SEMANTICS-DRIVEN PORTRAIT CARTOON STYLIZATION

Ming Yang[a,b] *Shu Lin*[a,b] *Ping Luo*[a,b] *Liang Lin*[a,b] *Hongyang Chao*[a, *]
[a]Sun Yat-Sen University, Guangzhou 510275, PR China
[b]Lotus Hill Research Institute, Wuhan 430074, PR China

ABSTRACT

This paper proposes an efficient framework for transforming an input human portrait image into an artistic cartoon style. Compared to the previous work of non-photorealistic rendering (NPR), our method exploits the portrait semantics for enriching and manipulating the cartooning style, based on a semantic grammar model. The proposed framework consists of two phases: a *portrait parsing* phase to localize and recognize facial components in a hierarchic manner, and further calculate the portrait saliency with the facial components; a *cartoon stylizing* phase to abstract and cartoonize the portrait according to the parsed semantics and saliency, in which the regions and structure (edges/boundaries) of the portrait are rendered in two layers. In the experiments, we test our method with different types of human portraits: daily photos, identification photos, and studio photos, and find satisfactory results; a quantitative evaluation of subjective preference is presented as well.

Index Terms— cartoon stylization, portrait parsing, NPR

1. INTRODUCTION

Image stylization is a long-standing and well-studied topic in the NPR and Graphics community, in that it provides stylistic and painterly visual effects while expressing artist's understanding. This paper presents a novel cartoonization framework for rendering a high resolution human portrait into the cartoon style. Two representative examples by our framework are shown in Fig.1.

Fig. 1. Two representative cartoonized portraits by our method.

Related work. In the literature, early cartoon stylization techniques mainly focused on simulating the effects of drawing tools (e.g. sketch and pencil strokes), and developing user interfaces for easy and flexible editing, such as [7, 8]. Recently, in order to achieve more painterly cartoon, many works exploit the ways to extract useful image contents, which will guide the rendering to embody the artist's intention and interpretation. For example, DeCarlo et al. [5] transformed an image into a line drawing by deriving the user areas of interest with eye-tracking. Image saliency was used to abstract imagery in terms of luminance and color opponency contrasts in [9]. Bhat et al. [1] further optimize the saliency-based rendering approach and pose an efficient framework. However, these methods still lack of using high-level (or semantic) understanding of images (e.g. scene and object recognition), since automatically parsing a generic images is quite challenging and time consuming. Hence, some stylization systems allowed user interaction for extracting semantic content from images, such as [2, 3].

In this paper, we study a semantics-driven portrait cartoonization approach according to extracted portrait semantics, which generates expressive and vivid cartoonized portraits.

Overview. The method consists of two phases: a *portrait parsing* phase and a *cartoon stylizing* phase. In the first phase, by employing a recent proposed probabilistic grammar model [10], each input portrait is hierarchically parsed into a graphical representation of facial components, namely "Parse Graph" (see Fig.3(b)). The image saliency is further calculated with the parse graph. We argue that this integrating of saliency with semantics is more reasonable and compatible with the human vision system, based on the observation that the visual attention often spreads in different magnitude with different facial components. The calculated saliency can be viewed as the proposal cues for the cartoon stylization. In the second phase, the stylizing method is inspired by the painting procedure of human artists, in which different styles are painted or manipulated with different parts in the portrait. We render the portrait in region layer and structure layer respectively. In the layer of region stylizing, we abstract and exaggerate the region areas of the portrait by utilizing a gradient-based filter guided by the proposal image saliency. In the layer of structure stylizing, the strong edges and boundaries (named as *sketches*) of portrait are cartoonized by stroke rendering. The proposal sketch saliency is treated as the opacity of the stroke. We also simulate the hardness of the strokes. In the both layers, the parameters of stylizing for different areas of the portraits are adapted with the semantic parse graph.

This work was partially supported by National Natural Science Foundation of China (Grant No.60970156) and Guangdong Natural Science Foundation Council (Grant No.07003728).

