Region Extraction of Multiple Moving Objects with Image and Depth Sequence

Katsuya Sugawara[1], Ryosuke Tsuruga[2], Toru Abe[2] and Takuo Suganuma[2]
[1]*NTT Resonant Inc., Granparktower, 3-4-1 Shibaura, Minato-ku, Tokyo 108-0023, Japan*
[2]*Cyberscience Center, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai 980–8577, Japan*

Keywords: Region Extraction, Multiple Moving Objects, Image Sequence, Depth Sequence.

Abstract: This paper proposes a novel method for extracting the regions of multiple moving objects with an image and a depth sequence. In addition to image features, diverse types of features, such as depth and image-depth-derived 3D motion, have been used in existing methods for improving the accuracy and robustness of object region extraction. Most of the existing methods determine individual object regions according to the spatial-temporal similarities of such features, i.e., they regard a spatial-temporal area of uniform features as a region sequence corresponding to the same object. Consequently, the depth features in a moving object region, where the depth varies with frames, and the motion features in a nonrigid or articulated object region, where the motion varies with parts, cannot be effectively used for object region extraction. To deal with these difficulties, our proposed method extracts the region sequences of individual moving objects according to depth feature similarity adjusted by each object movement and motion feature similarity computed only in adjacent parts. Through the experiments on scenes where a person moves a box, we demonstrate the effectiveness of the proposed method in extracting the regions of multiple moving objects.

1 INTRODUCTION

Determining the regions of individual moving objects is a requisite preprocessing step for various applications including visual surveillance, control (e.g. managing something by object movements), and analysis (e.g. diagnosing something from object movements). Thus, many object region extraction methods have been proposed for use in such applications. Most of the existing methods determine individual object regions in an image sequence according to the spatial and temporal similarity of image features, i.e., these methods regard a spatial-temporal area of uniform image features as a region sequence corresponding to the same object (Grundmann et al., 2010; Galasso et al., 2012; Xu and Corso, 2012).

Recently, along with the popularization of image and range sensing cameras such as Kinect (Microsoft, 2015), not only image sequences but also depth sequences can be easily acquired. Consequently, besides image features, depth features have been used in several methods for improving the accuracy and robustness of object region extraction (Fernández and Aranda, 2000; Çiğla and Alatan, 2008; Xia et al., 2011; Abramov et al., 2012; Bergamasco et al., 2012).

As is the case in image features, the spatial-temporal similarity of depth features is usually used for determining individual object regions. However, as shown in Figure 1, the depth in a moving object region varies with frames; therefore, depth features cannot be effectively used in such cases for determining a region sequence corresponding to the same object.

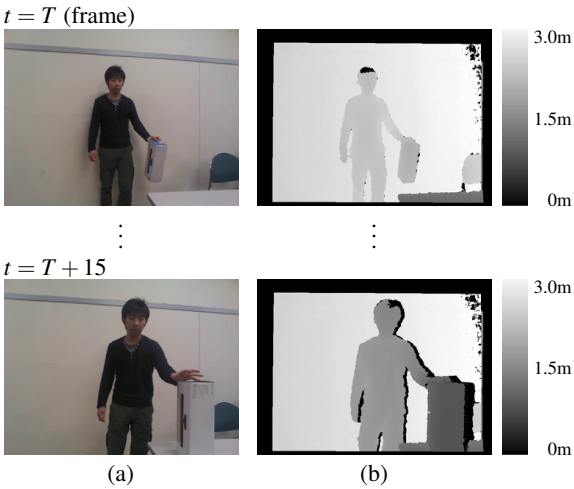


Figure 1: Example of (a) Image sequence and (b) Depth sequence acquired by Kinect.

Sugawara, K., Tsuruga, R., Abe, T. and Suganuma, T.
Region Extraction of Multiple Moving Objects with Image and Depth Sequence.
DOI: 10.5220/0005782402550262
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 255-262
ISBN: 978-989-758-175-5

Meanwhile, 2D motion (optical flow) features are derived from an image sequence, and 3D motion (scene flow) features are derived from an image and a depth sequence as shown in Figure 2. The similarities of such motion features also can be used as clues for extracting object regions. Unlike the 2D motion features, the 3D motion features are less subject to the difference in the distances between a camera and objects; therefore the similarity of 3D motion features is more suitable than that of 2D motion features for determining a region sequence corresponding to the same object. However, in a nonrigid or articulated object region, motion varies with part, and then it decreases the effectiveness of both 2D and 3D motion features for object region extraction.

To deal with these difficulties, this paper proposes a novel method for extracting the region sequences of multiple moving objects with an image and a depth sequence. The proposed method employs the similarities of image, depth, and image-depth-derived 3D motion features as clues for extracting object regions. In this method, depth feature similarity is adjusted by each object movement for adapting to the change of depth features in moving object regions, and 3D motion feature similarity is computed only in adjacent parts for adapting to the nonuniformity of motion features in nonrigid or articulated object regions.

