2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Madrid, Spain, October 1-5, 2018

Robot Artist Performs Cartoon Style Facial Portrait Painting

Ren C. Luo[1], Yu Jung Liu[2]

Abstract— **This paper presents a face portrait with cartoon stylization painting and associated algorithms with the visual feedback system to paint like a human cartoonist. The robot cartoonist creates the artwork in two stages-cartoon style transformation and robot artist for colorful painting. In the cartoon style transformation stage, it transfers human portrait photos to cartoon style by face detection and alignment, which can effectively decompose the face into individual components then replace by cartoon facial components. In the second stage, the robot uses an eye-in-hand system to obtain five basic colors (cyan, magenta, yellow, white and black) to automatically mix a variety of colors automatically. For painting strategy, we start with the outline of the face, which we use non-photorealistic rendering (NPR) to generate hand-painted strokes. After that, the robot artist will implement painting the facial features. We also demonstrate the success of this proposed research.**

Index Terms— **Cartoon face, robot painting**

I. INTRODUCTION

In recent years, most of the robots around the world have been shipped to factories. However, the application of robotics is not just limited to the industrial field [1]. A growing number of robotic researchers plunge into the educational developments, healthcare, household services, entertainment and so on [2] [3]. They all move towards the same goal to imitate human behaviors such as dancing, playing instruments, painting, and so forth [4]. Compared all these applications, painting is probably the most challenging because it depends on the knowledge of art and analytical skills involving multiple painting styles. In this work, we focus on painting human portraits with cartoon stylization, using more advanced technology to transform.

"Cartoon no.1: Substance and Shadow" was the first hand-painted cartoon created by John Leech in 1843. It parodied the preparatory cartoons for grand historical frescoes in the then-new Palace of Westminster. Leech occasionally uses cross-hatching and other realistic drawing techniques, illustrated in a black-and-white cartoon style. After a period of time, cartoons have also found their place in the world of science, mathematics and technology. From the 1920s to 1960s, dramatic cartoons were produced in enormous quantities and usually shown before a feature film in a movie theater. Disney, Fleisher, Warner Bros were the largest studios producing these 5 to 10-minute shorts. Nowadays, the number of mobile phone users in the world has increased rapidly; more and more companies have invested much capital in app development. For example, Bitstrips from Facebook, providing a template for users to choice facial features to create a new face with cartoon stylization. Another app called MomentCam, developed by Real Portrait Studio can easily transfer human portrait to cartoon style by using image-processing techniques. Nevertheless, the facial features part still looks the same as input photos. In order to show the innovation, we use components-based shape modeling and synthesis techniques to produce a "knowing smile" portrait. Furthermore, we look back to the classical roots of painting and create cartoon style portrait similar to human cartoonists draw.

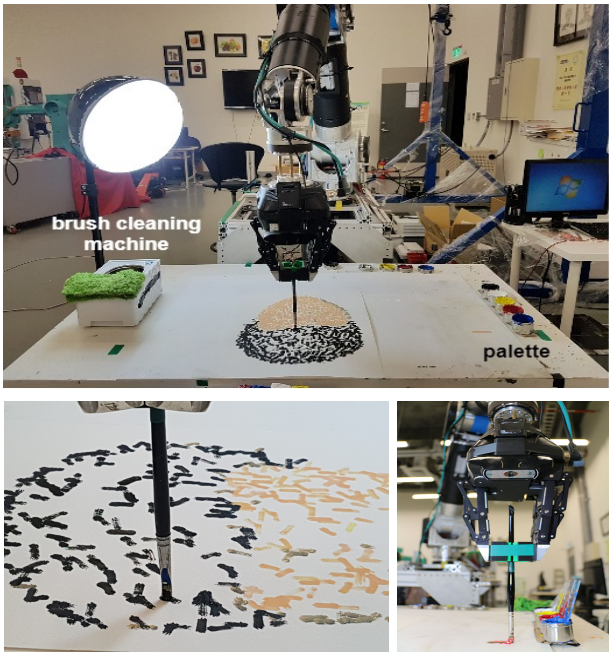

Fig. 1. (a) Top: robot painting configuration setup. (b) Bottom-left: color painting. (c) Bottom-right: visual feedback camera.

Cartoon style: Researchers studying in cartoon face often use ASMs (Active Shape Models) to analyze human faces [5], and process with a simple unshaped mask filter. In this paper, we consider providing a cartoon facial components dataset that matches the facial features. Besides, the facial features that decomposed from the faces were training through our machine learning methods.