*Corresponding author:
Email address: isschhy@mail.sysu.edu.cn (Hongyang Chao)

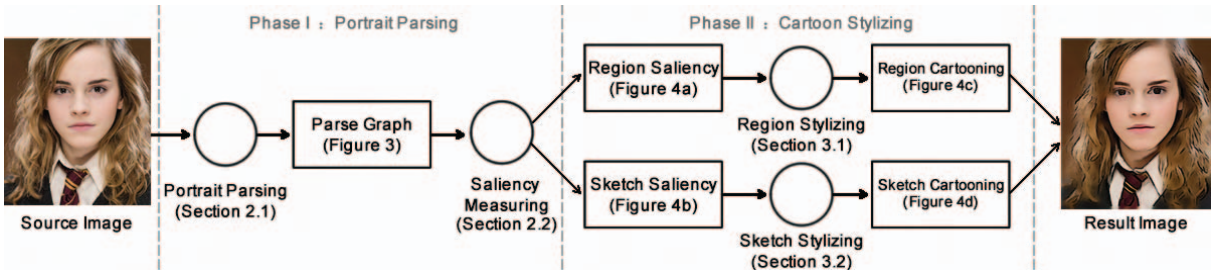

Fig. 2. Framework

The proposed framework is summarized in Fig.2.

The key contribution of this paper is in two aspects: extracting portrait semantics by employing the grammar model, and simulating the artists drawing for semantics-driven cartooning.

2. PORTRAIT PARSING

Given an input high resolution portrait image, we first decompose it into constituent components in a hierarchical structure, in that the image saliency is then calculated.

2.1. Hierarchical Decomposition and Recognition

A compositional grammar model for face representation [10] is learned off-line from a set of manually annotated face instances that includes various appearance and pose. This grammar model is in form of an And-Or graph in the computer vision, as illustrated in Fig.3(a), which consists of And-nodes, Or-nodes and Leaf nodes. The And-node represents the decomposition, which divides a face into parts from coarse to fine. The Or-node represents the alternatives to account for the diversity of face appearance. The Leaf-node represents a component or sub-template. Spatial relations and constraints are imposed between the nodes at the same level to ensure the validity of the properties (symmetry of eyes and spatial relationships among facial parts, etc.). Following [10, 6], the compositional face model can be defined as

$$\mathcal{G}_{And-Or} =< S, V_N, V_T, \mathcal{R}, \mathcal{P} >, \tag{1}$$

where S denotes the human face category, V_N and V_T denote non-terminal nodes and terminal(leaf) nodes of the graph respectively, $\mathcal{R}$ represents a set of pairwise relations defined on the edge between two graph nodes, $\mathcal{P}$ denotes the probability model defined on the graph structure.

Using the And-Or graph model with a recursive parsing algorithm, an input portrait image can be recognized into a hierarchic parse graph, as shown in Fig.3(b). In each node of the graph, the algorithm integrates two closely coupled process, bottom-up detection of parts/primitives from the image and top-down verification with learned dictionary. These two processes form an iterative loop. Please refer [10, 6] for more details. Then a portrait image I^{obs} is parsed as,

$$\mathcal{PG} =< V, E >, \tag{2}$$

where E represents the vertical edges and the horizontal edges, $V = \{v_i\}_{i=1}^{N}$ denotes a set of leaf nodes (the number $N = 7$). We define each leaf node as a 2-tuple $v_i =< \Lambda_i, l_i >$, where Λ_i and l_i denote the image domain and the label of each facial component.

Compared with the traditional face localization methods with the AAM model, our model can handle large-scale structural variations and facial details.

2.2. Portrait saliency computation

We present a semantics-driven approach for measuring both region saliency and sketch saliency.

Firstly, we adopt a long-edge detector [1] to compute the length and local orientation of the underlying dominant edge at each pixel. This process returns a two-channel image $\Psi = \{O, \Gamma\}$, where O and Γ denote orientation map and edge length map respectively.

