OBJECT SEGMENTATION BASED ON COMMON INFORMATION BETWEEN IMAGES

Wei-Ta Chu[1] (朱威達), *Chien-Ta Hung*[1] (洪建達), *Jen-Yu Yu*[2] (游人諭)

[1]Dept. of Computer Science and Information Engineering, National Chung Cheng University
wtchu@cs.ccu.edu.tw, hct96m@cs.ccu.edu.tw

[2]Information and Communication Research Labs, Industrial Technology Research Instutite
KevinYu@itri.org.tw

ABSTRACT

We exploit common information between images to construct data models and background models, and accordingly segment major objects in images without human intervention. This method can be applied to images that consist of same foreground objects in varied backgrounds, such as a person dressing the same in different scenes, or a major object appearing with different backgrounds. Experimental results show the effectiveness of the automatic segmentation method, and we provide discussion about the influence of common information in object segmentation.

1. INTRODUCTION

Object segmentation in images has been an age-old problem in computer vision and image processing communities. One of the challenges in object segmentation is the semantic gap between visual features and human perception. An object may contain connected pieces that have different visual appearance. Therefore, the methods that cluster pieces with homogenous color just segment an image into various regions rather than meaningful objects. Nonetheless, more than a dozen of studies have been proposed to conduct image segmentation, such as the ones based on the mean shift algorithm [1] and the ones based on graph cut [5][6][7].

Because automatic image segmentation is a notorious problem that hasn't been solved for several decades, some researchers turn to develop friendly tools to facilitate efficient manual segmentation. Li et al. [2] developed a tool such that users can draw a few strokes to roughly indicate foreground and background, and then the system segments an image into foreground part and background part. Wang et al. [3] further improved this work to make the tool more reliable.

Recently, automatic segmentation based on common information between images [8] or automatic transduction based on a manual segment result [9] draw contiguous attention. Gallagher and Chen [8] exploit the graph cut framework to segment clothes regions from background and parts of torso that are not covered by clothes. Based on the fact that the same person wears the same in a short period, their system automatically constructs a clothes model and a background model from a set of the same individual's photos. Costs about observed data and spatial discontinuity are respectively calculated to be fed into the graph cut framework. In contrast to automatic model construction, Cui et al. [9] started from a manual segmentation result, and therefore construct object models based on a more accurate foundation. Given a photo that contain similar object that has been manually segmented before, their system propagates segmentation effect to the new photo.

In this work, we conduct automatic object segmentation based the graph cut framework, which can be formulated as an energy minimization problem. We investigate how to automatically derive costs of observed data and spatial discontinuity that are specific to different applications. This work is developed under the idea proposed in [8]; however, we try to extend its feasibility in terms of more effective features and general object segmentation. The goal of this work is to utilize common information between images to achieve meaningful object segmentation.

The contributions of this work are summarized as follows.

- Using common information from a set of images to construct foreground/background models, and accordingly derive data cost and spatial discontinuity cost.
- More effective features than that used in [8].
- Extending the method that is originally designed for clothes segmentation to general object segmentation, under some constraints.

The rest of this paper is organized as follows. Section 2 describes the kernel component of this work, i.e., the graph cut framework. We describe object model construction, which is specially designed with the consideration of common information. Two applications, i.e., clothes segmentation and main object segmentation, will be stated in Sections 3 and 4, respectively. Discussions about model construction and limitation of this work are given in Section 5, and Section 6 concludes this paper.

2. GRAPH-CUT FRAMEWORK

Segmenting foreground from background or segmenting main object from others can be viewed as a binary labeling problem. From this viewpoint, many studies have been proposed to transform an image into a graph, and the targeted problem is transformed to labeling nodes of this graph. Finding the optimal solution of this labeling problem can be formulated in terms of energy minimization.

