Object-Based Spatial Segmentation of Video Guided by Depth and Motion Information

Jaime S. Cardoso
Universidade do Porto / INESC Porto
Porto, Portugal
jaime.cardoso@inescporto.pt

Jorge C. S. Cardoso
Escola das Artes, UCP
Porto, Portugal
jccardoso@porto.ucp.pt

Luís Corte-Real
Universidade do Porto / INESC Porto
Porto, Portugal
lreal@inescporto.pt

Abstract

Automatic spatial video segmentation is a problem without a general solution at the current state-of-the-art. Most of the difficulties arise from the process of capturing images, which remain a very limited sample of the scene they represent. The capture of additional information, in the form of depth data, is a step forward to address this problem. We start by investigating the use of depth data for better image segmentation; a novel segmentation framework is proposed, with depth being mainly used to guide a segmentation algorithm on the colour information. Then, we extend the method to also incorporate motion information in the segmentation process. The effectiveness and simplicity of the proposed method is documented with results on a selected set of images sequences. The achieved quality raises the expectation for a significant improvement on operations relying on spatial video segmentation as a pre-process.

1. Introduction

In a conventional camera, the only recorded information for each pixel is position and colour values. The fact of captured images remaining very limited samples of the scene they represent leads to problems in many growing application fields such as film and television post-production, object based video formats and human computer interaction, to name just a few. All the approaches to surmount these problems are based on the estimation of data that is simply not included in the digital images, and so are limited in the quality they can achieve.

The capture of additional data is a step forward to address these problems. Time-of-flight principle cameras with low-cost sensors are now becoming available, e.g., [8]. The european project MetaVision carried out a survey into possible methods for capturing depth and motion information, targeting enhanced pos-production operations [4].

We aim at investigating the use of depth and motion data to allow the development of better spatial segmentation algorithms. In section 2 we start by presenting our framework for spatial segmentation assisted by depth information. In section 3 the proposal is generalized to incorporate also motion information in the spatial segmentation process. Finally, we discuss the attained results in section 4, draw some conclusions and motivate future work in section 5.

2. Depth assisted image segmentation

A natural approach to incorporate the depth information into the segmentation process of a colour image would be to use symmetric models with respect to colour and depth information. However, it is known that the two sources have different degrees of reliability, i.e., in practice depth information is noisier and with lower resolution than colour information. To account for this a "dataset reliability" could be associated with each dataset so that a less reliable dataset has less effect on the fusion process. Even if that was implemented — by modifying a standard segmentation algorithm — we were still left with the problem of estimating the number of segments in the image. Moreover, the possible misalignment between colour and depth information had also to be taken into account.

As such, we suggest that a practical framework for hybrid image segmentation should:

1. use the depth information to automatically estimate the number and localization of objects in the image. This process should produce markers ('hints') to guide the segmentation of the colour image. The colour information may be used together with the depth to assist this process;
2. perform a guided image segmentation, using *essentially* the colour information, starting from the markers previously created.

We will now elaborate each of these two steps.

IEEE Workshop on Motion and Video Computing (WMVC'07)
0-7695-2793-0/07 $20.00

2.1. Marker extraction

In most real-life images, objects have large vertical sections. In order to exploit this property for object segmentation, and as others before [6, 5, 9], a density image is defined by transforming the depth information on the XY plane to the XZ plane. The value at position (x, z) of the density image denotes the number of points in the depth image at position x, taking the value z (by 'integrating' along the Y direction): let $D(x, y)$ be the depth information value at position (x, y) and $d(x, z)$ the value of the density image at position (x, z); then

$$d(x,z) = \sum_y \delta\left(D(x,y) - z\right) \qquad \text{with } \delta(n) \begin{cases} 1 \text{ if } n = 0 \\ 0 \text{ otherwise} \end{cases}$$

Early efforts have exploited the XZ image segments to infer bounding boxes for objects in the XY image. Our experiments with these models provided limited results, as the objects were not completely inside the extracted bounding boxes. This scenario suggested the use of the XZ image for object marker extraction; that can be accomplished using a simple threshold technique. More generally, we incorporated in this phase

1. a *pre-processing step*, which can include a low-pass filter, morphological operations, histogram equalisation or other preparatory operations. Our implementation performs a centered morphological opening operation, with a rectangular structuring element of size $(2o_h + 1) \times (2o_v + 1)$, followed by a centered morphological closing operation, with a rectangular structuring element of size $(2c_h + 1) \times (2c_v + 1)$;

