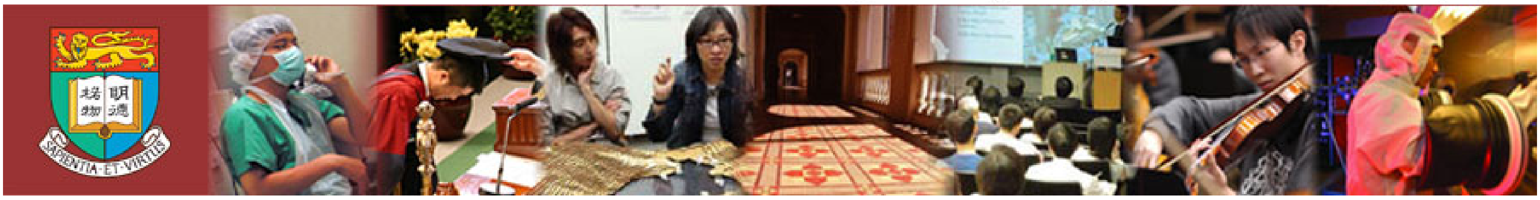
CORE

Metadata, citation and similar papers at core.ac.uk

Provided by HKU Scholars Hub

The HKU Scholars Hub The University of Hong Kong 香港大學學術庫



Title	Sub-scene generation: a step towards complex scene understanding
Author(s)	Zhu, S; Yung, NHC
Citation	The 2011 IEEE International Conference on Multimedia and Expo (ICME 2011), Barcelona, Spain, 11-15 July 2011. In Conference Proceedings of ICME, 2011, p. 1-6
Issued Date	2011
URL	http://hdl.handle.net/10722/158735
Rights	IEEE International Conference on Multimedia and Expo. Copyright © IEEE.

SUB-SCENE GENERATION: A STEP TOWARDS COMPLEX SCENE UNDERSTANDING

Shan-shan Zhu and Nelson H. C. Yung
Department of Electrical and Electronic Engineering
The University of Hong Kong
Hong Kong, China
Email: {sszhu, nyung}@eee.hku.hk

ABSTRACT

This paper proposes a new method for generating sub-scenes from images of natural scenes. It is based on over-segmented image patches initially and uses rules derived from human psychology to merge the patches into reasonable sub-scenes. This merging process focuses on covering semantic gaps by exploring semantic connections among the patches, such as the influence area of each patch where the potential interaction with other patches may exist; and applying visual organization rules of proximity property and color harmony property. The proposed method is simple yet efficient computationally. It does not have any limitation on the input image class and does not include a training process. Our experiments indicate that it handles complex scenes with clutters and occlusions very well, and outperforms general segmentation methods by its meaningful semantic grouping and self determined stop criteria.

Index Terms—sub-scene, semantic gap, complex scene

1. INTRODUCTION

Natural scene understanding is inherently challenging as it demands a high degree of machine intelligence in discerning the specific details of a scene, prior to understanding it. Outdoor scenes are somehow relatively easier to analyze than indoor scenes. This is probably because of the nature of the objects seen outdoor, e.g., buildings, mountains, beaches, and sky, and the way in which the scene is composed, such as the sky is above the sea. Indoor scenes are quite different in the sense that there are often far too many items cluttered together and they do not necessarily bear a strong spatial relationship with each other.

To tackle the scene understanding problem, three major research directions have so far been proposed in literatures. The first direction focuses on recognizing the objects that compose the scene, and can be considered as a local approach. This approach is intuitive, and obviously contextual information can be derived from recognized objects. For instance, Siddiquie and Gupta [1] recently proposed a smart question aided method for this purpose. An example of smart questions is: after the sea at the lower part of the scene has been detected, what are the things usually found above the sea. This increases the likelihood of objects such as boat. If a boat is detected then it can be argued that it is a boating scene. Otherwise, it continues to search for other alternatives. However, object recognition is not without problems, especially when complex scenes are concerned. The first issue is that object detection usually requires objects to be in full view. Unfortunately, complex scenes are almost always crowded with objects occluding each other. The sheer number of object is hard to handle and the coupling of occlusion could result in significant detection ambiguity. The second issue is that as objects in complex scenes are free to move around, changes in illumination across the field-of-view, object orientation and scale could introduce substantial appearance changes to confuse detection and recognition. The third issue is that most recognition approaches need liberal quantity of labeled training procedure for each object. Building the complete encyclopedic knowledge of all objects that may appear in a complex scene is an enormous task, let alone the complexity of recognizing each class under conditions of all possible perspectives and the appearance variations within them [2].