Diversity: For the applications converting photos into cartoon style, most researchers utilize image-processed skills to create streamline and monotonous palette [6]. This makes the image more interesting and reduces the time of painting. In our work, we use a completely different idea based on

The authors are with the International Center of Excellence in Intelligent Robotics and Automation Research (iCeiRA), National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, Taiwan 106.
[1]Ren C. Luo (email: renluo@ntu.edu.tw)
[2]Yu Jung Liu (email: yrliu@ira.ee.ntu.edu.tw)

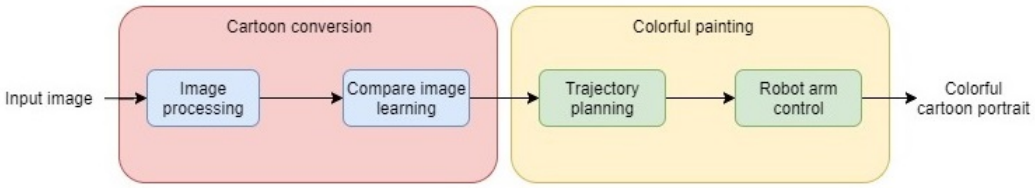

Fig. 2. System structure overview

recent machine learning techniques for image processing. It contains a variety of distinct cartoon facial features in the dataset. Moreover, it compares the difference between the two features and then replaces the previous one with similar cartoon style facial features. In addition, we can give a direct order to emerge as common facial expression to bring the diversity of the cartoon stylization.

Unlimited palette: Most colorful painting robots use predefined colors to paint their artwork, because of an algorithm and running time limitation. However, a colorful painting must retain all available colors. We use an eye-in-hand system and five basic colors to mix a variety of colors, which has the same behavior as human artists. The proposed algorithm had been illustrated in our previous work [7].

Efficiency: Frequently, photorealistic rendering algorithms need enough time to complete. Through simulation, tens of thousands of strokes can be created. To reduce the required time, we propose an effective strategy to determine the orientation of the strokes based on color gradient directions. Afterward, the robot will go through the same colors in the painting. For the darker shade of colors, we will stay at the end of the process.

The rest of the paper is organized as follows: In section II, we introduce the experimental setup such as the tools, media, converting and drawing process. Then we cover some of the implementation details of the overall system including machine learning, image processing, trajectory planning and robot arm control in section III. In section IV, we discuss the results and show the possible extensions. Eventually, we conclude in section V.

II. EXPERIMENTAL SETUP

The experimental setup is divided into two segments as shown in Fig. 1(a). The first part consists of an offline phase and runtime phase for converting portraits into cartoon style. We use Convolution Neural Network (CNN) to compare facial features images and then replace with the closest match. Fig. 1(b) shows the second part, a 50-70cm Arches paper is taped to the painting board as our canvas for its air-dried property. Different thickness brushes that used to draw distinct segments are located in an acrylic penholder. Five palette cups with acrylic paint are installed beside the painting board. When using acrylic paints, the opaque paints allow new strokes to cover the previous one for achieving the effect of the feedback system.

A seven Degree-of-Freedom (DOF) robot arm have been developed in our lab performed this experiment. The robot is equipped with a three-finger gripper actuated by a single motor at the end of the manipulator. A 3D printed clamp is attached to two fingers to help the gripper fetch the tools stably. As the robot paints different layers, the brush size varies. Once the color segment layer is finished, we provide the robot with thinner brushes to paint the facial features part, so it can describe finer details of the artwork. Additionally, an external camera is also placed at the top of the gripper as shown in Fig. 1(c). The configuration of the eye-in-hand system gives the robot an exact view of the scene and allows it to offer feedback when mixing colors. During the drawing process, the robot cartoonist will mix colors and paint on the canvas repeatedly. Moreover, a brush washing machine was set beside the robot to clean the painting brush when the chromatic error was too large for mixing colors.

Before starting the experiment, we perform calibration to provide the robots with a suitable environment. First, the position of the equipment must be fixed relative to the robots such as paper, painting board, palette cups and brush washing machine. Since the lighting of the room may affect the saturation and brightness of the color, we adjust the camera parameters so that it is less sensitive to the environment.