The goal is to find a labeling f that assigns each pixel $p \in \mathcal{P}$ a label $f_p \in \mathcal{L}$ [5], where the labeling is considered to be consistent with the observed data and should conform to smoothness of neighborhood. We consider energies of the form

$$E(f) = \sum_{p\in\mathcal{P}} D_p(f_p) + \sum_{\{p,q\}\in\mathcal{N}} V_{p,q}(f_p, f_q), \quad (1)$$

where $D_p(f_p)$ denotes the cost of assigning the pixel p the label f_p, $f_p \in \{0,1\}$. The label $f_p = 1$ means the pixel p is labeled as a foreground pixel, and $f_p = 0$ otherwise. The value $D_p(f_p)$ is called *data cost* in this paper. The notation $\mathcal{N}$ is the set of pairs of pixels that are spatially adjacent to each other. The value $V_{p,q}(f_p, f_q)$ denotes the penalty of assigning adjacent pixels p and q the labels f_p and f_q, respectively. The value $V_{p,q}(f_p, f_q)$ is called *smoothness cost* in this paper. Finding the optimal solution of this formulation is NP-hard. Fortunately, after decades of studies, fast approximate algorithms have been developed [5]. Therefore, we can efficiently find satisfactory solutions to conduct binary labeling.

Design of data cost and smoothness cost is the most central work in the energy minimization framework. Generally, if the pixel p is more similar to a foreground pixel, the data cost $D_p(f_p = 1)$ is smaller, and the cost $D_p(f_p = 0)$ is larger. For smoothness cost, more similar p and q are, higher penalty is assigned if p and q are assigned different labels, i.e., $V_{p,q}(f_p, f_q)$ is higher if $f_p \neq f_q$.

With the general guidelines described above, we have to develop appropriate data model and smoothness model for our targeted applications. In this work, we focus on automatically segmenting main objects from a set of images without human intervention. Two applications, i.e., clothes segmentation and main object segmentation, are developed.

3. CLOTHES SEGMENTATION

Figure 1 shows the flowchart of utilizing the graph cut algorithm to conduct clothes segmentation. We collect training data consisting of the same individual's images, in which the individual dresses the same. Therefore, the foreground object (the individual and his/her clothes) keeps the same and the background changes. In this work, training and test images are normalized into 180 × 240 or 240 × 180 without exception.

The mask determination module explores the common information between images, i.e., the regions that are similar in different images, to determine regions that roughly cover the foreground object. From the masked regions, data characteristics of the pixels in the regions are collected to build a data model, which describes how likely a pixel belongs to the foreground object. On the other hand, characteristics of pixels not in the masked regions are collected to build the background model.

Given a test image, assuming that there is only one face in it without loss of generality, we first detect the face and accordingly expand a region that covers the face and his/her upper body. After resizing the cropped image, visual features are extracted from regions in the image and are compared with the data model and background model to calculate data cost and smoothness cost. With the energy function specially designed for clothes segmentation, the optimal labeling solution is found based on the graph cut framework, and finally the foreground object (clothes) is determined. Details of important modules in this framework are described in the following subsections.

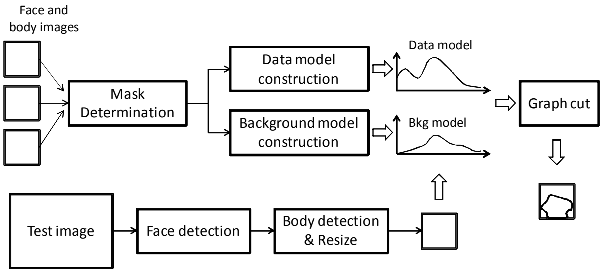

Figure 1. Flowchart of clothes segmentation based on common context information.

3.1. Mask Determination

Given a set of images in which the same individual dresses the same but would have different poses and lighting, we first adopt the normalized cut algorithm [4] to segment each image into regions. Each region consists of connected pixels with similar color features, and we call such kind of region a *superpixel* [8] in the following. Superpixels are the basic processing units in this work. The ultimate goal of this work is to assign the most appropriate label to each superpixel in the test image.

