Emerging Images

Niloy J. Mitra* Hung-Kuo Chu† Tong-Yee Lee† Lior Wolf§ Hezy Yeshurun§ Daniel Cohen-Or§

*IIT Delhi / KAUST †National Cheng Kung University, Taiwan §Tel Aviv University

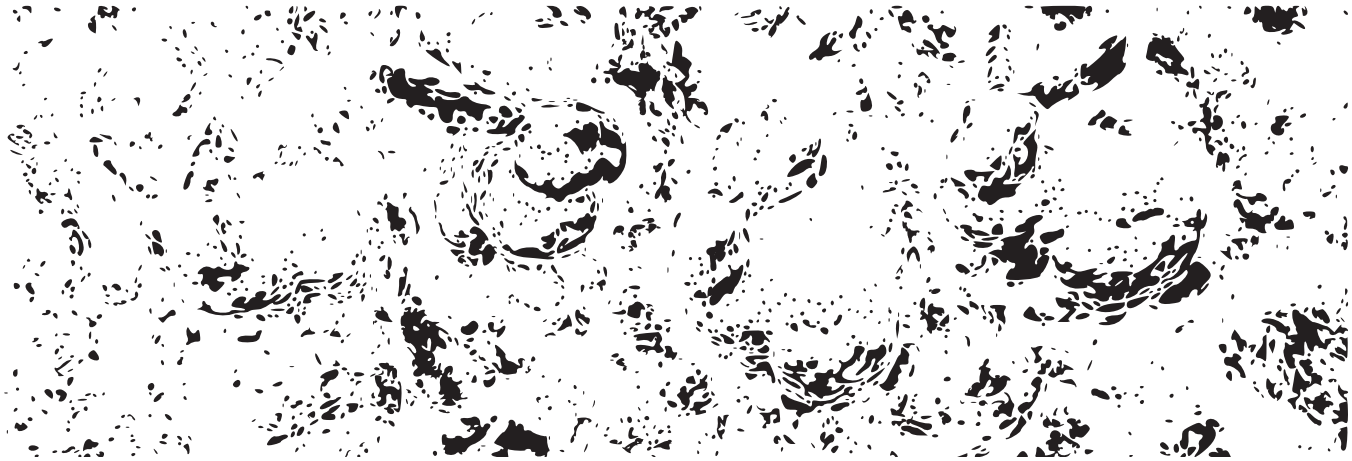

Figure 1: *This image, when stared at for a while, can reveal four instances of a familiar figure. Two of the figures are easier to detect than the others. Locally there is little meaningful information, and we perceive the figures only when observing the whole figures.*

Abstract

Emergence refers to the unique human ability to aggregate information from seemingly meaningless pieces, and to perceive a whole that is meaningful. This special skill of humans can constitute an effective scheme to tell humans and machines apart. This paper presents a synthesis technique to generate images of 3D objects that are detectable by humans, but difficult for an automatic algorithm to recognize. The technique allows generating an infinite number of images with emerging figures. Our algorithm is designed so that locally the synthesized images divulge little useful information or cues to assist any segmentation or recognition procedure. Therefore, as we demonstrate, computer vision algorithms are incapable of effectively processing such images. However, when a human observer is presented with an emergence image, synthesized using an object she is familiar with, the figure emerges when observed as a whole. We can control the difficulty level of perceiving the emergence effect through a limited set of parameters. A procedure that synthesizes emergence images can be an effective tool for exploring and understanding the factors affecting computer vision techniques.

1 Introduction

Emergence is the phenomenon by which we perceive objects in an image not by recognizing the object parts, but as a whole, all at once. In small local neighborhoods the image parts look meaningless, complex and random. However, when observed in its entirety, the main subject in the image suddenly *pops out* and is perceived as a whole. Although this phenomenon, originally popularized by the Gestalt school, has been well studied, the exact process of how we perceive such objects is not known. Lack of understanding of how humans perceive such forms, means that currently it is extremely challenging, if not impossible, to automate the recognition process. This makes emergence a good blind test, also known as Captcha [von Ahn et al. 2004], to distinguish between a human and an automated agent, commonly referred to as a *bot*.

Motivated by the unique mental skill of humans to perceive figures from meaningless parts and the need for more reliable captcha schemes, we investigate the problem of synthesizing images containing subjects that can be recognized by a human, but at the same time extremely difficult for a bot. Striking examples of emergence occur when there are no long, coherent boundaries that separate an object from its background. Humans cannot instantaneously detect the object in such images, and can probably recognize it only after several iterations that take into account numerous relationships between hypothetical objects and their context. The computational complexity of this human processing is believed to be extremely high [Tsotsos 1992], leading us to hypothesize that emergence images are hard for automatic algorithms to segment, identify, and recognize. Taking into account the complexity of the task, and the

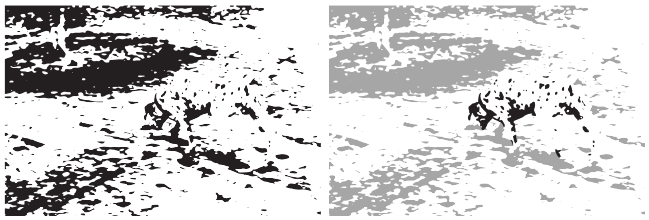

Figure 2: *A classic example of an emergence image. Although at first sight the left image looks meaningless, suddenly we perceive the central object as the Dalmatian dog* pops out.

ACM Reference Format
Mitra, N., Chu, H., Lee, T., Wolf, L., Yeshurun, H., Cohen-Or, D. 2009. Emerging Images. *ACM Trans. Graph. 28*, 5, Article 163 (December 2009), 8 pages. DOI = 10.1145/1618452.1618509 http://doi.acm.org/10.1145/1618452.1618509.

Copyright Notice
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or direct commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701, fax +1 (212) 869-0481, or permissions@acm.org.
© 2009 ACM 0730-0301/2009/05-ART163 $10.00 DOI 10.1145/1618452.1618509 http://doi.acm.org/10.1145/1618452.1618509

lack of a clear understanding of how humans solve the problem, it is highly unlikely, if not impossible, that these type of tasks could be carried out by bots in the near future.