III. METHODOLOGY

The overall procedure of colorful cartoon portrait drawing is shown in Fig. 2. The entire painting process can be divided into two stages-cartoon conversion and colorful painting. Once the robot receives an input image, the input image is then processed by a series of image processing techniques and compare learning steps to create a cartoon style image. Next, we apply a concrete method of painting for trajectory planning and assemble all of these trajectories. The trajectories in the image domain will be transformed to robot coordinates and be painted by the robot arm. The remaining part of this section will go through the whole system in details.

A. Image Processing

The structure of image processing is shown in Fig. 3. The face detection algorithm captures a human face and is then separated into three main parts: facial decomposition, color segments and contours. Each part goes through different processes (red, green and blue in Fig. 3) and combines together to create a cartoon style portrait. We will illustrate every component in this section.

1) **Face detection:** While receiving a human portrait, we would like to make sure the face is always at the center of the image. Therefore, a common Haar features tracking algorithm [8] is used to detect our faces.

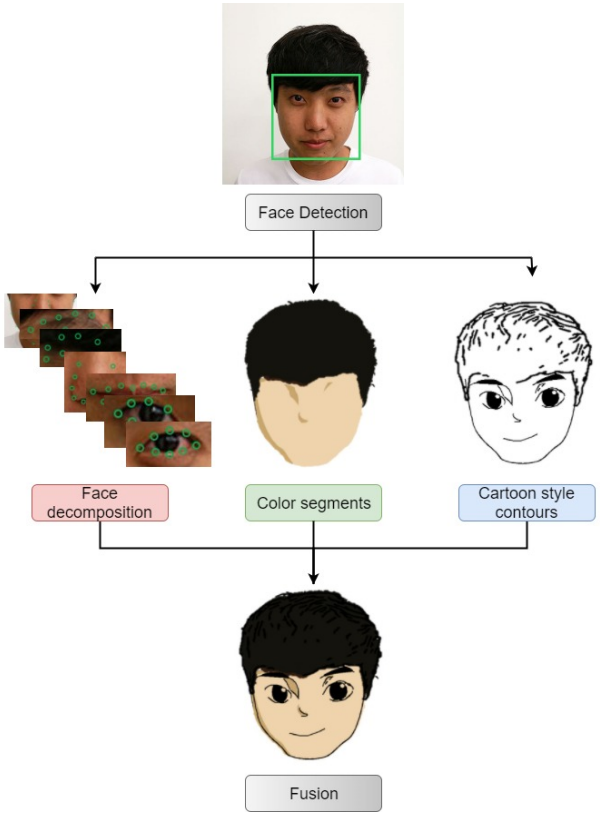

Fig. 3. The structure of image processing. (a) face detection, (b) face decomposition, (c) color segments, (d) cartoon style contours, (e) fusion.

For each feature, the optimal threshold classification function depends on a weak learner, such that the smallest number of instances being misclassified. A weak classifier $w_i(x)$ consists of a feature f_i, a threshold θ_i and a parity p_i demonstrating the direction of the inequality sign:

$$w_i(x) = \begin{cases} 1, & \text{if } p_i f_i(x) < p_i \theta_i \\ 0, & \text{if } otherwise \end{cases} \quad (1)$$

where x is a 24∗24 pixel sub-window of an image. As soon as detected, we pick out the region centered at the face and extend to the hair and shoulders for the further processing.

2) **Facial decomposition:** In this stage, we utilize the regression-based face alignment method to locate facial features [9]. The general idea of this method is to try to decompose the face into its basic components, including eyebrows, eyes, nose, mouth and chin. The bounding boxes defined by these landmarks surrounding the region of eyebrows, eyes, nose, mouth and face shape.

Face alignment by Explicit Shape Regression (ESR) [9] used boosted regression and two level cascade regression in the framework. T weak regressors $(R^1, ..., R^T)$ are combined in a superimposed manner using boosted regression. For a given face sample I and input shape S, each of the regressors computes a shape increment S^0 according to the sample features and updates the shape in a joint way,

$$S^t = S^{t-1} + R^t(I, S^{t-1}), \quad t = 1, ..., T \quad (2)$$

where $R^t(I, S^{t-1})$ are the regressors and depend on both image I and previous estimated shape S^{t-1}. We use shape index features to generate R^t as shown in Fig. 4.

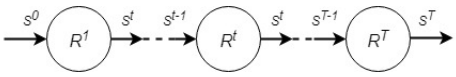

Fig. 4. Boosted regression structure.

