

Image Segmentation with Perceptual Guidance

Xiaofeng Mi
Department of Computer Science
Rutgers University
xmi@cs.rutgers.edu



Figure 1. Left top: the original David head image, left bottom, segmented image with mean shift on $L^*a^*b^*$ space powered with edge confidence guidance, Right: segmented image with mean shift on $L^*a^*b^*$ space plus perceptual measure

Abstract

This paper proposes an alternative way of embedding confidence information into the segmentation procedure. Two different types of perception measures and how to evaluate the measures are described. Instead of using these perception measures for pixel weight during segmentation, the proposed approach simply adds the perception measure as an additional feature space dimension. Taking human perception information into account makes the segmentation results match human visual perception more accordingly and thus are more consistent with actually human segmented results like rotoscoping. Experiment results show that this approach produces aesthetic pleasing segmentation results, which is useful for further image based depiction.

1 Introduction

Image segmentation, which aims to partition the image into semantically meaningful regions, is arguably considered as the most important low level vision operations.

During recent years, image segmentation has found its applications in computer graphics researches. Barrett and

Cheney proposed an interactive image editing system based on image segmentation [Barrett 02]. DeCarlo and Santella proposed an image abstraction technique by segmenting images on the Gaussian pyramids then selecting appropriate segmentation scales according to human interaction data [DeCarlo 02]. Zhou *et al* advocated the use of a hybrid focal region-based volume renderer that offers an alternative for visualization of internal structures of medical imaging data [Zhou 02], where segmentation information was introduced to enhanced rendering. This technique is also applied recently in video segmentation to produce highly abstracted, cartoon style video with temporary coherence [Wang 04]

However, the result directly from image segmentation is usually undesirable for applications in rendering. One of the biggest problems is that the segmentation boundaries are too arbitrary. Also, segmentation scales are usually not determined automatically. For example, in Figure 1, the image at left bottom is the segmented result from the original image at left top, directly using mean shift segmentation on $L^*a^*b^*$ space with careful chosen bandwidths such that important features like human eyes get kept. However, to keep such features as further rendering information will result in a significant clutter appears. Actually, without human interaction, in the traditional segmentation approach in Euclidian $L^*a^*b^*$



a



b

Figure 2. a is from “spirited away” by Miyazaki from Ghibli studio, whose style tends to use two colors, light color and shadow color, to illustrate characters. In rotoscoped images like the one shown in b, the artists partition the area into various kind of part while keeping the partition meaningful in some sense.

space, keeping details and keeping from non-salient regions are always a pair of conflict.

On the other hand, from an aesthetical viewpoint, stylization is a most important issue and the key problem in a non-photorealistic rendering system is to direct these resources of style to preserve meaningful visual form, while reducing extraneous detail. Visual form describes the relationship between pictures of objects and the physical objects themselves. A good example is the cartoon style abstraction, where human faces are depicted in very limited kinds of colors and the boundaries between image segments are determined uniformly – for example, given a particular cartoon style, most of the artists will follow a common way to place the region boundaries to depict human faces according to the shapes and the colors. Figure 2 shows some pictures from actual movies. In the left image, the child’s face is rendered in a much uniform color, as is the case with the pig (her parents according to the gut though). While in the right, which was taken from the famous movie “waking life” and was produced based on human rotoscoping, is a kind of more painting-style abstraction.

Segmenting images directly from pixel color information is not a good approach as shown in figure 1. And a lot of researches have been focused on combining the outputs of image segmentation and edge detection to improve the quality of the segmented image. Freixenet *et al* surveyed seven different strategies to combining similarity (region) and discontinuity (edge) information, in either embedded or post-processed approach [Freixenet 02]. However, as is show in figure 3(b), such kind of approach still suffers from the dilemma between scale selection and getting rid of nonsense region partitions.