The remainder of this paper is organized as follows. Section 2 presents the existing object region extraction methods based on several types of features.

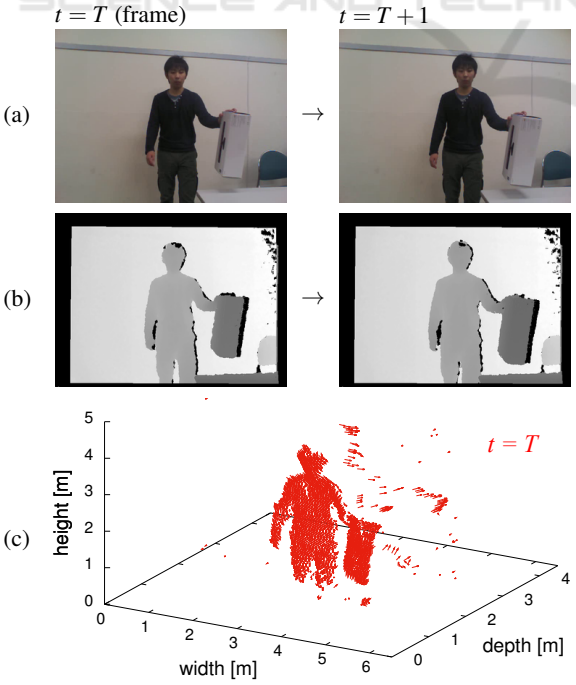


Figure 2: Example of (a) Image sequence, (b) Depth sequence, and (c) 3D motion features derived from them.

Section 3 explains the details of our proposed method for extracting the region sequences of multiple moving objects with an image and a depth sequence. Section 4 presents the results of object region extraction experiments. Finally, Section 5 concludes the paper.

2 RELATED WORK

Many methods have been proposed for extracting object regions from an image. Most of them regard a uniform area in the image as a subregion corresponding to the same object, and determine individual object regions according to the spatial similarity of image features, such as intensities and colors in the image (Comaniciu and Meer, 1999; Felzenszwalb and Huttenlocher, 2004; Achanta et al., 2012). To improve the accuracy and robustness of object region extraction, diverse types of features are used together with the image features. For example, 2D motion features (Çiğla and Alatan, 2008), which can be computed from an image sequence, and depth features (Fernández and Aranda, 2000; Çiğla and Alatan, 2008; Bergamasco et al., 2012), which can be acquired by a range sensing camera, are widely used.

Subregions extracted in each individual image frame are not corresponded with those in the other frames. Thus, various methods have been proposed for extracting a subregion sequence corresponding to the same object in a frame sequence. Their approaches are classified into two main types. One approach performs subregion extractions in each frame, and then carries out subregion matching between successive frames based on the temporal similarity of features (Xia et al., 2011; Abramov et al., 2012; Couprie et al., 2013). The other approach regards the frame sequence as 3D data, and extracts a 3D region (volume) corresponding with the same object according to the spatial-temporal similarity of features (DeMenthon and Megret, 2002; Grundmann et al., 2010). In both the approaches, to improve the accuracy and the robustness, 2D motion features (DeMenthon and Megret, 2002; Grundmann et al., 2010), depth features (Xia et al., 2011; Abramov et al., 2012), and 3D motion features (Xia et al., 2011) are also used in combination with image features.

The entire region of an object is commonly composed of different subregions with dissimilar properties. For such a object region, a number of different subregion sequences are constructed. Accordingly, several methods have been proposed for merging subregion sequences and extracting the sequence of an entire object region (Lezama et al., 2011; Trichet and Nevatia, 2013). Those methods use mainly motion

features for subregion sequence merging.

As mentioned above, in the existing methods, spatial-temporal similarities are computed from diverse types of features, and employed for extracting object region sequences. However, some types of features cannot contribute to the object region extraction in particular conditions. For example, the depth in a moving object region varies with frames; therefore, even if a series of subregions corresponds to the same moving object, its depth features are not always temporally similar under such a condition. Furthermore, the motion in a nonrigid or articulated object region varies with parts; therefore, although in the same object region, the spatial similarity of motion features is not always kept under such a condition.

3 REGION EXTRACTION OF MULTIPLE MOVING OBJECTS

We proposes a novel method for extracting the region sequences of multiple moving objects with an image and a depth sequence. A brief overview of the proposed method is depicted in Figure 3.