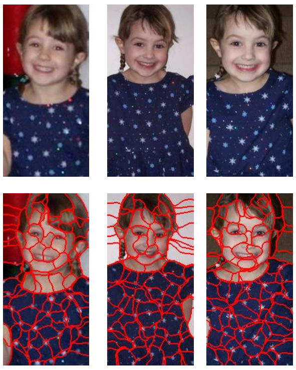

Figure 2. Example results of normalized cut.

Figure 2 shows some results of normalized cut. Parts of images in Figure 2 and in experiments are downloaded from the open photo collection [10].

To characterize each superpixel, we extract HSV (hue, saturation, value [intensity]) color histogram and edge histogram of each superpixel. In the HSV histogram, there are 16 bins for hue, 4 bins for saturation, and 4 bins for value. After detecting gradient of each pixel in a superpixel, an edge histogram is constructed on the basis of five bins, which roughly indicate orientation corresponding to $0°$, $45°$, $90°$, $135°$, and $180°$. Comparing with the features used in [8], we extract more elaborate edge features, and represent pixels in HSV color space.

Given a pair of training images p_i and p_j, we would like to find rough regions that may present clothes. Because the individual in both images wears the same, this task is achieved by finding common information between images. Before we enter the main process, we first detect pixels with skin colors and put the aside from the main process. Because faces and arms of the same individual in different images are absolutely common characteristics, we filter out them to avoid noises.

To check whether the pixels at the position (x, y) in p_i and p_j has "common" characteristics, we represent the pixels at (x, y) by the HSV color histogram and edge histogram of the superpixel where the pixels belong to. Accordingly, the ℓth histogram of two (x, y) pixels in p_i and p_j are denoted as $s_i^\ell(x, y)$ and $s_j^\ell(x, y)$, respectively. This important trick attenuates noisy pixels and sensing errors in a connected region (superpixel). All pixels in the same superpixel share the same data characteristics.

Based on this representation, the distance $d_{i,j}(x, y)$ between these two pixels are calculated as

$$d_{i,j}(x,y) = \sum_\ell \chi^2(s^\ell_{i(x,y)}, s^\ell_{j(x,y)}), \quad (2)$$

where $\chi^2(s^\ell_{i(x,y)}, s^\ell_{j(x,y)})$ denotes the χ^2 distance between the ℓth histogram, and the overall distance is calculated by summing χ^2 distance in terms of different histograms.

Without loss of generality, we can conduct the same process to any pair of images in the training data $\mathcal{P} = \{p_1, p_2, ..., p_N\}$. We have to emphasize again that any two images p_i and p_j present the same individual in the same dress. The overall χ^2 distance at the position (x, y) is calculated by averaging all possible pairs in $\mathcal{P}$:

$$\bar{d}_{i,j}(x,y) = \frac{1}{\binom{N}{2}} d_{i,j}(x,y),\ i \neq j,\ p_i \in \mathcal{P},\ p_j \in \mathcal{P}. \quad (3)$$

After calculating the overall distance between correspond pixels, we can construct a discrete distance distribution $B_\mathcal{P}[\cdot]$ to show common characteristics between images in $\mathcal{P}$. Figure 3 shows examples of distance distributions from different training set.

The mask region is determined by finding the distance d^t such that

$$\sum_{k=1}^{d^t} B_\mathcal{P}[k] = \gamma \times \sum_k B_\mathcal{P}[k], \quad (4)$$

where the value γ is set as 0.5 in this work.

To determine the mask covering clothes, we find the positions in which average χ^2 distances are smaller than d^t and assign them as the pixels in these positions as foreground. Figure 4 shows examples of the obtained foreground masks from different training data. In this figure, the first three columns are training data, and the fourth column shows the estimated foreground regions (masks), which are displayed in white. Note that the number of training images is not limited to three. The same process can be applied generally.