This paper proposes an alternative method of embedding confidence information into the segmentation procedure. Instead of using these perception measures for pixel weight, this paper simply adds the perception measure into the feature space as an additional channel. This is



a



b

Figure 3. a is the original image and b is the segmentation result by EDISON [Christoudias 02] with spatial bandwidth 7 pixels and color bandwidth (uniformly in $L^*u^*v^*$ space) 7, with synergetic embedded edge confidence. As show in b, although the area around eyes begins to get rough, there are still a lot of undesired region partitions.

based on the observation that human perception has some sorts of continuity on a particular feature, and this continuity will also guide human's visual perception. That is, human will feel natural if an image is segmented in accordance with their perception to the world. That also explains why artists with the same artistic abstraction style will segment a given image in a roughly same way.

2 Mean Shift Image Segmentation

A large class of current image segmentation algorithms are based on feature space, by analyzing the images in some carefully defined space – typically color spaces or texture spaces. The later are of particular significance in texture segmentation. In segmentation this way, image segmentation is simply a partition of an image into contiguous regions of pixels that have similar appearance, such as color or texture [Trucco and Verri 98]. Each region has aggregate properties associated with it, such as its average color. The algorithm described by Comaniciu and Meer [Comaniciu 02] is a paradigm of robust segmentation of color images, as it produces relatively cleaner results despite of some sort of image noises. Within this algorithm, colors are represented in the perceptually uniform color space $L^*u^*v^*$ [Foley et al. 97] which produces region boundaries that are more meaningful for human observers. The parameters of this algorithm include a spatial radius h_s (similar to the radius of a filter), a color difference threshold h_r , and the size of the minimum acceptable region M .

The mapping of real images to feature spaces often produces a very complex structure. Salient features whose recovery is necessary for the solution of the segmentation task, correspond to clusters in this space. As there is no *a priori* information is typically available, the number of clusters/classes and their shapes/distributions have to be discerned from the given image data.

Let $\{X_i\}_{i=1..N}$ be the set of N image vectors in the d -dimensional Euclidean space R^d . The multivariate kernel density estimate obtained with kernel $K(x)$ and window radius h (the universal bandwidth here), computed at point x is defined as the well known expression:

$$\hat{f}(x) = \frac{1}{Nh^d} \sum_{i=1}^N K\left(\frac{x - X_i}{h}\right) \quad (1)$$

The radially symmetric kernel function $K(x)$ satisfies:

$$K(x) = c_{k,d} k(\|x\|^2) \quad (2)$$

then, we rewrite the kernel density estimator as:

$$\hat{f}_{h,K}(x) = \frac{c_{k,d}}{Nh^d} \sum_{i=1}^N k\left(\left\|\frac{x - X_i}{h}\right\|^2\right) \quad (3)$$

The modes of the density are located among the zeros of the gradient $\nabla f(x) = 0$. Define:

$$g(x) = -k'(x), \quad G(x) = c_{g,d} g(\|x\|^2) \quad (4)$$

We have

$$\widehat{\nabla} \hat{f}_{h,K}(x) = \hat{f}_{h,G}(x) \cdot m_{h,G}(x) \quad (5)$$

where $\hat{f}_{h,G}(x)$ is proportional to the density estimate at x computed with the kernel G , and

$$m_{h,G}(x) = \frac{\sum_{i=1}^N x_i g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)}{\sum_{i=1}^N g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)} - x \quad (6)$$

is known as the mean shift, which shows that the estimation density gradient at location x is proportional to the offset of the mean vector computed in a window, and recursive application of the mean shift property make the pixels converge to their local density modes and thus pixels can be partitioned with local modes. This is the basis of mean shift segmentation.

3 Visual Perception

In traditional color space based segmentation, $L^*a^*b^*$ or $L^*u^*v^*$ color spaces are regarded as the best suitable for vision research because they have a metric a satisfactory approximation to Euclidean, thus allowing the use of spherical windows.

Though human perception towards an image is determined by the colors of the pixels, the perception of each pixel is not determined solely with that pixel. For example, while a lot of color space model have an explicit or implicit representation of color luminance, such as $L^*a^*b^*$, $L^*u^*v^*$, HSV and grayscale color space as well, human does not likely percept brightness just according to local vision area, instead, the surrounding visual area has significant impact on the human perceptions and human tends to discard certain properties of light, based on the principle, comes the first perception measure used as the image feature space for segmentation.

