On Human Inspired Semantic SLAM's Feasibility

Dominik Maximilián Ramík, Christophe Sabourin and Kurosh Madani

Images, Signals, and Intelligent System Laboratory
(LISSI / EA 3956), PARIS-EST Creteil (UPEC)
Senart Institute of Technology, Avenue Pierre Point, Lieusaint, 77127, France

Abstract. Robotic SLAM is attempting to learn robots what human beings do nearly effortlessly: to navigate in an unknown environment and to map it in the same time. In spite of huge advance in this area, nowadays SLAM solutions are not yet ready to enter the real world. In this paper, we observe the state of the art in existing SLAM techniques and identify semantic SLAM as one of prospective directions in robotic mapping research. We position our initial research into this field and propose a human inspired concept of SLAM based on understanding of the scene via its semantic analysis. First simulation results, using a virtual humanoid robot are presented to illustrate our approach.

1 Introduction

In mobile robotics, the ability of self-localization with respect to the environment is crucial. In fact, knowing precisely where the robot is, and what kind of objects surround it in any given moment of the time enables it to navigate autonomously and to interact with an unknown environment in a conscious manner. An informal definition of the Simultaneous Localisation And Mapping (SLAM) describes it as a process, in which a mobile robot explores an unknown environment, creates a map of it and uses it simultaneously to infer its own position on the map. In the real world SLAM applications, data association has often to be done under large amount of uncertainty. Moreover, the real environment is usually very complex and dynamic and it is not easy for a robot to interpret it in a reliable way. It is this complexity, what makes SLAM a challenging task. A comprehensive list and principle explications of nowadays most common SLAM techniques can be found in [1], [2] and [3]. Although from its beginning until present days the research community achieved a significant advance on the field of SLAM [4], it is not yet a solved problem. Autonomous navigation in dynamic environment [5] or understanding the mapped environment by including semantics into maps [6] are the actual challenges of this research field..

In this paper, the state of the art in robotic mapping is investigated. We identify a relatively new field of research within the field of SLAM, which is attempting to perform simultaneous localization and mapping with the aid of semantic information extracted from sensor readings. One of the research interests of our laboratory (LISSI) is autonomous robotics notably in relation to humanoid robots. We are convinced that the research on semantic SLAM will bring a useful contribution on this

Ramík D., Sabourin C. and Madani K. (2010).
On Human Inspired Semantic SLAM's Feasibility.
In *Proceedings of the 6th International Workshop on Artificial Neural Networks and Intelligent Information Processing*, pages 99-108

topic. We position our initial research into this field, drawing our inspiration from human way of navigation and place description. In fact, contrary to most of current SLAM techniques, which tend to infer precisely and globally the navigation environment, the human way of doing is based on very fuzzy description of the environment and it gives preference to local surroundings of the navigation backdrop. A simulation using a virtual humanoid robot (Nao robot) is presented to demonstrate some of the proposed ideas. The real Nao will be used in our further work.

The paper is organized in the following way: section 2 focuses on the state of the art in semantic SLAM. In the third section, we are discussing our approach to image segmentation and scene interpretation. Section 4 gives an overview of the robotic humanoid platform we use. The fifth section presents our initial results and the paper concludes with section 6.

2 Semantic SLAM

In this section, one of the latest research directions on the field of SLAM, the so-called semantic SLAM, is discussed. The concept itself may be perceived as a very important and pertinent one for future mobile robots, especially the humanoid ones, that will interact directly with humans and perform tasks in human-made environment. In fact, it is the human-robot interaction, which is probably one of important motives for employing semantics in robotic SLAM as humanoids are particularly expected to share the living space with humans and to communicate with them.

Semantics may be incorporated into the concept of robotic SLAM in many different ways to achieve different goals. One aspect of this may be the introduction of human spatial concepts into maps. In fact, humans usually do not use metrics to locate themselves but rather object-centric concepts and use them for purposes of navigation ("I am in the kitchen near the sink" and not "I am on coordinates [12, 59]") and fluently switch between reference points rather than positioning themselves in a global coordinate system. Moreover, presence of certain objects is often the most important clue for humans to recognize a place. An interesting work addressing the mentioned problems has been published in [7], in which the world is represented topologically with a hierarchy of objects and place recognition is performed based on probability of presence of typical objects in an indoor environment. A part of the work shows a study based on results of questioning about fifty people. The study was aimed to understand human concepts of self-localization and place recognition. It proposes that humans tend to understand places in terms of significant objects present in them and in terms of their function. A similar way (i.e. place classification by presence of objects) has been taken by [8] where low-level spatial information (grid maps) is linked to high-level semantics via anchoring. During experiments, the robot has interfaced with humans and performed tasks based on high-level commands (e.g. "go to the bedroom") involving robots "understanding" of the concept of bedroom and usage of low-level metric map for path planning. However, in this work, object recognition is black-boxed and the robot is not facing any real objects in the experiments but only boxes and cylinders of different colours representing different real-world objects.