The second direction focuses on generalizing the characteristics of a scene globally, without having to recognize individual objects. SIFT is one of common features considered by many of these methods to extract distinctive features globally [3]. It has been widely used primarily in scene categorization problem, and enjoyed some degree of success. From Qin and Yung [4], it can be seen that high categorization accuracy can be achieved for many outdoor scenes, but the same cannot be assumed for the more complex indoor scenes. This reveals the weakness of the global feature-based methods in that similar appearances of objects and the comparable scales between them (e.g. bedroom and living room) tend to reduce discrimination abilities.

Due to the shortcomings of the aforementioned two directions, the third direction combines the object level and global level together. In Quattoni and Torralba [5], they employed both the local and global representations to deal with the poorly recognize situation of indoor scenes. Some researches even generate new representation levels and

978-1-61284-350-6/11/$26.00 ©2011 IEEE

combine those together into a hierarchical structure as discussed in Parikh et al [6]. In their work, they build a structure from the root node of the entire scene through to objects and object parts, and finally to the leaves of features. The advantage of the hierarchical method is that it produces sufficient information for image understanding. However, it also introduces the complex problems of weighting among those levels and adopting different strategies towards different tasks. In this paper, we adopt the hybrid local-global approach and propose a new method that generates *sub-scenes* from a natural scene. Thus, the sub-scenes generated exhibit the merits of both levels, but avoid the weighting dilemma as in hierarchical structures. The sub-scene level reflects the way human comprehend a scene: break up the whole scene into several meaningful parts and by understanding these parts and the spatial relationship between them, one can understand the original scene. These parts are the sub-scenes we refer to in this paper. It is independent and robust enough to be a step towards further scene analysis and its embedded semantic information also outperforms other approaches.

The contributions of this paper are listed below:

1) The concept of sub-scene is introduced;

2) Sub-scenes are characterized by their properties; and

3) A novel adaptive method is proposed for generating sub-scenes from complex scenes, which is simple but robust.

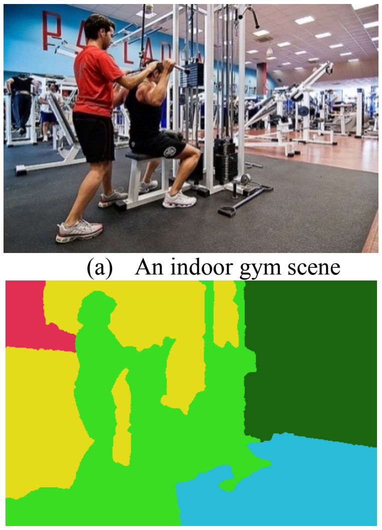
The organization of this paper is as follows: in Section 2 the definition of sub-scene is presented and its properties are explored. Section 3 illustrates the proposed method. Evaluation of the proposed method is given in Section 4, while conclusion is drawn in Section 5.

2. SUB-SCENE DEFINITION AND PROPERTIES

In general, sub-scenes can be viewed as meaningful entities that associate with some common knowledge classes and when joined together form the entire scene. The number of sub-scenes and their scales in the composition of a scene may be different from scene to scene. For simple scenes, sub-scenes may refer to the foreground and background, while in more complex scenes such as the one depicted in Figure 1(a), sub-scenes may mean the two people working out with the gym apparatus, the other apparatus at the background, the ceiling, the side wall and the floor. From Figure 1(b), the sub-scene generated by the proposed method consists of five classes in total: the two people together with the gym apparatus, the floor, the side wall, the distant objects and a small false segmented patch at the top left hand corner. We will use this example later on to discuss the definition and properties of the sub-scene. For the convenience, we will use the concept of *patch* to denote any segmented regions before it ended up as part of a sub-scene.