Our proposed method firstly extracts subregions from each frame, secondly constructs subregion sequences through subregion matching between successive frames, and finally merges subregion sequences into the region sequences of individual moving objects. To effectively make use of depth features and motion features in these processes, the proposed method employs depth feature dissimilarity adjusted by each object movement and motion feature dissimilarity computed only in adjacent parts.

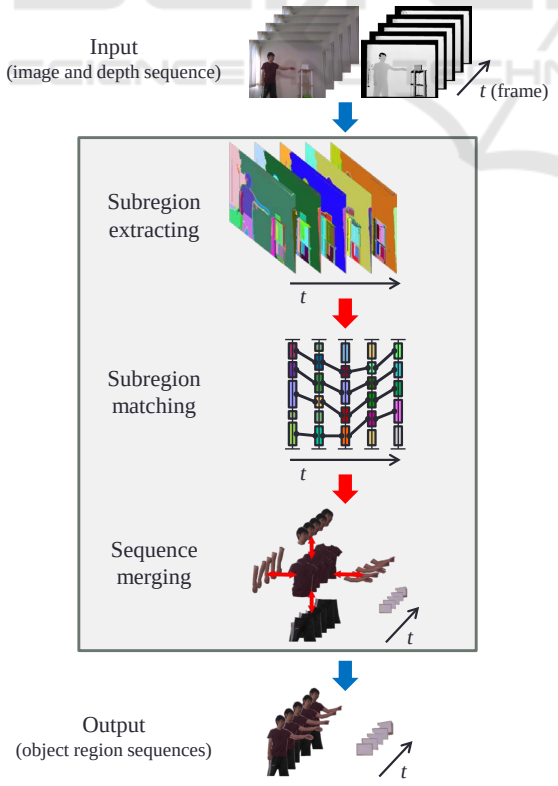


Figure 3: Overview of the proposed method.

3.1 Subregion Extracting

To extract subregions from each frame, our proposed method uses a graph-based image segmentation method (Felzenszwalb and Huttenlocher, 2004) with image and depth features. Example of subregion extraction from an image and a depth frame is shown in Figure 4. In this figure, (c) shows extracted subregions using (a) and (b), where each extracted subregion is assigned a random color.

In graph-based image segmentation methods, an entire image is represented by a graph $G = (V, E)$ with vertices $v_k \in V$ corresponding to pixels and edges $(v_k, v_l) \in E$ corresponding to pairs of neighboring vertices. Each edge (v_k, v_l) has a weight $w_S(v_k, v_l)$ based on some property of neighboring vertices v_k and v_l. A segmentation S is a partition of V into components (subregions) $C_i \in S$ such that each C_i corresponds to a connected component in a graph $G' = (V, E')$, where $E' \subseteq E$.

The method of (Felzenszwalb and Huttenlocher, 2004) uses the dissimilarity of features between neighboring vertices v_k and v_l as their edge weight $w_S(v_k, v_l)$. From such edge weights, the internal difference $Int(C_i)$ of a component $C_i \subseteq V$ is determined to be the largest edge weight in the minimum spanning tree $MST(C_i, E)$ of C_i as

$$Int(C_i) = \max_{(v_k, v_l) \in MST(C_i, E)} w_S(v_k, v_l), \quad (1)$$

and the difference $Dif(C_i, C_j)$ between two components $C_i, C_j \subseteq V$ is determined to be the minimum edge weight connecting C_i and C_j as

$$Dif(C_i, C_j) = \min_{v_k \in C_i, v_l \in C_j, (v_k, v_l) \in E} w_S(v_k, v_l). \quad (2)$$

If $Dif(C_i, C_j)$ is small compared with $Int(C_i)$ and $Int(C_j)$, then those C_i and C_j are merged into one

(a) (b) (c)

Figure 4: Example of (a) Image frame, (b) Depth frame, and (c) Extracted subregions using (a) and (b).

component. Through the iteration of this procedure, an image is segmented into final components, each of which is an area of spatially uniform features, which are extracted from a frame.

Our proposed method defines each edge weight $w_S(v_k, v_l)$ between neighboring pixels v_k and v_l from the dissimilarities of their image features I (RGB intensities) and depth features D by

$$w_S(v_k, v_l) = \|I(v_k) - I(v_l)\| + \sigma_S \|D(v_k) - D(v_l)\|, \quad (3)$$

where σ_S is a coefficient for the dissimilarity of depth features.

3.2 Subregion Matching

For constructing subregion sequences, the proposed method carries out graph-based matching (Couprie et al., 2013) on extracted subregions in successive frames. Example of subregion matching is shown in Figure 5, where random colors are assigned to each extracted subregion in (a) and each constructed sequence (moving object sequence only) in (b).