Our method is inspired by the well known image of the Dalmatian dog by R. C. James, shown in Figure 2. This image is probably the best known demonstration of the emergence effect. Detecting the dog in this image is quite hard for humans, but definitely by far harder for a bot. We present an automatic method that synthesizes such images by rendering 3D scenes that include a subject, which is recognized using the principle of emergence. Since emergence mechanism as perceived by humans is not well understood and unclear how to model, we perform a user study to validate the effectiveness of our results. We also conduct experiments to evaluate the difficulty of the problem for state-of-the-art computer vision methods.

Contribution. We present an automatic algorithm for creating emergence images (and videos) with controllable level of difficulty. Such a synthesized image locally appear as noise, while revealing itself to a human observer when viewed as a whole. This holds potential for generating controlled test data for computer vision algorithms, as well as creating puzzles to distinguish humans from bots.

2 Background

Humans are highly skilled in detecting and identifying 3D objects even from a monocular view. Given prior knowledge and experience, we can effortlessly recognize shapes while factoring out distortions due to camera projection, lighting, occlusion, etc. Although the computer vision and pattern recognition community has taken giant steps towards imitating human performance, we are still far from a general purpose vision machine that can carry out the task even for moderately difficult problems. Further, the task becomes increasingly difficult for humans (and prohibitively difficult for machines) as the object, along with its context, is revealed only in parts. According to Gestalt theory, the object *emerges* only when the relevant parts are exposed together giving the impression of the whole object.

In the computer vision and pattern recognition, significant research has been devoted towards detection and recognition of objects in images. The general approach is to consolidate low-level image analysis, such as segmenting the image or detecting local cues, into higher-level models. Absence of information in the local windows hinders the effort of a computer vision algorithm to model a shape prior for the recognition task. Specifically, apparently meaningless jagged patterns, called splats, scattered in the local windows as shown in Figure 3 can seriously hinder boundary detection, silhouette extraction, and shape from shading algorithms, which are common fundamental low- and mid-level ingredients for modeling.

The main advantage that allows humans to outperform computer vision algorithms for emergence, is probably how our life experience is represented in our brain in a way that allows very efficient and highly parallel top-down (global) and bottom-up (local) search [Epshtein et al. 2008]. Any attempt to implement such an approach in computer vision algorithms would have to address the extremely high computational complexity of the problem [Tsotsos 1992]. Using 3D geometry for generating emergence images allows us to vary object size, location, pose, and thus significantly increases the dimension of the search space.

Related work. Emergence, and in general Gestalt principles, are closely related to fundamental models in human vision [Kanizsa 1979], and computer vision [Kim 2000]. The Gestalt principles deal not only with relationships between the parts and the whole,

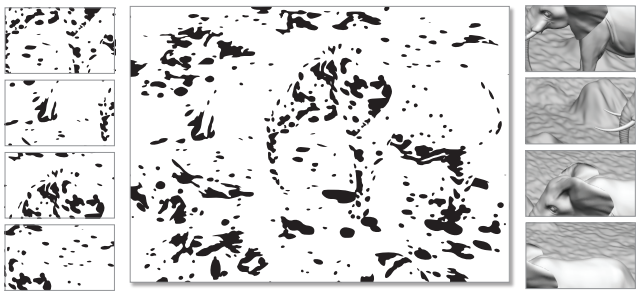

Figure 3: *(Left) Emergence images, when observed through small windows, look meaningless. Although we perceive the subject in the whole image (middle), the smaller sized segments, in isolation, look like random patches. (Right) In contrast, the elephant can be recognized through similar windows of the normal shaded scene.*

but also with more specific principles, like redundancy in information via symmetry, proximity, or continuity which have been extensively studied and explored [Zakia 2001]. Although theories have been proposed to explain these phenomena, we are still far from understanding the exact mechanism guiding our visual perception.

The principle of *emergence* can be described as the ability to segment and detect objects in absence of clear intensity, texture, or color boundaries. It exemplifies the limits of image analysis tools, where biological systems, including humans, extract meaningful information. Although large bodies of research work exist, tasks like image segmentation and object recognition, are still considered very challenging and unsolved in the general setting [Ullman 2000]. Emergence images, which are perceived as a whole and not from individual parts, are unlikely to be understood by direct inference based on low level vision primitives. As humans we possibly make use of well developed visual system, capable of integrating even the *most* subtle cues, coupled with prior knowledge and experience, to perceive such images.

Arcimboldo's paintings illustrate interesting emergence by taking natural elements such as vegetables and fruits, and juxtaposing them to suggest human and other shapes. Gal et al. [2007] present an interactive system for creating such 3D compound shapes or collages. Digital image mosaics are also built upon emergence by rendering large target image by arranging a collection of small source images which can be regular or irregular tiles. In computer graphics, many research efforts have been devoted to generate such mosaics [Kim and Pellacini 2002; Orchard and Kaplan 2008]. Although such efforts illustrate emergence, they are not meant to confuse bots, or prevent the use of computer vision techniques to potentially recover coherent object boundaries.

In making camouflage images, or in the art of concealing objects, the goal is to disguise a subject in plain sight and hide it from the viewer. Such images, which in a sense are counterparts of emergence ones, are popularly used in posters, puzzles and artistic illustrations. Yoon et al. [2008] present a hidden-picture puzzles generator which converts image of the background and objects into NPR stylized line drawing, and then finds suitable places to hide the objects. Our emergence synthesis algorithm is marginally related to NPR stippling and pointillism stylization algorithms [Deussen et al. 2000; Yang and Yang 2008]. However, unlike our technique, NPR techniques are neither meant to be locally meaningless, nor are they designed to be bot-secure. Recently, Shlizerman et al. [2008] have investigated the hypothesis that recognition precedes reconstruction in a top-down procedure for two-tone or *Mooney* images, which despite their sparse content seem to arouse vivid 3D perception of faces, both familiar and unfamiliar.

