition an image sequence $I(\boldsymbol{x},t)$, $\boldsymbol{x}$ being a spatial position and t being time, into six regions (volumes): moving object, moving or static background, background area that is going to be occluded by the object in subsequent frames (background occlusion volume), background area that was exposed by the object in preceding frames (background exposed volume), object area that is going to be occluded by a feature in the background (object occlusion volume), and object area that was exposed in preceding frames (object exposed volume). Furthermore, we want, jointly with the segmentation, to estimate motion parameters of the object and background. Using the multiphase level-set framework recently proposed by Vese and Chan [9], we partition the spatio-temporal volume of $I(\boldsymbol{x},t)$ into six volumes using three parameterized surfaces, $\vec{\varsigma}_1$, $\vec{\varsigma}_2$, and $\vec{\varsigma}_3$:

Object volume $\mathcal{V}_1$	$(\boldsymbol{x},t)$ inside $\vec{\varsigma}_1, \vec{\varsigma}_2, \vec{\varsigma}_3$,
Object occluded vol. $\mathcal{V}_5$	$(\boldsymbol{x},t)$ inside $\vec{\varsigma}_2, \vec{\varsigma}_3$, outside $\vec{\varsigma}_1$,
Object exposed vol. $\mathcal{V}_6$	$(\boldsymbol{x},t)$ inside $\vec{\varsigma}_1, \vec{\varsigma}_3$, outside $\vec{\varsigma}_2$,
Backg. volume $\mathcal{V}_2$	$(\boldsymbol{x},t)$ outside $\vec{\varsigma}_1, \vec{\varsigma}_2, \vec{\varsigma}_3$,
Backg. occluded vol. $\mathcal{V}_3$	$(\boldsymbol{x},t)$ inside $\vec{\varsigma}_1$, outside $\vec{\varsigma}_2, \vec{\varsigma}_3$,
Backg. exposed vol. $\mathcal{V}_4$	$(\boldsymbol{x},t)$ inside $\vec{\varsigma}_2$, outside $\vec{\varsigma}_1, \vec{\varsigma}_3$.

Since three surfaces can partition the image sequence domain into 8 volumes, additional volumes, $\mathcal{V}_7$ and $\mathcal{V}_8$, are defined as "don't care" volumes that will be eliminated during optimization. An example of cross-sections of various volumes is shown in Figs. 2(c-d). Part of the object visible throughout the sequence is shown in white, and following regions are represented in shades of gray, from light to dark gray: part of the object that is going to be occluded in subsequent frames, background, part of the background

that is going to be occluded in subsequent frames, and background region exposed in preceding frames.

Let $\boldsymbol{p}, \bar{\boldsymbol{p}}$ be motion parameters (e.g., affine) associated with the object and background, respectively. Using the multiphase representation [9], we propose the following variational formulation:

$$\begin{aligned}\min_{\vec{\varsigma}_1,\vec{\varsigma}_2,\vec{\varsigma}_3,\boldsymbol{p},\bar{\boldsymbol{p}}} & \iiint_{\mathcal{V}_1} \xi(\boldsymbol{x},t;\boldsymbol{p})d\boldsymbol{x}dt + \omega_2 \iiint_{\mathcal{V}_2} \xi(\boldsymbol{x},t;\bar{\boldsymbol{p}})d\boldsymbol{x}dt + \\ & \omega_3 \iiint_{\mathcal{V}_3} \xi_1(\boldsymbol{x},t;\bar{\boldsymbol{p}})d\boldsymbol{x}dt + \omega_4 \iiint_{\mathcal{V}_4} \xi_2(\boldsymbol{x},t;\bar{\boldsymbol{p}})d\boldsymbol{x}dt + \\ & \omega_5 \iiint_{\mathcal{V}_5} \xi_1(\boldsymbol{x},t;\boldsymbol{p})d\boldsymbol{x}dt + \omega_6 \iiint_{\mathcal{V}_6} \xi_2(\boldsymbol{x},t;\boldsymbol{p})d\boldsymbol{x}dt + \\ & \iiint_{\mathcal{V}_7} K_{pen}d\boldsymbol{x}dt + \iiint_{\mathcal{V}_8} K_{pen}d\boldsymbol{x}dt + \\ & \lambda_1 \iint_{\mathcal{S}_1} d\vec{\varsigma}_1, + \lambda_2 \iint_{\mathcal{S}_2} d\vec{\varsigma}_2, + \lambda_3 \iint_{\mathcal{S}_3} d\vec{\varsigma}_3,\end{aligned} \tag{1}$$