An approach treating this "gap" between object recognition and semantic SLAM is presented in [9]. Here, a system based on a mobile robotic platform with an omni-directional camera is developed to map an outdoor area. The system generates a semantic map of structures surrounding the robot. Buildings and non-buildings labelled on the map. In [6], a more general system is presented, employing a wheeled robot equipped with a laser 3D scanner. Authors show ability of their robot to evolve in an indoor environment constructing a 3D semantic map with objects like walls, doors, floor and ceiling labeled. The process is based on Prolog clauses enveloping pre-designed common knowledge about such an environment (i.e. the doors are always a part of a wall and never a part of the floor). This enables the robot to reason about the environment. Further in the paper, an object detection method using the laser range data is shown with a classifier able to distinguish and to tag different objects surrounding the robot including humans and other robotic platforms. In [10] active object recognition is performed by a mobile robot equipped by a laser range finder and a camera with zoom. A semantic structure is extracted from the environment and integrated to robots map. It allows the robot to reach previously detected objects in an indoor environment. Another object recognition technique is shown in [11] including an attention system. Based on recognized objects a spatial-semantic map is built.

3 Image Segmentation and Scene Interpretation

Section 2 showed the pertinence of semantic SLAM in the frame of state of the art robotic mapping. It is exactly this field, on which we are focusing our research. Our motivation comes from the natural ability of human beings to navigate seamlessly in complex environments. Obviously, the way we are orienting ourselves in the space is very different from what contemporary robots do. To describe a place, we use often very fuzzy language expressions and approximation (as shown in [7]). This is in contrast with most of the current SLAM algorithms. In navigation or place description people also rarely use "global coordinates" but rather divide the scene into some kind of hierarchic clusters around distinctive objects, which then act as local origin of coordinates. Another interesting point is that people are able to infer distance of an object according to its apparent size and their experience of object's true size. From what has been mentioned so far it is clear that recognition of objects and understanding of their nature (semantic treatment) is an integral part of human navigation or "human SLAM" and not just an extra layer of it. We believe that application of semantics and human inspired scene description could bring a considerable benefit in development of robust SLAM applications for autonomous robotics.

To initiate a semantic scene interpretation, the image has to be segmented first. Although there exist many approaches to image segmentation (see [12]) for a reference), not all of them are suitable for purposes of mobile robotics, because it requires image treatment being done in real time. Our segmentation algorithm breaks the input image into parts containing similar colors present in different brightness levels. This should reflect that different shades do not usually mark different objects but only different light conditions on the same object.

We have chosen to use the YCbCr color model within our algorithm. This color model consists of three channels. The Y channel is dedicated to the luminance component of the image and stores the information about light and dark parts of the image. The other two channels Cb and Cr contain respectively the blue and the red chrominance component of the image. The YCbCr color model is more practical for purposes of our color segmentation algorithm, than classical RGB. It is because YCbCr separates the luminance of color and the color itself into different channels, while in RGB both color and its lightness are mixed together. The algorithm works in two stages, coarse to fine. In the first one, the Cr and Cb components of the image are acquired, their contrast is stretched and median filter is applied on both of them to remove noise. Then a single scan is made through rows and columns of the image and the first position that is not already occupied by a detected component is chosen as a seed point with coordinates x_{seed} and y_{seed}.