Captcha, a mechanism to distinguish between humans and bots,

was originally explored by Ahn et al. [2004]. Subsequently, Mori and Malik [2003] use object recognition techniques for solving text based Captchas. Using shape matching techniques, possible regions where letters might be embedded are identified, and then plausible words are extracted from the candidate letters. In an attempt to reduce machine accuracy, more elaborate methods of text Captcha have been developed, by adding destructing lines and texture thus making the Captcha puzzles harder to segment. Such efforts, which can be quite taxing for human users, have also been overridden [Yan and Ahmad 2008]. In an effort to develop more secure puzzles, researchers have turned to general object recognition based Captcha [Elson et al. 2007]. However, bots have been shown to solve such puzzles with non-trivial probability [Golle 2008].

3 Synthesis of Emerging Images

In this section, we present the key principles behind our emergence synthesis procedure as we balance between two conflicting requirements: The generated images should remain recognizable to a human; while to a bot, the images should appear as a collection of meaningless patches. We explore the effect of the algorithm parameters on the difficulty of the emergence images. While alternative approaches are conceivable, the guiding principle behind each stage is more important than the specifics.

Guiding principle. Inspired by the image of the Dalmatian dog (Figure 2), our emergence image synthesis algorithm renders 3D models by texturing them with large dots, which we call *splats*. Locally, such images are complex in sense of the amount of edges and contains no local structure from which one can reliably extract meaningful information. In particular, algorithms such as shape from shading, silhouette detection and image segmentation, will yield meaningless results from images that consists of scattered splats. The most relevant approach to our emerging images is based on shape from texture (SfT) algorithms. While there are several SfT algorithms that work well in simplified settings [Lobay and Forsyth 2006], and even though recently, a model have been suggested for the human ability [Massot and Herault 2008], such algorithms are yet to match human ability.

Generating a locally noisy image is easy. The challenge is to ensure that the whole can still emerge and be recognized by humans. Thus, the scattering of the splats cannot be completely random, but should respect the subject shape, pose, and its silhouette. We observe that such a semi-random splat distribution looks meaningless when viewed through small windows (see Figure 3).

To enable humans to detect and recognize the subject, the splats are generated by rendering the model and scene geometry. These splats generally follow and respect the silhouettes and shape of model just enough to produce the emerging effect. When all the pieces are put together humans can somehow detect consistency of silhouettes and other cues to apply a cognitive process that allows recognizing familiar models. In our studies, we observed that the emergence effect as perceived by humans depends on their familiarity with the subject and its pose.

Harder for bots and easier for humans. We employ two post processing steps to make the synthesized images harder to learn using any vision algorithm: First, we explicitly make sure that local windows have similar statistics, and do not stand out. Second, we perturb or remove splats along the subject's silhouette to make it hard to extract useful information by looking for continuity along splat boundaries. Further, to make the task harder for bots equipped with learning methods, we increase the dimensionality of the search space by selecting, as subjects, 3D articulated models in arbitrary poses viewed from different directions. Another factor that can confuse a potential algorithm is the inclusion of outlier shape parts, in unusual poses, as part of the background scene.

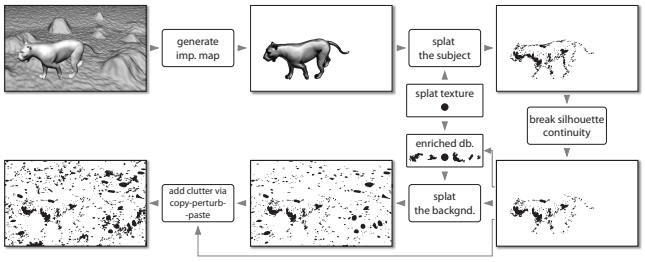

Figure 4: *Given 3D geometry of a scene, an importance map for the subject is computed. The subject is rendered using the importance map to modulate the texture density (see Figure 5). Patches along the silhouettes are perturbed or deleted, making it difficult to trace the subject's contours. Optionally, to add contextual cues, the remaining scene can be rendered using an enriched set of texture patches. Finally, we add clutter using a cut-perturb-paste approach to* hide *the location of the emergence figure (see Section 4).*

To assist the human in the emergence process and to easily recognize the subject, we prefer the main subject to be a familiar one. Objects with intricate surface geometry as characteristic or defining features are unsuited as main subjects, since our sparse splat rendering only captures low frequency geometry. As an additional aid to humans, we give higher priority to salient view directions [Lee et al. 2005], instead of selecting arbitrary poses and view directions.

Increased algorithm-security is achieved at the cost of marginally reduced accessibility. We provide intuitively control parameters to generate progressively difficult emergence images. All the examples in the paper are rendered using fixed sets of parameters for three difficulty levels: easy, medium, and difficult.

4 Algorithm

Given a 3D scene consisting of a subject, a ground plane, and optionally, a collection of auxiliary objects, we now describe our emergence image synthesis algorithm (see Figure 4). Later we discuss how to automatically construct and setup the scene.

Rendering the subject. For a subject, specified as a mesh M, we create an importance map I_M by taking into account its surface geometry, light positions, and the view direction. The importance map assigns importance values to each of the mesh vertices, and is constructed using two intensity maps capturing silhouette and shading information. We assign unit weight to any mesh vertex where

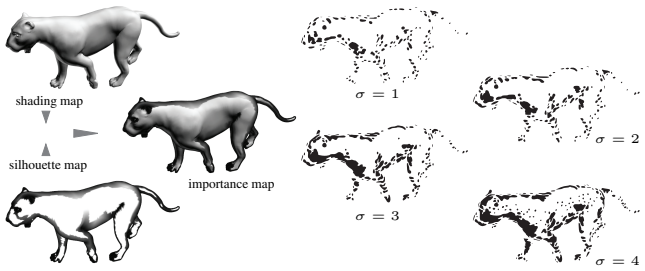

Figure 5: *Given the subject geometry and the scene parameters, a combination of shading and silhouette maps yields an importance map for the subject (darker regions denote higher importance). The importance map is used to control both the spacing and size of the textures on the subject. As seen with the lion textured at four different levels, emergence effect is easier to perceive at higher densities.*

Figure 6: *(Left) Texturing the subject, the horse in this case, results in long complex splats along the silhouette, which may reveal important information about the subject. (Middle) We erode and remove parts of long complex silhouettes, and then, (right) locally perturb the remaining splats using small rotations and translations. To highlight the changes, we show the original splats in gray.*

the corresponding view vector and the surface normal are nearly orthogonal, and diffuse the weights to adjacent vertices. In our experiments, we use a margin of ± 5 degrees for orthogonality testing. For any vertex if $\mathbf{n}$ and $\mathbf{l}$, respectively denote the unit surface normal and unit light vector, we store $(1 - \mathbf{n}.\mathbf{l})$ in the shading map. Finally, the importance map I_M is constructed as the vertex-wise maximum of the silhouette and the shading maps (see Figure 5).