where $\vec{\varsigma}_1 = \partial(\mathcal{V}_1 \cup \mathcal{V}_3 \cup \mathcal{V}_6 \cup \mathcal{V}_7)$, $\vec{\varsigma}_2 = \partial(\mathcal{V}_1 \cup \mathcal{V}_4 \cup \mathcal{V}_5 \cup \mathcal{V}_7)$, $\vec{\varsigma}_3 = \partial(\mathcal{V}_1 \cup \mathcal{V}_5 \cup \mathcal{V}_6 \cup \mathcal{V}_8)$, and $\cup_{i=1}^{8}\mathcal{V}_i = \Omega \times \mathcal{T}$, while $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ are areas of the three surfaces. The constant weights ω_i reflect the uncertainty as to how accurately motion parameters can explain the dynamics occurring in individual volumes, while constants λ_1, λ_2, and λ_3 associate a cost with the Euclidean lengths $d\vec{\varsigma}_1$, $d\vec{\varsigma}_2$, and $d\vec{\varsigma}_3$, respectively. Penalty terms, K_{pen}, are introduced to discourage assigning a point to the two unused volumes ($\mathcal{V}_7$ and $\mathcal{V}_8$). The individual terms of the above energy measure cumulative consistency of image sequence voxels with models corresponding to different volumes we are estimating:

- *object volume* term ($\mathcal{V}_1$): measures intensity variation (sample variance) along object motion trajectories (small only for voxels for which a set of motion parameters exists that induces small intensity variations, e.g., object);
- *background volume* term ($\mathcal{V}_2$): measures intensity variation along background motion trajectories (small only for voxels for which the another set of motion parameters exists that induces small intensity variations, e.g., background);
- *background occlusion volume* term ($\mathcal{V}_3$): unexplained image dynamics ahead of the voxel (future), i.e., occlusion at a later time (small only for voxels for which intensity variation along background motion trajectories is small up to that voxel and large afterward),
- *exposed background volume* term ($\mathcal{V}_4$): unexplained image dynamics prior to the voxel (past), i.e., newly-exposed pixels (small only for voxels for which intensity variation along background motion trajectories is large up to that voxel and small afterward),
- *object occlusion volume* term ($\mathcal{V}_5$): similar to the background occlusion term, only here we measure intensity variation along object motion trajectories;
- *exposed object volume* term ($\mathcal{V}_6$): similar to the exposed background term, only here we measure intensity variation along object motion trajectories.

The above measures are implemented through the following expressions:

$$\begin{aligned}\xi_0(\boldsymbol{x},t;\boldsymbol{p};k,l) &= \frac{1}{k-l+1}\sum_{i=k}^{l}(\widetilde{I}(\boldsymbol{c}(t_i;\boldsymbol{x},t),t_i) - \mu_{k,l}(\boldsymbol{x},t;\boldsymbol{p}))^2, \\ \xi(\boldsymbol{x},t;\boldsymbol{p}) &= \xi_0(\boldsymbol{x},t;\boldsymbol{p};1,N), \\ \xi_1(\boldsymbol{x},t=t_j;\boldsymbol{p}) &= \xi_0(\boldsymbol{x},t;\boldsymbol{p};1,j) + \frac{\alpha_1}{\xi_0(\boldsymbol{x},t;\boldsymbol{p};j,N)+1}, \\ \xi_2(\boldsymbol{x},t=t_j;\boldsymbol{p}) &= \frac{\alpha_2}{\xi_0(\boldsymbol{x},t;\boldsymbol{p};1,j)+1} + \xi_0(\boldsymbol{x},t;\boldsymbol{p};j,N),\end{aligned} \tag{2}$$