Given N training samples $(I_i, \hat{S}_i)_{i=1}^N$, $\hat{S}_i$ represents the true shape of the i-th sample I_i. We keep training $(R^1, ..., R^T)$ until the training error is minimized by using (3).

$$R^t = \underset{R}{\operatorname{argmin}} \sum_{i=1}^{N} ||\hat{S}_i - (S_i^{t-1} + R(I, S_i^{t-1})|| \quad (3)$$

where S_i^{t-1} is the i-th estimation for the previous shape, and the output of R^t is a shape increment. However, the single regressor is too weak and the training converges slowly. To converge faster and more stably while training, a two level cascade regression is used as shown below.

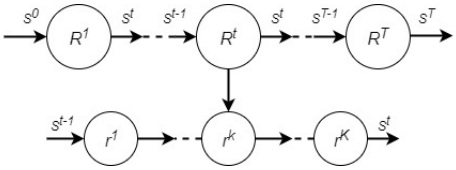

Fig. 5. Two level cascade regression structure.

The difference between the first and second levels is that the input S^{t-1} of each regressor R^t in the first level is completely different. But the input to each regressor r^k in the second layer is the same. For example, all regressor inputs in the second layer of R^t are S^{t-1}. Thus, the numerical value of the two cascades (T, K) has a significant impact on the final result. Incidentally, we tried the classical ASM method presented by Cootes and Taylor [10] before. The algorithm seems very sensitive to noise and was seriously affected when the photos were not frontal faces.

3) **Color segments:** For the cartoon of style portrait painting, we propose a strategy to simplify the original image by partitioning it into multiple segments having similar colors, which is similar to authentic cartoon portrait. In this paper, we apply mean shift algorithm described by Comaniciu and Meer [11] for the robust segmentation of color images. Mean shift is a versatile and powerful non-parametric iterative algorithm that can be used for lots of purposes like tracking, clustering and finding modes, and so on. For each input data point, mean shift associates it with the nearby peak of the probability density function of the dataset. Mean shift defines a window around a data point and computes the mean of it, then shifting the center of the algorithm until convergence. After every iteration, we can consider that the window shifts to a denser region of the dataset. According to the above description, the procedure of mean shift can be summarized as follows:

- Start with a random point and fix a window around it.
- Compute the mean of data within the window.
- Shift the window to the mean.
- Repeat until convergence.

For the application of color segmentation, colors are initially represented in the perceptually uniform color space $L*u*v*$ [12]. Mean shift is then applied to the L, u, v, x, y 5-dimensional space. The parameters of this algorithm include a spatial radius hs (similar to the radius of a filter), a color different threshold hr, and the size of the minimum acceptable region M. In this paper, the size of the original image is rescaled into 900∗600 or 600∗900 pixels. The output of this segmentation algorithm on our test image is shown in Fig. 3(c) with hs = 20 (in pixel units), hr = 6 (in $L * u * v*$ units), and M = 1500 (pixels). After the color segmentation, we recomputed the average color of each region and took it as a representation for the segment.

4) **Contours:** To draw the contour of the cartoon portrait, we use Flow-based Difference-of-Gaussians (FDoG) filtering method [13] for the extraction of facial features and apply morphological operations to remove tiny and trivial edges.

a) Coherent lines Extraction: Cannys method [14] is generally considered the standard for edge detection and is usually used to extract the contours of the human portrait in other works of robot drawing. It detects discontinuities of intensity by checking the zero-crossing od second directional derivative of a smoothed image. Since it doesn't consider the flow of the image, the relation between an edge and its neighbors, the detected contours may be trivial and discontinuous. In this paper, FDoG filter is used to improve the quality of the contours. The main idea is to take into account the direction of the local edge tangent flow (ETF) and apply a DoG [15] filter in the direction perpendicular to the local edge flow. DoG is done based on the flow-based kernel to find enough pixels that fit the important lines. First, we apply each pixel along the perpendicular of ETF to do one-dimensional filter in $[-T, T]$ (Fig. 6):