Before rendering the subject with simple patterns, we first generate splat centers on the mesh M according to its importance map I_M. Ideally, the resulting distribution in the image space should be adaptive with blue-noise property. To achieve this, we adopt a Penrose tiling-based sampling technique introduced by Ostromoukhov et al. [2004] to obtain an importance sampling over an image, which attracts dense samples to regions of high importance value. While such a pseudo-random distribution of center locations reveals little to bots, it also makes the result meaningless for humans. To reveal subtle hints, we modulate the randomness using the importance image, i.e., the frame buffer image of the importance map I_M. The density of the generated centers relates to the difficulty of the emergence figures. We control the density of the resulting image space centers using a scaling factor σ in the tiling-based sampling algorithm. Subsequently, with a standard ray-mesh intersection based inverse lookup, we *unproject* the image space centers to 3D positions on the mesh M.

At the end of the tiling step, we have a collection of 3D splat locations on the subject. Texture mapping [Schmidt et al. 2006] the subject geometry using input pattern(s) centered around these selected locations results in overlapping splats in the image space. We refer to such groups of splats as *complex splats*. Figure 5 shows the results of texturing the lion at four different densities, where σ denotes the density control parameter. Typically, silhouette boundaries, which have high importance values along them, get rendered as long complex splats. Such splats may reveal important information about the subject's silhouette when analyzed using algorithms employing boundary extrapolation and continuity analysis. Hence, we break long complex splats along silhouettes into smaller parts, delete a few, and disturb continuity information by perturbing the others using small rotations and translations in the image space (see Figure 6). More specifically, each complex splat, an ensemble of smaller splats, is iteratively broken by randomly removing the atomic splats until the corresponding composite splats separate. Subsequently we apply a 2D screen space rotation by a random angle to each image space splats. Silhouettes are perturbed by randomly jittering the corresponding atomic splat centers.

Complex splats on the subjects form visually interesting patterns and are good candidates for introducing clutter in the remainder of the image. Hence, we insert these complex splats into the texture

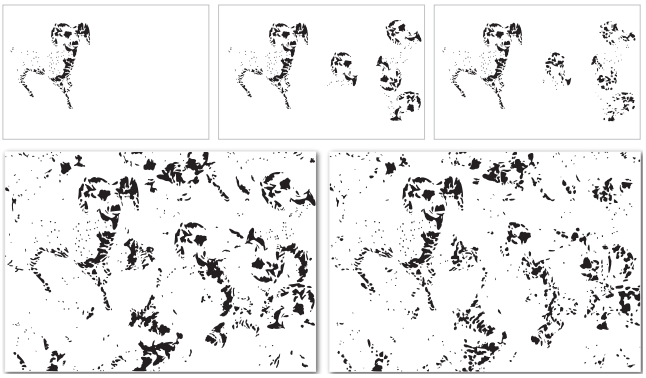

Figure 7: *We copy splats within random windows on the rendered subject (top-left) to fill in other parts of the image. The copied parts, when simply translated and rotated (top-middle), generate clutter. To prevent bots from using feature descriptors for detecting such repetitions, we apply 3D perturbations to the splat centers (top-right). (Bottom) We show the results of such cut-paste (left) and cut-perturb-paste (right) approach using multiple window sizes and locations.*

database. The enriched database is then used to render the background scene and the auxiliary objects (or parts) resulting in splats, similar to those on the subject, appearing in other parts of the image. As a result the subject splats do not stand out in the final image.

Scene generation and setup. Before describing the rest of the synthesis procedure, we briefly describe how the 3D scene is setup. From a database of 3D models, we randomly select a subject. We use a database of common animals, placed upright, to help humans to use their prior knowledge to factor out other distortions, or camouflaging effects. We select a viewpoint that maximizes the visible saliency of the mesh, as proposed by Lee et al. [2005]. Subjects in non-standard poses or viewpoints may result in synthesized images that appear as meaningless clutter (see Figure 8). We place the subject on a (bumpy) ground plane to impart a sense of orientation to the humans. The ground is rendered similar to the subject using the enriched texture database, previously constructed. The complex splats from the subjects make it difficult for any segmentation based approach to successfully isolate out the subject. Finally, the scene is lit from a default light position. The examples in the paper are generated using this automatic setup procedure.

Copy-perturb-paste. Having rendered the subject and the ground plane, we add controlled clutter to make it harder for the bots to identify regions where the subject may lie, i.e., we want to make the subject and the rest of the image look similar when ob-

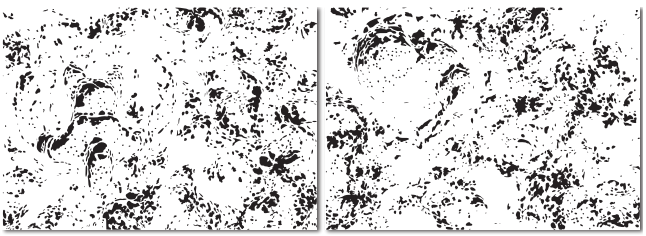

Figure 8: *We often fail to perceive an emergence image when the subject is in an uncommon pose. Among the users who were shown the above images, the average success rate was only* 54% *and* 4%*, respectively. When the inverted versions of these images were shown, the success rates went up to* 96% *and* 91%*, respectively.*

