J. Vis. Commun. Image R. 25 (2014) 994–1005

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Sub-scene segmentation using constraints based on Gestalt principles

Shan-shan Zhu *, Nelson H.C. Yung

Laboratory for Intelligent Transportation Systems Research, Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong, China

ARTICLE INFO

Article history:
Received 2 July 2013
Accepted 23 February 2014
Available online 6 March 2014

Keywords:
Unsupervised image segmentation
Semantic gap
Gestalt principles
Proximity grouping
Pattern clustering
Sub-scene segmentation
Sub-scene property
Mahalanobis distance

ABSTRACT

In this paper, an unsupervised sub-scene segmentation method is proposed. It emphasizes on generating more integrated and semantically consistent regions instead of homogeneous but detailed over-segmented regions usually produced by conventional segmentation methods. Several properties of sub-scenes are explored such as proximity grouping, area of influence, similarity and harmony based on psychological principles. These properties are formulated into constraints that are used directly in the proposed sub-scene segmentation. A self-determined approach is conducted to get the optimal segmentation result based on the characteristics of each image in an unsupervised manner. The proposed method is evaluated over three datasets. For quantitative evaluation, the performance of the proposed method is on par with state-of-the-art unsupervised segmentation methods; for qualitative evaluation, the proposed method handles various sub-scenes well, and produces neater results. The sub-scenes segmented by the proposed method are generally consistent with natural scene categories.

1. Introduction

Image segmentation aims to partition an image into non-overlapping homogeneous regions and is fundamental for all kinds of image processing and computer vision applications such as object and saliency detection [1,2], semantic annotation [3,4], event detection [5], and hierachical scene understanding [6].

Despite years of research, image segmentation remains a very challenging problem because it is inherently an ill-posed and ambiguous problem [7]. There are various possibilities to perceive and segment an image because people have different preferences. Besides, the "correct" segmentation may be different according to different visual tasks. To address the problem of ambiguity in segmentation, Arbelaez et al. proposed to collect human labeled boundaries as ground truth and perform the segmentation in a supervised manner [8]. Following this trend, the supervised methods usually emphasize on estimating the boundary probabilities rather than achieving integrated regions [9,10]. The drawback is that the boundary may be in a discontinuity state and the disjointed edges affect visual perception when closed contours are preferred [11].

It is of course more challenging and demanding in unsupervised image segmentation. As the general purpose of unsupervised image segmentation is to derive segments which are suitable for human perception, relying on human perceptual rules from psychology is inevitably one of the major directions. Perceptual rules have been carefully studied and are used in many unsupervised segmentation methods [7,12–16]. The most widely used is the Gestalt principles [17]. Gestalt is a psychology term that means unified whole. It refers to the theory which describes how people tend to group visual elements when certain principles are fulfilled. It concludes principles such as continuity, closure, similarity and proximity. However, there are difficulties to quantize them in mathematics since these principles are abstract psychology concepts. Actually, only a few principles such as similarity and proximity are used in literature, and they are interpreted in a simplified way. According to the similarity principle, regions with the most similar appearances are considered to be merged [7,12–16]. In realizing of the proximity principle, only neighboring regions are actually merged [6,16,18]. It is necessary to accomplish the perceptual rules more deeply to further improve image segmentations.

Besides, it has long been identified that there is a "semantic gap" between the segmented patches and the semantic entities that can be readily used. Both Malisiewicz and Efros [1] and Jianping et al. [6] stated that homogenous segmented patches may not correspond to physical objects in the real world. The fundamental reason for this semantic gap roots in the limitation of current objectives of image segmentation which focuses on detecting precise boundaries [10] and producing homogenous regions.

* Corresponding author.
E-mail addresses: sszhu@eee.hku.hk (S.-s. Zhu), nyung@eee.hku.hk (N.H.C. Yung).

http://dx.doi.org/10.1016/j.jvcir.2014.02.017
1047-3203/© 2014 Elsevier Inc. All rights reserved.

Therefore, any slight change in the image is captured and objects that are segmented into several parts are acceptable. However, these parts are needed to be integrated together to meet human expectations. Use global image context or apply corresponding perceptual rules to piece together the segmented parts and form semantically consistent regions is necessary [4,6].

In this paper, an unsupervised sub-scene segmentation method is proposed to narrow the semantic gap. The notion of the sub-scene is intuitively derived from human perception towards a scene. When a person sees a scene, he may partition the scene into several sub-scenes, where the sub-scene fulfills certain "function" and the meaning of the entire scene is probably derived by combining the functions of the sub-scenes. The notion of sub-scene used in this paper may appear to be similar to semantic segmentation such as [19,20]. However, the major difference is that the sub-scene here is not confined to fixed categories set beforehand and it does not need to go through a training step neither. Several perceptual rules are explored based on human psychology such as proximity grouping, area of influence by objects and harmony, and they are transformed into constraints which can be applied to low level features. With a self-determined retrieval approach, sub-scenes can be generated automatically. The contributions of the proposed method are:

1. Proximity grouping is formulated more appropriately using influence areas instead of being restricted to neighboring pairs;
2. Balancing between proximity grouping and similarity grouping is achieved by a self-determined optimal retrieval strategy; and
3. The unimportant details are ignored and a more integrated segmentation result is achieved.