2.1. Sub-scene definition

The purpose of defining sub-scene is to facilitate effective human-scene interaction eventually. When a new scene is being viewed, people tend to divide it into several patches, some of which they are familiar with, whereas others are new or uncommon. Typically, it is those new and uncommon patches that attract attention. Sometimes, it is also the patch that is the largest or at the center of view that catches our eyes. As depicted in Figure 1(a), the focus of the scene is likely the two people working out with the gym apparatus, which can be interpreted as people working out with an apparatus. If identification is performed at an object level, then the likely interpretation would be the people in front of the apparatus and the contextual information of the relationship between people and apparatus will disappear. This leads to the function of sub-scenes: a sub-scene groups the objects with direct and strong interactions and separates those with weak interaction. It focuses on contextual information more than other representations and thus covers the semantic gap. What is more, if the side wall does not have red letters painted on it, it may be recognized as sky and contradiction results in the co-existence of the sub-scene sky (outdoor) and the sub-scene gym apparatus (always indoor according to common sense). In that particular case, both the sky-like side wall and the work-out pair would need further interrogation.



(a) An indoor gym scene

(b) Generated sub-scenes (denoted by different color)

Figure 1: An example of generated sub-scenes from a complex indoor scene.

As contextual information is conveyed by sub-scenes, it is possible to consider the problem from an Artificial Intelligence (AI) point of view. To start with, each scenario may be described in the Disjunctive Normal Form (DNF) which is a disjunctive of conjunctive clauses. Take $\mathcal{D}\mathrm{escrb}(\cdot)$ as the describe function then there is $\mathcal{D}\mathrm{escrb}(\cdot) = \mathrm{clause1} \vee \mathrm{clause2} \vee \cdots \vee \mathrm{clause\ n}.$ Take the scenario in Figure 1(a) as an example, then

$\mathcal{D}\text{escrb}(\text{Fig1. a}) =$ ‘two people working out with gym apparatus’ $\vee$ ‘the background apparatus’ $\vee$ ‘the ceiling’ $\vee$ ‘the side wall’ $\vee$ ‘the floor’. Furthermore, sub-scenes are corresponding instances of these clauses.

Definition: Sub-scenes is a division of the original scene that its semantic information forms the conjunctive clauses in the DNF representation of the scene.

Let I denotes the given scene and Ω denotes the one possible division, $\Omega = 1,2,\cdots,\mathrm{n}$, and let $S(k)$ as the corresponding sub-scene k, $k \in \Omega$. Let $\mathcal{S}\text{em}(\cdot)$ as the conjunctive semantic information generalized by sub-scenes, then we have

$$\mathcal{D}\text{escrb}(I) = \cup_{k\in\Omega} \mathcal{S}\text{em}[S(k)]. \qquad (1)$$

Among all the possible combinations, the S^* we are looking for refers to the set of sub-scenes that achieves the most compact result under the conjunctive restriction, which means

$$\mathcal{D}\text{escrb}(I) = \min_n \cup_{k=1,2,\cdots,n} \mathcal{S}\text{em}[S^*(k)], \qquad (2)$$

in which

$$\begin{cases} \cap_{k\in\Omega} S(k) = \emptyset \\ \cup_{k\in\Omega} S(k) = I \\ \mathcal{S}em(\cdot)\ limited\ to\ conjunctive\ restriction \end{cases}$$

Contextual problems are always difficult to be resolved directly through the definition, it will be more fruitful to explore its properties and from there to determine the set of S^* in (2).

2.2. Sub-scene properties

The properties of sub-scene are defined as follows:

Property 1: Each sub-scene has an area of influence (AOI).

Actually, besides sub-scene, each object or patch also has its own AOI. Within the AOI, there is a potential interaction with the neighboring patches. If indeed neighboring patches have a strong interaction, the semantic information of them must have a strong conjunction as well, which fulfills the condition in (2); and those patches are likely to be in the same sub-scene. So the AOI indicates the possible existence of interaction though it can only be confirmed later. To demonstrate the influence area, without loss of generality, a Normal distribution is employed. The AOI is assumed to have a Normal distribution $\mathcal{N}(\mu, \sigma^2)$, in which μ is the centre and σ^2controls its radius of influence.

Property 2: Patches of the same sub-scene tend to be in close proximity.

This nearness property is based on a psychology perspective. It is derived from Gestalt principles of visual perception [7]. It reveals the tendency that human brain groups spatially neighboring objects together and assumes they belong to the same sub-scene. Therefore if a patch has an intense relationship with a surrounding sub-scene it would be classified as part of the sub-scene even though there is slight difference between them. This is guaranteed by the proximity properties. But this does not mean sub-scene must be a non-separable region. If the two separate regions of the same sub-scene do not have strong joint influence on other patches they would stay separated. In other word, if they have a weak interaction in between or the surrounded region is unique then they would not be grouped together. This is illustrated in Figure 2.

