Sensors and Materials, Vol. 33, No. 4 (2021) 1363–1373
MYU Tokyo

S & M 2540

Household Goods Recognition Using Hierarchical Multi-object Segmentation

Wenjuan Wang,[1] Jinlian Zhuang,[1] Xiaomei Zhang,[1]
Chih-Hsien Hsia,[2*] Chun-I Li,[3] and Cheng-Fu Yang[4,5**]

[1]College of Mathematics and Information Engineering, Longyan University, Fujian 364012, China
[2]Department of Computer Science and Information Engineering, National Ilan University, Ilan County 260, Taiwan
[3]Department of Electrical and Computer Engineering, Tamkang University, New Taipei City 251, Taiwan
[4]Department of Chemical and Materials Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan
[5]Department of Aeronautical Engineering, Chaoyang University of Technology, Taichung 413, Taiwan

(Received October 21, 2020; accepted February 12, 2021)

Keywords: hierarchical, multiple objects, object segmentation method, scale-invariant feature transform

Nowadays, the algorithms most used for object recognition are based on a constructed database or on training and learning processes with many samples, allowing robots to effectively perform object recognition. If objects in a home environment do not appear in a database, a system cannot recognize and segment items from household goods. In this study, we proposed an algorithm to reduce the processing complexity of object recognition that combines a depth image, object segmentation, and model construction with the GrabCut algorithm, and uses a hierarchical design for the segmentation of items. This algorithm uses the depth image to find the approximate locations and sizes of multiple objects in a coarse layer, then it uses GrabCut as a fine segmentation technology to segment the edges of objects and construct the models. First, we use the inputs of binocular vision to generate an anaglyph image, which is used as the base to perceive the environment's 3D information. At the same time, the too distant background is filtered, then histogram segmentation of the analysis image is used to partition each object. Next, GrabCut is used to find a convergent partition on the masking image to generate complete object edges. Finally, the scale-invariant feature transform (SIFT) is used for the extraction and recognition of feature points, and the database is updated.

1. Introduction

With the increasing maturity of robot technology, robots can be applied in a wider range of fields. The first generation of robots were industrial robots, the second generation of robots had the technology of sensing, and the third generation of smart robots have smaller volumes and are integrated with computers. Smart robots that can carry out services (service-type robots) are key development projects in developed countries, and robots are being developed for home care, security, environment cleaning, interactive learning, medical care, and fire prevention and rescue. Recently, the requirements and applications of service-type robots have become increasingly important. The key technologies being investigated and developed for service-

*Corresponding author: e-mail: chhsia625@gmail.com
**Corresponding author: e-mail: cfyang@nuk.edu.tw
https://doi.org/10.18494/SAM.2021.3174

ISSN 0914-4935

type robots include mechanical drive technology, environmental perception technology, smart control software, robot vision processing, and embedded systems. For robots to be used in a home environment, the main research areas include improving the ability of robots to interact, communicate, and cooperate with humans so that robots can adapt to home environments, understand the tasks that people want them to do, and complete the tasks rapidly and safely.

Home robots are no longer limited to futuristic TV shows or movies: robot companions, personalized assistants, and home management aids have been steadily improving since Roomba first hit the store shelves in 2002. Liu *et al.* reviewed the evolution of robotic research and development over the past 50 years, and they defined home service robots in terms of three major categories: robot manipulators, mobile robots, and biologically inspired robots.[(1)] Zachiotis *et al.* provided a thorough overview of state-of-the-art (SOA) solutions available in home service robotics, and their detailed analysis of consumer-oriented robots suggested future demand for robots in the areas of entertainment, education, social purposes, gaming, and households. They also found that research-oriented robots were focused on the purposes of entertainment, development platforms, security, and household/rehabilitation.[(2)] Because of the rapidly growing population of elderly people, the need for healthcare is on the rise. Ramoly *et al.* investigated a framework for service robots in smart homes. This framework included robotics and smart environments, and provided a promising solution for monitoring in which a robot interacts with and provides companionship to users. They found that sensor data is not perfect in real scenarios because the environment changes over time, and they tackled these problems in order to improve the autonomy and efficiency of robots in smart environments.[(3)]

