2004

Visual perceptual process model and object segmentation

Wanqing Li
University of Wollongong, wanqing@uow.edu.au

P. Ogunbona
University of Wollongong, philipo@uow.edu.au

Lei Ye
University of Wollongong, lei@uow.edu.au

Igor Kharitonenko
University of Wollongong, igor@uow.edu.au

Visual perceptual process model and object segmentation

Abstract

Modeling human visual process is crucial for automatic object segmentation that is able to produce consistent results to human perception. Based on the latest understanding of how human performs the task of extracting objects from images, we proposed a graph-based computational framework to model the visual process. The model supports the hierarchical nature of human visual perception and consists of the key steps of human visual perception including pre-attentive (pre-constancy) grouping, figure-and-ground organization, and attentive (post-constancy) grouping. A divide-and-conquer implementation of the model based on the concept of shortest spanning tree (SST) has demonstrated the potential of the model for object segmentation.

Disciplines

Physical Sciences and Mathematics

Publication Details

This paper originally appeared as: Li, W, Ogunbona, Lei, Y et al, Visual perceptual process model and object segmentation, Proceedings. 7th International Conference on Signal Processing, 31 August - 4 September 2004, vol 1, 753-756.

VISUAL PERCEPTUAL PROCESS MODEL AND OBJECT SEGMENTATION

Wanqing Li, Philip O. Ogunbona, Lei Ye and Igor Kharitonenko

School of Information Technology and Computer Science
University of Wollongong, Australia
{wanqing,philip_ogunbona, lei, igor}@uow.edu.au

ABSTRACT

Modeling human visual process is crucial for automatic object segmentation that is able to produce consistent results to human perception. Based on the latest understanding of how human performs the task of extracing objects from images, we proposed a graph-based computational framework to model the visual process. The model supports the hierarchical nature of human visual perception and consists of the key steps of human visual perception including pre-attentive (pre-constancy) grouping, figure-and-ground organization, and attentive (post-constancy) grouping. A divide-and-conquer implementation of the model based on the concept of shortest spanning tree (SST) has demonstrated the potential of the model for object segmentation

1. INTRODUCTION

The extraction of semantically meaningful objects from single image or a sequence of images has recently become an active research topic in multimedia signal processing [1, 2, 3]. The algorithms developed so far for object segmentation, in general, fall into in three categories, each category in fact adopts different definition of "object". The first one, known as moving object segmentation from a sequence of images [2, 3], defines the object as a group of pixels moving in a same or coherent manner. Motion is the key feature used for grouping pixels that belong to the same object.

The second category an extension of traditional region or edge-based segmentation with some heuristic about the formation of the objects [1] or with some extra information such as depth or range [3]. The region or edge based segmentation usually serves as low-level processing and a rule based system is applied to group the segmented regions into objects. Objects are often defined in this case by their geometric or chromatic formation.

The third category employs visual perception theory established by gestalt psychologists and cognitive scientists [12, 13], especially, the principles of perceptual grouping [13]. Research focus in this category has been in the past on how to quantize the qualitative rules of perceptual grouping and how to incorporate these rules into the traditional segmentation scheme [4, 5, 6]. Little literature was published about how to model the entire visual perceptual process that consists of three major stages: pre-attentive grouping, figure-ground organization and attentive grouping.

Cognitive study [12] has shown that human tend to divide the world into coherent visual units and the process involved is more than physics, i.e. signal processing, that traditional image processing techniques are most capable of. To automatically segment an image with results consistent to human perception, it appears indispensable to model the visual perceptual process. This paper presents a graph-based approach to modeling the human visual perception.

The paper is organized as follows. Section 2 presents a feedback theory on the visual perception process based on the recent research outcome in cognition science. Section 3 proposes a graph based computational framework to model the visual process. Dynamic grouping are introduced to form a hierarchical representation. In Section 4, we present a divide-and-conquer implementation of the model based on the concept of Shortest Spanning Tree (SST) with preliminary results. The paper is concluded with some discussion and further work.

2. HUMAN VISUAL PERCEPTION MODEL

The perception of the world around us entails the segmentation or grouping into identifiable objects or coherent visual units. The grouping problem is well known to be part of the generic figure-ground separation problem. It is useful to consider the object segmentation task as a stream of information processing in which the input is the visual scene and the expected output is the identified objects within the scene. Through a process of perceptual organization humans utilize the information presented in the visual stimuli to separate and eventually aggregate features into objects. It is held that the perceptual process proceeds in two stages - preattentive and attentive. The preattentive process is parallel in nature, operating across the visual field and pertains to the perception of those elementary features that do not require attentional resources. In fact visual attributes including orientation, colour and size differences are supposed to be perceived pre-attentively [8, 9]. The second stage, attentive process, feeds on the output of the preattentive stage. There has also been experimental evidence challenging the validity of a sequential two-stage process. In [10], it was demonstrated that there are some "'preattentive' information that cannot be overtly perceived without attention"'. Perhaps, the two process operate iteratively and there is a third process in between, wherein the figure-ground ambiguity is resolved.