3.1 Brightness Measure

As is surveyed by Gooch *et al* [Gooch 04], brightness perception can be modeled using operators such as differentiation, integration and thresholding. These methods model lateral inhibition which is one of the most pervasive structures in the visual nervous system [Palmer 1999]. Lateral inhibition is implemented by a cell's receptive field having a center-surround organization. Thus cells in the earliest stages of human vision respond most vigorously to a pattern of light which is bright in the center of the cell's receptive field and dark in the surround, or vice-versa. Such antagonistic center-surround behavior can be modeled using neural networks, or by computational models such as Difference of Gaussians, Gaussian smoothed Laplacians and Gabor filters.

Like what Gooch *et al* did in perceptual measure based human face illustration [Gooch 04], this paper follows Blommaert and Martens' [Blommaert 1990] model of human brightness perception. The aim of the Blommaert model is to understand brightness perception in terms of cell properties and neural structures. For example, the scale invariance property of the human visual system can be modeled by assuming that the outside world is interpreted at different levels of resolution, controlled by varying receptive field sizes. Blommaert and Martens demonstrated that, to a first approximation, the receptive fields of the human visual system are isotropic with respect to brightness perception, and can be modeled by circularly symmetric Gaussian profiles R_i :

$$R_i(x, y, s) = \frac{1}{\pi(\alpha_i s)^2} e^{-\frac{x^2 + y^2}{(\alpha_i s)^2}} \quad (7)$$

where s is the number of pixels involved in the corresponding level of brightness reception field. $\alpha_1 = 1/(2\sqrt{2})$ and $\alpha_{i+1} = 1.6 * \alpha_i$. The neural response V_i as the function of image location, scale and luminance distribution L can be computed by convolution:

$$V_i(x, y, s_i) = L(x, y) \otimes R_i(x, y, s_i) \quad (8)$$

The firing frequency evoked across scales by a luminance distribution L is modeled by a center-surround mechanism:

$$V^i(x, y, s_i) = \frac{V_i(x, y, s_i) - V_{i+1}(x, y, s_i)}{2^\phi / s_i^2 + V_i(x, y, s_i)} \quad (9)$$

where V_i is specified by (8) and the term $2^\phi / s^2$ is introduced to avoid singular cases when V_i is approaching zero. And in the imaging application, $\phi = 1$ is an appropriate value.

An expression of brightness B is now deriving by summing V^i over all scales:

$$B = \sum_{s_i=s_0}^{s_{\max}} V^i(x, y, s_i) \quad (10)$$

Practically, choosing the total number of scales up to 8 is



Figure 4. brightness measure of figure 3(a)

actually good enough.

The result of these computations is an image which could be seen as an interpretation of human brightness perception. For example, the brightness perception of figure 3(a) is shown as figure 4, with low perception value shown in black and high value in white.

Adding the calculated brightness image into the original $L*a*b*$ color space can help guiding image segmentation given that appropriate bandwidth at this dimension is chosen. The image shown on the right of figure 1 is such an example and figure 5 shows yet another example, which is the segmented result for figure 3(a).



Figure 5. Segmentation result of image shown in figure 3(a) with color space bandwidths 0.015 and perception bandwidth 0.1. (both color channel and perception values take domain in [0, 1])

3.2 Photometric Invariant Measure

As shown in figure 2(a), in a broad range of applications, one tends to omit the surface orientation and the shading, highlight or shadow effects and treats the whole object as a single segment, which makes it obvious that photometric invariance is essential for image segmentation. This subsection, with the example, argues that by adding the hue value as the measure of photometric invariance, photometric invariant segmentation could be retrieved.