where N is the number of frames in the image subsequence we are currently processing, $\boldsymbol{c}(t_i;\boldsymbol{x},t)$ is the object or background motion trajectory (i.e., $\boldsymbol{c}$ is a spatial position at time t_i of a feature that was at position $\boldsymbol{x}$ at time t [7]). α_1 and α_2 are weighting coefficients used to adjust the influence of the two terms in ξ_1 and ξ_2. The average intensities along trajectories $\boldsymbol{c}$, between two time instants t_i and t_j, are computed as follows:

$$\mu_{i,j}(\boldsymbol{x},t;\boldsymbol{p}) = \frac{1}{j-i+1}\sum_{k=i}^{j}\widetilde{I}(\boldsymbol{c}(t_k;\boldsymbol{x},t),t_k), \tag{3}$$

where $\widetilde{I}$ denotes the interpolated intensity (e.g., bicubic interpolator) because $(\boldsymbol{c}(t_i;\boldsymbol{x},t),t_i)$ need not belong to Λ.

We decompose minimization (1) into two interleaved minimizations: estimation of motion parameters given segmentation surfaces, and estimation of segmentation surfaces with fixed motion parameters. We describe the latter minimization; details of motion parameter estimation can be found in [8].

Following the multiphase level set method formalism [9], we represent energy minimization in (1) using three level set functions, $\phi_1(\boldsymbol{x},t)$, $\phi_2(\boldsymbol{x},t)$, and $\phi_3(\boldsymbol{x},t)$. Regions we are estimating are now defined through level set functions as follows:

$$\begin{array}{ll}Object\ volume & \phi_1(\boldsymbol{x},t)>0,\ \phi_2(\boldsymbol{x},t)>0,\ \phi_3(\boldsymbol{x},t)>0, \\ Object\ occl.\ vol. & \phi_1(\boldsymbol{x},t)<0,\ \phi_2(\boldsymbol{x},t)>0,\ \phi_3(\boldsymbol{x},t)>0, \\ Object\ exp.\ vol. & \phi_1(\boldsymbol{x},t)>0,\ \phi_2(\boldsymbol{x},t)<0,\ \phi_3(\boldsymbol{x},t)>0, \\ Backg.\ volume & \phi_1(\boldsymbol{x},t)<0,\ \phi_2(\boldsymbol{x},t)<0,\ \phi_3(\boldsymbol{x},t)<0, \\ Backg.\ occl.\ vol. & \phi_1(\boldsymbol{x},t)>0,\ \phi_2(\boldsymbol{x},t)<0,\ \phi_3(\boldsymbol{x},t)<0, \\ Backg.\ exp.\ vol. & \phi_1(\boldsymbol{x},t)<0,\ \phi_2(\boldsymbol{x},t)>0,\ \phi_3(\boldsymbol{x},t)<0.\end{array}$$

In order to carry out minimization (1), level set surfaces should be evolved along the direction of steepest descent, which is the direction of negative gradient of the total energy with respect to ϕ_1, ϕ_2, and ϕ_3. As a result of minimization, we obtain the following level set evolution equations (valid for all $(\boldsymbol{x},t)$ that are omitted for brevity), with τ being the algorithmic evolution time, and κ_{m_1}, κ_{m_2}, and κ_{m_3} are mean curvatures of level set surfaces ϕ_1, ϕ_2, and