Eq. 1 captures how seed point is used to extract a segment of interest (S) from the image. The P stands for all the pixels in the image, whereas p is the actually examined pixel. Predicate C is true only if its two arguments (p, p_{seed}) are in four-connectivity and I stands for intensity of pixel. Seed pixel is denoted by p_{seed}. A pixel of the image belongs to the segment S under the following conditions: the difference of intensities of the current and the seed pixel is smaller than a threshold and there exists a four-connectivity between it and the seed pixel. Applying this on both chroma sub-images (Cr and Cb) we obtain segments denoted as S_{Cr} and S_{Cb}. A new segment S is then determined following Eq. 2 as the intersection of segments found on both chroma sub-images without pixels already belonging to an existing segment.

$$\forall p \in P;\ C(p, p_{seed}) \,\&\, |I(p) - I(p_{seed})| < \varepsilon \rightarrow p \in S \tag{1}$$

$$S = S_{Cr} \cap S_{Cb} - S_{all}. \tag{2}$$

At the end of the scan, a provisory map of detected segments is available leaving out components whose size is below certain threshold. At this stage, the image is often over segmented due to the method of selection of seed points. However, it serves as the first guess about the positions of regions. In the second step, all the provisory segments are sorted according to their area and beginning with the largest one the algorithm of segmentation is run again. This time the seed points are derived from centers of segments defined by Eq. 3 and Eq. 4.

The seed point k_{seed} is determined as such a pixel from the skeleton whose distance from its closest contour pixel is maximal. Here, K is the set of pixels of skeleton belonging to segment S and C is the set of contour pixels of S. D_i denotes the minimal distance between the given pixel k_i and the contour. In this step, similar segments from the previous step are effectively merged. At this point, found segments may contain distinctive areas of different brightness having similar chroma. The ultimate step of the algorithm is in constructing a histogram of luminance values of each segment. The histogram is then polished by application of sliding average. If multiple significant clusters are found in the histogram, the segment is broken-up accordingly to separate them. Having finished this step, found segments are stored for further use.

$$D_i = \arg\min_{k_i} \left| k_i - c_j \right| \qquad where$$
$$i \in \{0,...,|K|\} \ and \ k_i \in K = \{k_0,...,k_m\} \ and \ c_j \in C = \{c_0,...,c_n\} \tag{3}$$

$$k_{seed} = \max_{i \in \{0,...,|K|\}} D_i \tag{4}$$

$$Q = 4\pi n / o^2. \tag{5}$$

In the next processing step, the segments are labeled with linguistic terms describing their horizontal and vertical position and span with respect to the image frame. Both average color and its variance are computed for each segment along with the number of pixels forming the segment. The compactness (Q) of the segment is computed following Eq. 5, where n denotes the number of pixels of the segment and o the number of pixels forming the contour of the element. These features, which represent each segment, are then used in a set of linguistic rules, acting as a prior knowledge about the world. The aim is to determine the nature of each segment and its appurtenance to an object of the perceived environment. For example, a compact segment found in mid-height level surrounded by the wall is considered as a "window".

4 Humanoid Robot Platform

The robotic platform we use for simulation and experiments is described in this section. It is based on Nao, a humanoid robot manufactured by Aldebaran Robotics[1]. The robot is about 58cm high with height slightly exceeding 4kg. Its degrees of freedom are as follows: 2 DOF for the head, 5 DOF for each arm, 1 DOF for the pelvis, 5 DOF for each leg and 1 DOF for hands to control the grasp. Concerning the available sensors, it is equipped with two CMOS cameras with resolution up to 640x480px. One camera is on the front of the head and the other is covering the space around the feet of the robot (this one was added specially because of the usage of Nao in RoboCup robotic football matches). Two channel sonar and 2 IR sensors are in robot's chest. It also possesses a tactile sensor, bumpers and inertial sensors. To interact with humans, robot is equipped with voice synthesizer and a speech recognition unit.

The robot can operate in fully autonomous mode using its AMD Geode 500MHz processing unit to run programs and behaviors stored in its memory. Alternatively, it can be operated remotely from another computer via a WiFi (or Ethernet) connection. To perform simulations, a virtual version of Nao is available for the Webots simulation program developed by Cyberbotics[2]. The program allows us to create a virtual world and to simulate robots interaction with it including gathering of sensor data from cameras and IR/sonar sensors. Nao can be programmed in different manners.