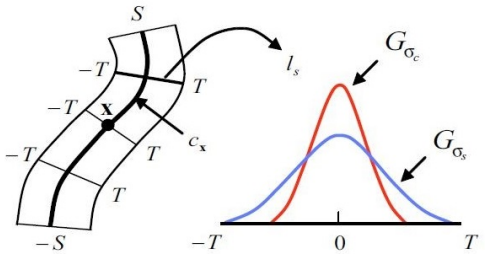

Fig. 6. Flow-based DoG filtering. Left: Kernel enlarged. Right: Gaussian components for DoG.

$$F(s) = \int_{-T}^{T} I(l_s(t)) f(t) dt \tag{4}$$

where $f(t) = G_{\sigma_c}(t) - \rho G_{\sigma_s}(t)$ is the DoG, then do one-dimensional Gaussian in $[-S, S]$:

$$G(x) = \int_{-S}^{S} G_\sigma(s) F(s) ds \tag{5}$$

Such an approach not only makes the edges more consistent (without too many short lines) but suppresses noise so as to improve line drawing characteristics.

b) Morphological Operation: To accomplish simple trajectories from the extracted contours, we apply the first thinning algorithm [16] to reduce all lines to single pixel thickness. We also perform close operation [17] to remove tiny and isolated edges which may be too trivial for robot painting. Afterward, the lengths will guide their edges. The result of contours extraction is shown in Fig. 3(d) and Fig. 7. As far as we can see, FDoG filtering method conveys more coherent and important features of faces than Cannys, which is a suitable solution for us to enhance the quality of our portrait.

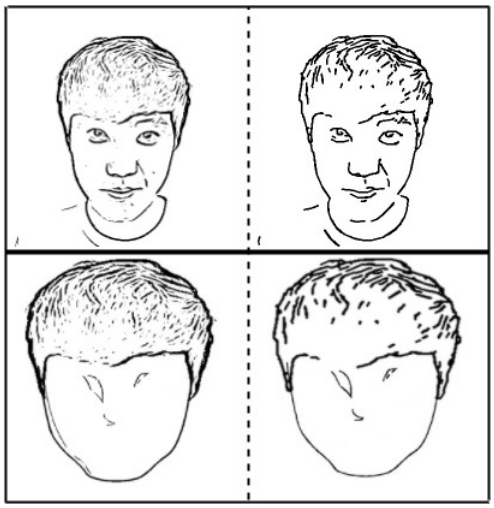

Fig. 7. Contour extraction. Left: Cannys method. Right: FDoG filtering method.

5) **Fusion:** Finally, we combine color segments with no facial feature and contours of the cartoon facial features to create a non-photorealistic rendering cartoon style portrait as shown in Fig. 3(e).

B. Compare Image Learning

In order to create cartoon style facial features. We used compared image learning method [18] to deal with the features we extracted and the cartoon style database. The method was based on Siamese network [19] [20] and used Spatial Pyramid Pooling (SPP) [21] to achieve different size pictures of the input network. Our purpose is to compare the similarity of two images in Fig. 8.

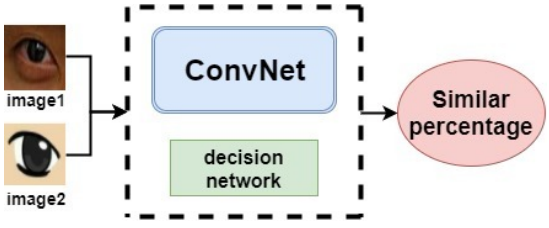

Fig. 8. Learn a general similar percentage for image.

Therefore, we construct the convolution neural network model input as two pictures. The Siamese network consists of two branches, image 1 and image 2 are two independent single-channel gray scale images. The main idea of this method is to combine two images into a two-channel image. Two $(1, 64, 64)$ single-channel data are concatenated together as a $(2, 64, 64)$ dual-channel matrix, and then this matrix is used as the network input. The second improvement is Central-surround two-stream network. Suppose we enter a $64*64$ image, the Central-surround two-stream network [18] will transfer the image into two $32 * 32$ pictures for the input of the network. It turns the image into multi-scale input, which can improve the match effect and accuracy of two images. The final baptism is, the existing convolutional neural network, the input layer of the picture size is generally fixed such as $32 * 32$ or $96 * 96$. However, SPP algorithm solves the problem while the training data is a variety of image sizes and also improve network availability, robustness and so on. After the above three evolutions, the final structure of the network, is the highest precision algorithm: 2-channel + Central-surround two-stream + SPP network structure.

C. Trajectory Planning

In this part, a set of trajectories is extracted from the caricatured image for robot painting.