Secondly, we introduce two factors (i.e. c^{Sk}, c^{Rg}) to drive measure of sketch saliency and region saliency respectively. Thus our property for saliency measuring is a 2-tuple $< c_i^{Sk}, c_i^{Rg} >, i = 1, \ldots, N$. These two factors for each Leaf-node of $\mathcal{G}_{And-Or}$ are specified emprically by the professional artists, and each of them is permitted with minor fluctuation (see Table 1).

We denote the sketch saliency image by S^{Sk} and the region saliency image by S^{Rg}. Note that $S^{Rg} = \{S_x^{Rg}, S_y^{Rg}\}$ have two channels, one estimating the saliency of the gradient in the x-direction and the other estimating the saliency of the gradient in the y-direction. For each pixel $j \in \Lambda_i (i = 1, \ldots, N)$,

$$S_{ij}^{Sk} = c_i^{Sk} \cdot \Gamma_{ij}, \tag{3}$$

$$\begin{cases} S_{x,ij}^{Rg} = c_i^{Rg} \cdot \cos^2 O_{ij} \cdot \Gamma_{ij}; \\ S_{y,ij}^{Rg} = c_i^{Rg} \cdot \sin^2 O_{ij} \cdot \Gamma_{ij}. \end{cases} \tag{4}$$

One example of our saliency measure is shown in Fig.4(a)(b).

3. CARTOON STYLIZING

In the stylizing phase, regions and sketches are stylized independently. We denote the stylized region image and the stylized sketch image by F^{Rg} and F^{Sk} respectively. The final result F is generated by overlaying F^{Sk} on top of F^{Rg}.

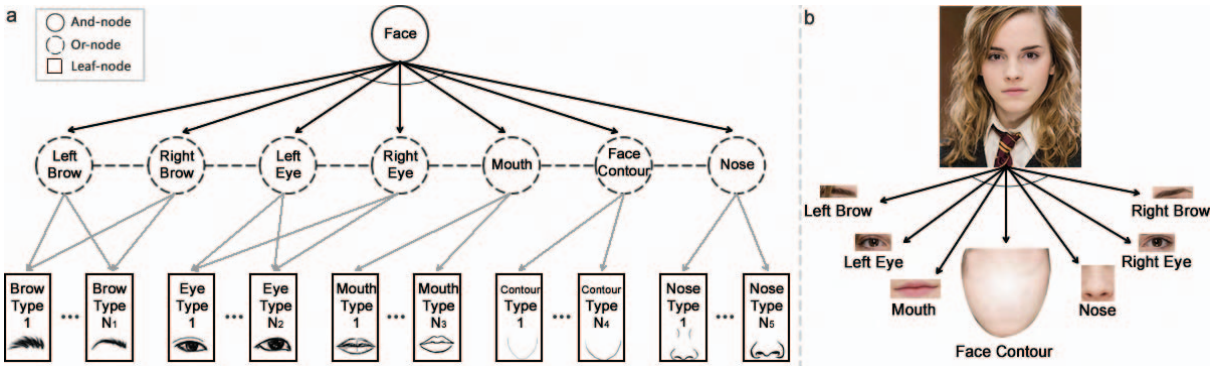

Fig. 3. (a) shows the And-Or graph for portrait. An Or-node (dashed) is a "switching variable" for possible choices of the components. Only one child is assigned for each object instance. An And-node (solid) represents a composition of children with certain spatial and appearance relations. (b) lists an instance and its parsing graph.

Comp.	c^{Sk}	c^{Rg}	ω^d	ω^g	c^{ex}	c^{ab}	c^{hd}
Left Brow	0.5	1.5	0.03	0.9	1.5	0.8	0.6
Right Brow	0.5	1.5	0.03	0.9	1.5	0.8	0.6
Left Eye	0.2	1.7	0.08	1.1	1.5	0.9	1.2
Right Eye	0.2	1.7	0.08	1.1	1.5	0.9	1.2
Mouth	0.6	1.6	0.02	1	1.4	0.9	0.7
Contour	0.7	1.4	0.02	1	1.3	1	0.9
Nose	0.6	1.3	0.04	1	1.4	0.9	0.4

Table 1. Parameter specification for an instance. Note that each Leaf-node of And-Or graph corresponds to a specification of these seven parameters. Here we just list the specification for the N Leaf-nodes of a parse graph.