Owing to the rapid progress of the robot industry and Taiwan's aging society, there has been increased interest in employing robots to perform some of the healthcare and domestic tasks in the home. In the home, many tasks can be performed by robots, making machine vision very important because of its use in image analysis to allow robots to make judgments from input images. In the home environment, the recognition of household goods is very important. In this study, we used the home environment as the main axis of technological development, which could help housekeeping robots to identify items correctly in the home. Nowadays, most object identification algorithms are dependent on a constructed database or training and learning processes using many objects. When household items are not contained in the database, then robots need off-line algorithms to manually construct models of these items, which will make it more difficult for robots to identify objects in the environment.

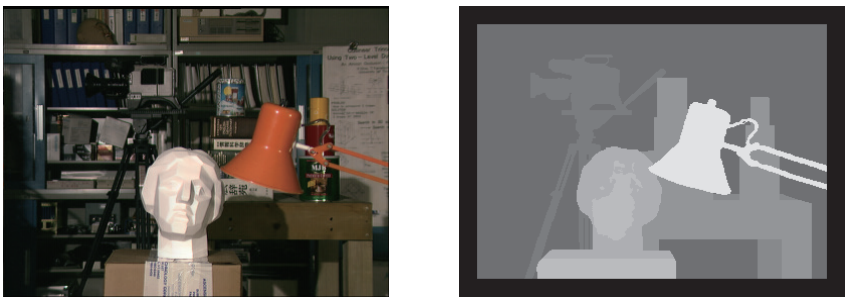
The theory of graph cuts (GrabCut) was first used as an optimization method in the computer vision field, and it is an object segmentation algorithm based on graph cuts. GrabCut starts with a user-specified bounding box around objects to be segmented, then estimates the color distributions of target objects and the background using Gaussian mixture modeling (GMM). Basavaprasad and Ravindra investigated an improved GrabCut algorithm for object segmentation that combined the technologies of statistics and graph cuts, and their algorithm accomplished detailed object segmentation with a suitable input.[(4)] Kang *et al.* proposed an object segmentation method based on an improved non-interactive GrabCut algorithm, in which they used bilateral filtering to preserve edges and for noise reduction.[(5)] In this study, we proposed an algorithm that combines a depth image, household goods segmentation, and

model construction with GrabCut, and we used a hierarchical design for the segmentation of items. The proposed algorithm uses hierarchical multiobject segmentation technology to sense household goods and recognize them for further applications, and the recognition results can be used in different fields, for example, robot applications, virtual reality, automated tracking systems, and 3D movies. Thus, this work can also be applied to optical sensors and imagers. The depth image is used to find the approximate locations and sizes of multiple objects in the coarse layer, then GrabCut is used as fine segmenting algorithm to extract the edges of objects. This means that the proposed algorithm can be used to segment multiple objects and construct models. Our novel object recognition algorithm can automatically construct models for static household items in a non-stationary background. Also, the information completeness of recognized objects is close to that obtained with manually built models, and when the database is upgraded, the images can be used to achieve an acceptable object recognition rate.

2. Related Research

2.1 Stereo vision

The application fields of stereo vision are wide; for example, it can be used in robot applications, virtual reality, security monitoring systems, automated tracking systems, 3D human–computer interaction interfaces, and 3D movies. Before stereo vision is applied in these fields, target objects must be sensed and 3D information of the object and the environment must be acquired. A depth image can give us "depth" or "z" information of objects in the real world, and the intensity values in the image represent the distances of the objects from a viewpoint. Generally speaking, when we acquire a depth image, we can obtain the depth information at the same time, with the brightness of the image pixels expressing the parallax; higher pixel values suggest that objects are closer and lower pixels suggest that objects are farther away, as shown in Fig. 1.