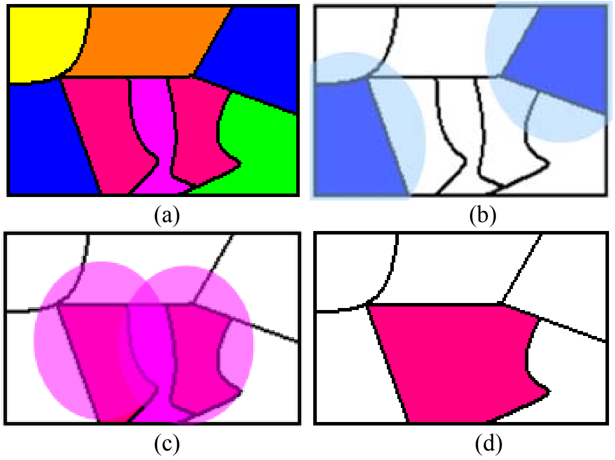


Figure2: Illustration of Properties 1 and 2.

Figure 2(a) depicts the patches of a scene. Figure 2(b) illustrates two distant patches (left bottom and right top) whose interaction is weak due to their non-overlapping AOI (expressed by the shadow). So they stay separated. Figure 2(c) and (d) illustrate the two patches at mid-bottom of the imagethat are not neighbors but near enough to each other. As their AOI overlap, and assume for illustration purpose that they have intense interaction and the middle patch has limited difference with them, these patches are grouped into one sub-scene, according to Property 2.

Properties 1 and 2 explore the relationship between sub-scenes. However, to generate a sub-scene we also need to know the attribute of the sub-scene itself. Furthermore, this attribute should enable the low level features to be combined under the condition of (2). The semantic gap between concrete features and abstract semantic relationships must be covered by it. Property 3 as described below tackles this problem.

Property 3: Both natural and artificial sub-scenes follow the discipline of harmony.

This is a hidden but deeply embedded rule in the world around us and has been a guideline for some of our behavior. For example, people tend to design furniture set with uniform or repeated color patterns. These harmony appearances are not only for aesthetic purpose but also assist in our human cognition towards our surroundings.The same effect applies to shape and texture as well. In this paper, we will focus on making use of the aspect of color harmony.

3. PROPOSED METHOD

The outline of the proposed method is depicted in Figure 3. Initially, any given image is over-segmented and then the proposed method is applied to merge and generate sub-scenes. First, it performs neighborhood merging based on similar semantics. After that, it globally explores semantic conjunctive patches by using the combination of AOI and evaluation functions. Before describing the specific algorithm of the proposed method, we will discuss the conjunction achievement to fulfill the conditions in sub-scene definition.

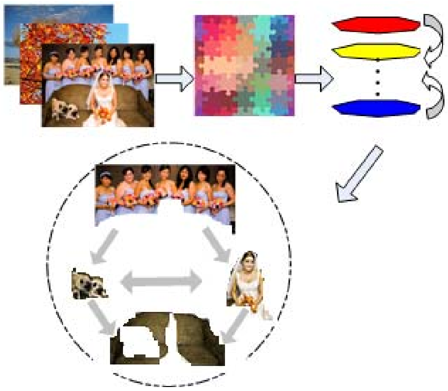


Figure3: Outline of proposed method.

3.1. Semantic conjunction achievement

To fulfill the conditions in (2) the patches which form the same sub-scene must have strong semantic conjunctions with one another. According to Property 3 we choose color harmony as the criteria. The evaluation function used for the semantic conjunction is the Bhattacharyya distance of normalized histograms of the hue of two patches. Take M, N as two patches, and take $\mathrm{Histo}[M(i)]$ as the i^{th} bin of the normalized histogram of patch M and the semantic conjunctive evaluation is expressed by $E(M,N)$. Then we have

$$E(M,N) = \sqrt{1 - \sum_{\mathrm{i}} \sqrt{\mathrm{Histo}[M(i)] \cdot \mathrm{Histo}[N(i)]}}. \qquad (3)$$

$E(M,N)$ is between 0 and 1, and is smaller when the two patches M and N are more similar. So if $E(M,N)$ is smaller than a threshold then it can be assumed that M and N have intense semantic conjunctions.