Fig. 4. Given an input image, both sketch saliency and region saliency are calculated. Then regions and sketches are stylized independently. (a) Region saliency map. (b) Sketch saliency map. (c) Stylized Regions. (d) Stylized Sketches.

3.1. Region stylizing

Let $G^{obs} = \{G_x^{obs}, G_y^{obs}\}$ be the gradients of the input image I^{obs} and $G^{Rg} = \{G_x^{Rg}, G_y^{Rg}\}$ be the gradients of the stylized region image F^{Rg}. Note that G^{Rg} should approximate to our desire gradients so as to simulate the cartoon style. On the other hand, F^{Rg} cannot drift too much from I^{obs}.

We compute F^{Rg} that fully holds these constraints. Therefore, our objective is,

$$(F^{Rg})^* = \arg\min_{F^{Rg}}(E_d + E_g), \tag{5}$$

where E_d and E_g represent data cost and gradient cost, respectively. These two functions can be defined as follows:

$$E_d = \sum_{i=1}^{N}\sum_{j\in\Lambda_i}\omega_i^d(F_{ij}^{Rg} - I_{ij}^{obs})^2, \tag{6}$$

$$E_g = \sum_{i=1}^{N}\sum_{j\in\Lambda_i}\omega_i^g(G_{ij}^{Rg} - c_i^{ex}\cdot(1-\exp\{-\frac{S_{ij}^{Rg}}{c_i^{ab}}\})\cdot G_{ij}^{obs})^2. \tag{7}$$

Note that there are four parameters that influence the effect of the region rendering in the cost functions: ω^d and ω^g specify weights of data cost and gradient cost, respectively. $c^{ex}(c^{ex} \geq 1)$ control the amount of exaggeration, while $c^{ab}(c^{ab} > 0)$ control the amount of abstraction.

Thus our property for region stylizing is a 4-tuple $< \omega_i^d, \omega_i^g, c_i^{ex}, c_i^{ab} >, i = 1, \ldots, N$. Each Leaf-node of $\mathcal{G}_{And-Or}$ corresponds with such a property, which is also specified by the professional artist (see Table 1).

The result from this step is shown in Fig.4(c).

3.2. Sketch stylizing

The long-edge detection result Γ can be visualized to generate sketches so as to make the result look as if the artist outlined the salient edges using brush strokes [1]. In our method, such visualization is guided by proposal sketch saliency.

Suppose the sketch color value is denoted by C. We treat S^{Sk} as an opacity map. For each pixel $j \in \Lambda = \Lambda_1 \cup \Lambda_2 \cup ... \cup \Lambda_N$,

$$I_j^{SE} = S_j^{Sk} \cdot C, \tag{8}$$

where I^{SE} denotes the visualized image.

To simulate strokes of different hardness in different components, a gradient-based filter is applied. Thus the computation of sketch stylizing can be formulated as

$$(F^{Sk})^* = \arg\min_{F^{Sk}}(\sum_{i=1}^{N}\sum_{j\in\Lambda_i}(F_{ij}^{Sk} - I_{ij}^{SE})^2 + \sum_{i=1}^{N}\sum_{j\in\Lambda_i}(G_{ij}^{Sk} - c_i^{hd}G_{ij}^{SE})^2), \tag{9}$$

where G^{SE} denotes the gradients of I^{SE}, c^{hd} specifies the hardness of the stroke. Therefore our property for sketch stylizing is 1-tuple $< c_i^{hd} >, i = 1, \ldots, N$, with which a Leaf-node of $\mathcal{G}_{And-Or}$ corresponds. The property specification for sketch stylizing is specified by the artist as well (see Table 1).