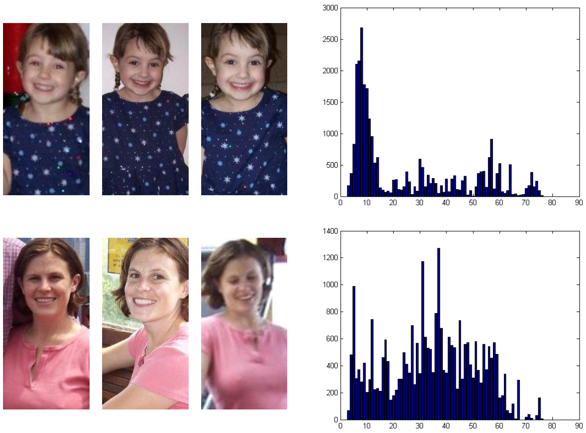

Figure 3. Examples of distance distributions.

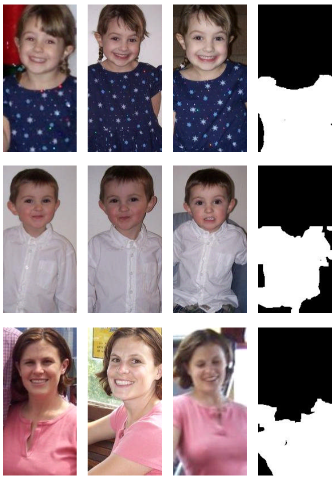

Figure 4. Examples of foreground masks.

3.2. Model Construction

Let's denote the regions in foreground mask K^f, and denote the ones not in K^f as K^b to be background regions. Based on K^f and K^b, we construct a clothes model and several background models as follows.

Give a set of test images $\mathcal{Q} = \{q_1, q_2, ..., q_T\}$, we first collect data characteristics of the pixels covered by K^f, and construct the clothes model $M_i^{(1)}$ in terms of HSV histogram and edge histogram for each image q_i, $i = 1, 2, ..., T$. The overall clothes model is equally contributed by each image:

$$M^{(1)} = \frac{1}{T} \sum_i M_i^{(1)}. \quad (5)$$

The overall clothes model presents the average features in suspected clothes regions. Because clothes in the set of test images are the same, this model captures common information across images. On the other hand, background in different images would vary. Therefore, we construct a background model $M_i^{(0)}$ for the test image q_i by collecting data characteristics of pixels in q_i and are covered by K^b.

3.3. Graph Cut for Segmentation

With the clothes model and background models, we can calculate data cost and smoothness cost and utilize the graph cut framework defined in Eq. (1) to segment an image into clothes regions and others. One thing worth mentioning is that the unit of labeling in this work is a superpixel s_p. We assign a label to each superpixel rather than a pixel, while the energy formulation is same as defined in Eq. (1). Accordingly, the data cost term is defined as:

$$D_{s_p}(f_p) = \exp(-\alpha d(s_p, M^{(f_p)})), f_p \in \{0,1\}, \quad (6)$$

where $d(s_p, M^{(f_p)})$ denotes the χ^2 distance between the superpixel s_p and the model $M^{(f_p)}$ when s_p is assigned the label f_p.

For the smoothness cost, it is defined as:

$$V_{p,q}(f_p, f_q) = (f_p - f_q)^2 \exp(-\beta d(s_p, s_q)). \quad (7)$$

Note that s_p and s_q are superpixels adjacent to each other.

According to the suggestion in [8], the parameters α and β are set as 1 and 0.01, respectively. These two parameters control weights of data cost and smoothness cost. Because clothes regions are often occluded by other objects, such as hands, β is set much smaller than α to attenuate the influence of smoothness cost.