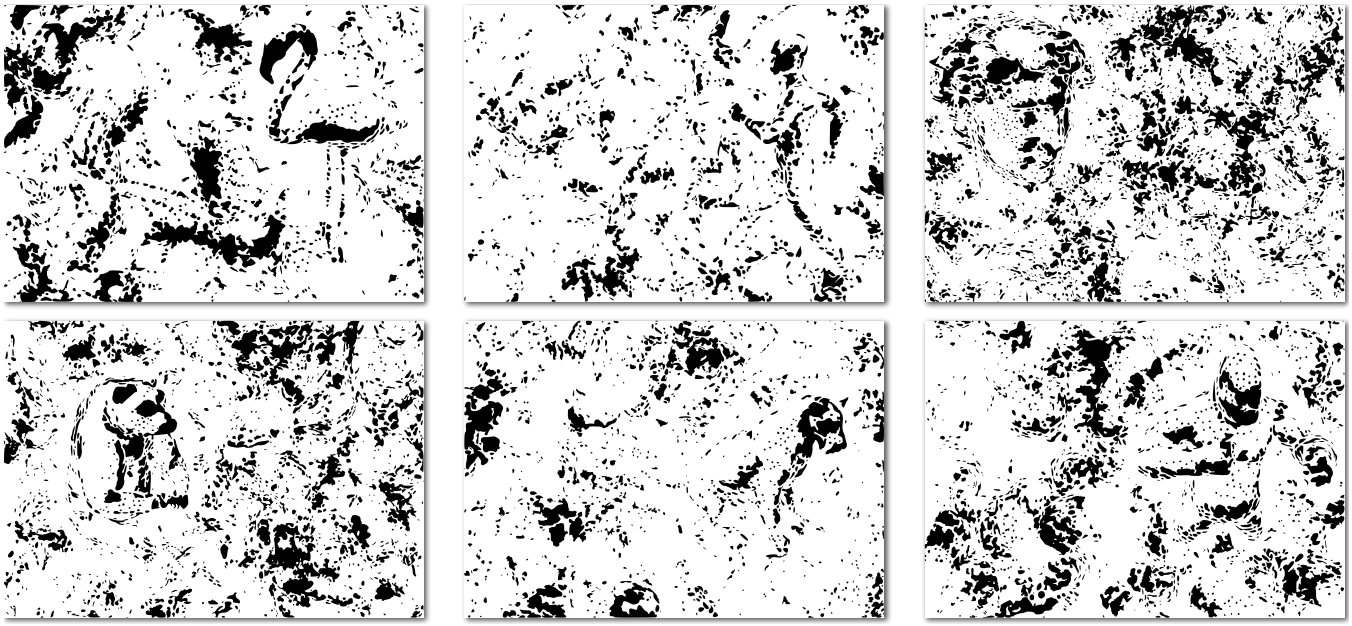

Figure 9: *Typical emergence images generated by our synthesis algorithm. We generate a range of examples on various subjects synthesized at different difficulty levels. Each example contains exactly one subject. (Please refer to supplementary material for other examples.)*

served through small windows. To achieve this, we simply copy small regions of the subject and place them in other regions of the image after arbitrary translations/rotations. Such copy-pasted regions are easy to filter out by humans since we do not perceive the subject seen through isolated small windows in absence of a global context (see Figure 3). If we use such a simple scheme, bots can use a robust feature detector, like Scale Invariant Feature Transform (SIFT) [Lowe 2004], to identify such potential cut-paste operations, and derive valuable information about the location of the subject. To prevent this, we can either copy patches from a separate set of images, or as we choose to do here, perturb in 3D the center locations of the splats, regenerate image space splats, and then rotate them before pasting, making the process robust to SIFT based detectors. More specifically, windows selected uniformly randomly on the subject in the screen space, are rotated about their centers by a random angle. The selected splats are then perturbed in 3D and pasted back into the background image. This cut-perturb-paste cloning operation, summarized in Figure 7, is performed using windows of varying sizes and locations. For all examples, we use windows of sizes $5, 10, 20\%$ of the image window, and finally merge the three layers.

Algorithm 1 Generate **Emergence Image** (Mesh M, view direction $\mathbf{v}$, light position $\mathbf{l}$)

$I_M \leftarrow$ ImportanceMap(M, $\mathbf{v}$, $\mathbf{l}$).
complex splats $\leftarrow$ SplatCreate(I_M).
Break complex splats and disturb continuity information along silhouettes.
for each of three scales (window sizes) **do**
  repeat
    Copy-perturb-paste windows from the subject to add background clutter.
  until # iterations exceeds threshold
end for

5 Results and Discussion

In this section, we evaluate our synthesis method (see Algorithm 1*) based on the two requirements for emergence figures: (1) being recognizable by humans, and (2) being unrecognizable by present-day bots. Since the performance of the biological system and the silicon one are not independent, there is a tradeoff between these two requirements. The following parameters of the proposed method provide a direct control over this tradeoff (see Figure 11):

- **Splat density:** Boosting the density of splats on the subject makes it easier to perceive the object as its silhouette becomes prominent (see Figure 5).
- **Silhouette perturbation:** Long complex splats on the object silhouette are eroded and perturbed (see Figure 6) to break their continuity. More perturbation creates a more challenging image for bots, and a less prominent emergence effect for humans.
- **Background clutter:** After perturbing the subject splats, we copy paste compound splats to other regions of the scene to add clutter. The ground plane, if present, is also rendered using complex splat patterns. This helps to better hide the subject (see Figure 7).

Table 1 lists the default values used in this paper for generating emergence images at easy, medium, and difficult settings.

Validation. Our method makes it hard to design an algorithm to automatically recognize emerging figures. We assume the bot knows our algorithm, but not the chosen parameters, scene layout, or selected object poses, which are automatically set at runtime. In a possible line of attack, the bot may employ low-level segmentation techniques to identify potential subject locations. Our cut-perturb-paste approach effectively hides the subject, preventing it from standing out from its surroundings. In Figure 10, using a multi-scale Canny edge detector, we identify persistent edges, and then string them together based on spatial and curvature continuity. While this clearly extracts the object curves from the original scenes, the method reveals little on emergence images.

*Please refer to the project webpage for demo application/code, supplementary results, and emergence videos.