The perceptual organization theories fall into two camps namely, sequential and interactive. The sequential theorists [7] postulate that segmentation is a process of grouping low-level features into regions which are further grouped into foreground and background based on higher-level cues [11]. The possibility of interactivity among high-level object knowledge, intermediate figure-ground cues and low-level features is suggested by the interactive theorists (see Figure 1).

Both theories have inspired some of the well known segmentation algorithms. The sequential theory has guided the bottom-up approach to segmentation in which low-level features are initially grouped together based on constraints such as similarity and con-

0-7803-8406-7/04/$20.00 © 2004 IEEE.

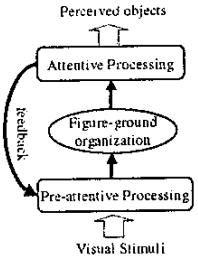

Fig. 1. Three stages of human visual process: pre-attentive processing, figure-ground organization and attentive processing

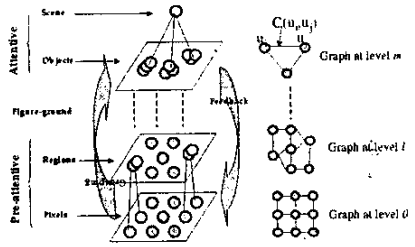

Fig. 2. A graph-based model of visual perceptual process

tinuity. Further grouping into perceptual units is then based on higher-level knowledge which often include context information. This raises the question of "What is a visual object?" [12]. Unfortunately, there is no succinct definition that suits all situations. This inability to provide a generic working definition of an object becomes problematic when a quantitative model is to be developed for computational purposes. Simply using one grouping cue does not define an object. A hierarchy of grouping cues suggests a hierarchical perceptual organization model in which there is an interplay, at various stages in the process, of the various cues. In [11, 12], a hierarchical approach that lends itself to graph-theoretic representation was adopted. In this paper we follow [12] and propose a graph-based model of perceptual organization. Our model differs in the sense that we employ all the grouping cues throughout the perceptual grouping process. The weight associated with each cue determines its relative importance at different stages of the grouping process. This model defers commitment to a given interpretation of the scene until the influence of all available information (grouping cues) has been selectively utilized.

3. A GRAPH BASED MODEL

The three stages, pre-attentive grouping, figure-and-ground organization and attentive grouping, in the visual perceptual process and the hierarchical nature of the process presents the minimum requirements of modeling the process. First, the model has to be able to efficient represent the visual units and their relationship or coherency at different spatial scale. As a result of this multiple scale representation, the hierarchical decomposition of visual context can be guaranteed. Secondly, the grouping (Gestalt) rules can be quantified under the representation and easily recalculated at difference scale. As seen in [13], some rules are more applicable to edges than regions and others are vice versa. Thirdly, grouping inference is applied to visual units at different scale levels such that a hierarchical decomposition can be achieved.

3.1. Graph representation

Figure 1 shows a graph-based model of the visual process. According to the visual process described in Section 2, pixels are grouped into regions in the pre-attentive stage. These regions are then classified into foreground (figure) and background (ground). Foreground regions are further grouped into objects or coherent visual units in the attentive stage. The grouping from pre-attentive to attentive forms a hierarchical decomposition of the image. At any given scale level in the hierarchy, we adopt a weighted non-directional graph to represent its structure.

A weighted non-directional graph $G(V, E)$ is composed of a set of nodes, V, and a set of links, E, that bound the nodes together. Each link, e, has an attribute or weight associated with it. Nodes and links in the graph represent respectively visual units and the connectivity or spatial relationship among the visual units. The weight of a link is a measurement of the coherency of the two units connected by the link.

3.2. Dynamic coherency

The coherency of two visual units is reflected by the weight of the link that connected the two units. It is calculated based on the grouping rules [13, 4, 5, 6].