The result from this step is shown in Fig.4(d).

Fig. 5. More results

4. EXPERIMENTS

We apply our method on three categories of human portraits: daily photos, identification photos and studio photos. Ninety images (Thirty for each category) are selected from the public LHI dataset for test. Experiments on a PC with Core Duo 2.53GHZ CPU show that a parsing phase takes around $2.5 \sim 3.5$ minutes, while a stylizing phase takes around $4 \sim 15$ seconds. Some of our results are shown in Fig.1 and Fig.5.

In addition, we present a psychology experiment to further quantitatively demonstrate the effectiveness of our framework. For each object category, we apply Bhat [1]'s approach and Winnemöller [9]'s approach on our test images as well. Therefore each test portrait image has three stylized results (see Fig.6). We put all these results on a voting system and recruit twenty volunteers to vote for the best result of each test image subjectively. Here we define a favor rate δ to evaluate the experiment: $\delta = favor\ hits/(total\ category\ tests * total\ volunteers)$. Table 2 reports the evaluating result of the human perception experiment. It shows that our approach produces more satisfactory effects.

Fig. 6. A comparison of our result for cartoon stylization to Bhat's and Winnermoller's results. (a)Original image. (b)Bhat's result. (c)Winnemöller's result. (d)Our result.

Photo category	Favor Rate δ (%)		
	Bhat's	Winnemöller's	Ours
Daily photo	27.8	21.0	51.2
ID Photo	25.3	19.2	55.5
Studio Photo	34.7	18.2	47.1

Table 2. Favor rate

5. CONCLUSION

A semantics-driven approach for image-based cartoon stylization has been proposed. Compared with previous methods, this approach benefits from richer meaningful portrait semantic information, which leads to better simulation of artistic cartoon style. We'll develop an interactive system for professional artists to specify the parameters more easily and adopt learning-based algorithm to learn the parameters specified so as to extend our work in the future.

6. REFERENCES

[1] Pravin Bhat, Larry Zitnick, Michael Cohen, Brian Curless, "GradientShop: A Gradient-Domain Optimization Framework for Image and Video Filtering", *ACM Trans. on Graphics*, 2009.

[2] Hong Chen, Nanning Zheng, Lin Liang, Yan Li, Ying-Qing Xu, Heung-Yeung Shum, "PicToon: a personalized image-based cartoon system", *ACM Multimedia*, pp. 171–178, 2002.

[3] Fang Wen, Qing Luan, Lin Liang, Ying-Qing Xu, Heung-Yeung Shum, "Color Sketch Generation", *NPAR*, pp. 47–54, 2006.

[4] Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, Jim Graham, "Active shape models—their training and application", *Computer Vision and Image Understanding*, vol. 61(1), pp. 38–59, 1995.

[5] Douglas DeCarlo, Anthony Santella, "Stylization and abstraction of photographs", *ACM Trans. on Graphics*, vol. 21(3), pp. 769–776, 2002.

[6] Liang Lin, Tianfu Wu, Jake Porway, Zijian Xu, "A Stochastic Graph Grammar for Compositional Object Representation and recognition", *Pattern Recognition*, vol. 42(7), pp. 1297-1307, 2009.

[7] Frank Van Reeth, "Integrating $2\frac{1}{2}$D computer animation techniques for supporting traditional animation", *Computer Animation*, pp. 118–125, 1996.

[8] Zsö fia Ruttkay, Han Noot, "Animation CharToon faces", *NPAR*, pp. 91–100, 2000.

[9] Holger Winnemöller, Sven C. Olsen, Bruce Gooc, "Real-time video abstraction", *ACM Trans. on Graphics*, vol. 25(3), pp. 1221–1226, 2006.

[10] Zijian Xu, Hong Chen, Song Chun Zhu, Jiebo Luo, "A Hierarchical Compositional Model for Face Representation and Sketching", *IEEE Trans. on PAMI*, vol. 30(6), pp. 955–969, 2008.