ϕ_3, respectively:

$$\frac{\partial \phi_1(\tau)}{\partial \tau} = F_1 \|\nabla \phi_1(\tau)\| = \|\nabla \phi_1(\tau)\| \{\lambda_1 \kappa_{m_1} - [(\xi(\boldsymbol{p}) - \omega_5 \xi_1(\boldsymbol{p}))H(\phi_2(\tau))H(\phi_3(\tau)) + (\omega_6 \xi_2(\boldsymbol{p}) - K_{pen})(1 - H(\phi_2(\tau)))H(\phi_3(\tau)) + (K_{pen} - \omega_4 \xi_2(\bar{\boldsymbol{p}}))H(\phi_2(\tau))(1 - H(\phi_3(\tau))) + (\omega_3 \xi_1(\bar{\boldsymbol{p}}) - \omega_2 \xi(\bar{\boldsymbol{p}}))(1 - H(\phi_2(\tau)))(1 - H(\phi_3(\tau)))]\}$$

$$\frac{\partial \phi_2(\tau)}{\partial \tau} = F_2 \|\nabla \phi_2(\tau)\| = \|\nabla \phi_2(\tau)\| \{\lambda_2 \kappa_{m_2} - [(\xi(\boldsymbol{p}) - \omega_6 \xi_2(\boldsymbol{p}))H(\phi_1(\tau))H(\phi_3(\tau)) + (\omega_5 \xi_1(\boldsymbol{p}) - K_{pen})(1 - H(\phi_1(\tau)))H(\phi_3(\tau)) + (K_{pen} - \omega_3 \xi_1(\bar{\boldsymbol{p}}))H(\phi_1(\tau))(1 - H(\phi_3(\tau))) + (\omega_4 \xi_2(\bar{\boldsymbol{p}}) - \omega_2 \xi(\bar{\boldsymbol{p}}))(1 - H(\phi_1(\tau)))(1 - H(\phi_3(\tau)))]\}$$

$$\frac{\partial \phi_3(\tau)}{\partial \tau} = F_3 \|\nabla \phi_3(\tau)\| = \|\nabla \phi_3(\tau)\| \{\lambda_3 \kappa_{m_3} - [(\xi(\boldsymbol{p}) - K_{pen})H(\phi_1(\tau))H(\phi_2(\tau)) + (\omega_5 \xi_1(\boldsymbol{p}) - \omega_4 \xi_2(\bar{\boldsymbol{p}}))(1 - H(\phi_1(\tau)))H(\phi_2(\tau)) + (\omega_6 \xi_2(\boldsymbol{p}) - \omega_3 \xi_1(\bar{\boldsymbol{p}}))H(\phi_1(\tau))(1 - H(\phi_2(\tau))) + (K_{pen} - \omega_2 \xi(\bar{\boldsymbol{p}}))(1 - H(\phi_1(\tau)))(1 - H(\phi_2(\tau)))]\}$$

We implement these equations iteratively using standard discretization as described by Sethian [10]. In each iteration we calculate the forces F_1, F_2, and F_3 at zero level-set points of all surfaces, extend these forces using the fast marching algorithm by solving $\phi_i \cdot \nabla F_i = 0$ for F_i, $i = 1, 2, 3$, and update the surfaces ϕ_1, ϕ_2, and ϕ_3. Re-initialization of the surfaces using the fast marching algorithm by solving $\|\nabla \phi_i\|$=1 is performed every 100 iterations to keep surfaces as close as possible to a signed distance function.

3. EXPERIMENTAL RESULTS

We have tested the algorithm on several image sequences. First, we used a natural-texture, synthetic-motion, static background, test sequence. We initialized the algorithm with volumes and motion parameters resulting from our simple segmentation algorithm based on motion detection [3]. We used the following parameters: $\alpha_1 = \alpha_2 = 10$, $\alpha_3 = \alpha_4 = 4 * 10^4$, $\lambda_1 = \lambda_2 = \lambda_3 = 2.5$, $\omega_2 = 2, \omega_3 = \omega_4 = \omega_5 = \omega_6 = 10$, $K_{pen} = 10^2$. The algorithm converges after 1000 iterations; four of the final six tunnels are shown on Fig. 1. Fig. 2 shows two frames of the resulting segmentation labels. Clearly, object and background as well as occluded and exposed background are accurately estimated, but occluded and exposed parts of the object are less precise.