2. an *"hysteresis" thresholding operation*. If a value is not inferior to the upper threshold limit, t_h, it is immediately accepted; if the value lies below the low threshold, t_l, it is immediately rejected; points which lie between the two limits are accepted if they are connected to pixels which exhibit strong response (at least t_h);

3. a *connected component analysis*. Each connected component is identified as an object marker. Our system uses 8-connected neighbourhoods;

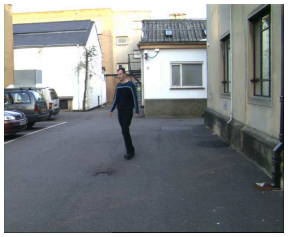
4. a *post-processing step*, with objects with z values less than a predefined value (10% of the maximum possible z value in our implementation) being ignored. Low z values correspond to points farthest away of the camera or points to which the depth could not be estimated.

The resulting object segmentations, in the XZ image, for the images in Figure 1, are presented in Figures 2(a) and 2(c); each object is represented with a unique colour.

The object markers can be transported to the XY plane by including a pixel (x, y) in the marker of the object $\mathcal{O}_i$ if the corresponding $(x, z) = (x, D(x, y))$ value in the density image lies in the object marker $\mathcal{O}_i$. The result is illustrated in Figures 2(b) and 2(d).

(a) 'chess' image.

(b) 'chess' depth map.

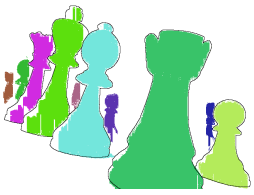


(c) 'walk street' image.

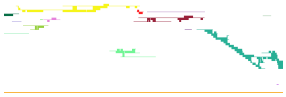
(d) 'walk street' depth map.

Figure 1. Selected test images, with perfect and noisy depth information.

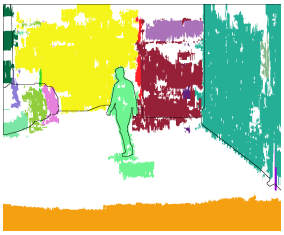
(a) Objects in the XZ image. $t_h = \quad t_l = \quad o_h = 2 \quad o_v = 1 \quad c_h = 0 \quad c_v = 3$; 10 objects detected.



(b) Markers in the XY image, superimposed on the ground truth segmentation.



(c) Objects in the XZ image. $t_h = 40 \quad t_l = 20 \quad o_h = 2 \quad o_v = 0 \quad c_h = 0 \quad c_v = 3$; 16 objects detected.

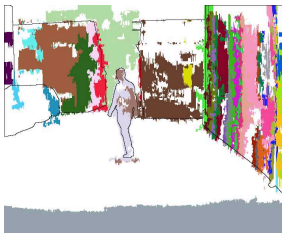


(d) Markers in the XY image, superimposed on the ground truth segmentation.

Figure 2. Automatic marker extraction for the test image set.

While the pre-processing by morphological opening was implemented mainly as a kind of noise removal, with the effect of eliminating small and thin objects, the pre-processing by closing has the effect of filling small and thin holes in objects, and *connecting nearby objects*. This last property is of major importance in the presence of real-life depth maps. The quantisation effect in depth information, when sufficiently severe, is responsible for the formation of small vertical strips, incorrectly identified as individual objects (see Figure 3(a)). Morphological closing, with a structure element large enough to overcome the quantization effect on depth, allows to restore the object connectivity, as seen in Figure 2(c).

(a) Objects in the XZ image without closing as a pre-process. $t_h = 40, t_l = 20, o_h = 2, o_v = 0, c_h = 0, c_v = 0$



(b) Objects in the XY image without closing as a pre-process, superimposed on the ground truth segmentation.

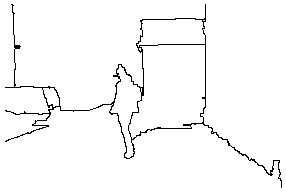
Figure 3. Importance of morphological closing as a pre-process.

2.2. Guided image segmentation

Many approaches are possible for the segmentation using the hints produced in the previous stage. A well-known choice for guided image segmentation algorithm is based on the watershed. That can be achieved by modifying the image so that it only has regional minima wherever the markers are nonzero, using the H-minima transform [7].