In the method of (Couprie et al., 2013), two successive frames t and $t+1$ are represented by a graph $G = (V, E)$. A set V of vertices comprises two vertex sets $V(t)$ and $V(t+1)$; vertices $v_i \in V(t)$ correspond to subregions extracted in the frame t, and vertices $v_j \in V(t+1)$ correspond to those in the frame $t+1$. Edges $(v_i, v_j) \in E$ correspond to pairs of vertices in different frames, and each (v_i, v_j) has a weight $w_T(v_i, v_j)$ associated with the dissimilarity of features between v_i and v_j. Using a minimum spanning forest algorithm based on these edge weights $w_T(v_i, v_j)$, the correspondences between vertices $v_i \in V(t)$ and $v_j \in V(t+1)$ are determined. Through successively applying this procedure to all frames, subregion sequences, each of which is a series of subregions with temporally uniform features, are constructed.

The method of (Couprie et al., 2013) defines each edge weight $w_T(v_i, v_j)$ from the differences in shape, position (centroid), and appearance (image) features between subregions v_i and v_j. In addition to those

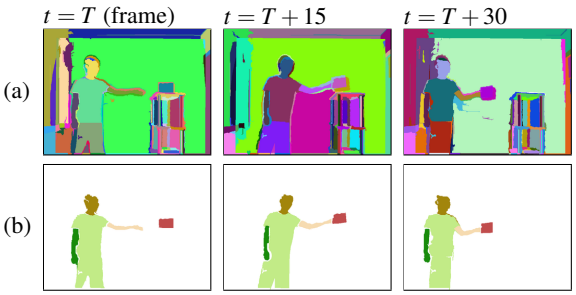


Figure 5: Example of (a) Subregions extracted in each frame, and (b) Subregion sequences constructed from (a).

differences, for improving the accuracy and robustness of subregion matching, our proposed method also takes into account the differences in depth features, and determines each edge weight $w_T(v_i, v_j)$ by

$$w_T(v_i, v_j) = d_s(v_i, v_j) d_p(v_i, v_j) + \sigma_{Ta} d_a(v_i, v_j) + \sigma_{Td} d_d(v_i, v_j), \quad (4)$$

where d_s, d_p, d_a, and d_d represent the differences in shape, position, appearance, and depth features, respectively, while σ_{Ta} and σ_{Td} denote coefficients for d_a, and d_d. The feature differences in Equation (4) are defined as

$$d_s(v_i, v_j) = \frac{|v_i| + |v_j|}{|v_i \cap v_j|}, \quad (5)$$

$$d_p(v_i, v_j) = \|(C(v_i) + M2(v_i)) - C(v_j)\|, \quad (6)$$

$$d_a(v_i, v_j) = \|I(v_i) - I(v_j)\|, \quad (7)$$

$$d_d(v_i, v_j) = \|(D(v_i) + M3_d(v_i)) - D(v_j)\|. \quad (8)$$

In Equation (5), $|v_i|$, $|v_j|$, and $|v_i \cap v_j|$ represent the number of pixels of subregions v_i, v_j, and overlapping area between v_i and v_j. In Equations (6)∼(8), C, I, and D denote the centroid of pixels, the mean of image features (RGB intensities), and the mean of depth features in each subregion, respectively.

As shown in Equation (8), the proposed method adds $M3_d(v_i)$ to the depth feature $D(v_i)$, and determines the depth feature difference $d_d(v_i, v_j)$ from $D(v_i) + M3_d(v_i)$ in the frame t and $D(v_j)$ in the frame $t+1$. Here, $M3_d(v_i)$ denotes the depth directional component of $M3(v_i)$, which is the median of 3D motion (scene flow) in v_i. This is intended to adapt for depth feature changes in moving object regions and then use depth features effectively for subregion matching. In the same way, as shown in Equation (6), the centroid $C(v_i)$ is adjusted by adding $M2(v_i)$, which is the median of 2D motion (optical flow) in v_i, and used for determining the position feature difference $d_p(v_i, v_j)$.

To obtain $M2(v_i)$ and $M3_d(v_i)$, the 2D motion of each pixel in the frame t is determined from the two successive image frames t and $t+1$ (Farnebäck, 2003). Using the 2D motion in the frame t, the median of 2D motion $M2(v_i)$ is computed for each subregion v_i. Besides, the correspondences between the pixels in the frame t and those in the frame $t+1$ are determined from the 2D motion in the frame t. By incorporating the pixel correspondences into the depth frames t and $t+1$, the 3D motion in the frame t also can be determined. Using the 3D motion in the frame t, the median of 3D motion $M3(v_i)$ and its depth directional components $M3_d(v_i)$ are computed for each subregion v_i.

3.3 Sequence Merging

The proposed method merges subregion sequences corresponding with the same object into a single object region sequence. Example of subregion sequence merging is shown in Figure 6, where random colors are assigned to each subregion sequence in (a) and each individual object region sequence in (b).

