Synthesizing Emerging Images from Photographs

Cheng-Han Yang
Dept. of Computer Science
National Tsing Hua University
Hsinchu 300, Taiwan
b9932028@gmail.com

Ying-Miao Kuo
Dept. of Computer Science
National Tsing Hua University
Hsinchu 300, Taiwan
jollytreeskuo@gmail.com

Hung-Kuo Chu
Dept. of Computer Science
National Tsing Hua University
Hsinchu 300, Taiwan
hkchu@cs.nthu.edu.tw

ABSTRACT

Emergence is the visual phenomenon by which humans recognize the objects in a seemingly noisy image through aggregating information from meaningless pieces and perceiving a whole that is meaningful. Such an unique mental skill renders emergence an effective scheme to tell humans and machines apart. A recent state-of-the-art work proposes to synthesize images of 3D objects that are detectable by human but difficult for an automatic algorithm to recognize. However, using 3D objects as inputs brings drawbacks. For instance, the quality of results is sensitive to the viewing and lighting conditions in the 3D domain. The available resources of 3D models are usually limited, and thus restricts the scalability. This paper presents a novel synthesis technique to automatically generate emerging images from regular photographs, which are commonly taken with decent setting and widely accessible online. We adapt the previous system to the 2D setting of input photographs and develop a set of image-based operations. Our algorithm is also designed to support the difficulty level control of resultant images through a limited set of parameters. We conducted several experiments to validate the efficacy and efficiency of our system.

Keywords

Emergence, Gestalt psychology, image segmentation, image synthesis

1. INTRODUCTION

Emergence refers to the unique mental skill of humans to perceive objects in a seemingly noisy image not by recognizing the object parts, but as a whole. A classic example of emerging image can be found in Figure 1. Typically, the local windows in such an image reveal only meaningless, complex and random patterns. Only when observed in its entirety, the main subject (e.g., The Dalmatian dog) suddenly *emerges* and is perceived as a whole. Although this special phenomenon has been studied extensively in Gestalt psychology with simple synthetic patterns, the process of how exactly the humans perceive complex objects is still unknown. It means that modeling an automatic recognition will be extremely challenging, or not even possible. This makes emergence an effective scheme for Turing test (e.g., Captcha [14]) to distinguish between humans and automated machines. Moreover, a system that can synthesize an infinite number of emerging images can be valuable for exploring and understanding the factors involving in both cognitive psychology and computer vision techniques.

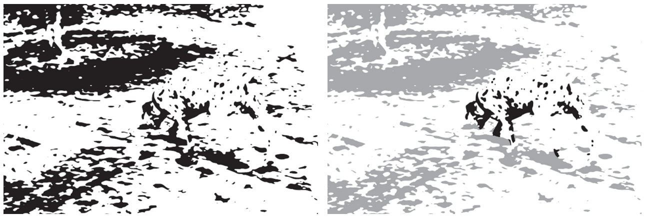

Figure 1: *A classic example of an emerging image by R. C. James. Although the left image reveals meaningless patterns in small local neighborhoods, we can still perceive a Dalmatian dog sniffing around as shown on the right.*

Inspired by the Dalmatian dog image (see Figure 1), Mitra *et al.* [12] presented a pioneer work to investigate the problem of synthesizing images containing subjects that can be recognized by humans, but are extremely difficult for automatic machines. Their system takes 3D objects with well-defined viewpoint and lighting as inputs, and renders the 3D scene according to two guiding principles. To generate images that are easy for human, the system renders complex splats that capture the silhouette and shading of subjects. Those long coherent splats are further broken and utilized to introduce clutter in the remainder of the image to make segmentation harder for machines. Results are validated by conducting experiments to test against humans and state-of-the-art computer vision methods. However, using 3D objects as inputs imposes several limitations: (i) the image quality may degrade quickly as the viewing and lighting conditions changing in the 3D domain; (ii) available resources of 3D model repository are limited, and (iii) operations on the 3D domain are computationally expensive. Thus, the existing system is still far behind a practical and scalable solution for a massive production of emerging images.