[1] http://www.aldebaran-robotics.com
[2] http://www.cyberbotics.com/

The choice of languages includes C, C++, Python and URBI and the code can be run locally on robot's CPU or distantly via a network connection. After having explored different ways of programming

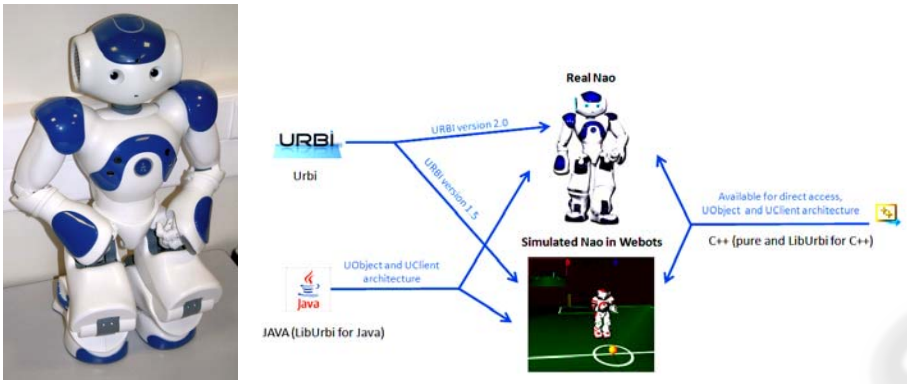

Fig. 1. A scheme describing our humanoid robotic platform, showing different possibilities of programming it. On the left a photo of the real Nao used in our laboratory.

Nao, we have chosen URBI developed by Gostai[3] for development purposes (see Fig. 1). There are several reasons for this choice. First, URBI is a specific language developed especially for robotics and by its nature allows simple and fast development of robotic behaviors. Moreover, it provides a simple way of managing parallel processes, which may be a complex task in other languages. Although programming in URBI involves writing in URBI script, which is then interpreted by an interpreter, URBI programs do not suffer from lose of performance. The code of its internal objects is written in C++ to keep high efficiency of the language. LibURBI connectors allow user to develop own objects using so called UObject architecture and to plug them into the language. These objects can be developed in C++ or Java code (a connector is available for Matlab as well). User-created objects can be run directly on the robot or transparently on a remote machine via CORBA technology. With these properties, URBI seems to us to be suitable for developing complex behaviors on robots as well as computationally intensive tasks as image processing.

For the demo simulation presented in the next sections, we used the simulated robot described above and we are going to use its real equivalent in our further research on the field of semantic SLAM. The task itself may be not perceived as being strictly specific for humanoid robots. However, the motivation to use a humanoid robots comes from the fact, that they are specially designed with the aim to interact with humans and to act in human-made environment. If they are already imitating humans in their physical form, why should not we enable them to do the same on the level of their software? The concepts we are exploiting here come from human approach to navigation and orientation in the space. Thus embedding such human inspired semantic SLAM capabilities onto a humanoid robotic platform seems pertinent to us.

[3] http://www.gostai.com/

5 Results

After having the image segmented, all segments are labeled and interpreted by a set of rules representing prior knowledge about objects. Following the mentioned rules segments can be even merged so that e.g. multiple fragments of floor partially occluded by objects laying on it are labeled as belonging to the same object of type "floor". Fig. 2 gives an example resulted from the left image (supposed as the original image acquired by robot's vision system). Fig. 3 depicts the intermediate steps of segmentation. This "semantic" information is subsequently used to approximate the actual distance of certain objects. Having an object of type "window", it is looked-up in a table containing usual sizes of different objects and once found the size information is used along with the pixel size of the object on the image and the field of view of the camera to compute the approximate distance of the window (see the right image in the Fig. 2). This is described by Eq. 6 (simplified for horizontal size only). The distance d to an object is the product of estimated real width w_{real} of the object and tangent of its width in pixels w_{px} on the image multiplied by fraction of the horizontal field of view φ and the width w_{image} of the image in pixels

$$d = w_{real} * \tan (w_{px} * \varphi / w_{image}) \quad (6)$$

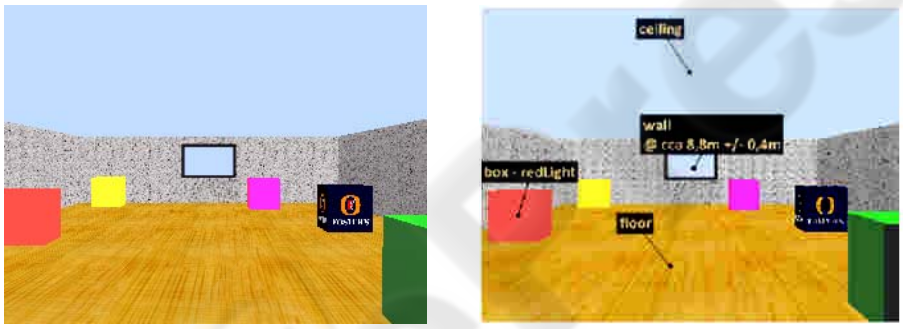

Fig. 2. A view of the robot's random walking sequence. The left image is the original one. The right image shows the result after the interpretation phase. Some of the detected objects are labeled. The opposing wall is labeled also with its approximate distance with respect to the robot.