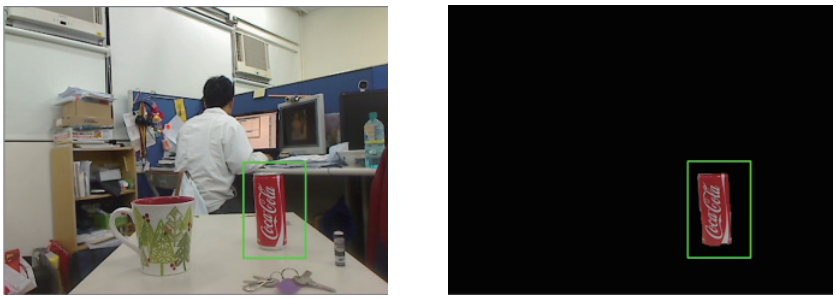
(a) (b)

Fig. 1. (Color online) Depth image. (a) Original image and (b) depth image demonstrating optical parallax.

2.2 Image segmentation

The purpose of object segmentation is to gather pixels of the same type into cluster regions, which represent different surfaces, objects, or parts of an object. There are many object segmentation technologies, including object recognition, mobile object detection, depth images, and template comparison. When an object is segmented, it can be subdivided into its constituting areas or objects, and the degree of subdivision is dependent on the problem to be solved. This means that once an object of interest has been segmented, the segmentation process should be stopped. An object segmentation algorithm is usually based on the intensity values of two basic characteristics: discontinuity and similarity. The first kind of algorithm uses sudden changes in the image gradient to segment images. The second kind of algorithm segments images into similar regions according to predefined criteria. The critical value method, seeding region growth method, seeding region segmentation method, and image-merging method are all examples of the second kind of algorithm. GrabCut is a 2D image segmentation algorithm used in general applications.[6–8] Users of GrabCut only need to drag the selected input images and then to roughly divide then into the foreground and background, as shown in Fig. 2. The main steps of the GrabCut algorithm are listed below:

(1) Users need to input two or three conditions: the foreground and/or background and the unknown regions. Generally speaking, the image in a box (region of interest) is marked as the unknown region and the image outside the box is marked as the background.
(2) In the initial object segmentation, the pixels of the unknown part in the box are classified as foreground and the other pixels are classified as background.
(3) The foreground and background are mixed using GMMs to construct models.
(4) The pixels classified as foreground are assigned to the most likely foreground using the Gaussian mixture and the pixels classified as background are assigned to the most likely background using the Gaussian mixture.
(5) The new Gaussian model is learned by the pixel sequence.
(6) Finally, a graph is constructed and GrabCut is used to sort the new foreground and background; then, steps 4–6 are repeated until the sorting result converges.



(a) (b)

Fig. 2. (Color online) Household goods segmentation using GrabCut algorithm: (a) original image and (b) segmentation image.

2.3 Image matching

Image matching is one of the key technologies used in many applications of computer vision including object recognition, analysis of 3D internal modeling, stereo matching, and motion tracking. The scale-invariant feature transform (SIFT) can clearly describe the feature points of an image and describe the images and objects in various situations.[9–11] The feature points include the image scale, image whirling, partial brightness, and multivision invariance. These feature points can have better distributions in the spatial and frequency domains, and they decrease the probability of matching failure, which is caused by masking and noise. When an effective algorithm is used, it can extract a large number of feature points from an image. Because the feature points have a high-level uniqueness, an effective algorithm can provide a large volume of feature points to obtain the correct similarities between objects in the images in a database. The following are the main steps in generating the feature points:
(1) Detection of the limited spatial dimensions.
(2) Confirmation of the feature points.
(3) Description of the feature points.
By following these steps, the feature points used in SIFT can be obtained.