Let $R = R_i, i = 1, 2, 3, \ldots, N$ be a set of grouping rules. $U = u_j, j = 1, 2, 3, \ldots, M$ be the set of visual units at a particular level in the hierarchy, and $r_i(u_j, u_k)$ be the coherency between node u_j and u_k that is calculated based on rule R_i, where $r_i \epsilon [0, 1]$. Then the total coherency between unit u_i and u_j is defined as the fusion of all $r_i, i = 1, 2, 3, \ldots, N$

$$C(u_i, u_j) = \Im(r_i), r_i = 1, 2, 3, \ldots, N \qquad (1)$$

where $\Im(*)$ is a fusion operation. A typical fusion technique is Dempster's rule of combination [14]. As described in Section 2, the contribution of each grouping rule varies in the entire visual process. For instance, color and intensity similarity contribute significantly at the low-level of the hierarchy where pre-attentive grouping happens, while, on the other hand, geometric similarity and co-linearity are much more important in the attentive grouping than pre-attentive. In the figure-and-ground resolution stage, similarity in depth and motion becomes influential. To deal with this dynamic nature, we introduce a weighting factor to each rule and this weighting factor changes dynamically from the low levels to the high levels of the hierarchy. Therefore, the corehence function becomes

$$C(u_i, u_j) = \Im(r_i, w_i), r_i = 1, 2, 3, \ldots, N \qquad (2)$$

where w_i is the weighting function for rule R_i. Theoretically, each rule would have its own weighting function, i.e. there are

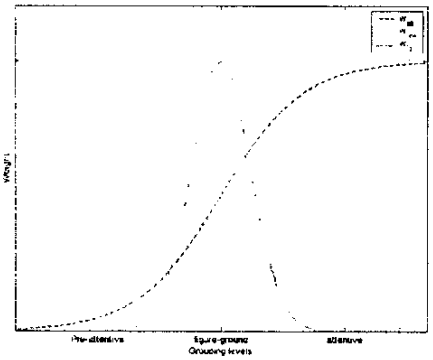

Fig. 3. Weighting functions for dynamic coherency

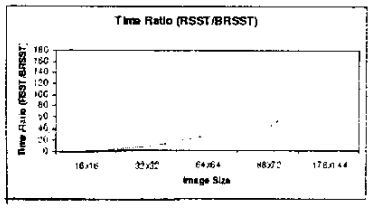

Fig. 4. Comparison of machine time used by RSST and BRSST on the same machine with the same code. RSST was treated as a special case where there is only one block for the entire image.

N weighting functions, $w_i, i = 1, 2, 3, \ldots, N$. However, rules can be divided into three categories: 1) rules that contribute significantly in pre-attentive grouping, 2) rules that are significant in figure-and-ground resolution, and 3) rules that are important for attentive grouping. Respectively, three weighting functions, w_{pre}, w_{fg} and w_{att}, as shown in Figure 3, are defined.

3.3. Computational process

The hierarchy is generated by iteratively grouping visual units. At the lowest level, each pixel is a visual unit. Two units are grouped into one unit if the coherency between them is high. When a new unit formed, the coherencies between the new unit and other units are calculated. The process stops when the entire image becomes one visual unit, the root of the hierarchy. Figure 2 demonstrates the grouping process. The hierarchy forms a full description of the scene. It can be cut into a fixed number of visual units or at the link with coherency less than a specific threshold. These units are expected to form a decomposition of the scene that is coherent with human perception.

For a given image, the computational process can be described follows

Computational Process:

Step 1: Create a graph where each pixel is represented as a node (visual unit) and a link is established between two visual units if they are spatially connected

Step 2: Calculate the coherency of any linked visual units

Step 3: Group the most coherent two visual units to form a new visual unit. Replace the two units with the newly formed one.

Step 4: Recalculate the coherencies between the new unit and its spatially connected ones.

Step 5: If there is more than one unit left in the graph, repeat Step 3 and 4. Stop otherwise.

4. IMPLEMENTATION

The process described in the previous section is equivalent to Recursive Shortest Spanning Tree (RSST) [18]. Its implementation needs to sort the coherencies for every selection of the most coherent two units to be grouped. It is a very high computationally demanding algorithm and its complexity is analyzed in [15]. Kwok and Constantinides devised a fast RSST algorithm [16] and proposed a parallel implementation of the algorithm. The authors proposed a divide-and-conquer approach [17], called block based RSST (BRSST) that suits both sequential and parallel implementation. In the approach, image is first divided into blocks. RSST is applied to each block to segmented the block into number of regions. Regions of all blocks are merged using RSST into the RSST of the original image. It is expected that the BRSST is equivalent to RSST as long as grouping of each block does not go beyond pre-attentive grouping. Figure 4 shows the comparison of the time spent on conventional RSST and BRSST. As the size of the image increases, BRSST performs substantially over RSST. BRSST is over 100 times faster than RSST for a QCIF image.

Figure 5 shows the results of RSSST (first row) and BRSST respectively where the following rules were employed

- Uniform connectedness and similarity in luminance
- Element connectedness
- Common region and relative size

The rules were applied uniformly across the visual process and no rules were applied for figure-and-ground selection. In other words, w_{pre} and w_{att} were set to 1 and w_{fg} was set to 0. The image was segmented into 9 visual units.