To begin with, subregions of large 3D motion features are determined in all frames, and each sequence including such subregions is chosen as a target subregion sequence S_i corresponding to part of a moving object. Our proposed method computes the dissimilarity between two target subregion sequences, and merges them into a single sequence if their dissimilarity is less than a given threshold. Iterating this process for all target subregion sequences until all similar sequences unify, our method obtains the region sequences of multiple moving objects individually.

Let subregion sequences S_i and S_j consist of subregions $s_i(t) \in S_i$ ranging in frame $Ts_i \leq t \leq Te_i$ and subregions $s_j(t) \in S_j$ ranging in frame $Ts_j \leq t \leq Te_j$, respectively. Suppose that the frames of these two sequences S_i and S_j overlap each other from Ts to Te as shown in Figure 7 (a). Our proposed method defines the dissimilarity between S_i and S_j by

$$DS(S_i,S_j) = \frac{1}{Te-Ts+1}\sum_{t=Ts}^{Te} ds(s_i(t),s_j(t)), \quad (9)$$

where $ds(s_i(t),s_j(t))$ is the dissimilarity, which is defined from the differences of 3D motion features, between subregions $s_i(t)$ and $s_j(t)$ in the frame t. Although motion features are not always spatially uniform in a nonrigid or articulated object region, adjacent pixels in the same object region are thought to be similar to each other in their motion features as shown in Figure 7 (b). Therefore, to improve the accuracy and robustness of sequence merging, the proposed method determines the dissimilarity $ds(s_i(t),s_j(t))$ from the differences of 3D motion features in adjacent pixels between subregions $s_i(t)$ and $s_j(t)$.

To determine adjacent pixels between subregions $s_i(t)$ and $s_j(t)$, for each pixel pair $(v_k(t),v_l(t))$ made

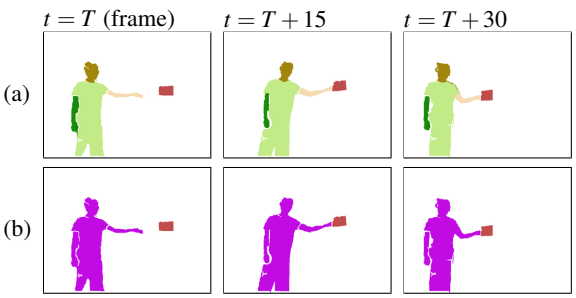


Figure 6: Example of (a) Subregion sequences, and (b) Individual moving object region sequences obtained from (a).

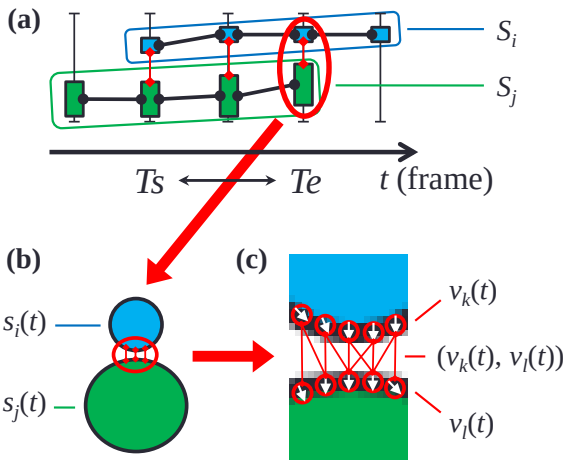


Figure 7: (a) Subregion sequences S_i, S_j, (b) Subregions $s_i(t)$, $s_j(t)$ in the frame t, (c) Pixels $v_k(t)$, $v_l(t)$, and their pixel pair $(v_k(t),v_l(t))$.

from pixels $v_k(t) \in s_i(t)$ and $v_l(t) \in s_j(t)$, its inter-pixel 3D distance $d_{P3}(v_k(t),v_l(t))$ is computed by

$$d_{P3}(v_k(t),v_l(t)) = \|P3(v_k(t)) - P3(v_l(t))\|, \quad (10)$$

where $P3(v_k(t))$ and $P3(v_l(t))$ are 3D positions corresponding to $v_k(t)$ and $v_l(t)$, respectively. Among all pixel pairs, N pixel pairs $(v_k(t),v_l(t)) \in AP_{ij}(t)$ with the shortest inter-pixel 3D distances $d_{P3}(v_k(t),v_l(t))$ are chosen for the pairs of adjacent pixels as shown in Figure 7 (c). For each adjacent pixel pair $(v_k(t),v_l(t)) \in AP_{ij}(t)$, the difference between its pixels $v_k(t)$ and $v_l(t)$ is defined as

$$dv(v_k(t),v_l(t)) = d_{P3}(v_k(t),v_l(t)) + \sigma_M d_{M3}(v_k(t),v_l(t)), \quad (11)$$

where $d_{M3}(v_k(t),v_l(t))$ represents the difference in 3D motion features defined as

$$d_{M3}(v_k(t),v_l(t)) = \|M3(v_k(t)) - M3(v_l(t))\|, \quad (12)$$

and σ_M is a coefficient for $d_{M3}(v_k(t),v_l(t))$. The median of $dv(v_k(t),v_l(t))$ computed for $(v_k(t),v_l(t)) \in AP_{ij}(t)$ is used as the dissimilarity $ds(s_i(t),s_j(t))$ between subregions $s_i(t)$ and $s_j(t)$ in the frame t.