An N -dimensional spectrum can be denoted as:

$$\vec{c} = m_b(\vec{n}, \vec{s})\vec{c}_b + m_s(\vec{n}, \vec{s}, \vec{v})\vec{c}_s \quad (11)$$

Where \vec{n} , \vec{s} , \vec{v} are three-dimensional, denoting the surface normal, the direction of the illumination source and the direction of the viewer respectively and where \vec{c} and \vec{c}_s are the surface albedo and Fresnel reflectance respectively, which are both N -dimensional with N the number of samples taken in the wavelength range. For example, in RGB color space, N will be 3.

Gevers argues that this color space can be transformed into a different polar coordinate representation as follows[Gevers 03]:

$$p_s = 1 - \min\{c(\lambda_1), \dots, c(\lambda_N)\} \quad (12)$$

$$\theta_h = \alpha[c(\lambda_i) - [1 - \rho_s], \phi(i, N)] \quad (13)$$

Where θ_h takes on values in the range $0 \leq \theta_h \leq 2\pi$ and where

$$\phi(i, N) = \frac{i-1}{N-1} \cdot \frac{4}{3}\pi \quad (14)$$

and

$$\alpha(w_i, \theta_i) = \arctan \left(\frac{\sum_{i=1}^N w_i \sin \theta_i}{\sum_{i=1}^N w_i \cos \theta_i} \right) \quad (15)$$

For *RGB* spectrums, we have

$$\begin{aligned} \theta &= \arctan \frac{(R - \rho_s) \sin(0) + (G - \rho_s) \sin(\frac{2}{3}\pi) + (B - \rho_s) \sin(\frac{4}{3}\pi)}{(R - \rho_s) \cos(0) + (G - \rho_s) \cos(\frac{2}{3}\pi) + (B - \rho_s) \cos(\frac{4}{3}\pi)} \\ &= \arctan \frac{\frac{\sqrt{3}}{2}G - \frac{\sqrt{3}}{2}B}{R - \frac{1}{2}G - \frac{1}{2}B} = \arctan \frac{\sqrt{3}(G - B)}{(R - G) + (R - B)} \end{aligned} \quad (16)$$

which, by the definition of hue value of color given by Levkowitz and Herman [Levkowitz 93], shows that the value of θ is exactly the hue value of the color.

As the brightness perception measure, adding the hue value as the perception measure to the feature space will help to produce photometric invariant segmentation result as long as the appropriate bandwidth is selected. Figure

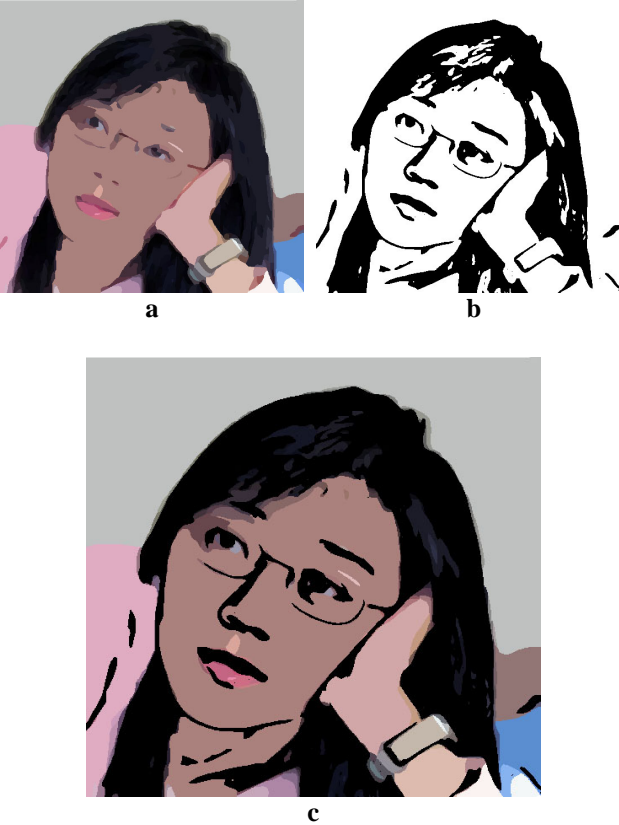


Figure 6. a shows the photometric invariant segmentation result, b is the binary image for brightness measure and c is the combination of the two images.