3. Hierarchical Multi-object Segmentation Algorithm

The main purpose of this study is to design an algorithm for home environments that can automatically detect and segment multiple home objects in an image at the same time. For this purpose, we will introduce the proposed hierarchical object segmentation algorithm to solve the problem of how to separate stationary home objects under a non-stationary background. In the proposed algorithm, the hierarchical model is divided into a coarse layer and a fine layer, and both are used to construct the image model and complete the segmentation automatically.

3.1 Modeling of coarse layer

The coarse layer uses the depth image as the base, then it removes the background and segments the independent objects via hierarchical statistics, as shown in Fig. 3. Through compensation of the morphology and conditional processing, we filter the non-required objects or foreground and enlarge the object's size range, which is beneficial for the process of back-end segmentation to result in the object's inward convergence.[12–14] The process of coarse-layer modeling segments the objects in the home one by one and marks them. The segmentation results are not the real edges of all segmented objects but they can cover the objects' information. The image's information is used in the proposed fine-layer segmentation for further edge convergence, in which the edges of objects are converged to the edges of real objects. The segmentation results are close to those obtained manually.

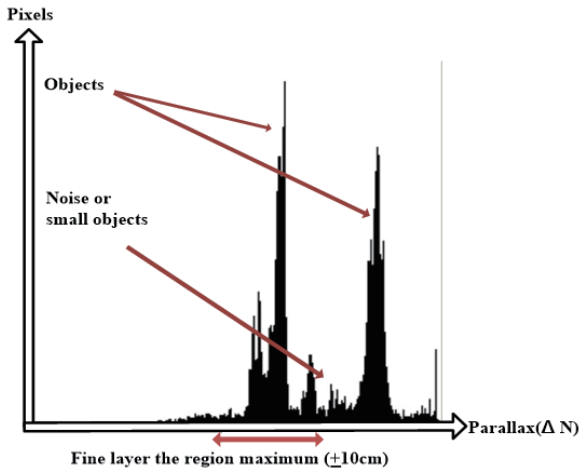


Fig. 3. (Color online) Statistics obtained from histogram of depth image.

3.2 Modeling of fine layer

In the fine layer, we use GrabCut as the algorithm and incorporate a suitable foreground and background to evaluate the performance of the segmentation process. The coarse layer obtains a suitable model of the foreground, although contours are not complete, and GrabCut is used for edge convergence on the foreground of interest. Using GrabCut, the pixels of fixed images can be set as the foreground or background, and it is possible to set the object as the foreground. When we set different attributes of pixels on the fixed image, the weights of edges between the foreground and background modeling and each point of pixels are influenced, which may influence the segmentation results. If the accuracy of foreground and background modeling is improved, the segmentation results will more closely match the optimum manual segmentation results. Figure 4(a) shows an original image before segmentation, Fig. 4(b) shows the result of manual segmentation of selected household goods, and Fig. 4(c) shows the image of selected household goods segmented manually then automatically segmented by the proposed algorithm. As shown in Fig. 4(c), the image obtained by automatic segmentation is similar to the manually segmented image and has a high level of perfection of about 99%.

3.3 Construction of image model

This purpose of this study is to investigate an algorithm for automatic segmentation of the target object and to apply this algorithm for image modeling. However, the most basic method of image modeling is to save the segmentation image and make a comparison, then the image-matching process is used for further recognition and to update or delete the duplicate image sample. There are many image-matching processes, including the template-matching method,[15] contour or shape comparison method, histogram comparison method,[16] and feature point matching method.[17,18] However, Mikolajczyk and Schmid have proven that in many object recognition algorithms, when the feature points are constructed by a SIFT-based algorithm, they

(a) (b) (c)

Fig. 4. (Color online) Results of object segmentation. (a) Original image before segmentation, (b) image of selected household goods after automatic segmentation, and (c) image of selected household goods after manual segmentation.