The paper is organized as follows. In Section 2, the proposed method is presented in details. Section 3 describes the experiments on three datasets, where each dataset emphasizes a different aspect of scenes. Comparison and discussion are given for each one of them. Section 4 concludes the paper.

2. Proposed method

2.1. Problem formulation

Consider I as a given image, one way to partition the image into M regions is $\Gamma_M(I) = \left\{R_M^i\right\},\ i = 1,\ldots,M$, where R_M^i represents region i. The common split-and-merge approach towards image segmentation is to first generate a number of superpixels and then gradually merge them until a stop criterion is satisfied; or complete the merging steps to the end as $\Gamma_M(I) \rightarrow \Gamma_{M-1}(I) \rightarrow \cdots \rightarrow \Gamma_1(I)$ and then select the optimal segmentation $\Gamma^*(I)$ from the entire process. The optimal $\Gamma^*(I)$ is the partition that minimizes the cost function which consists of two parts:

$$E(\Gamma_M) = \frac{1}{M}\sum_{i=1}^{M} J\left(R_M^i\right) - \lambda\frac{2}{M(M-1)}\sum_{i=1}^{M}\sum_{j=i+1}^{M}\delta\left(R_M^i, R_M^j\right). \tag{1}$$

The term $J\left(R_M^i\right)$ represents the dissimilarity within each region and the term $\delta\left(R_M^i, R_M^j\right)$ represents the dissimilarity between regions. The cost function is based on the similarity principle. Minimizing it can produce segments with large dissimilarity between regions but homogeneous within each region. However, the two terms in (1) are contradictory to each other during region merging. Merging most similar regions can generally increase the global inter-region dissimilarity but may increase the intra-region dissimilarity. As a result, λ is introduced to balance the tradeoff.

Similarly, the sub-scene segmentation which also models the semantic interactions minimizes a cost function as:

$$E(\Gamma_M) = \frac{1}{M}\sum_{i=1}^{M} K\left(R_M^i\right) - \lambda\frac{2}{M(M-1)}C(\mathbf{R_M}), \tag{2}$$

in which

$$K\left(R_M^i\right) = J\left(R_M^i\right) + \sum_{p\in R_M^i} S\left(R_M^i, p\right), \tag{3}$$

$$C(\mathbf{R_M}) = \sum_{i=1}^{M}\Delta\left(R_M^i, \mathbf{R_M^{-i}}\right). \tag{4}$$

The term $K\left(R_M^i\right)$ is the improved intra-region dissimilarity evaluation which combines the color and spatial interactions. Denote p as a pixel, the term $S\left(R_M^i, p\right)$ is to measure the spatial interaction between the pixel and the sub-scene region. Similarly, a high order term $C(\mathbf{R_M})$ is to evaluate the inter-region dissimilarity among all sub-scene regions and $\Delta\left(R_M^i, \mathbf{R_M^{-i}}\right)$ is to evaluate the influences. The sub-scene properties such as proximity grouping and area of influence can be incorporated. The expression of each term and the sub-scene properties will be studied in following sections. Before that, a brief comparison of sub-scenes and image segmentation is given.

2.2. Difference between sub-scenes and image segmentation

The major difference between sub-scene segmentation and common image segmentation is that the former attempts to bring semantic constrains into consideration and generate meaningful entities instead of detailed but broken components. Take Fig. 1 as an example. The proposed method is compared with Mean-shift [16], one of the traditional image segmentation method. It can be seen that details such as the texture of the trees, rocks on the road, the changes of the grassland and the different parts of the building are captured by Mean-shift. However, some of these segmented regions can be considered under the same category to give a better integrated semantic meaning. For instance, the proposed method generates neater sub-scenes that represent sky, grassland, trees, roads, buildings, sea and boat.

Section 2.3 presents the concept of the sub-scene. Following that, sub-scene properties based on psychological perspective are described in Section 2.4. Section 2.5 presents the constraints derived from sub-scene properties which are applied to the proposed framework. Section 2.6 discusses the detailed realization of the proposed method.

2.3. Concept of the sub-scene

The concept of a sub-scene comes from the psychological perspective when human confronts a scene [21]. Typically, a scene is visualized as a composition of several functional sub-scenes related to each other according to some visual rules, of which details are neglected initially. Then general understanding of the scene is achieved both by recognizing the function of each sub-scene and by interpreting their relationships. In this regard, some sub-scenes are inherently less distinct while other sub-scenes may play a more prominent role in determining the scene category. Fig. 2 depicts the sub-scene of chairs in different scenes such as living room, classroom, auditorium and so forth. As can be seen, the display of chairs plays a dominant role in distinguishing indoor scenes such as auditorium and living room. The scale and composition of a sub-scene together with its relationship with other sub-scenes are (1) more inspiring than mixing them together and treating the scene as a whole, which is the approach taken by most scene categorization methods, and (2) better to be understand than they are