6(a) shows such the results for figure 3(a).

Interestingly, if we combine the binary image of the brightness image computed as shown in figure 4 with a brightness threshold (shown in figure 6(b)) and the photometric invariant segmentation result, we can get a popular cartoon stylization result, as shown in figure 6(c). Another example of the same kind is shown in figure 7(d).

4 Discussion and Conclusion

This paper proposed a paradigm of adding perceptual measures into the feature space while performing segmentation to get more desirable results. Two perceptual measures are introduced and experiments shows that this method can help produce segmentation results which make much more sense from the perceptual and aesthetical point of view.

Such methods also provide an alternative for scale control while segmentation. Compare the result of figure 3 and figure 5, it is obvious that by taking brightness measure into account, figure 5 deals with the scale near subtle areas like eyes and others like face area more sophisticatedly.

The work described in this paper falls into the paradigm of applying vision techniques in computer graphics area. More examples with the proposed approach are shown in Figure 7 and figure 8

There are open problems remain in adding additional feature dimension though. One practical issue to deal with is the bandwidth determination of perceptual feature space. Since the measure of human perception is totally another story other than color space, the bandwidth selection at perceptual dimension is currently done interactively. How to select a proper bandwidth for the perceptual measure still needs more efforts. In fact, sometimes the bandwidths are not easy to choose. For example, to produce the image shown in figure 7(c), where the yellow ball on the torch being a separate region other than the sky or the green statue, the bandwidth lies in a small range of the feature space and it really requires one with great patience, which is not a good practice (As for pure color-feature based mean shift, as the conflict between fine scale for detail and course scale for abstraction, the system can not produce a similar result with both the yellow ball preserved and the shading effects eliminated).

The examples shown throughout this paper is segmented via mean shift approach. However, theoretically, any feature-space based segmentation should be applicable for the space with perceptual measure added as additional dimension.

Both the brightness and the hue perception measure have limitations due to the singular cases. As for brightness measurement, since it operates based on the luminance level of pixels, it might fall into wrong way when the luminance levels of pixels are similar while the colors are

different. Though this happens rarely, the problem with hue integration described in section 3.2 for photometric invariant segmentation occurs much common: it suffers from the instability of the hue value around regions where R , G , B values are similar. How to find a robust photometric invariant measure for digital images is still an open issue.

The approach described in this paper shows special advantages when segmenting human faces using brightness perceptual measure, as can be seen from the examples. While the effect on the segmentation of other type of objects remains a little bit controversial and need further exploration.

Acknowledgements

Thanks to Prof. Elgammal and Prof. DeCarlo and Anthony Santella for the discussions of the work which have benefited me a lot. The implementation of mean shift segmentation was largely borrowed from EDISON system.