5. DISCUSSION

We presented in this paper a theory of visual process based on the latest study in cognitive science and Gestalt psychology. The theory describes the three major grouping stages in the visual perception: pre-attentive, figure-and-ground and pattentive. The hierarchical nature of the visual grouping makes the graph as a natural choice. Quite different from existing graph-based segmentation techniques where features are used in a non-discriminating way throughout the levels in the hierarchy, we introduce weighting functions to address the fact that grouping rules should contribute differently to the grouping at various levels of the entire process. The weighting functions not only unify the three grouping stages into one framework, i.e. applying the same grouping strategy from

Fig. 5. Segmentation results by RSST (first row) and BRSST. The image was segmented into 9 units

lower levels to higher levels, but also provide a smooth and continuous transition between the stages. In this way, better coherent visual units will be achieved.

In the future, the model will be extended in two aspects. First, a mechanism will be added to allow feedback from higher levels to lower levels, which would enable a resolution of any ambiguity identified at the higher levels. Secondly, a concept of virtual connection will be introduced to allow visual units that are not spatially connected in the image to be grouped together. This virtual connection will enable the model to identify occlusion.

We have implemented the core part of the model and achieved promising preliminary results. The model shall be tested on a large set of images and results will be reported in the future.

6. REFERENCES

[1] S. Antania, R. Kasturi and R. Jainb "A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video", *Pattern Recognition*, Vol.35, Issue 4, 2002, pp.945-965.

[2] G. L. Foresti, "Object recognition and tracking for remote video surveillance"', *IEEE Trans Circuits and Systems for Video Technology*, vol.9, no.7, 1999, pp.1045-1062.

[3] J. Wang and S. Singh, "Video analysis of human dynamicsa survey", *Real-time imaging*, Vol.9, Issue 5,2003, pp.321-346 .

[4] J. Randall, L. Guan, X. Zhang and W. Li, "Hierarchical cluster model for perceptual image processing"', *ICASSP'02*, Orlando, Florida, May 13 - 17, 2002, vol. 1, pp.1041-1044.

[5] J. Randall, L. Guan, X. Zhang and W. Li, "Hierarchical cluster model for image segmentation", *ICASSP'04*, (accepted)

[6] T. Tamaki, T. Yamamura and N. Ohnishi, "'Image segmentation and object extraction based on geometric features of regions", *SPIE conference on visual Communications and Image Processing*, SPIE vol.3653, 1999, pp.937-954.

[7] D. Marr, *Vision: a computational investigation into the human representation and processing of visual information.* San Francisco, CA; W.H. Freeman.1982.

[8] A.M. Triesman, "Preattentive processing in vision," *Computer Vision, Graphics and Image Processing*, Vol. 31, pp 156-177, 1985.

[9] B. Julesz and J. R. Bergen, "Textons, the fundamental elements in preattentive vision and perception of textures," *Bell System Technical Journal*, vol 62, pp 1619-1645, 1983.

[10] J. S. Joseph, M. M. Chun and K. Nakayama, "Attentional requiremnets in a 'preattentive' feature search task", Nature, Vol. 387, pp 305-307, 1997.

[11] S. X. Yu, "Computational models of perceptual organization," Technical Report CMU-RI-TR-03-14, Carnegie Mellon University, May 2003.

[12] J. Feldman, "What is a visual object?", *TRENDS in Cognitive Sciences*, Vol. 7, No.6, pp 252-256, June 2003.

[13] S. E. Palmer, "Perceptual grouping: It's later than you think"' *Current Directions in Psychological Science*, vol.11, 2002, pp.101-106.

[14] G. Shafer, *A mathematical theory of evidence*,Princeton University Press, 1975

[15] S. H. Kwok and A. G. Constantinides, "A fast recursive shortest spanning tree for image segmentation and edge detection"', *IEEE Trans Image Processing*, vol.6, no.3, 1997, pp.328-332

[16] S. H. Kwok and A. G. Constantinides, "A parallel recursive shortest spanning tree algorithms for image segmentation in distributed computing environment", *Journal of Parallel and Distributed Computing*, vol.56, 1999, pp.181-207

[17] W. Li, P. Ogunbona, X. Zhang and J. Zhang, "Block-Based Image Segmentation Method and System", *US Patent Application US20020181771 A1*, Motorola Inc.

[18] O. J. Morris, M. de J. Lee and A. G. Constantinides, "Graph theory for image analysis: an approach based on the shortest spanning tree", *IEE Proceedings*, vol.133, no.2, 1986, pp.146-152.