It is always difficult to set the threshold, especially when the image is not known a priori. To tackle this issue, we develop a novel self-guaranteed threshold. When evaluating whether patch N should be merged with patch M, we take part of M and evaluate it against the whole M and take this result as a reference threshold. Some patches change violently while some are reasonably consistent. No matter what the situation is, if the evaluation result of M and N is within or close to that of M and M part, we can have the confidence to merge N to M. Take BOOLEAN expression of $Merge(M,N)$ as the indicator, if and only if it is true, then merge N to M.

$$Merge(M,N) = \begin{cases} 1, & \text{when } E(M,N) < E(M, M_{part}) + \xi; \\ 0, & \text{otherwise.} \end{cases} \qquad (4)$$

In this equation the slack variable ξ controls the tolerance. Since the likelihood of interaction increases within the AOI as discussed in properties 1 and 2, we explore the AOI of the patches and raise ξ to increase the tolerance. As the distribution of AOI is assumed to have a Normal distribution $\mathcal{N}(\mu, \sigma^2)$ and as 2σ from the mean accounts for 95%, and 3σ from the mean accounts for 99%, we can simplify the AOI by regarding the radius of AOI from the center of the region to be 3σ while the region itself expands 2σ. Thus the AOI extends from the region by its radius.

3.2. Objective formulation and algorithm

The objective formula is to get the minimization as follow.

$$g(I) = \min\left\{ \sum_{i=1}^{n} \left[\frac{E\big(S(i), S(i)_{\mathrm{part}}\big)}{\sum_{j=1\ j\neq i}^{n} E\big(S(i), S(j)\big) \cdot {}^{1}\!/_{n-1}} \right] \cdot \mathrm{e}^{\tau n} \right\} \qquad (5)$$

It aims at compact semantic conjunction within each sub-scene while keep the discrimination among them and penalize if the total number of sub-scenes is large by the $\mathrm{e}^{\tau n}$ part.

In the implementation, an over-segmented result is used as the input and gradually merged to achieve the above objective. In our experiment, we apply the method of Tan and Yung [8] to get the over-segmented image. In fact, these over-segmented patches can be generated from any segmentation methods.

The algorithm for generating sub-scenes is depicted in Figure 4. The algorithm does not have a training step as it endeavors to explore the attribution within each scene rather than building up models and then fitting the given image. It first merges semantic conjunction regions and then relaxes ξ to examine AOI of each patch. The slightly raised slack variable ξ is a tradeoff between the compactness of each sub-scene and the total sub-scene number. This total number is not set manually; it is derived automatically from the algorithm and is different from scene to scene although we limit it to within ten, which is a conservative estimate of the total sub-scene human can handle in reality.

```
Sub-scene generation algorithm
Require: Over-segmented image $I$.
Initialize:$I = \cup S(i)$,$i$=1,⋯,50. $n$=50. Set$\xi$ =0.1.
Repeat:
  For each patch $S(i)$, $i$=1,⋯,50:
    1: Calculate reference threshold: $E(S(i), S(i)_{part})$ as in Eqn. 3.
    2: Evaluate semantic conjunction of its neighboring patches $E(S(i), S(j))$.
    3: Merge neighbor patches follow Eqn. 4.
    4: $i \leftarrow i$+1 and go to step 1.
Until no more merging processed in Step 3.
Repeat:
    For each sub-scene $S(i)$, $i$=1,⋯,$n$:
    5: Dilate $S(i)$ to get its AOI.
    6: Evaluate semantic conjunction as in Step 1 to Step 3 with the current $\xi$.
    7: Relax the tolerance a little bit more, e.g. $\xi \leftarrow \xi + 0.05$.
    8: $i \leftarrow i$+1 and go to Step 5.
Check if $n$<10 fulfilled then stop repeating.
```

Figure4: Proposed sub-scene generation algorithm.

4. RESULTS AND DISCUSSION

We have evaluated the proposed method using the dataset of Indoor scene proposed in [5]. These scenes are complex and quite challenging. In our results, we use different colors to indicate different sub-scenes. In comparison, we again use the result of [8] but as their method needs the total number of segmentation to be set beforehand, we use the number generated by our method to achieve the equivalent experimental condition. We also compared with segmentation results generated by Normalized cuts [9] and Graph Cuts [10, 11, 12]. The results are depicted in Figure 5.