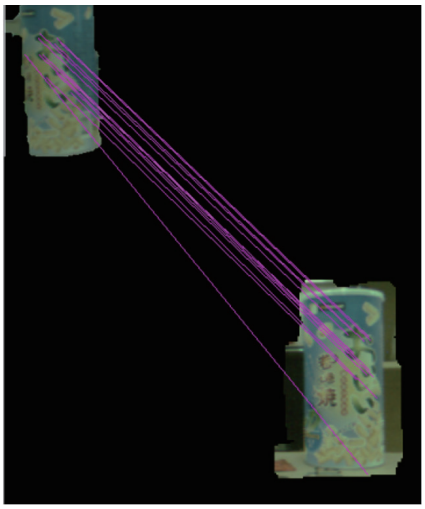


Fig. 5. (Color online) Result of image matching.

are the most stable in the cases of image interference, object rotation, and affine transformation.[19] In an experimental environment, an object is placed randomly, and the distances between different objects and two cameras are not stable. To overcome this problem, the SIFT algorithm extracts the feature points of a segmented object image and updates them in the database, and it uses the complete image or feature points to replace the incomplete ones or add the object image to the database. A characteristic of the SIFT algorithm is that it has reasonable robustness against changes in scale, rotation, vagueness, brightness, and affine transformation, and the extracted high-dimensionality feature points have improved robustness. Figure 5 shows the result of matching feature points using the SIFT algorithm.

4. Experimental Results

The equipment used in our experiment to evaluate the proposed algorithm included a desktop computer equipped with an Intel^{R} Core^{TM} 2 Q9550 (2.83 GHz, 2 GB RAM); the

running program was compiled by Visual Studio 2008, the operating system is Microsoft Windows XP SP3 (32 bits), the resolution of the sequence of images is 640 × 480, and the image format of the RGB color system is 24 bits. The proposed algorithm was evaluated in an indoor home environment. Five experimental scenes were constructed, as shown in Fig. 6, which were used to demonstrate the suitability of the proposed algorithm for identifying most household objects.

4.1 Correct coverage ratio

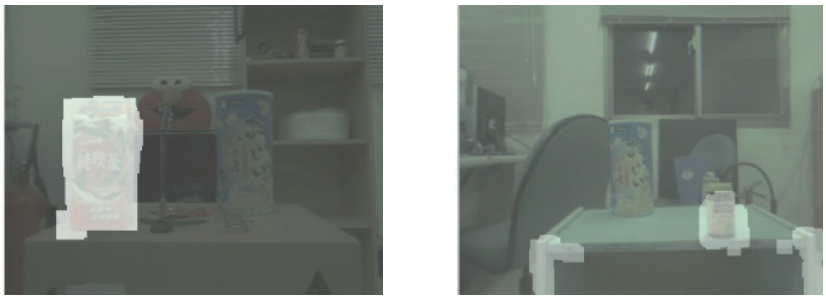
In this study, we use the possibly correct and correct coverage ratio of the foreground as the correct coverage ratio of the coarse layer. If the foreground is correctly covered, we call this result a true positive (TP), and if the foreground is not correctly covered, we call this result a false negative (FN). Figure 7 shows a schematic diagram of the coverage conditions. From Table 1, we can see that there are many causes of an FN: one is the mismatch on the foreground of the non-object, which will cause the recognition area or the amount of noise to be too large. As a result, the mechanism to filter objects with too small areas cannot remove the object's foreground or noise.

4.2 Coverage accuracy

The above demonstration has proven that an image processed by the coarse layer can achieve high coverage factors for most home objects. In this study, the use of GrabCut in the fine layer

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

Fig. 6. (Color online) Experimental scenes: (a) scene 1, (b) scene 2, (c) scene 3, (d) scene 4, and (e) scene 5.



(a) (b)

Fig. 7. (Color online) Conditions of coverage factor. (a) Possibly correct coverage of object in left of image and (b) possibly incorrect coverage of object.

Table 1
TP/FN ratios for different scenes.

Scene	No. of detected objects	TP	FN	TP rate (TP/No. of detect objects) (%)
1	17	17	0	100
2	25	24	1	96
3	21	21	0	100
4	27	24	3	88.88
5	22	22	0	100
Totals	112	108	4	96.42

Table 2
Table of definitions.