We also applied our algorithm to a natural sequence acquired with a static video camera (progressive, 30fr/s): a car enters the scene from the right, moves to left and disappears behind a wall on the left. This time we use the following parameters in our experiment: $\alpha_1 = \alpha_2 = 500, \alpha_3 = \alpha_4 = 2{*}10^4, \lambda_1 = \lambda_2 = \lambda_3 = 2.5$, $\omega_2 = 2, \omega_3 = \omega_4 = 10, \omega_5 = \omega_6 = 4, K_{pen} = 10^2$. Fig. 3 shows the object tunnel and occlusion/exposed volumes corresponding to four segmentation regions, obtained after the algorithm converged at 1000 iterations. For visualization reasons the shown tunnels start at frame #10. Fig. 4 shows two frames of the resulting segmentation labels. As can be seen from this figure, all background regions are well estimated: occlusion region is in front of the car, exposed region grows behind the car. Object occlusion and exposure regions are in the right place (front and rear of the car, respectively) but they are not precise. This is, we believe, due to inaccuracies in the estimated motion parameters and variations of object intensity over time.

4. CONCLUSIONS

We have proposed a novel framework for simultaneous segmentation of video sequences into moving objects and background, and detection of occlusion and newly-exposed areas. We have introduced a new concept of spatio-temporal volumes of occluded and exposed voxels. The framework is based on variational principles and uses the multiphase level-set methodology for solution. The initial results are very encouraging; quite accurate object and background tunnels, as well as occluded/exposed background volumes have been recovered. However, the accuracy of occluded/exposed object volumes is low. We believe this is due to errors in the estimated motion parameters and object's intensity variations in time, and we plan to address this issue.

5. REFERENCES

[1] R. El-Feghali, A. Mitiche, and A.-R. Mansouri, "Tracking as motion boundary detection in spatio-temporal space," in *Int. Conf. Imaging Science, Systems, and Technology*, June 2001, pp. 600–604.

[2] A.-R. Mansouri and A. Mitiche, "Spatial/joint space-time motion segmentation of image sequences by level set pursuit," in *Proc. IEEE Int. Conf. Image Processing*, Sept. 2002, vol. 2, pp. 265–268.

[3] J. Konrad and M. Ristivojević, "Joint space-time image sequence segmentation based on volume competition and level sets," in *Proc. IEEE Int. Conf. Image Processing*, Sept. 2002, vol. 1, pp. 573–576.

[4] F. Heitz and P. Bouthemy, "Multimodal estimation of discontinuous optical flow using Markov random fields," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, no. 12, pp. 1217–1232, Dec. 1993.

[5] M. Irani and S. Peleg, "Motion analysis for image enhancement: resolution, occlusion and transparency," *J. Vis. Commun. Image Represent.*, vol. 4, no. 4, pp. 324–335, Dec. 1993.

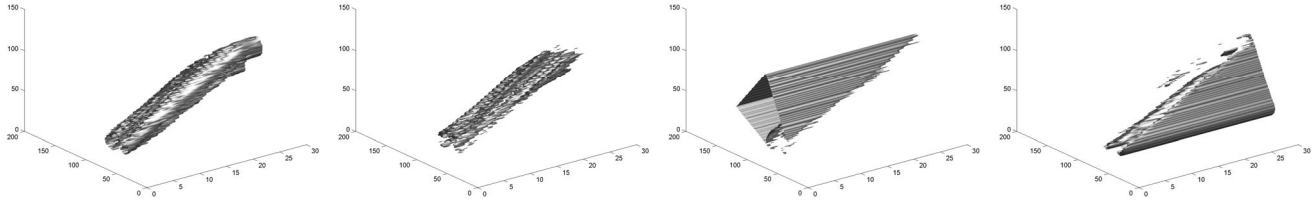
[6] K.P. Lim, A. Das, and M.N. Chong, "Estimation of occlusion and dense motion fields in a bidirectional Bayesian framework," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 5, pp. 712–718, May 2002.