The first two scenes in Figure 5 are the pool scene and the cloister scene which have obvious illumination change and shadows. Our method separated sub-scenes of the pool, the floor, the wall and the cloister well while due to the reflection interference there is a small patch error classified at the middle right in the pool scene. In comparison, the results from [8] and Normalized Cut separate the scene incorrectly like the right part of the cloister is set apart from the whole architecture in the result of [8] while the whole wall is split into three parts with Normalized Cut. The result of Graph Cut is good at the pool sub-scene but failed at the rest. The fine decoration and shadow influence the Graph Cuts method a lot, so that it separated the scene into small patches which do not provide a general idea of the whole scene. The following row is a hair salon scene. Our method grouped the same attribute sub-scenes together, such as the three sets of mirror, while result in [8] has some under segmented region at the left part and over segmented mirror patch at the right part. The Normalized Cut separated the three sets of mirror apart. The Graph Cuts segmented the wall well but fail to segment the mirrors. The next scenario is a garage scene. The motorcycle sub-scene, the outside trees and the garage had been segmented successfully while heavy occlusions greatly influence the result of [8]. The Normalized Cut and Graph Cuts segmented the motorcycle sub-scene well but both have problems on the outside view and the garage. Last two scenarios are the auditorium and shop scenes which also reveal the merits of the proposed method in dealing with clutters and in grouping sub-scenes of similar attributes. The Normalized Cut method had some false separation as in the seats and in the shelves. Graph Cuts segmented the seats into two classes and segmented the shelf into a mess.

From the results shown in Figure 5, it can be seen that the proposed method generates sub-scenes that are more suitable for complex scene than traditional segmentation. The traditional segmentation methods are more focus on segmenting each part of an object instead of treating it as one. Obviously, they fail to treat a group of objects as a whole sub-scene as well and the clutters and illumination problem greatly influence their behavior. On the other hand, the proposed method could achieve very promising result of sub-scenes through relatively simple steps unsupervised thus paves the way towards complex scene understanding.

5. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a new context-aware concept of ‘sub-scene’ and developed a method to generate them from a complex scene. The sub-scene concept is based on human psychology towards scene analysis, especially when confronting with a complex one. It groups the compact semantic conjunctions together and explores the spatial relationship among them. Based on the definition and properties of sub-scene, the proposed method searches the semantic connection to generate sub-scenes. It does not need heavy human labeling step because it does not include a training process. It also is able to handle complex scenes with clutters and occlusions. In our experiment, the proposed method was evaluated using the database of indoor scenes which are very challenging and complex enough that performances of general segmentation methods drop dramatically. Sub-scene generation is a concrete step towards scene understanding in the future. They reveal organization and contextual information of the scene and used to produce logical relationship description of the scene.

REFERENCES

[1] B. Siddiquie and A. Gupta, “Beyond active noun tagging: Modeling contextual interactions for multi-class active learning,” *IEEE Comput. Soc. Conf. on Comput. Vision and Pattern Recogn.,* pp. 2979-2986, 2010.

[2] G. Wang, D. Hoiem and D. Forsyth, “Building text features for object image classification,” *IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.,* pp.1367-1374, 2009.