4 EXPERIMENTAL RESULTS

To demonstrate the effectiveness of our proposed method, we conducted experiments on the region sequence extraction of multiple moving objects.

Scenes where a person moves a box were taken by Kinect (Microsoft, 2015), and three pairs (Scenes 1, 2, and 3) of an image and a depth sequence were used in the experiments. Both the image and depth sequences were 640×480 pixels in size and 30 fps

in frame rate. Scenes 1, 2, and 3 contained 500, 400, and 400 frames, respectively. In these sequences, pixel positions on the depth frame were converted to those on the image frame by using Kinect for Windows SDK (Microsoft, 2013).

The experiments were carried out on a PC (Intel Core i7-3770@3.40GHz, 8GB, Windows 7 Pro x64), and a program was implemented with Visual C++ 2010. Through preliminary experiments, the parameters in the proposed method were determined as $\sigma_S = 3.0$, $\sigma_{Ta} = 10.0$, $\sigma_{Td} = 30.0$, and $\sigma_M = 37.5$.

4.1 Experiments of Subregion Matching

Firstly, to illustrate the effectiveness of our proposed method in subregion matching, we carried out experiments on the construction of subregion sequences.

As shown in Figure 8 (a), (b), and (c), subregions are extracted using the image and the depth sequence of Scene 1. To these subregions extracted from each frame, three subregion matching methods are applied,

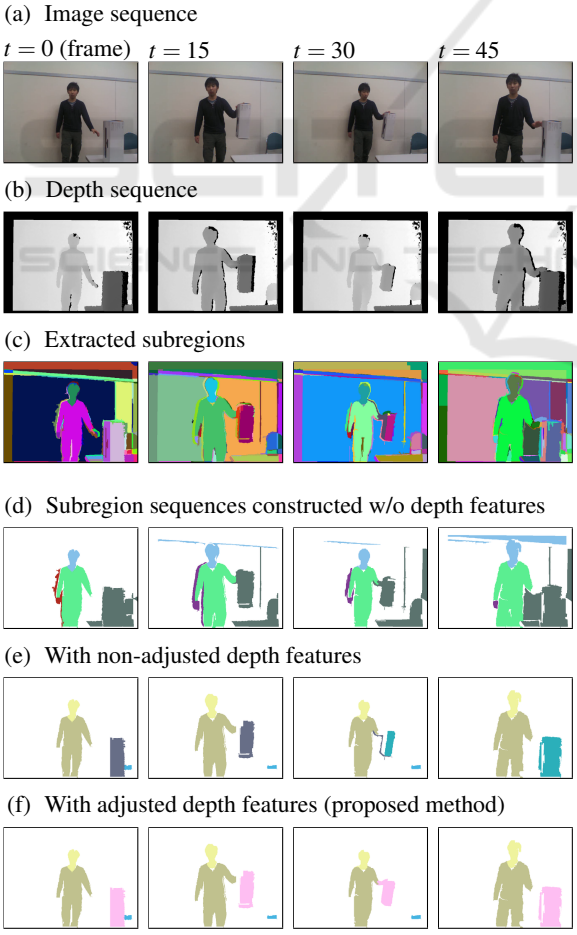


Figure 8: Experimental results of constructing subregion sequences (Scene 1).

and their results are shown in Figure 8 (d), (e), and (f), where random colors are assigned to each constructed sequence (moving object sequence only).

Figure 8 (d) shows the subregion sequences constructed by a subregion matching method without depth feature dissimilarity (with only image feature dissimilarity), whose approach corresponds to that of (Couprie et al., 2013). Figure 8 (e) shows the result by a method with non-adjusted depth feature dissimilarity, which corresponds to the approach of (Abramov et al., 2012). Figure 8 (f) shows the result by the proposed method, which uses adjusted depth feature dissimilarity.