This technique requires a marker to the outer of the detected objects. Toward that end the distance to the closest marked pixel is computed for all non-marked pixels — marked pixels are naturally at zero distance. Because it is likely that markers do not fill completely the corresponding object, an outside marker was defined as the set of pixels with a computed distance above a predefined threshold. Empirically the threshold was set at 0.25 the maximum distance. The final segmentation is presented in Figure 4. Appreciate the good control of the number of segments achieved, without compromising the quality of the segmentation. The objects were correctly identified as a whole.

(a) 'chess' segmentation image.



(b) 'walk street' segmentation image.

Figure 4. Results for the watershed with markers algorithm.

3. Spatial video segmentation assisted by depth and motion information

We have tackled the issue of improving the quality of image segmentations with depth information. However, depth is not the Graal of the segmentation problem. Consider the case when two objects moving in opposite directions cross with each other. If their depth is similar, even the method presented so far will fail to divide the two objects. Or, as already observed, it is still difficult to segment objects from the ground where they stand. This creates the interest for integrating more information in the segmentation techniques. Motion information is an exemplar candidate. Velocity information may be used to link adjacent but visually dissimilar surfaces or to divide surfaces not easily separable by static criteria alone. Often, ambiguous object boundaries in a single image frame are easily resolved when dynamic effects are evaluated based on a sequence of frames.

If it is true that for synthetic sequences motion values can be computed exactly, that is not the typical scenario, where motion is *estimated* from a sequence of images. Therefore, our approach should be robust against inaccuracies in the motion information, as it is against inaccuracies in the depth information. A key observation when addressing the problem of segmenting assisted by both depth and motion information is that these two cases of distinct information, which are often treated separately, have in fact much in common (see Figure 5):

- depth information is typically computed from stereo information, with two images acquired simultaneously.
- motion information is typically computed from sequential information, with two images acquired sequentially.

Akin to motion techniques, a class of stereo methods is based on the matching of small blocks. Therefore, techniques integrating depth and motion in the segmentation

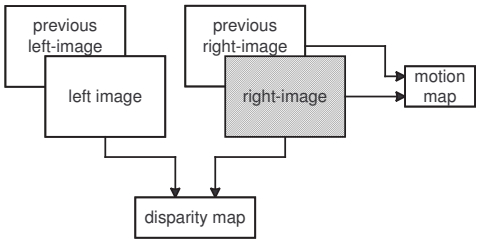


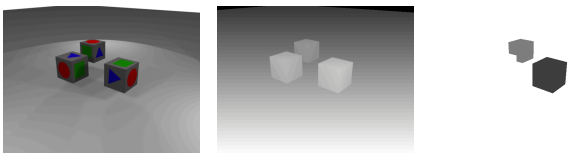
Figure 5. Motion and depth estimation for the right image.

process should be symmetric with respect to these two sources.

Depth information was used to automatically estimate the number and localization of objects in the image. The estimation was performed from markers derived from a density image $d(x, z)$. The same methodology can be adopted for the motion data, creating a density image $d(x, v)$ and extracting object markers. A sensible approach now is to produce a single marker-image, integrating both marker-images. However, note that they are defined over different domains. That hinders the direct merging of both marker-images. To surmount this problem, the integration can be performed in the (x, y) plane, by first transposing both marker-images to this plane:

$$\begin{aligned} d_{xz}(x, y) &= d(x, z) \quad \text{if } Z(x, y) = z \\ d_{xv}(x, y) &= d(x, v) \quad \text{if } V(x, y) = v \end{aligned} \tag{1}$$

Having depth-markers and motion-markers, we want them to cross-validate each other and to allow depth-markers to sometimes divide motion-markers (for objects with similar movement at different depths) and the opposite, motion-markers to divide depth-markers (to divide objects at similar depth but with different movements). A powerful tool to visualize such set-up is the intersection graph, as represented in Figure 9 for the 'cubes' image. The 'cubes' image is presented in Figure 6; the corresponding depth- and motion-maps are in Figures 7 and 8.



(a) 'cubes' colour image. (b) 'cubes' depth map. (c) 'cubes' motion map.

Figure 6. 'cubes' image.