The aim of this computation is absolutely not to infer the exact distance of an object, but rather to determine whether it is "far" or "near" in the context of the simulated world or if it is nearer to the robot in comparison to another object. This can help in the further process of creation of the map of the location.

This demo, however limited, gives a preliminary idea of principles of semantic mapping and more importantly, it gives a starting point for the research in the area of semantic SLAM. Resigning to precise metric position of every object in the mapped world and replacing it only by rough metric and human expressions like "near to" or "beside of" is believed to enable us to create faster and more robust algorithms for robotic SLAM. Using of "object landmarks" to navigate in an environment is certainly more meaningful that using e.g. simple points in case of classical SLAM. Knowing

the nature of an object gives an opportunity to distinguish between important and random objects. One can imagine a robot with an ability of choosing landmarks for purposes of its navigation by itself. With the knowledge about available objects, it could prefer to pick up the most important and stable objects that are unlikely to change their place or appearance in the lifetime of the map.

Fig. 3. The two segmentation steps relative to the result of Fig. 2.

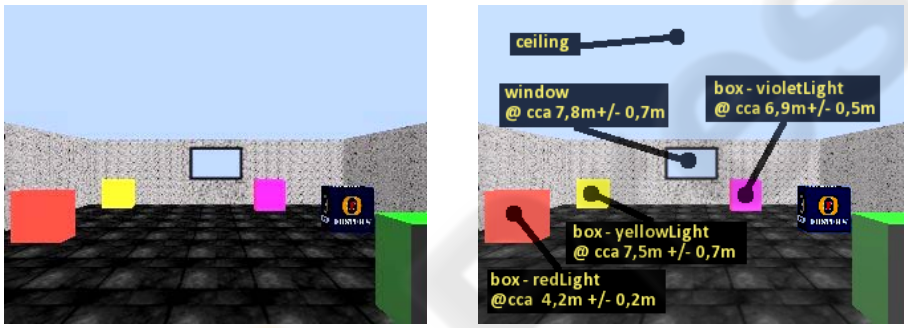

Fig. 4. The same room with different textures (left) and resulted interpreted image (right).

The seed point k_{seed} is determined as such a pixel from the skeleton whose distance from its closest contour pixel is maximal. Here, K is the set of pixels of skeleton belonging to segment S and C is the set of contour pixels of S. D_i denotes the minimal distance between the given pixel k_i and the contour. In this step, similar segments from the previous step are effectively merged. At this point, found segments may contain distinctive areas of different brightness having similar chroma. The ultimate step of the algorithm is in constructing a histogram of luminance values of each segment. The histogram is then polished by application of sliding average. If multiple significant clusters are found in the histogram, the segment is broken-up accordingly to separate them. Having finished this step, found segments are stored for further use.

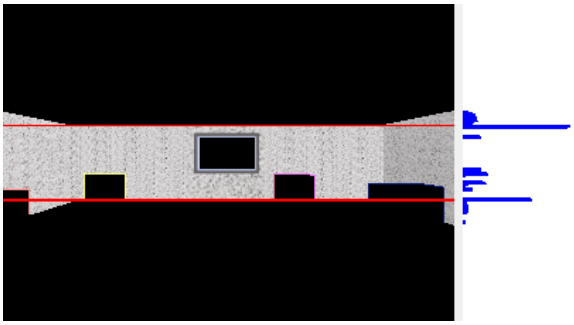

Fig. 5. Detection of the wall as a overall (macro) object in robot's environment.

It is important to notice the robustness of the proposed approach and the fact that the estimated distances, even if approximated, are relevant enough for extracting pertinent features relating environmental information. Fig. 4 gives results obtained from the right image showing the above-mentioned purpose. Fig. 5 gives an example of extended possibilities of the technique in detecting environmental information. In fact, it allows potentiality of a higher-level semantic labeling of objects constituting the robot's environment. Here, one can notice that in the given example (detection of the room's wall including the associated objects as window, etc…) allows the possibility to link previously labeled objects (for example the window) to the "wall" in term of "room's wall with the window".