1) **No facial feature painting:** After color segmentation, we paint our colorful cartoon portrait without the facial feature. We determine several characteristics to generate a set of hand-painted stroke, including its orientation, length and radius of brushes. Practically, the stroke orientation can be automatically painted by gradient direction that has the least change. The idea behind this concept follows the applications of Non-photorealistic rendering (NPR) [22] [23]. The gradients G_x,G_y along x and y axes are obtained from the input image to get the orientations of the strokes. Therefore, the orientations are computed as $arctan(G_x, G_y)+/2$. Where 90 is added to make the strokes perpendicular to the gradient direction. For making artworks looks normal, we give the stroke a basic length plus a certain amount of disturbance. In every iteration of refinement, we slowly use a thinner brush so as to descript more details of the face.

2) **Cartoon facial feature painting:** Once the no facial features painting is finished, the contours of cartoon face are drawn with a black pigment or marker, starting from a rough silhouette, and add iteratively smaller details. The contours are interpolated with a third order spline to smooth the trajectories.

D. Robot Arm Control

The experiment is performed on a 7-DoF robot arm equipped with a three-finger gripper actuated by a single motor at the end of the manipulator. The control architecture of robot arm is shown in Fig. 9. A PC running RTOS is the control center of the whole system. It communicates with the hardware components such as the motor driver and the motor encoder through a DAQ card installed on the PCI bus. The RTOS is required since we would like to do the motor position control on the PC running at 1kHz. The DAQ card is installed on the PCI bus of the PC, endowing the PC with the capability of sending an analog voltage signal to control the motor driver and reading the encoder counted from the motor. When the analog voltage signal from the DAQ card is received, the motor driver will output a corresponding current to drive the motor. This implies that the current control loop is done internally by the motor driver, outputting a torque to actuate the robot arm. The output torque equals to the product of the input current and the torque constant of the motor. The rotating angle will be count by the encoder-equipped on the motor and then feedback to the PC through the DAQ card. This closes the loop so that the accurate position or velocity control can be done on the PC.

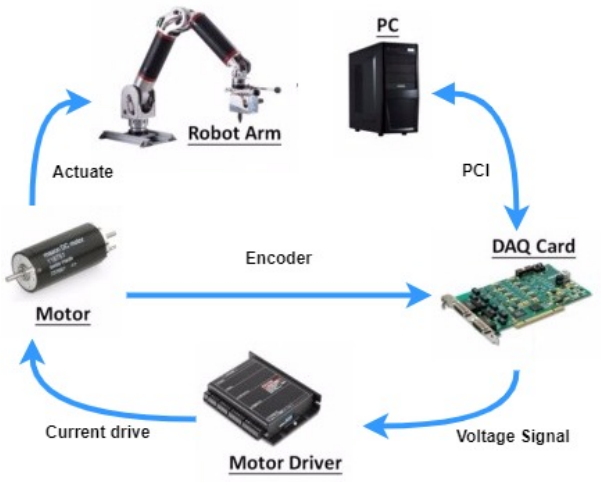

Fig. 9. Control architecture

IV. RESULT, DISCUSSION AND FUTURE WORK

Several simulations with different facial portraits and expressions composed of the facial features of our database are shown in Fig. 10. The painting progress is shown in Fig. 11. The top-left one is the first layer and the continued images are subsequent layers created in the refinement stage. Then, the robot will end up with cartoon facial contours. The first and the refinement stage takes approximately one to two hours which depend on the complexity of the pictures. We successfully present a hand-painted style, which is more similar to the artworks created by human beings. The scheme for color processing and chosen facial features gives pleasing results. The advantages of using mixed media lie in the complementarity of different media. However, it takes effort to completely cover the whole area.