Each node d_i representes the set of pixels belonging to the same marker in the depth map (with an extra node, d_3,

(a) 'cubes' XZ markers. (b) 'cubes' XY markers.

Figure 7. Depth markers in the XV and XY planes.

(a) 'cubes' XV markers. (b) 'cubes' XY markers.

Figure 8. Motion markers in the XV and XY planes.

to represent the unmarked pixels); each node m_i representes the set of pixels belonging to the same marker in the motion map (with an extra node, m_3, to represent the unmarked pixels); and the weight of each edge represents the number of pixels in the intersection of a marker in the depth map with a marker in the motion map. We would like to come up with a sensible procedure yielding three markers in this example, represented thicker in figure 9, translating into the markers shown in figure 10(a).

The problem now is to define a procedure for choosing which intersection should give rise to a new independent marker, and which will be aggregated under the unmarked

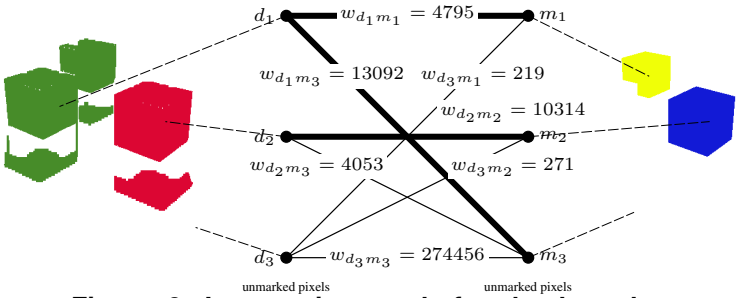


Figure 9. Intersection-graph for depth and motion maps.

(a) Ideal merged markers. (b) Real merged markers.

Figure 10. Merging for the 'cubes' image.

pixels. An empirical evaluation of several alternatives, led us to suggest the following rule: associate a new marker to an intersection if the intersection weight is a substantial part of any of the two incident nodes[1]. Mathematically

$$\text{if } \max\left(\frac{w_{d_i,m_j}}{\sum_\ell w_{d_\ell,m_j}}, \frac{w_{d_i,m_j}}{\sum_\ell w_{d_i,m_\ell}}\right) \begin{cases} > \epsilon \text{ mark intersection } d_i, m_j \\ \leq \epsilon \text{ unmark intersection } d_i, m_j \end{cases} \quad (2)$$

Stated equivalently

$$\text{if (local inconsistency)} \begin{cases} < (1-\epsilon) \text{ mark intersection } d_i, m_j \\ \geq (1-\epsilon) \text{ unmark intersection } d_i, m_j \end{cases}$$

where the local inconsistency is given by

$$\min\left(\frac{\sum_\ell w_{d_\ell,m_j} - w_{d_i,m_j}}{\sum_\ell w_{d_\ell,m_j}}, \frac{\sum_\ell w_{d_\ell,m_j} - w_{d_i,m_j}}{\sum_\ell w_{d_i,m_\ell}}\right)$$

Adopting this procedure, the fused marker would yield ($\epsilon = 0.4$) as represented in Figure 10(b). Note that three markers were indeed created. However, the noise present in the depth markers was not completely removed. In fact, because this noise is jointed in edge $w_{d_1 m_3}$ with pixels corresponding to the leftmost cube, it impossible to recover from it with this framework (without loosing the marker of the leftmost cube). Even adopting a generic approach based on fuzzy logic, it would have been difficult to eliminate this noise, as the density values at the leftmost cube and at the noisy pixels are essentially the same.

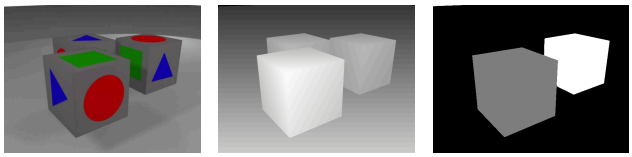
We propose then to extend the technique presented in section 2 in the following way:

- create the XZ density image, operate on it to extract depth marker and transport them to the XY plane, as in the basic proposal.
- repeat the above procedure, now using the motion information, resulting in a motion marker image in the XY plane.
- merge both XY marker-maps using the local refinement error to prune markers in non-concordant pixels.
- apply the colour image segmentation guided by the fused markers.