6 Conclusions and Perspectives

Simultaneous localization and mapping is an important ability for an autonomous mobile robot. State of the art techniques have been discussed here giving an idea about the current state on the field of SLAM. In spite of a great advance in SLAM techniques in past years, most of the existing SLAM solutions can accommodate only a particular case or environment. A stable and generally usable SLAM solution is still missing. Given the state of the art of SLAM, one of the basic directions, which are expected to play a key role in future development of SLAM is so called "semantic SLAM": adding a semantic level into robotic mapping should help robots to go beyond simple "structural" information about the world that surrounds them. It should enable them to "understand" it.

In this paper, we identify the pertinence of semantic SLAM for the future development in mobile robotics and we present our initial research on this field. Our research is strongly inspired by the human way of navigation and place description. The semantic information about objects in the scene may improve mapping capabilities of robots. It should enable them to reason about their environment as well as to share their knowledge with humans and receive commands using human concepts and categories in a seamless way.

For description of a scene by semantic means, a good algorithm for image segmentation is an important starting point. Preferably, it should perform segmentation using both color and texture information. For real time use, fast and efficient algorithms are

required. A part of our future work will be dedicated to further development of such an algorithm. Another part of our future work will be focused on development of algorithms of semantic SLAM we outlined in this paper. They will be consequently implanted and verified in an indoor environment on the real Nao robot.

References

1. Durrant-Whyte, H., et al. Simultaneous Localisation and Mapping (SLAM): Part I The Essential Algorithms. Robotics and Automation Magazine, Vols. 13, No 2, pp. 99-110, (2006).
2. Durrant-Whyte, H., et al., Simultaneous Localisation and Mapping (SLAM): Part II State of the Art. Robotics and Automation Magazine, Vols. 13, No 3, pp. 108-117 (2006)
3. Muhammad, N., Fofi, D. and Ainouz, S. Current state of the art of vision based SLAM. Image Processing: Machine Vision Applications II, Proceedings of the SPIE, Vol. 7251, pp. 72510F-72510F, (2009)
4. Thrun, S. and Leonard, J. J. Simultaneous Localization and Mapping. [ed.] B. Siciliano and O. Khatib. Springer Handbook of Robotics. Berlin Heidelberg: Springer-Verlag, 37, (2008)
5. Hahnel, D., et al. Map Building with Mobile Robots in Dynamic Environments. Proceedings of the IEEE International Conference on Robotics and Automation. Taipei: IEEE, Vol. 2, pp. 1557-1563, (2003)
6. Nüchter, A., Hertzberg, J. Towards semantic maps for mobile robots. Robotics and Autonomous Systems. Amsterdam : North-Holland Publishing Co., Vol. 56, pp. 915-926, (2008)
7. Vasudevan, S., et al. Cognitive maps for mobile robots-an object based approach. Robotics and Autonomous Systems. Amsterdam: North-Holland Publishing Co., Vol. 55, pp. 359-371, (2007)
8. Galindo, C., et al. Multi-Hierarchical Semantic Maps for Mobile Robotics. Internati. Conf on Intelligent Robots and Systems (IROS 2005). Edmonton : IEEE, pp. 2278- 2283, (2005)
9. Persson, M., et al. Probabilistic Semantic Mapping with a Virtual Sensor for Building/Nature detection. International Symposium on Computational Intelligence in Robotics and Automation. Jacksonville : IEEE, pp. 236-242, (2007)
10. Ekvall, S., Jensfelt, P. and Kragic, D. Integrating Active Mobile Robot Object Recognition and SLAM in Natural Environments. International Conference on Intelligent Robots and Systems, 2006 IEEE/RSJ. Beijing: IEEE, pp. 5792-5797, (2006)
11. Meger, D., et al. Curious George: An attentive semantic robot. Robotics and Autonomous Systems. Amsterdam : North-Holland Publishing Co., Vol. 56, pp. 503-511, (2008)
12. Lucchese, L. and Mitra, S. K. Color image segmentation: A state-of-the-art survey. Proc. Indian Nat. Sci. Acad. (INSA-A). Vols. 67-A, pp. 207-221, (2001)