In this work, we present an automatic method that synthesizes emerging images from regular photographs. The idea is mainly motivated by two observations. First, it is much more easier to acquire photographs than 3D models. For example,

MM '16, October 15-19, 2016, Amsterdam, Netherlands
© 2016 ACM. ISBN 978-1-4503-3603-1/16/10...$15.00
DOI: http://dx.doi.org/10.1145/2964284.2967304

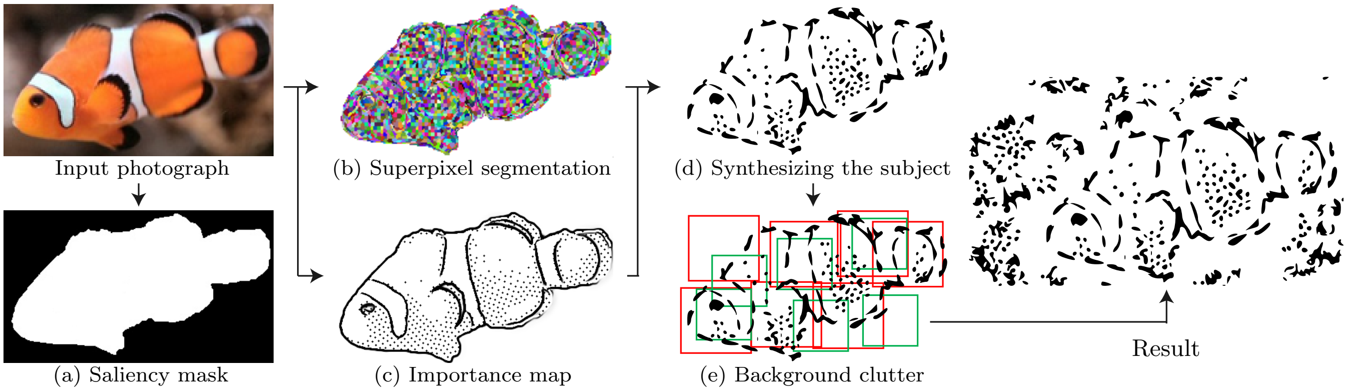

Figure 2: *System overview. Given an input photograph, our system first extracts the main subject based on the detected image saliency. Then the foreground subject is over-segmented into superpixels, which correspond to unit splat used for rendering. An importance map that captures edge features and luminance distribution of subject is computed. To synthesize the subject, we use the importance map to determine which superpixels should be colored in black, followed by breaking long coherent splats and performing spatial perturbation. Lastly, we adopt a simple copy-perturb-paste algorithm to copy splats within random windows on the subject to fill the remainder parts of image, and obtain the final result.*

we can effortlessly take the photographs via smart phone or download from public photo sharing spaces such as flickr. Second, most of photographs, especially those taken by professionals, are composed of clear subjects with decent viewpoint and lighting. To generate effective emerging images, our system follows the model of previous system, and lies the technical novelty in developing a set of operations tailored for 2D image setting. The algorithm is carefully designed to guarantee the synthesized subjects can be recognized using the principle of emergence, at controllable difficulty levels.

We tested our system on a wide variety of input photographs, ranging from human portrait, animals, cartoon, etc, and generated over 900 emerging images so far. Please refer to the supplementary file for a subset of 100 examples. For validation, we performed a user study to evaluate the effectiveness of our results. We also conducted experiments to evaluate how the performance of current state-of-the-art object recognition methods is impaired by our results.

2. RELATED WORK

How the humans perceive subjects in the emerging images is closely related to Gestalt theory, which consists of several principles to describe the relationships between the parts and the whole [10]. Among all the principles, our system is based on two principles, *closure* and *continuity*, to design algorithm for synthesizing emerging images. Other principles such as symmetry, proximity, which have been extensively studied, are worthy exploring to provide additional cues for recognition.