Although we have implicitly assumed throughout this discussion that the motion information is in the scalar form, yielding a scalar motion map, that is not typically the case, as motion information is typically available in the X and Y directions or in any other equivalent form such as (intensity, angle). In this case we would have two density images, from which two motion marker-images could be created and merged with the depth marker-image. Observing that, although the local inconsistency is commutative it is not associative, the above-defined procedure would have to be conveniently extended to handle three or more marker-images. A possibility would be to generalize the local inconsistency itself to three or more images. Because in this work we will restrict to one motion marker-image, this generalization will not be further pursued here.

4. Results

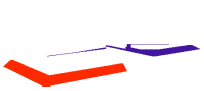
A synthetic image sequence was created having three cubes in the center surrounded by a background, see Figure 11. The sequence was rendered with Maya 7; the depth maps were generated using Maya's renderer; the motion maps were manually generated.



(a) Colour image. (b) Depth map. (c) Motion map.

Figure 11. Frame 6 of 'cubes' sequence.

Following the procedure already formulated, depth markers in the XZ plane and motion markers in the XV plane were extracted, as depicted in Figure 12.

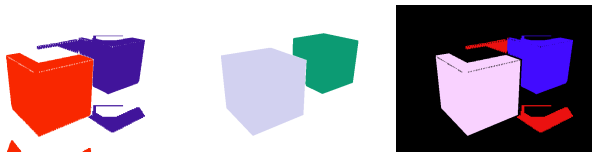


(a) Depth markers in the XZ plane. (b) Motion markers in the XV plane.

Figure 12. Extracted markers for frame 6 of 'cubes' sequence.

[1]If the edge is connecting a node corresponding to a marker with a node corresponding to unmarked pixels, the test should be made only with the node corresponding to the marker. If an edge is connecting two nodes corresponding to unmarked pixels, it is always unmarked.

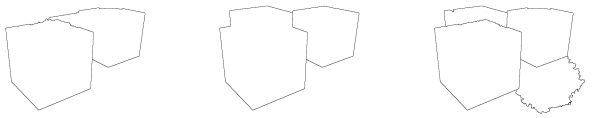
Then, markers were transported to the XY plane and merged following the method adopted in the previous section, creating three different marker-images, depicted in Figure 13. Finally, the comparisons were done using three



(a) XY markers for depth information. (b) XY markers for motion information. (c) XY markers by merging depth and motion markers.

Figure 13. Extracted markers in the XY plane, for frame 6 of 'cubes' sequence.

versions of the guided watershed segmentation: starting from depth markers we obtained segmentation 14(a), starting from motion markers alone, the segmentation 14(b) is attained; from the merged markers resulted the segmentation 14(c). When both depth and motion information is used to extract markers we are able to correctly divide the three cubes (although with a extra spurious region as side effect).



(a) Guided by depth markers. (b) Guided by motion markers. (c) Guided by depth and motion markers.

Figure 14. Watershed image segmentation guided by the extracted markers.

4.1. Results with real-life sequences

This section provides experimental results obtained on the 'Indoor' stereo sequence acquired with a monochrome MEGA-D digital stereo head (by Videre Design) equipped with a pair of 4.8 mm lenses.[2] The image size is 640×480. Two temporally consecutive stereo pairs of the sequence are shown in Figure 15. The depth information, in the form of a disparity map, obtained with the Single Matching Phase (SMP) stereo algorithm [3], is also freely available[2]. It is depicted in Figure 16(a) for frame 113.

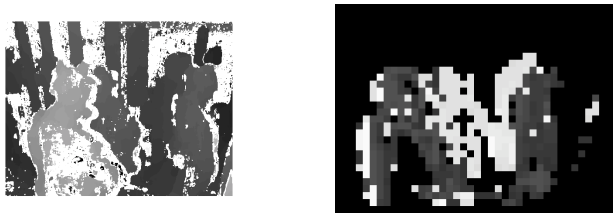
[2]The sequence was downloaded from `http://labvisione.deis.unibo.it/~smattoccia/stereo.htm`.

(a) Left frame 112. (b) Right frame 112.

(c) Left frame 113. (d) Right frame 113.

Figure 15. 'Indoor' stereo sequence.

The motion information was computed with a basic block matching algorithm, as implemented in the OpenCV software, with a block size of 16×16 and a search region of 65×65. The obtained motion information in the X-direction is depicted in Figure 16(b).