As can be seen from Figure 8 (d), some parts of the background are mistakenly regarded as subregions of the person or box by the subregion matching method without depth feature dissimilarity. Consequently, subregions of the person's head and the background are incorporated into the light blue sequence, and subregions of the box and the background are incorporated into the dark green sequence. Compared to this, as shown in Figure 8 (e), such inaccurate correspondences between subregions can be reduced by using depth feature dissimilarity. However, subregions of the box in $t = 0, 15$ and $t = 30, 45$ are assigned different colors. This is because, subregions corresponding to the same object are regarded as corresponding to different objects according to the change of their depth features.

In contrast to those results, as shown in Figure 8 (f), the proposed method constructs three different sequences from subregions of the person's head, the person's body, and the box individually. These experimental results indicate the effectiveness of our proposed method, which uses the temporal dissimilarity of depth features adjusted by each object movement, in subregion matching.

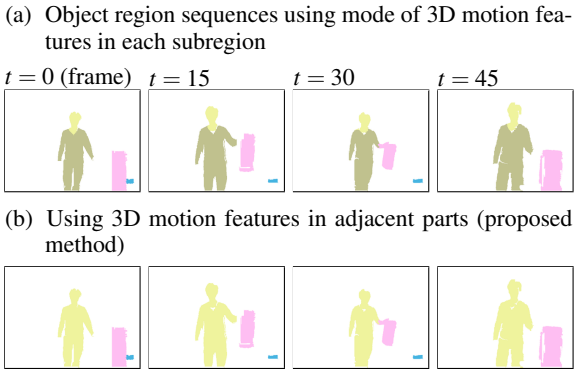


Figure 9: Experimental results of merging subregion sequences (Scene 1).

4.2 Experiments of Sequence Merging

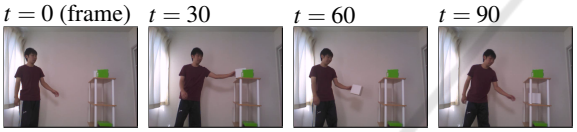
Secondly, to illustrate the effectiveness of our proposed method in subregion sequence merging, we carried out experiments on the extraction of individual object region sequences.

From Scenes 1, 2, and 3, subregion sequences are constructed by the proposed method with adjusted depth feature dissimilarity as shown in Figures 8 (f), 10 (d), and 11 (d). To these subregion sequences, two sequence merging methods are applied, and their results are shown in Figures 9 (a), (b), 10 (e), (f), 11 (e), and (f), where random colors are assigned to each individual object region sequence.

Figures 9 (a), 10 (e), and 11 (e) show the region sequences extracted by a sequence merging method which computes the mode of 3D motion features in each subregion and uses the dissimilarity of 3D motion feature modes between subregions. This approach corresponds to that of (Trichet and Nevatia, 2013). As can be seen from those results, although the region sequences of a person and a box are extracted separately in all scenes, person's head and body are also extracted as different region sequences because the movement of the head is substantially different from that of the body.

Figures 9 (b), 10 (f), and 11 (f) show the results by the proposed method using 3D motion feature dissimilarity computed only in adjacent parts. Almost the entire region of the person is correctly extracted as a single region sequence in all scenes. These experimental results indicate the effectiveness of our proposed method, which uses the spatial dissimilarity of 3D motion features computed only in adjacent parts, in subregion sequence merging.

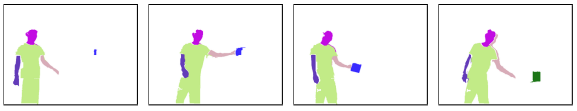
(a) Image sequence



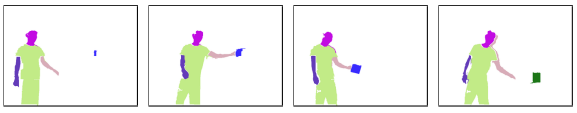
(b) Depth sequence

(c) Extracted subregions

(d) Subregion sequences constructed with adjusted depth feature



(e) Object region sequences merged using mode of 3D motion features in each subregion



(f) Using 3D motion features in adjacent parts (proposed method)

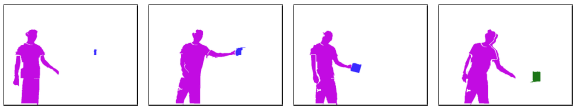
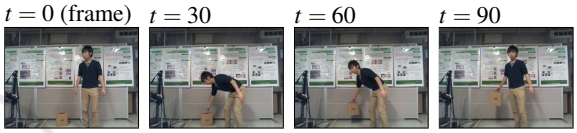


Figure 10: Experimental results of merging subregion sequences (Scene 2).