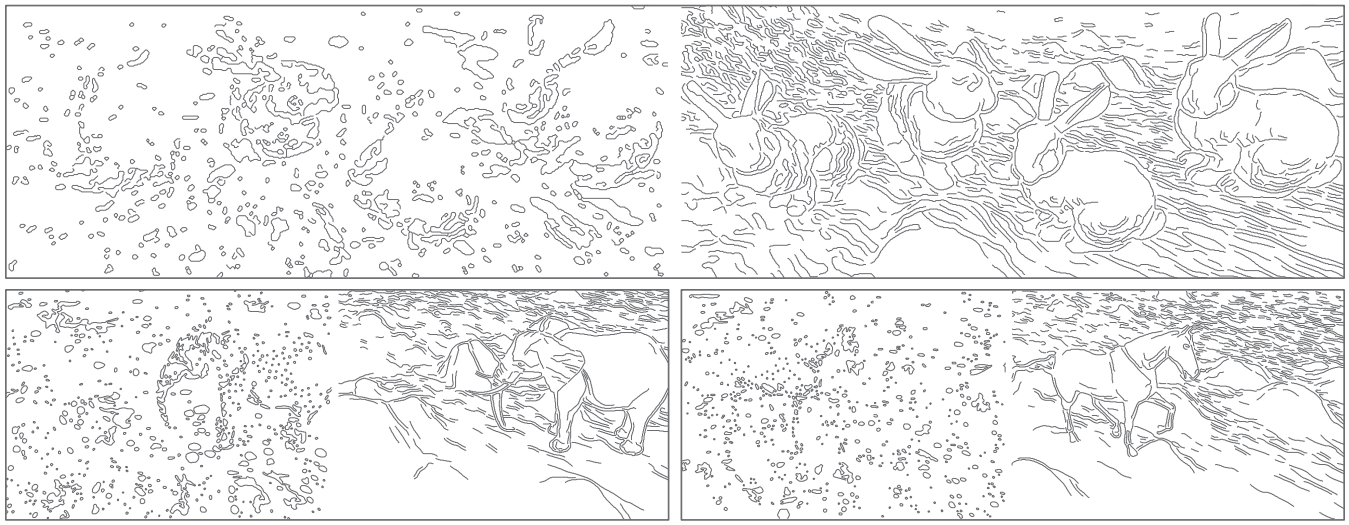

Figure 10: *In many computer vision recognition or segmentation algorithms, the first stages comprise of multi-scale edge detection or other means of bottom-up region processing. At multiple-scales, we detect edges using standard Canny edge detector, and retain the ones that persists scales. Such curves are then linked together based on spatial proximity and curvature continuity. We observe that while on the original renderings the method successfully extracts the feature curves (right image in each box), on the emerging images the results can mostly be seen as noise. This indicates the difficulty that bottom-up algorithms face when detecting objects in the emergence images.*

Recent attempts to compare human and machine vision, have demonstrated that learning-based computerized systems can closely mimic human performance when the visual system is constrained to use only its feedforward mechanisms [Serre et al. 2007b]. Emergence images, which are not perceived instantly, are possibly recognized through the use of top-down feedback connections. To understand how well learning-based vision systems distinguish between emergence images, we experimented in a simplified setting. We tested three modern computer vision systems to discriminate between only two classes of objects based on rendered images, which are either standard images or emergence ones. The first system comprised of a variant of bag of SIFT features, based on hierarchical K-means for building a dictionary of visual features [Nister and Stewenius 2006]. The second one, called C2, is based on a model of the human visual system [Serre et al. 2007a], and employs a battery of Gabor filters followed by collecting similarity statistics for a large set of image fragments. Both the systems employ a Support Vector Machine [Cortes and Vapnik 1995] for recognition. The third system is a hierarchical fragment based system that is trained in a top-down manner [Epshtein and Ullman 2005].

For testing we used a model database consisting of articulated poses of horses and humans. Emergence images were automatically generated in the easy mode (see Section 4) using 30 scenes with horses, and 30 with humans (see supplementary material). In each phase, we used half of the images (selected randomly) per class for training, and the remaining for testing. The process was repeated 100 times, and the results were averaged over all the runs.

	easy	medium	difficult
splat density	1.5	1.2	1.0
silhouette perturbation (frac. wrt. bbox diag.)	1.00	0.20	0.08
rand. perturb. angle (degree)	[-5,5]	[-10,10]	[-15,15]
rand. perturb. displacement (along x, y direction)	0.005	0.005	0.001
cut-perturb-paste frequency	400, 100, 25 times for small, medium, large window sizes, respectively.		

Table 1: *Default values used to generate emergence images at easy, medium, difficult levels.*

While on the standard images, all the systems reliably distinguished horses from humans, their performance sharply degraded on the emergence images (see Table 2). Obtained results are slightly better than chance since the properties of the rendered objects do influence the statistics of the image. This weak statistical link is decreased when harder image synthesis settings are used, and is becoming harder to exploit when more than two classes are presented. A set of 50 users, when shown a random set of images from the same data set, could mark them as horses or humans with close to perfect accuracy (users were asked if they see a horse or a human).

system	standard image		emergence image	
	accuracy %	SD	accuracy %	SD
bag-of-SIFT [Nister and Stewenius 2006]	88.6	8.5	60.4	13.3
C2 [Serre et al. 2007a]	93.7	4.7	51.7	7.5
frag.-based [Epshtein and Ullman 2005]	75.0	6.3	59.0	7.9

Table 2: *Performance comparison of three learning based computerized systems on standard and emergence images (easy mode).*

User study. A good emergence image synthesis algorithm should generate images reliably perceived by humans, at an easily controllable difficulty level. In order to evaluate our method, we conducted a user study involving 310 participants spread across three continents ranging in age from 14 to 60. Users were shown images synthesized from a collection of 31 scenes generated using a database of 15 familiar objects. Rigs or skeletons, if available, were used to introduce articulated pose variations. For each scene, emergence images were automatically synthesized at three levels of difficulty, using default parameter settings (see Table 1). There were also three scenes with inverted subjects (non-familiar pose) generated at easy setting. Thus, the user study comprised of a total of 96 images of which each user was shown a random selection of ten images. Elapsed time starting from showing of the image to the user starting to type her answer was recorded as *response time*. Each image was shown (on screen) for a maximum of one minute. Users were not given any practice trials, but were shown the Dalmatian image (see Figure 2) as an example of emergence image. In order to prevent contamination, no user was shown the same scene

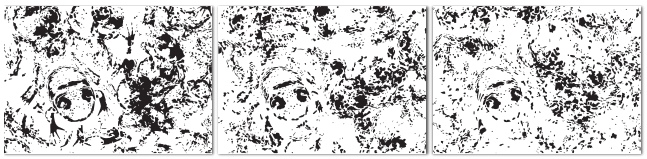

Figure 11: *Emerging frog at various difficulty levels, increasing from left to right. We control the difficultly by controlling the sampling density, breaking the silhouette continuity, perturbing silhouette patches, and adding clutter using cut-perturb-paste.*

at two different difficulty settings.