The second row in Figure 5 shows data costs with respect to the clothes models, and the third row shows data costs with respect to the background models. Brighter superpixels denote smaller data costs. From the second row, we clearly see that clothes regions generally have lower data costs with respect to corresponding clothes models. On the other hand, the third row shows that non-clothes regions have noisy distributions of data costs respect to corresponding background models. It's reasonable because background in different images varies drastically, and the background models don't capture data characteristics very well.

Figure 6 shows results of clothes segmentation based on datasets with different number of images and with different capturing environments. From Figure 6(b), (c), (e), (g), (f), and (h), we can see that the proposed method well tackles with clothes with sophisticated texture, which cannot be easily achieved by conventional image segmentation methods. Figure 6(d) and (e) show clothes regions that are significantly occluded by hands and other objects can be effectively determined.

Lighting conditions and characteristics of background would influence segmentation performance. In Figure 6(e), the obtained clothes regions come to pieces, and some superpixels are not successfully labeled due to lighting variations or wrinkles. In the second result of Figure 6(g), parts of background are similar to the person's clothes, and are erroneously labeled as clothes regions.

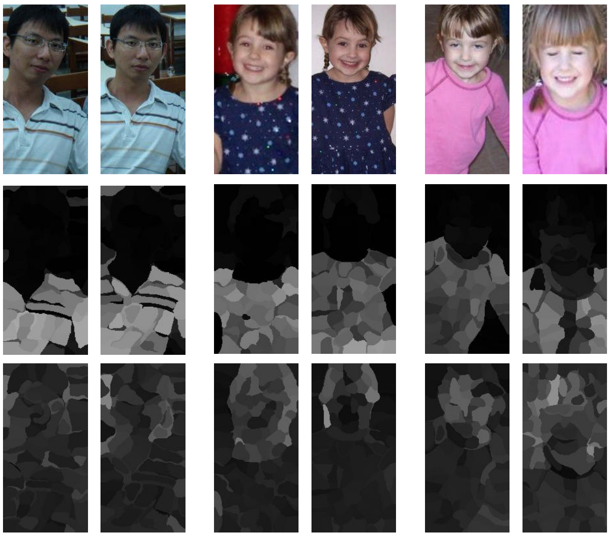

Figure 5. Data costs of superpixels.

4. MAIN OBJECT SEGMENTATION

The essential idea of this work is to discover common information between images and then construct appropriate data models and background models to facilitate binary labeling based on the graph cut framework. Therefore, we extend this method to general object segmentation.

If the test images conform to the assumption of this work, i.e., there is one main objects appearing in different images, we can utilize the same framework to find it. Figure 7 shows some results of main object segmentation. In contrast with clothes segmentation, we don't have to filter out pixels with skin colors.