Predicted	Actual	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

is proposed to further process the image, for example, to delete incorrect image information and retain correct image information. Table 2 shows the definitions of TP, FN, false positive (FP), and true negative (TN) in this study. TP represents the pixel of a real foreground object that is a pixel of the foreground of the output after segmentation. FN represents the pixel of a real foreground object that is mistaken for a pixel of the background of the output after segmentation. FP represents the pixel of a real background object that is mistaken for a pixel of the foreground of the output after segmentation. TN represents the pixel of a real background object that is the pixel of the background of the output after segmentation. The accuracy rate is defined as

$$Accuracy\,rate = \frac{TP}{TP + \left(FN + FP\right)}. \tag{1}$$

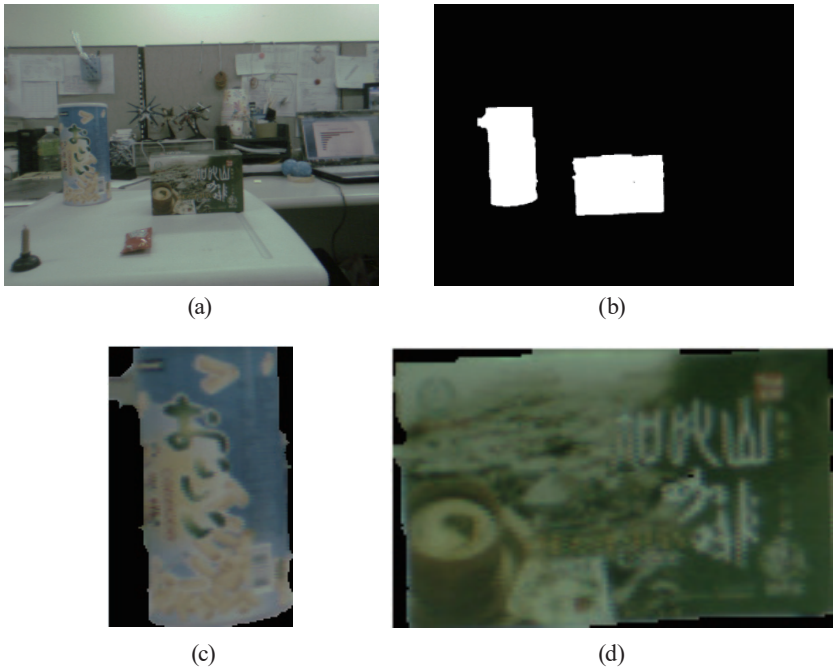
In this study, we use a manually segmented image as the standard, and through observation, we find that the accuracy rate of automatically segmented images is up to 99% when the manually segmented image is considered as the correct output. The accuracy rate in Eq. (1) is obtained by a comparison of a manually segmented image with an automatically segmented image. Figure 8 shows schematic diagrams of TP and FN + FP. Table 3 gives the average accuracy rate for the five scenes shown in Fig. 6, and the segmentation result of scene 1 is shown in Fig. 9. Comparing the results of automatic and manual segmentation of the image,

(a) (b)

Fig. 8. Acquisition of images of objects. (a) TP is the area of the image with pixel value 255 (comparison of automatic segmentation and manual segmentation) and is defined as the same output region. (b) FN+FP is the area of the image with pixel value 255 and is defined as the different output region.

Table 3
Comparison of recognition accuracy rates.

Algorithm	Accuracy rate (%)	Processing time (640 × 480) (ms)
Manual object segmentation with SIFT	96.48	1628
Proposed algorithm with SIFT	93.05	1630.5



(a) (b)

(c) (d)

Fig. 9. (Color online) Real images in experimental setup. (a) Original scene, (b) segmentation mask of output region, (c) segmentation object A with recognition accuracy rate of 95.12%, and (d) segmentation object B with recognition accuracy rate of 9.54%.

we found that the feature points generated by the SIFT algorithm confirmed the sufficiently correct information and sufficient robustness of the segmentation image for object recognition applications. The recognition in this study is dependent on the algorithm proposed by Silva.[14] If the data of the test object is input, because the segmentation object of the coarse layer is

compared with the data saved in the database, more than three feature points (which are recognized as having a sufficiently close geometry relationship) must be recognized for the same object. In contrast, a segmented object is recognized as an object different from those in the database.