The creation of emerging images is fast enough to allow for automatic creation of a very large set of figures. With our unoptimized implementation the synthesis time for each emergence image is less than 5 seconds. Figure 9 shows some of the emergence images employed in the user-study, while the full collection is available as supplementary material.

The user response, Figure 12, indicates that we can generate perceivable emergence images at controllable difficulty levels. To evaluate the statistical significance of the correlation between the independent variables, i.e., easy/medium/difficult conditions, we use one-way analysis of variance (ANOVA) [Watson et al. 2001]. A first analysis shows that the effect of the rendering parameters on recognition accuracy is *significant* at the $p < 0.01$ level for the three conditions, $F(2, 90) = 15.2$. A second analysis shows that the effect of the rendering parameters on the recognition time is *significant* at the $p < 0.05$ level for the three conditions, $F(2, 90) = 5.6$. While the success rates are consistently high for the easy/medium levels, we also observe that the rate of false positive is low. Interestingly, when people fail to identify the subjects, they usually report not seeing anything meaningful, rather than perceiving a different subject. For the majority of the images, the rate of false positive is less than 5% (for exceptions see Figure 13). Colleagues in our laboratory can now reliably recognize easy/medium emergence images, synthesized using new subjects, in only a few seconds. This hints that humans can probably quickly learn to *read* emergence images. However, a proper user study, with guidance from psychologists, is required to judge the real implications.

The user study indicates that emerging images are good candidates for Captcha puzzles. In an example instance, the user is asked to select an object out of several options, and then to indicate its image location. For reasonable parameters, a random guess (by a bot) for such puzzles would be successful once every several tens of images. The puzzle can then be repeated to decrease exponentially the probability of random success.

Extensions and discussion. It is widely believed in psychology that humans are highly skilled in processing motion cues, e.g., [Barlow and Tripathy 1997]. Our experiments in which we found that

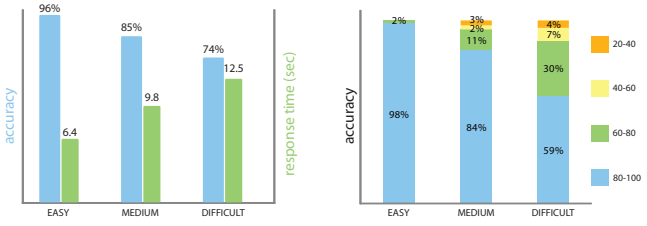

Figure 12: *(Left) Difficulty level as perceived by users and as predicted by our synthesis parameters. (Right) Perceived difficulty level in each category changes gradually. For example, 98% of the easy images were recognized by at least 80% of the observers.*

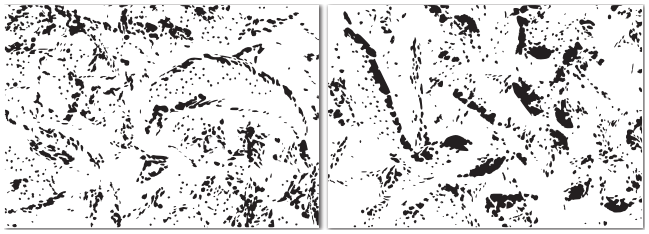

Figure 13: *Emergence images with high false positive responses. (Left) Many confused the dolphin, synthesized at hard setting, with a lizard. (Right) The eagle, synthesized at medium setting, was overlooked and a face was perceived towards the top-right corner.*

simple motion in a 3D scene is perceived using a very basic synthesis approach, supports this belief. Independently working with each frame, we sparsely splat the subject at *random*, and clutter the background using our cut-perturb-paste approach. Unlike single images, we skip any importance map computation. In any single frame most of the observers fail to see the subject (average recognition rate, per frame, is 9.6% averaged over 50 users). However, when the frames are presented as a video sequence, the underlying motion is revealed as a strong emerging motion (quickly recognized by all the test users). Available motion tracking code [Stauffer and Grimson 2000], which reliably detected the subjects in the original video, extracted only garbage from the emergence sequence. While we perceive the moving subjects in presence of motion, the subjects *disappear* as soon as the frame is frozen (see supplementary video).

Limitations. Currently, our algorithm performs best for objects with low-frequency features or with characteristic silhouette curves. It will be interesting to further understand the theoretic limits that restrict transmission of high frequency details using binary emergence images. In keeping with our intuition, our user study indicates that the pose of the subject has a significant effect on the ease of perception (see Figure 8). Our approach of using mesh saliency [2005] for guiding pose selection is just a first step. Better understanding of our perception process might lead to a better method, specially for models with low geometric details.

We observed that the emergence effect as perceived by humans depends on their familiarity with the subject. For example, our experiments indicate that elephants are easily perceived by Asians compared to others. Although one can improve users' familiarity by exposing them to a pool of candidates, we are looking for alternate ways to cope with this difficulty.

6 Conclusions

We presented an algorithm to synthesize emerging figures of 3D objects that are perceivable by humans, but, as our analysis and experiments indicate, are hard, if not impossible, for current bots to recognize. Based on our user study, we note a strong correlation between the difficulty level predicted by our algorithm, and that perceived by users. We also presented interesting results for emerging motion which increase the gap between humans and bots.