There are other kinds of drawings that illustrate interesting emergence in different contexts. Gal *et al.* [7] present a 3D collage system for creating compound shapes by juxtaposing different fruit models to mimic Arcimboldo's painting. The idea is further extended by Huang *et al.* [8] who utilize the internet images to compose a 2D figure that resembles the input image. Yoon *et al.* [8] present a hidden-picture puzzle generator that aims at finding suitable places to hide small objects in a clutter background image. Chu *et al.* [4] propose to camouflage a target subject by synthesizing its appearance using the textures sampled from the surrounding. Rather working in the spatial domain, Tong *et al.* [13] show that comparable camouflage images can be created through the composition in the frequency domain. Note that while above works all aim to conceal the subjects in the images to a certain degree, they are not designed to model the special phenomenon of emergence where the local neighborhoods reveal little.

Captcha, a kind of Turning test to distinguish between humans and automated machines, was originated from the work of Ahn *et al.* [14]. Not surprisingly, the intriguing nature of emergence leads to a potential new Captcha scheme that might outperform existing text-based systems. With the supports of publicity available photo database, our system can benefit the Captcha applications in terms of puzzles numbers and instant feedbacks. However, there are other system-level considerations for building a mature Captcha scheme, which is beyond the scope of this work.

3. ALGORITHM

Preprocessing. Given an input photograph, our system first extracts the foreground subject using automatic saliency cut [3] (see Figure 2(a)). The user can optionally provide scribbles to refine the saliency mask. Then it performs an image over-segmentation [1] to obtain superpixels on the foreground subject (see Figure 2(b)). We refer to each superpixel as a *splat*, which is used as a basic rendering unit. The problem of synthesizing an emerging foreground is formulated as assigning binary color (black or white) to these splats. Note that the size (d_s) and compactness (d_c) of superpixels might vary as changing the image resolution. For the sake of simplicity, we resize the input image to fit the resolution of 640 × 480 and use the setting of $d_s = 30$ pixels and $d_c = 4$ across the experiments.

Importance map. The importance map captures the necessary visual cues of subject for humans to recognize, and is constructed using two feature maps (see Figure 2(c)). The first one represents the salient edge features computed by applying the flow-based image filter [11]. The other one models the luminance distribution (e.g., shading) with a set of sampling points. This is done by applying the Poisson Disk Sampling [2] on the luminance map of subject. The density of sampling points is controlled by α and β, indicating respectively the minimum and maximum valid distance among sampling points.

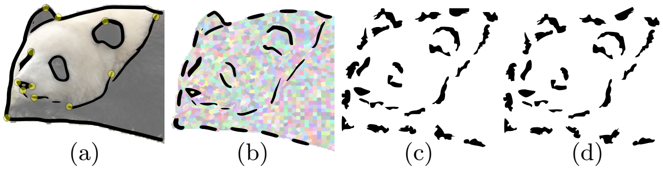

Figure 3: *Edge perturbation. (a) Original image and edge map. The yellow circles indicate the detected corners. (b) The edges are broken into line segments and (c) the corresponding complex splats. (d) Applying 2D space perturbation to complex splats.*

Synthesizing the subject. Once we have the superpixels and importance map, an initial rendering of subjects is obtained by superimposing two maps and assigning black to those superpixels hit by the black pixels in the importance map. We refer to the grouping of superpixels as *complex splats*. However, such a naive approach reveals important information about subject's silhouette, which can possibly be identified by machines using boundary extraction (see Figure 3(a)). Hence, we break long coherent complex splats along silhouettes into smaller parts in a manner such that the machines can not restore the original shape via interpolating and extrapolating between parts. To this end, we first compute central line segments of the complex splats using image thinning and detect corners on the line segments. Then, starting from each corner point, we move along the line segments and iteratively break the complex splat by removing the atomic splats in a controlled manner. More specifically, we use two parameters a and b to control the length of retained complex splats and the gap between consecutive complex splats, respectively. An example of edge breaking can be found in Figure 3(b,c). To null the chance of using interpolation and extrapolation algorithm, we further apply a 2D space perturbation with random rotation and positional jitters to the complex splats (see Figure 3(d)).