Although the robot cartoonist has not yet been able to paint like a professional cartoonist, we demonstrate the innovation and more applications for compare learning algorithm. We also show generality and robustness of our approach by applying it to a variety of human portrait and compare our output with stylized results created by cartoonist via a comprehensive user study. Nevertheless, we planned to work on many possible issues that can be improved and look

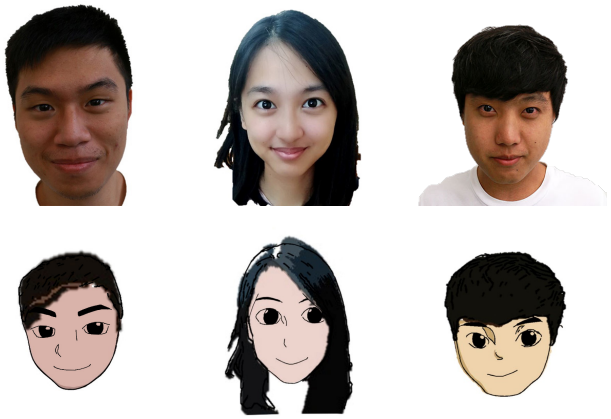

Fig. 10. Simulations.

forward to integrating with multiple databases and generate visually convincing facial cartoon images with various styles. In the future, we hope that this application is not only for entertainment, but also can be applied to educational purposes. For example, robots can systematically demonstrate painting techniques to teach beginners in a scientific way.

Fig. 11. Painting in different layers

V. CONCLUSION

This paper presents a realization of a robot capable of painting colorful cartoon portrait in the artistic way. There are two main contributions in this work. First, we break the limitation that robot creates just monochromatic works and use the only single medium. Second, we integrate multidisciplinary knowledge, involving image processing, computer vision, machine learning and robotics, and aesthetics, is unique and complete. The integration of different fields enables us to deal with problems with different perspectives and create unexpected results.

REFERENCES

[1] I. Sassi, A. Benabdelhafid, and S. Hammami, "Industrial ecosystem of the territory: Strategies and perspectives," in *Service Operations And Logistics, And Informatics (SOLI), 2015 IEEE International Conference on*, pp. 216–219, IEEE, 2015.

[2] T. Shibata, "An overview of human interactive robots for psychological enrichment," *Proceedings of the IEEE*, vol. 92, no. 11, pp. 1749–1758, 2004.

[3] J. Schulte and D. Schulz, "Mlnerva: A second generation mobile tour guide robot," *Roc. of*.

[4] Y. Kuroki, M. Fujita, T. Ishida, K. Nagasaka, and J. Yamaguchi, "A small biped entertainment robot exploring attractive applications," in *Robotics and Automation, 2003. Proceedings. ICRA'03. IEEE International Conference on*, vol. 1, pp. 471–476, IEEE, 2003.

[5] C. A. Behaine and J. Scharcanski, "Enhancing the performance of active shape models in face recognition applications," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 8, pp. 2330–2333, 2012.

[6] O. Deussen, T. Lindemeier, S. Pirk, and M. Tautzenberger, "Feedback-guided stroke placement for a painting machine," in *Proceedings of the Eighth Annual Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging*, pp. 25–33, Eurographics Association, 2012.

[7] R. C. Luo, M.-J. Hong, and P.-C. Chung, "Robot artist for colorful picture painting with visual control system," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pp. 2998–3003, IEEE, 2016.

[8] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[9] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[10] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[11] 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.

[12] J. F. Hughes, A. Van Dam, J. D. Foley, M. McGuire, S. K. Feiner, D. F. Sklar, and K. Akeley, *Computer graphics: principles and practice*. Pearson Education, 2014.

[13] H. Kang, S. Lee, and C. K. Chui, "Coherent line drawing," in *Proceedings of the 5th international symposium on Non-photorealistic animation and rendering*, pp. 43–50, ACM, 2007.

[14] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.

[15] B. Gooch, E. Reinhard, and A. Gooch, "Human facial illustrations: Creation and psychophysical evaluation," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 1, pp. 27–44, 2004.

[16] T. Y. Zhang and C. Y. Suen, "A fast parallel algorithm for thinning digital patterns," in *Communication of the ACM 27*, pp. 236–239, 1984.

[17] J. Serra, "Image analysis and mathematical morphology," in *New York: Academic Press*, 1982.

[18] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4353–4361, 2015.

[19] J. Bromley, I. Guyon, Y. LeCun, E. Sickinger, and R. Shah, "Signature verification using a siamese time delay neural network," in *Advances in Neural Information Processing Systems*, pp. 737–744, 1993.

[20] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively with application to face verification," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 539–546, 2005.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1904–1916, 2015.

[22] P. Litwinowicz, "Processing images and video for an impressionist effet," in *SIGGRAPH*, 1997.

[23] J. P. Collomosse and P. M. Hall, "Painterly rendering using image salience," in *Proceedings of the 20th UK conference on Eurographics*, p. 122, 2002.