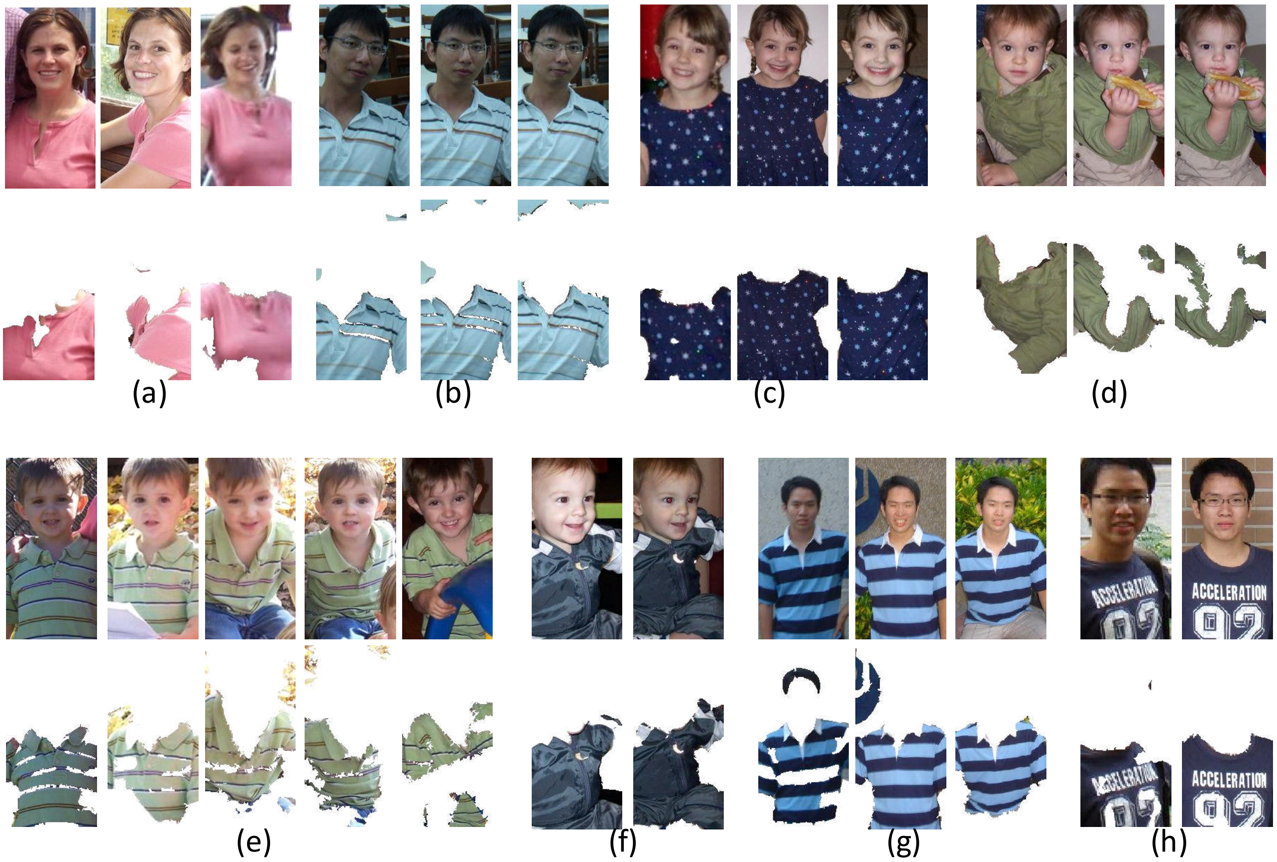

Figure 6. Results of clothes segmentation.

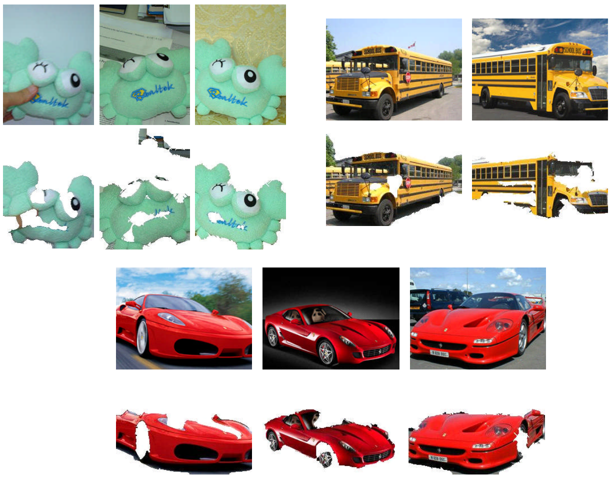

Figure 7. Results of main object segmentation.

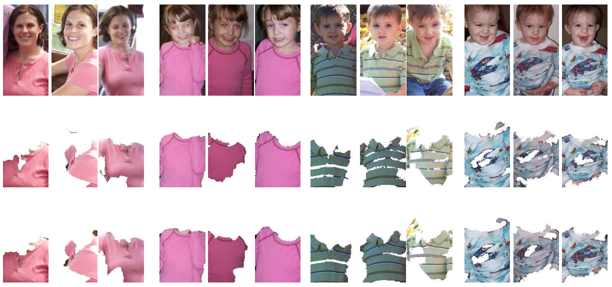

Figure 8. Comparison of segmentation results based on 100 superpixels and 50 superpixels.