Background clutter. In this step, we employ the *copy-perturb-paste* procedure by Mitra *et al.* [12] to fill the remainder parts of image. In brief, we add controlled clutter to make it harder for machines to identify where the subject may lie by making the rest of the image look similar when observed from a local window. We simply copy small regions of the subjects, randomly perturb the interior splats, and paste the them to other regions of image. During the procedure, we further apply a repel field to cancel the paste of background splats if they are close to the subject contour. A parameter d', representing the minimum proximity distance, is used to control the difficulty level.

Difficulty level control. As we can see our algorithm involves several stages, some of which can be used to control the difficulty level of synthesized emerging images. They are: (i) sampling density of importance map ($[\alpha, \beta]$); (ii) degree of edge perturbation ($[a, b], \theta, \delta$); and (iii) background clutter (d). Table 1 lists the default values used for generating emerging images at easy, medium, and difficult settings.

4. EXPERIMENT AND RESULTS

We have tested our system with a wide variety of input photographs, ranging from human portrait to man-made objects. Our algorithm is efficient, which takes less than 0.3 second to generate an emerging image from a photograph with resolution 640 × 480 using a moderate PC. This enable us to interactively craw photographs from internet and generate corresponding emerging images. As a result, we have generated over 900 emerging images at three difficulty levels. Figure 5 shows six examples generated by our system. Please refer to the supplementary file for a collection of 100 examples. In the following, we conducted experiments to validate the effectiveness of our system.

	easy	medium	difficult
edge perturbation $[a\text{:}b]$*D (preserve:break length(pixels))	[0.06:0.04]	[0.04:0.06]	[0.02:0.08]
sampling density distance $[\alpha\text{:}\beta]$*D (minimum:maximum (pixels))	[0.01:0.10]	[0.02:0.08]	[0.03:0.06]
rand. perturb. angle θ (degree)	[-10,10]	[-15,15]	[-20,20]
rand. perturb. displacement δ*D (along x, y direction (pixels))	0.01	0.02	0.03
rand. repel field distance d*D (pixels)	[0.03:0.02]	[0.02:0.01]	[0.01:0.00]

Table 1: *Parameters used to generate emerging images at easy, medium and difficult levels. D is the diagonal length of subject's bounding box.*

Exp I: user study. A good emerging image synthesis algorithm should generate images reliably perceived by humans, at an easily controllable difficulty level. To evaluate our algorithm, we conducted a user study involving 20 participants. Users were shown a sequence of 20 images with different subjects (only one subject per image) and are synthesized at different difficulty levels. During the trial, the users were asked to type what they saw. Elapsed time starting from showing the image until the user type her answer is recoded as response time. Each image was shown for a maximum of one minute.

As shown in Figure 4 (Left), the user response indicates that we can generate perceivable emerging images at controllable difficulty levels. We use one-way analysis of variance (ANOVA) [15] to evaluate the statistical significance of the results. The analysis shows that the recognition accuracy is significant at the $p < 0.05$ level, $F(2,398) = 3.51$. The response time is significant at the $p < 0.01$ level, $F(2,398) = 494$. As far as the recognition accuracy and the response time are concerned, the correlation is significant with controllable difficulty levels.

Exp II: validation with machines. We test our results against two popular learning based systems, Caffe [9] and Bag-of-features [5], to evaluate how difficult the current state-of-the-art to crack the emergence.

Caffe [9] is a deep learning framework made with expression, speed, and modularity. We prepared 50 examples, which are used as queries to Caffe for recognizing the sub-

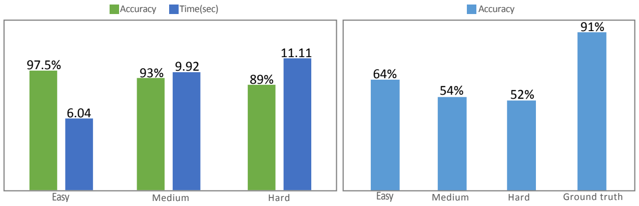

Figure 4: *(Left) Recognition accuracy and response time on three difficulty levels of generated emerging images as observed in course of our user study. (Right) Performance of a learning based image recognition systems operating on regular and emerging images.*