References

[1] William A. Barrett, Alan S. Cheney, Object-based image editing, *Proceeding of SIGGRAPH 2002*, Pages: 777 - 784
 [2] F. J. J. Blommaert, and J.-B. Martens1990. An object-oriented model for brightness perception. *Spatial Vis.* 5, 1, 15-41.
 [3] Christopher M. Christoudias, Bogdan Georgescu, Peter Meer, Synergism in Low Level Vision, 16 th International Conference on Pattern Recognition (ICPR'02)
 [4] Rorin Comaniciu, Peter Meer, Mean Shift: A Rubust Approach Toward Feature Space Analysis, *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, Vol 24, No. 5, May 2002.
 [5] Doug DeCarlo, Anthony Santella. Stylization and Abstraction of Photographs. In *SIGGRAPH 2002*
 [6] J. Freixenet, X. Munoz, D. Raba, J. Marti, and X. Cufi. Yet another survey on image segmentation. In 7th European Conference on Computer Vision, volume III, pp. 408-422, Copenhagen, Denmark, May 2002
 [7] James D. Foley, Andries van Dam, Steven K. Feiner, John F. Hughes, *Computer Graphics: Principles and Practice in C (2nd Edition)*, Addison Wesley, 1996.
 [8] Th. Gevers and H.M.SG. Stokman, Robust photometric invariant region detection in multispectral images. *International Journal of Computer Vision* 53(2), pp.135-151, 2003
 [9] Bruce Gooch, Erik Reinhard, Amy Gooch, *Human Facial Illustrations: Creation and Psychophysical Evaluation*, *ACM Transactions on Graphics*, Vol. 23, No. 1, January 2004, Pages 27-44
 [10] H. Levkowitz and G.T. Herman. Glhs: A generalized lightness, hue and saturation color model. *Computer Vision, Graphics, and Image Processing: Graphical Models and Image Processing*, 55(4), 271-285
 [11] P. Meer, B. Georgescu: Edge detection with embedded confidence. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23, 1351-1365, 2001
 [12] S. E. Palmer. *Vision Science: Photons to Phenomenology*. The MIT Press, Cambridge, Mass.
 [13] E. Trucco., and A. Verri, *Introductory Techniques for 3-D Computer Vision*. Prentice-Hall. 1998.
 [14] Jue Wang, Yingqing Xu, Heung-Yeung Shum and Michael Cohen. *Video Tooning*. *SIGGAPH2004*
 [15] Jianlong Zhou and Klaus D. Tönnies. Focal region-based volume rendering. In *Proceedings of the 9th International Workshop on Systems, Signals and Image Processing*, pages 394-397, Manchester, U.K., 7.-8. November 2002

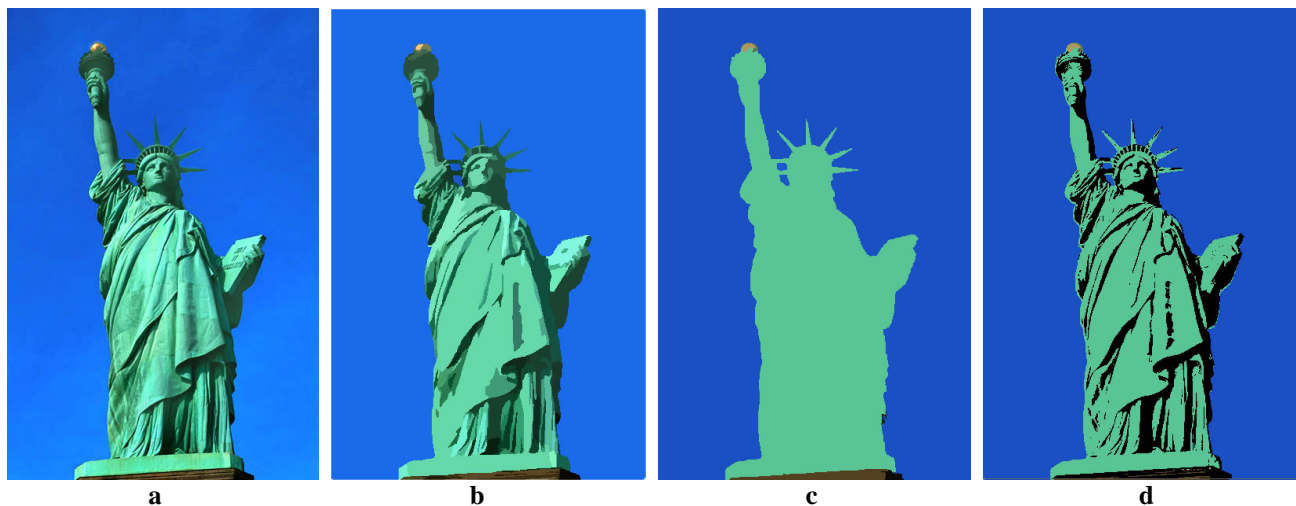


Figure 7. The statue of liberty under guided segmentation. The source image is 400×600 and the segmentation result shown in b takes brightness as the perceptual guidance and the segmentation result c takes hue value as the photometric invariant perceptual guidance, d is the combination of c and the binary image of the brightness measure obtained with a user specified threshold.