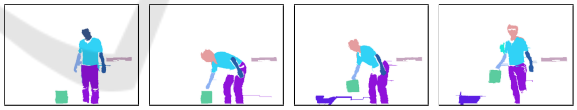
(a) Image sequence



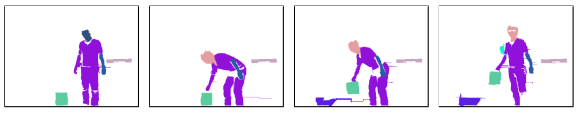
(b) Depth sequence

(c) Extracted subregions

(d) Subregion sequences constructed with adjusted depth feature



(e) Object region sequences merged using mode of 3D motion features in each subregion



(f) Using 3D motion features in adjacent parts (proposed method)

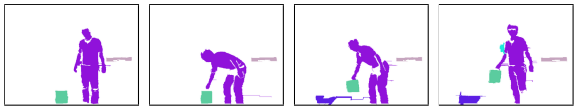


Figure 11: Experimental results of merging subregion sequences (Scene 3).

5 CONCLUSIONS

In this paper, we have proposed a method for extracting the region sequences of multiple objects using an image and a depth sequence. The proposed method extracts subregions from each frame, constructs subregion sequences through subregion matching between successive frames, and merges subregion sequences into the region sequences of individual objects. To effectively make use of depth features and 3D motion features in these processes, our proposed method employs depth feature similarity adjusted by each object movement and 3D motion feature similarity computed only in adjacent parts. Through the experiments, we demonstrated the effectiveness of our proposed method in extracting the region sequences of multiple moving objects, where the depth varies with frames, and articulated objects, where the motion varies with parts.

Currently, our proposed method extracts object region sequences from a whole input sequence (i.e. it cannot process every input frame serially), and the average processing time of every frame is more than two seconds. In future work, we would like to investigate extending our method not only to improve the accuracy of object region sequence extraction but also to process every set of a few input frames or every input frame serially in real time. Furthermore, we plan to conduct quantitative evaluation of the proposed method for various scenes.

ACKNOWLEDGEMENT

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) under a Grant-in-Aid for Scientific Research (C) (No.15K00171).

REFERENCES

Abramov, A., Pauwels, K., Papon, J., Wörgötter, F., and Dellen, B. (2012). Depth-supported real-time video segmentation with the Kinect. In *Proc. IEEE Workshop Appl. Comput. Vision*, pages 457–464.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Susstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Machine Intell.*, 34(11):2274–2282.

Bergamasco, F., Albarelli, A., Torsello, A., Favaro, M., and Zanuttigh, P. (2012). Pairwise similarities for scene segmentation combining color and depth data. In *Proc. 21st Int. Conf. Pattern Recognit.*, pages 3565–3568.

Çiğla, C. and Alatan, A. A. (2008). Object segmentation in multi-view video via color, depth and motion cues. In *Proc. IEEE Int. Conf. Image Process.*, pages 2724–2727.

Comaniciu, D. and Meer, P. (1999). Mean shift analysis and applications. In *Proc. Int. Conf. Comput. Vision*, volume 2, pages 1197–2003.

Couprie, C., Farabet, C., LeCun, Y., and Najman, L. (2013). Causal graph-based video segmentation. In *Proc. IEEE Int. Conf. Image Process.*, pages 4249–4253.

DeMenthon, D. and Megret, R. (2002). Spatio-temporal segmentation of video by hierarchical mean shift analysis. Technical Report TR-4388, Center for Automat. Res., U. of Md, College Park.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Proc. Scand. Conf. Image Anal.*, pages 363–370.

Felzenszwalb, P. F. and Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59(2):167–181.

Fernández, J. and Aranda, J. (2000). Image segmentation combining region depth and object features. In *Proc. 15th Int. Conf. Pattern Recognit.*, volume 1, pages 618–621.

Galasso, F., Cipolla, R., and Schiele, B. (2012). Video segmentation with superpixels. In *Proc. 11th Asian Conf. Comput. Vision*, volume 1, pages 760–774.

Grundmann, M., Kwatra, V., Han, M., and Essa, I. (2010). Efficient hierarchical graph-based video segmentation. In *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pages 2141–2148.

Lezama, J., Alahari, K., Sivic, J., and Laptev, I. (2011). Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pages 3369–3376.

Microsoft (2013). Kinect for Windows SDK v1.8. http://www.microsoft.com/en-us/download/details.aspx?id=40278. Online; accessed 1–Sep.–2015.

Microsoft (2015). Kinect – Windows app development. https://dev.windows.com/en-us/kinect. Online; accessed 1–Sep.–2015.

Trichet, R. and Nevatia, R. (2013). Video segmentation with spatio-temporal tubes. In *Proc. IEEE Int. Conf. Adv. Video Signal Based Surv.*, pages 330–335.

Xia, L., Chen, C.-C., and Aggarwal, J. K. (2011). Human detection using depth information by Kinect. In *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, pages 15–22.

Xu, C. and Corso, J. J. (2012). Evaluation of super-voxel methods for early video processing. In *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pages 1202–1209.