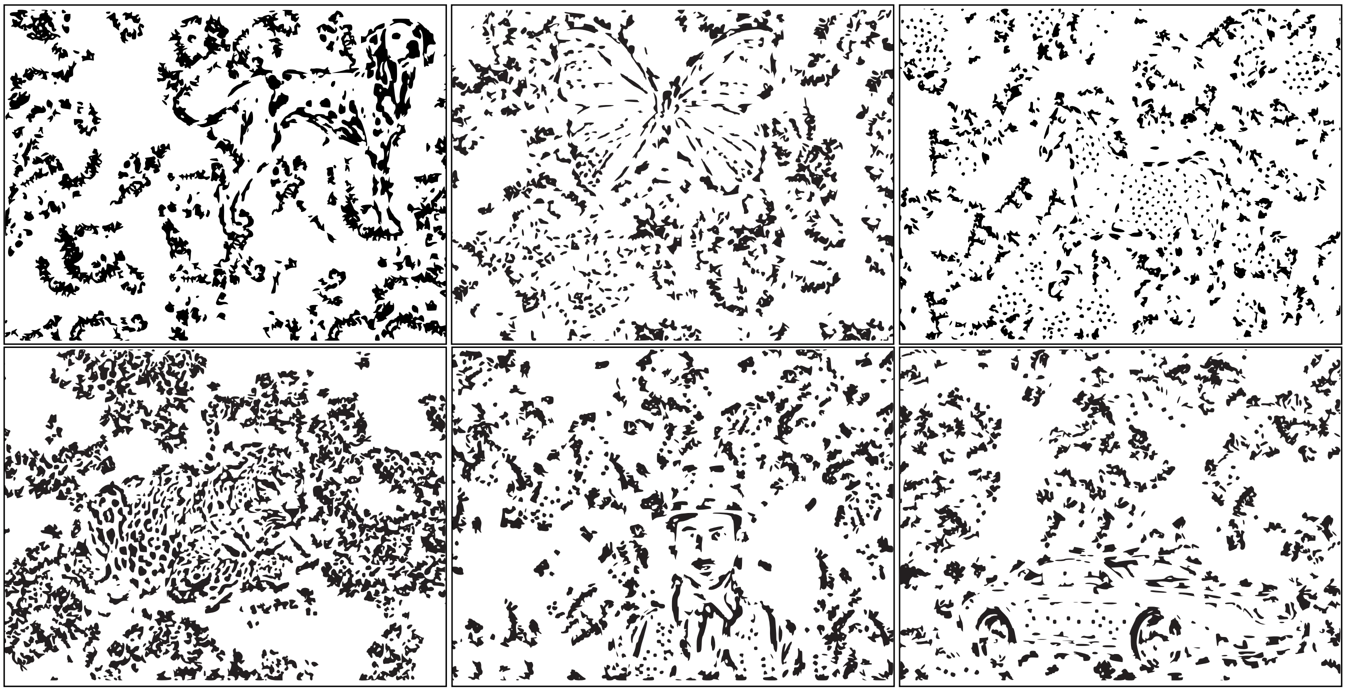

Figure 5: *Six emerging images generated by our system at, from left column to right column, easy, medium, and hard difficulty levels.*

jects therein based on the pre-defined models. The result indicates that while a 100 percent accuracy was returned by Caffe when using the original image as queries, Caffe failed to identify any of subjects in the emerging image even at easy difficulty level.

Bag-of-features is a kind of visual image categorization that assigns a category label to a query image. To learn bag of features used for image category classification, we used the most widely used Caltech 101 [6] as target image data set. We selected the three categories (cat, hawksbill and leopards), each of which contains 100 images using [training: verification] with [4:6] ratio, randomly picked for extraction and classification features. For each image, we generated corresponding emerging images at three three difficulty levels. The verification results are shown in Figure 4 (Right), indicating that the performance of system drops significantly (30% ∼ 40%) when dealing with our results.