(a) Right frame 113 depth image. (b) Right frame 113 motion image.

Figure 16. Computed depth and motion images for right frame 113 of 'Indoor' sequence.

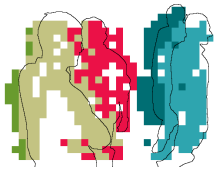
Note that the depth map is smaller than the original image (due to the stereo depth algorithm). Depth and motion markers were extracted (see Figures 17(a) and 17(b)) and transported into the XY plane, shown in Figures 17(c) and 17(d). We considered only the X component of the motion vectors. Observe that, unlike the motion information, the depth information was unable to create distinct markers for each person.

Finally, we proceeded with the guided image segmentation step, experimenting three different possibilities for the initial markers: depth-markers only, motion-markers only, and the merging of depth- and motion-markers. Depth and motion markers were merged with $\epsilon = 0.4$. Due to the high

(a) XZ markers. $t_h = 20$ $t_l = 3$ $o_h = 2$ $o_v = 1$ $c_h = 0$ $c_v = 2$; 3 objects detected.

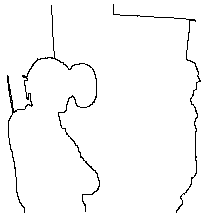
(b) XV markers. $t_h = 45$ $t_l = 3$ $o_h = 2$ $o_v = 0$ $c_h = 0$ $c_v = 8$; 5 objects detected.

(c) XY markers for depth data, superimposed on the ground truth segmentation.

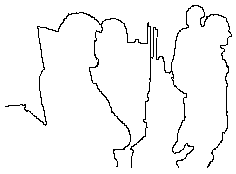


(d) XY markers for motion data, superimposed on the ground truth segmentation.

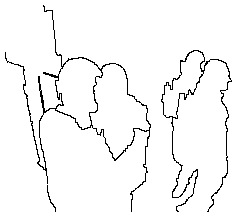
Figure 17. Extracted markers for left frame 113 of 'Indoor' sequence.



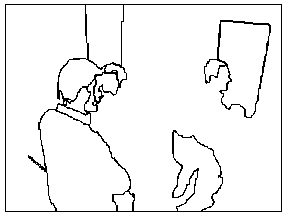
(a) Watershed segmentation guided by depth markers alone.



(b) Watershed segmentation guided by motion markers alone.



(c) Watershed segmentation guided by depth and motion markers together.



(d) Mean-shift based method.

Figure 18. Results for frame 113 of the 'Indoor' sequence.

level of noise present in the depth and motion information, only the colour information was used to compute the gradient fed to the watershed algorithm. Results are depicted in Figure 18. As also noted previously in the synthetic example, here too the integration of motion and depth information in the segmentation process leads to a decent division of the two men present in the image. It is important to stress that this was achieved with very low quality depth and motion information.

This experiment shows the strengths of the system presented here. The combination of motion and depth information in the marker extraction step leads to a more reliable and consistent segmentation — observe the incapability of separating the men from each other and from the background when using only depth to guide the watershed algorithm. Motion information improves segmentation results, without assuming motion continuity. In this aspect, the system is general and performs well.

4.1.1 Image sequence processing

This study was completed by segmenting a set of 12 consecutive frames of the 'Indoor' sequence, from frame 102 to frame 113, shown in Figure 19. All parameters of the algorithm were kept constant for the whole set. Results relative to a manual ground truth are summarized in Table 1. The proposed algorithm was compared with standard segmentation algorithms. Results with the Mean Shift algorithm [2] are also reported in Table 1. Figure 18(d) presents a tipical segmentation obtained with the Mean Shift algorithm. The comparison of the methods rely on the quantitative measures introduced in [1], namely d_{sym} and d_{mut}. The partition-distance d_{sym} is a strict discrepancy measure between two segmentations of the same image; under- and over-segmentations are appropriately penalised. This measure attains the zero value only when the two segmentations coincide exactly. However, for some purposes, it is important to have measures tolerant to mutual refinements, relaxing the conditions for proximity between two segmentations. Intuitively, two segmentations are consistent if they are partially a over-segmentation, partially a under-segmentation of each other. This consistency is effectively measured with the mutual partition-distance [1].