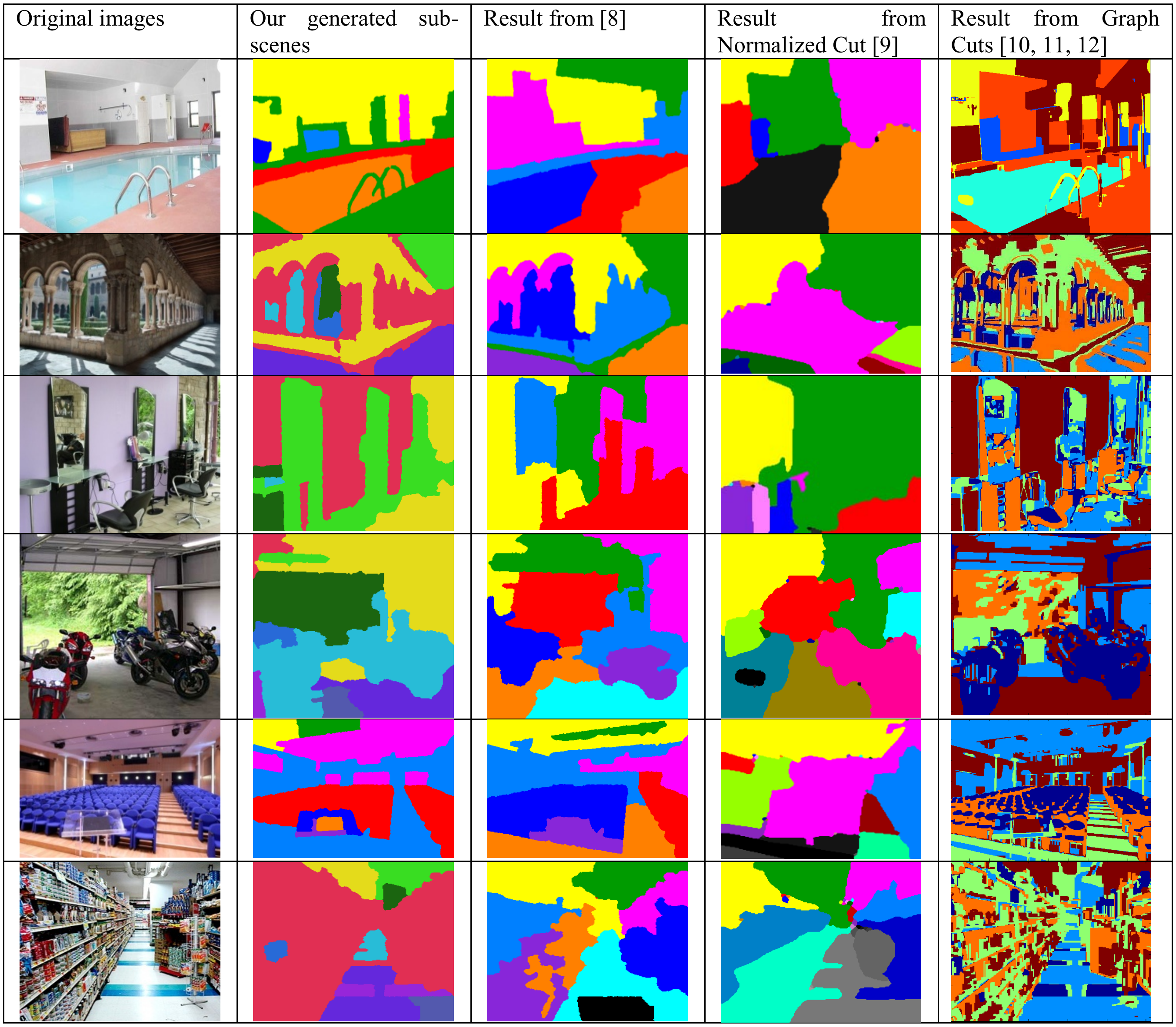

Original images
Our generated sub-scenes
Result from [8]
Result from Normalized Cut [9]
Result from Graph Cuts [10, 11, 12]

Figure5: Results of our method and comparison with [8], Normalized Cut [9] and Graph Cuts [10, 11, 12].

[3] D. G. Lowe, “Object recognition from local scale-invariant features,” *IEEE* I*nternational Conf. on Computer Vision*, vol. 2, pp.1150-1157, 1999.

[4] J. Qin and N. H. C. Yung, “Scene categorization via contextual visual words,” *Pattern Recognition*, vol. 43, Issue 5, pp.1874-1888, May 2010.

[5] A. Quattoni and A. Torralba, “Recognizing indoor scenes,” *IEEE Conf. on Computer Vision and Pattern Recognition*, pp.413-420, 2009.

[6] D. Parikh, C. L. Zitnick, and T. Chen, “Unsupervised learning of hierarchical spatial structures in images,” *IEEE Conf. on Computer Vision and Pattern Recogn.*, pp.2743-2750, 2009.

[7] R. Kowalski and D. Westen, Psychology, 5th edition, USA: John Wiley & Sons, pp.135-136, 2009.

[8] Z. G. Tan and N. H. C. Yung, “Image segmentation towards natural clusters,” *IEEE* I*nternational Conf. on Computer Vision*, pp. 1-4, 2008.

[9] S. Jianbo and M. Jitendra, “Normalized cuts and image segmentation,” *IEEE Comput. Soc. Conf. on Computer Vision and Pattern Recogn.*, pp.731-737, 1997.

[10] Y. Boykov, O. Veksler and R. Zabih, “Efficient approximate energy minimization via Graph Cuts,” *IEEE Trans. PAMI*, pp. 1222-1239, 2001.

[11] V. Kolmogorov and R.Zabih, “What energy functions can be minimized via Graph Cuts?” *IEEE Trans. PAMI*, pp.147-159, 2002.

[12] Y. Boykov and V. Kolmogorov, “An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision,” *IEEE Trans. PAMI*, pp. 1124-1137, 2004.