Limitation. Our system still has several limitations. (i) The quality of synthesized image depends on the quality of extracted saliency mask (see Figure 6(a)). (ii) If the foreground subject occupies a large portion of image regions, the difficult control is less effective due to the central-fixation (see Figure 6(b)). (iii) Our system can not generate effective results for the subjects with textureless appearance (see Figure 6(c)).

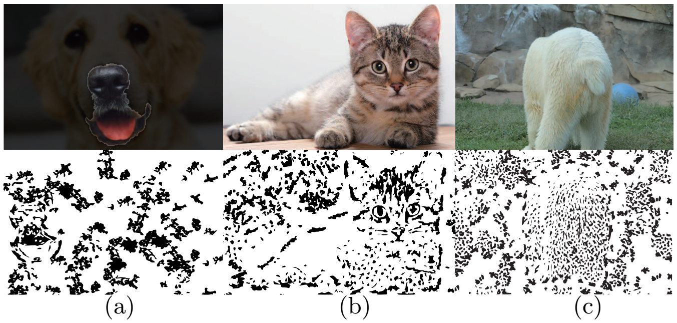

Figure 6: *Limitations. (a) Saliency cut fails to extract correct foreground mask. (b) Central-fixation problem. (c) Textureless subject.*

5. CONCLUSIONS

We present an automatic method to synthesize emerging images from photographs at controllable difficulty level. We conducted several experiments to evaluate the efficiency and efficacy of our system. The experimental results shows that our system is efficient and is capable of generating a massive production of emerging images, which may benefit to future development of new Captcha system.

Acknowledgements

We thank the anonymous reviewers for their invaluable comments and suggestions. The project was supported in part by the Ministry of Science and Technology of Taiwan (102-2221-E-007-055-MY3, 103-2221-E-007-065-MY3 and 105-2221-E-007-104-MY2).

6. REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. SÃijsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, 2012.

[2] R. Bridson. Fast poisson disk sampling in arbitrary dimensions. In *ACM Trans. Graph. (Proc. SIGGRAPH)*. ACM, 2007.

[3] M. M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S. M. Hu. Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(3):569–582, March 2015.

[4] H.-K. Chu, W.-H. Hsu, N. J. Mitra, D. Cohen-Or, T.-T. Wong, and T.-Y. Lee. Camouflage images. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 29:51:1–51:8, 2010.

[5] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV*, volume 1, pages 1–2. Prague, 2004.

[6] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[7] R. Gal, O. Sorkine, T. Popa, A. Sheffer, and D. Cohen-Or. 3D collage: Expressive non-realistic modeling. In *Proc. of NPAR*, page 14. ACM, 2007.

[8] H. Huang, L. Zhang, and H.-C. Zhang. Arcimboldo-like collage using internet images. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 30(6):155:1–155:8, Dec. 2011.

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[10] G. Kanizsa. *Organization in Vision: Essays on Gestalt Perception*. Praeger New York, 1979.

[11] J. E. Kyprianidis and J. Döllner. Image abstraction by structure adaptive filtering. In *Proc. EG UK Theory and Practice of Computer Graphics*, pages 51ÂřV–58, 2008.

[12] N. J. Mitra, H.-K. Chu, T.-Y. Lee, L. Wolf, H. Yeshurun, and D. Cohen-Or. Emerging images. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 28(5):163:1–163:8, 2009.

[13] Q. Tong, S.-H. Zhang, S.-M. Hu, and R. R. Martin. Hidden images. In *Proc. of NPAR*, pages 27–34. ACM, 2011.

[14] L. von Ahn, M. Blum, and J. Langford. Telling humans and computers apart automatically. *Commun. ACM*, 47(2):56–60, 2004.

[15] B. Watson, A. Friedman, and A. McGaffey. Measuring and predicting visual fidelity. In *ACM Trans. Graph. (Proc. SIGGRAPH)*, pages 213–220. ACM, 2001.