It is visible that the marker based algorithms produce less over-segmented results (smaller number of regions and inferior values of d_{sym}), while maintaining the consistency of the segmentation (d_{mut} value).

5. Discussion

This study presents a new approach for object-based spatial segmentation of video. The main idea is to use depth and motion information to guide a segmentation using essentially the colour information. This method is likely to produce simpler and less over-segmented segmentations.

The system presented here differs significantly from the established techniques for segmentation from motion and depth. However, most of the components used in this system are techniques known in the literature. One strength

(a) Frame 102. (b) Frame 103. (c) Frame 104.

(d) Frame 105. (e) Frame 106. (f) Frame 107.

(g) Frame 108. (h) Frame 109. (i) Frame 110.

(j) Frame 111. (k) Frame 112. (l) Frame 113.

Figure 19. Twelve frames from the 'Indoor' sequence.

Frame	marker based			mean shift		
	regions	d_{sym}	d_{mut}	regions	d_{sym}	d_{mut}
102	5	18.38	13.55	9	30.59	13.4
103	5	25.81	15.73	10	27.73	9.96
104	3	16.24	7.85	9	26.48	9.25
105	3	23.84	13.63	10	29.95	15.62
106	4	17.7	12.24	11	33.81	16.95
107	4	21.47	11.51	10	34.5	14.91
108	4	16.09	13.62	12	35.57	17.68
109	6	22.19	15.68	11	40.08	21.83
110	5	20.35	19.61	16	49.93	13.59
111	5	20.02	19.4	13	42.59	20.44
112	5	19.18	17.74	18	50.95	17.96
113	5	18.25	8.67	13	32.17	15.96
mean	4.5	20.0	14.1	11.8	36.20	15.63

Table 1. Results for frames 102–113 of the 'Indoor' sequence, for marker and mean shift based methods.

of this system is that it performs satisfactorily under severe conditions of noise in the auxiliary metadata. This was demonstrated by using the output of a simple block motion estimation as the source of motion data, with its block effect (the block size was 16×16) and spatial instability. The flexibility of this system to integrate additional metadata should also not be underestimated. An additional strength is its simplicity, making it suitable for real-time applications.

The type of segmentation performed by the proposed system should be distinguished from those obtained with systems using a sequence of frame with memory instead of a simple pair of consecutive frames. Because no motion continuity is assumed, this system is more general and copes transparently with camera motion, video shot transitions or illumination changes; on the other hand, it expectedly performs worst when motion continuity is verified. The proposed segmentation technique could in fact be used as a building block of a complete tracking system or memory-based segmentation system.

Future work include the use of active contours in the guided segmentation phase. Active contours might improve further the results, particularly when markers extend beyond the objects. A solution of this kind eliminates the need for a refinement operation between the marker extraction and the guided image segmentation stages, due to its ability to expand or contract as appropriate.

Acknowledgments

This work has been partially supported by VISNET II, a Network of Excellence funded by the sixth Framework Programme of the European Commission.

References

[1] J. S. Cardoso and L. Corte-Real. Toward a generic evaluation of image segmentation. *IEEE Transactions on Image Processing*, 14:1773–1782, nov 2005.

[2] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:603–619, may 2002.

[3] L. Di Stefano, M. Marchionni, and S. Mattoccia. A fast area-based stereo matching algorithm. *Image and Vision Computing*, 22(12):983–1005, Oct 2004.

[4] O. Grau, S. Minelly, and G. A. Thomas. Applications of depth metadata. In *Proceedings of International Broadcasting Convention (IBC 2001)*, pages 62–70, sep 2001.

[5] Y. Huang, S. Fu, and C. Thompson. Stereovision-based object segmentation for automotive applications. *EURASIP Journal on Applied Signal Processing*, pages 2322–2329, 2005.

[6] N. H. Kim and J. S. Park. Segmentation of object regions using depth information. In *Proc. of the IEEE Int. Conf. on Image Processing ICIP 2004*, pages 231–234, 2004.

[7] P. Soille. *Morphological image analysis*. Springer-Verlag, 1999.

[8] The canestavision electronic perception development kit (ep devkit). `http://www.canesta.com/`, 2005.

[9] F. Tsalakanidou, S. Malassiotis, and M. G. Strintzis. Face localization and authentication using color and depth images. *IEEE Transactions on Image Processing*, 14:152–168, 2005.