[7] M. Chahine and J. Konrad, "Estimation and compensation of accelerated motion for temporal sequence interpolation," *Signal Process., Image Commun.*, vol. 7, no. 4–6, pp. 503–527, Nov. 1995.

[8] M. Ristivojević and J. Konrad, "Joint space-time motion-based video segmentation and occlusion detection using multi-phase level sets," in *Proc. SPIE Visual Communications and Image Process.*, Jan. 2004.

[9] L. Vese and T. Chan, "A multiphase level set framework for image segmentation using the Mumford and Shah model," *Intern. J. Comput. Vis.*, vol. 50, no. 3, pp. 271–293, 2002.

[10] J.A. Sethian, *Level Set Methods*, Cambridge University Press, 1996.

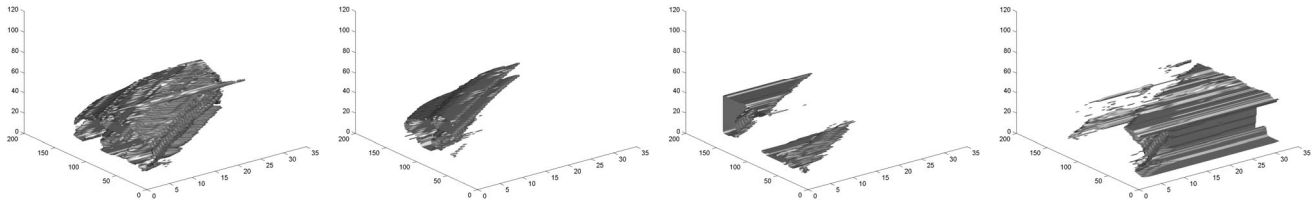


(a) (b) (c) (d)

Fig. 1. Tunnels obtained using the multiphase segmentation algorithm : (a) object, (b) object occlusion, (c) background occlusion, and (d) background exposed volume.

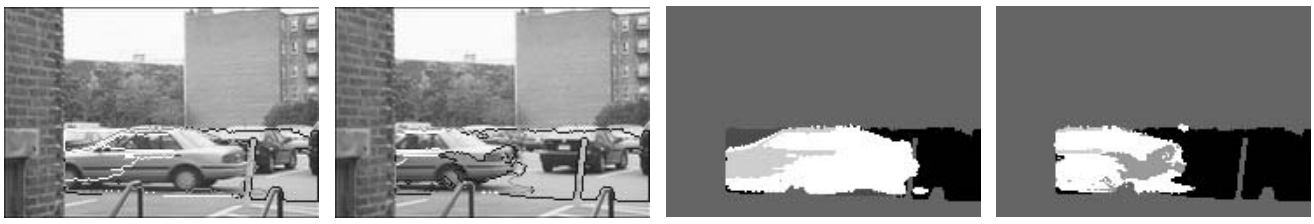
(a) (b) (c) (d)

Fig. 2. Frames (a) #15 and (b) #25 from the synthetic test image sequence overlaid with final level-set contours, and (c-d) corresponding frames from the sequence of labels (white – object, light gray to dark gray: object occlusion region, background, background occlusion, and background exposed region) derived from these results.



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

Fig. 3. Tunnels obtained using the multiphase segmentation algorithm : (a) object, (b) object occlusion, (c) background occlusion, and (d) background exposed volume.



(b) (c) (b) (c)

Fig. 4. Frames (a) #20 and (b) #30 from the car image sequence overlaid with final level-set contours, and (c-d) corresponding frames from the sequence of labels (white – object, light gray to dark gray: object occlusion region, object exposed region, background, background occlusion, and background exposed region) derived from these results.