5. Conclusion

In this study, we proposed a segmentation algorithm that is suitable for the recognition of most household objects and the construction of their models. In the proposed hierarchical model, the coarse layer can roughly segment the foreground object and background, and then the GrabCut algorithm is used in the fine layer to drastically reduce the amount of segmentation information and then segment the correct object. The proposed algorithm can segment objects in an environment with a suitable distance and appropriate size, and the segmented object information can be used in the modeling of back-end images. In experiments, the average accuracy rate of the proposed algorithm reached 93.05%. As compared with the algorithm of manual object segmentation with the SIFT algorithm, although the proposed algorithm has a lower recognition accuracy rate, it has the advantage of automated recognition, which can greatly reduce the recognition time and increase the range of applications of robots.

Acknowledgments

This work was supported by the Educational Research Projects of Young and Middle-Aged Teachers in Fujian Province (JAT200593) and Project of Education and Teaching Reform in Fujian Province (FBJG 20170073). This work was also supported by project Nos. MOST 109-2221-E-390-023 and MOST 109-2221-E-390-023.

References

1 L. Liu, S. H. Yang, Y. Wang, and Q. Meng: Meas. Control **42** (2009) 12.
2 G. A. Zachiotis1, G. Andrikopoulos, R. Gomez, K. Nakamura, and G. Nikolakopoulos: 15th IEEE Int. Conf. Robotics and Biomimetics (ROBIO 2018) 1999.
3 N. Ramoly, A. Bouzeghoub, and B. Finance: IRBM **39** (2018) 413.
4 B. Basavaprasad and M. Ravi: Int. J. Eng. Res. Technol. **3** (2014) 310.
5 F. Kang, C. Wang, J. Li, and Z. Zong: Adv. Multimedia **2018** (2018) 1083876.
6 C. Rother, V. Kolmogorov, and A. Blake: ACM Trans. Graphics **23** (2004) 309.
7 N. An and C. Pun: 10th Int. Conf. Computer Graphics, Imaging and Visualizatio (CGIV 2013) 79.
8 D. Yang and T. Deng: Int. Cong. Image and Signal Processing Applications (ICSIPA 2011) 999.
9 D. G. Lowe: Inter. J. Compu. Vision **60** (2004) 91.
10 S. L. Al-khafaji, J. Zhou, A. Zia, and A. W. Liew: IEEE Trans. Image Process. **27** (2018) 837.
11 H. R. Kher and V. K. Thakar: 2014 5th Int. Conf. Signal and Image Processing (ICSIP 2014) 50.
12 A. Sharma, M. D. Ansari, and R. Kumar: 4th Int. Conf. Signal Processing, Computing and Control (ISPCC, 2017) 246.
13 H. Zhao, J. Tang, and B. Luo: 5th Int. Conf. Computer Science & Education (ICCSE 2010) 1599.
14 A. C. Silva: Int. Conf. Systems, Signals and Image Processing (IWSSIP, 2020) 19.
15 G. Wu, W. Liu, X. Xie, and Q. Wei: IEEE Int. Conf. Image Processing (2007) VI-169.
16 P. Chang and J. Krumm: 1999 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (1999) 1063.
17 Y. Bastanlar, A. Temizel, and Y. Yardimci: IET Electron. Lett. **46** (2010) 346.
18 P. Azad, T. Asfour, and R. Dillmann: IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS 2009) 4275.
19 K. Mikolajczyk and C. Schmid: IEEE Trans. Pattern Anal. Mach. Intell. **27** (2005) 1615.