The ability to create emergence figures using our scheme, or variations thereof, gives us a tool to synthesize images at varying levels of difficulty by controlling the generation parameters. We hope that this will help psycho-analysts perform targeted tests and better understand the workings of the human visual system, including the largely undiscovered role of top-down feedback. In addition, the same images can be used to test computerized models of the human visual system, and as a benchmark for new approaches in machine vision. This will require a multi-disciplinary research endeavor backed by a comprehensive user study.

Acknowledgements

This work is supported in part by the Landmark Program of the NCKU Top University Project (contract B0008), the National Science Council (contracts NSC-97-2628-E-006-125-MY3 and NSC-96-2628-E-006-200-MY3) Taiwan, the Israeli Ministry of Science, and the Israel Science Foundation. Niloy was supported by a Microsoft Outstanding Young Faculty Fellowship. We are grateful to the members of Computer graphics Group/Visual System Lab, National Cheng-Kung University, in particular Shu-Hau Nien, for helping to conduct the user evaluation, and the various users who participated in the user study. We are grateful to the anonymous reviewers for their comments and suggestions.

References

BARLOW, H., AND TRIPATHY, S. P. 1997. Corresp. noise and signal pooling in the detection of coherent visual motion. *J. Neurosci. 17*, 20.

CORTES, C., AND VAPNIK, V. 1995. Support vector networks. In *Machine Learning*, 273–297.

DEUSSEN, O., HILLER, S., OVERVELD, C. V., AND STROTHOTTE, T. 2000. Floating points: A method for computing stipple drawings. *Computer Graphics Forum 19*, 40–51.

ELSON, J., DOUCEUR, J. R., HOWEL, J., AND SAUL, J. 2007. Asirra: a captcha that exploits interest-aligned manual image categorization. In *ACM Conf. on CCS*, 366–374.

EPSHTEIN, B., AND ULLMAN, S. 2005. Feature hierarchies for object classification. In *IEEE ICCV*, vol. 1, 220–227.

EPSHTEIN, B., LIFSHITZ, I., AND ULLMAN, S. 2008. Image interpretation by a single bottom-up top-down cycle. *Proc. of the National Acad. of Sc. 105*, 38, 14298–14303.

GAL, R., SORKINE, O., POPA, T., SHEFFER, A., AND COHEN-OR, D. 2007. 3D collage: Expressive non-realistic modeling. In *Proc. of NPAR*, ACM Press, 7–14.

GOLLE, P. 2008. Machine learning attacks against the asirra captcha. In *ACM Conf. on CCS*, 535–542.

KANIZSA, G. 1979. *Organization in Vision: Essays on Gestalt Perception*. Praeger New York.

KIM, J., AND PELLACINI, F. 2002. Jigsaw image mosaics. In *ACM SIGGRAPH Trans. Graph.*, 657–664.

KIM, B. 2000. *Perceptual Org. for Artificial Vision Sys.* Springer.

LEE, C. H., VARSHNEY, A., AND JACOBS, D. W. 2005. Mesh saliency. In *ACM SIGGRAPH Trans. Graph.*, 659–666.

LOBAY, A., AND FORSYTH, D. A. 2006. Shape from texture without boundaries. *Int. J. of Computer Vision 67*, 1, 71–91.

LOWE, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision 60*, 91–110.

MASSOT, C., AND HERAULT, J. 2008. Model of frequency analysis in the visual cortex and the shape from texture problem. *Int. Journal of Computer Vision 76*, 2 (February), 165–182.

MORI, G., AND MALIK, J. 2003. Recognizing objects in adversarial clutter: Breaking a visual captcha. In *IEEE CVPR*, 134–141.

NISTER, D., AND STEWENIUS, H. 2006. Scalable recognition with a vocabulary tree. In *IEEE CVPR*, 2161–2168.

ORCHARD, J., AND KAPLAN, C. S. 2008. Cut-out image mosaics. In *Proc. of NPAR*, 79–87.

OSTROMOUKHOV, V., DONOHUE, C., AND JODOIN, P.-M. 2004. Fast hierarchical importance sampling with blue noise properties. *ACM SIGGRAPH Trans. Graph. 23*, 3, 488–495.

SCHMIDT, R., GRIMM, C., AND WYVILL, B. 2006. Interactive decal compositing with discrete exponential maps. In *ACM SIGGRAPH Trans. Graph.*, 605–613.

SERRE, T., WOLF, L., BILESCHI, S., RIESENHUBER, M., AND POGGIO, T. 2007. Robust object recognition with cortex-like mechanisms. *Trans. on PAMI 29*, 3 (March), 411–426.

SERRE, T., OLIVA, A., AND POGGIO, T. 2007. A feedforward architecture accounts for rapid categorization. *PNAS 104*, 15 (April), 6424–6429.

SHLIZERMAN, I. K., BASRI, R., AND NADLER, B. 2008. 3D shape reconstruction of Mooney faces. In *IEEE CVPR*, 1–8.

STAUFFER, C., AND GRIMSON, W. E. L. 2000. Learning patterns of activity using real-time tracking. *Trans. on PAMI 22*, 8, 747–757.

TSOTSOS, J. 1992. On the relative complexity of active vs. passive visual search. *Int. Journal of Computer Vision 7*, 2, 127–141.

ULLMAN, S. 2000. *High Level Vision: Object Recognition and Visual Cognition*. MIT Press.

VON AHN, L., BLUM, M., AND LANGFORD, J. 2004. Telling humans and computers apart automatically. *Commun. ACM 47*, 2, 56–60.

WATSON, B., FRIEDMAN, A., AND MCGAFFEY, A. 2001. Measuring and predicting visual fidelity. In *ACM SIGGRAPH Trans. Graph.*, ACM, 213–220.

YAN, J., AND AHMAD, A. S. E. 2008. A low-cost attack on a microsoft captcha. In *ACM Conf. on CCS*, 543–554.

YANG, C., AND YANG, H.-L. 2008. Realization of seurat's pointillism via non-photorealistic rendering. *The Visual Computer 24*, 5, 303–322.

YOON, J., LEE, I., AND KANG, H. 2008. A hidden picture puzzles generator. In *Computer Graphics Forum*, vol. 27, 1869–1877.

ZAKIA, R. D. 2001. *Perception and Imaging*. Focal Press.