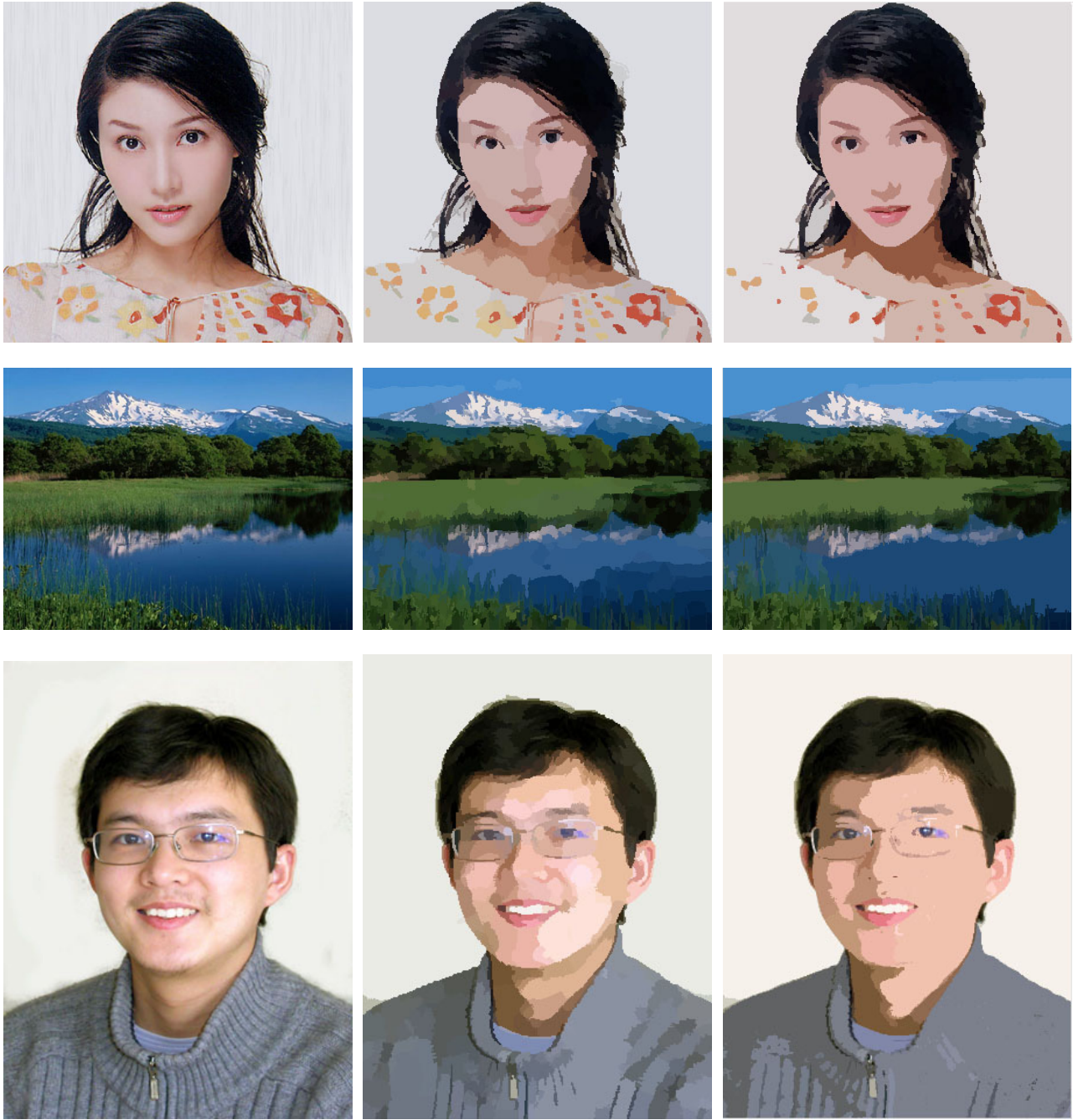


Figure 8. More examples, left column are the original images, mid column are the images segmented by color spaces only, the right column are segmentation results with perceptual guidance.