It's not surprising that accurate object segmentation is harder to be achieved, especially appearance of the main object may vary significantly, and objects in different images may be captured from different viewpoints.

5. DISCUSSION

5.1. Influence of Superpixels

The basic unit for labeling is a superpixel. Based on the normalized cut algorithm, we segment an image into arbitrary regions. If we segment an image into a large number of small pieces (superpixels), each superpixel itself has high self similarity. However, the clothes may be over-segmented into many small pieces because of wrinkles or slight lighting variations. On the contrary, if the number of extracted superpixels decreases, each superpixel covers a larger region and may contain content with larger variation. However, viewing the clothes as a combination of a few large pieces matches human perception because of the smoothness nature of the human vision system. Therefore, setting of the number of superpixels may influence construction of data models and background models.

In the segmentation results shown above, we segment each image into 100 superpixels, while Figure 8 shows the comparison of segmentation results based on 100 superpixels (the second row) and 50 superpixels (the third row). Generally, the results with smaller number of superpixels, i.e., larger-area superpixels, are slightly better, though larger superpixels may contain parts of non-clothes objects, such as neck and hair.

5.2. Limitation

The central step for discovering common information between images is using pixels with corresponding coordinates in different images to estimate the main object's regions. If there are few pixels that are with the same coordinates and simultaneously fall into the main object region, area of the estimated mask would be small, and therefore we cannot capture main object's characteristics well. This problem becomes worse especially when the main object occupies very different regions in different images. This is the reason that segmentation results in Figure 7 are worse than that in clothes segmentation.

In addition, features are always key elements of image matching. In this work, we simply use HSV color histogram and edge information to characterize each superpixel. More advanced features that more accurately capture data characteristics should be exploited to facilitate mask determination and model construction.

6. CONCLUSION

We have presented a method to automatically segment the main object based on common information in different images. Given a set of images that contain the same object in different backgrounds, this system automatically estimates regions covering the main object, and accordingly constructs data models and background models. Segmenting the main object from background is viewed as a binary labeling problem, and we utilize the graph cut framework to achieve optimal labeling in an efficient way. We provide extensive experimental results that include segmentation of clothes in different poses and lighting variations, and segmentation of primary objects in a sequence of images.

In the future, we would like to break the limitation of pixel correspondence in mask determination. Moreover, more elaborate features and extended applications would be investigated.

ACKNOWLEDGEMENT

This work was partially supported by the National Science Council of the Republic of China under grants NSC 97-2221-E-194-050.

REFERENCES

[1] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 5, pp. 603-619, 2002.

[2] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum, "Lazy snapping," *Proceedings of ACM SIGGRAPH*, pp. 303-308, 2004.

[3] C. Wang, Q. Yang, M. Chen, X. Tang, and Z. Ye, "Progressive cut," *Proceedings of ACM Multimedia*, pp. 251-260, 2006.

[4] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp. 888-905, 2000.

[5] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximation energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 11, pp. 1222-1239, 2001.

[6] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 9, pp. 1124-1137, 2004.

[7] Z. Hu, H. Yan, X. Lin, "Clothing segmentation using foreground and background estimation based on the constrained Delaunay triangulation," *Pattern Recognition*, Vol. 41, No. 5, pp. 1581-1592, 2008.

[8] A.C. Gallagher and T. Chen, "Clothing cosegmentation for recognizing people," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[9] Cui, J., Q. Yang, F. Wen, Q. Wu, C. Zhang, L. Von Gool, and X. Tang, "Transductive object cutout," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[10] Gallagher, A. Consumer image person recognition database, http://amp.ece.cmu.edu/downloads.htm