Vismantic: Meaning-making with Images

Ping Xiao, Simo Linkola
Department of Computer Science and Helsinki Institute for Information Technology HIIT
University of Helsinki, Finland
{ping.xiao, slinkola}@cs.helsinki.fi

Abstract

This paper presents *Vismantic*, a semi-automatic system generating proposals of visual composition (*visual ideas*) in order to express specific meanings. It implements a process of developing visual solutions from 'what to say' to 'how to say', which requires both conceptual and visual creativity. In particular, Vismantic extends our previous work on using conceptual knowledge to find diverse visual representations of abstract concepts, with the capacity of combining two images in three ways, including *juxtaposition*, *replacement* and *fusion*. In an informal evaluation consisting of five communication tasks, Vismantic demonstrated the potential of producing a number of expressive and diverse ideas, among which many are surprising. Our analysis of the generated images confirms that visual meaning-making is a subtle interaction between all elements in a picture, for which Vismantic demands more visual semantic knowledge, higher image analysis and synthesis skills, and the ability of interpreting composed images, in order to deliver more ideas that make sense.

Introduction

Aesthetics and meaning are two main concerns of art. The work presented in this paper focuses on meaning-making in image generation. Particularly, we are interested in conveying specific meanings, in contrast to vague or divergent interpretations. A common way of constructing meanings in images is combining objects, where meanings arise from the objects (denotation and connotation) and the relations between them. Such combination involves two main decisions: *which objects to combine* and *how to combine them*.

Contemporary print advertisements offer abundant examples of combining objects to express specific meanings. In general, an ad tells about a desirable attribute of a product. Hence, usually two objects are combined, the product (or something closely related) and another thing that embodies the attribute. For example, an ad for promoting dairy products shows a bone made of milk. Regarding how to visually combine two objects, Phillips and McQuarrie (2004) identified three ways (*visual operations*): *juxtaposition* (two objects side by side), *fusion* (two objects merged together), and *replacement* (only one object is present, which occupies the usual place of the other object).

Obviously, the above visual operations do not appear only in ads, and the relations between objects are not limited to attribute. The news collage in (Krzeczkowska et al. 2010) (see Related Work) is an example of juxtaposing more than two objects. Dalí's liquid clock[1] is an example of fusion, and Duchamp's urinal[2] surrounded by artworks in an exhibition can be seen as an example of replacement.

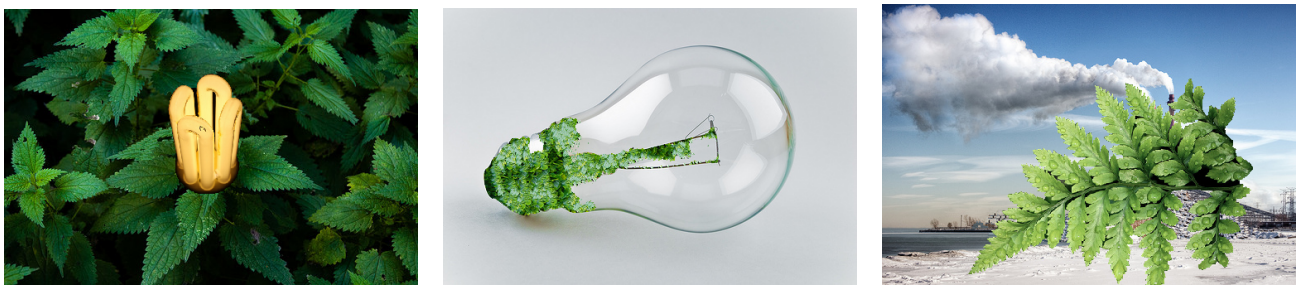
In this paper, we present *Vismantic*, a semi-automatic system combining pictures of objects to express simple meanings described by pairs of a subject word and a message word. A message may be an attribute of the subject, or have a causal or an opposite relation to the subject. Vismantic first searches for photos that represent the subject and the message respectively and are as diverse as possible. It then applies juxtaposition, fusion and replacement to the photos found. We provide the formalization and computational implementation of the three visual operations. Nevertheless, Vismantic is not yet fully automatic; it needs user filtering at intermediate stages.

Vismantic is a workflow of integrating conceptual and visual creativity in making images. Such integration is necessary, since both kinds of creativity are required in common visual communication tasks and there do not exist many such systems. We present here the first version of Vismantic and the results of an informal evaluation, which functions as identifying the problems in the field. Another important objective of the present work is using computational modeling for studying *visual compositional semantics*. The semantics of an image is a synergy of every element in it, including the subtle details. But, there has not been much formal study on it. Formalization and computational implementation are great tools for testing rules and hypotheses. Our newly gained insights are presented in the Evaluation and Analysis section.

Vismantic focuses on the *variety* and *novelty* of compositions (*visual ideas*), rather than generating perfect images. As an example, Fig. 1 shows some of the ideas generated by Vismantic in order to say "electricity is green (sustainable)".

In the remainder of this paper, related work is introduced first, followed by the details of how Vismantic works. We then present the experiment we conducted to evaluate its

[1] http://en.wikipedia.org/wiki/The_Persistence_of_Memory
[2] http://en.wikipedia.org/wiki/Fountain_%28Duchamp%29



(a) (b) (c)

Figure 1: Example visual ideas generated for Task 1 “electricity is green (sustainable)”. 1a: a light bulb replaces a tuft of green leaves; 1b: green leaves are fused with the screw base and wire filament of a light bulb; 1c: a branch of leaves replaces a power station.

ideation capacity, as well as our analysis of the test results. Finally, we give conclusions and propose future work.

Related Work

Within the Computational Creativity community, the bulk of work on visual creativity has concentrated on aesthetics, while meaning creation has only come into focus lately.

Krzeczkowska et al. (2010) created a computer visual artist, which has a basic level of intention and expresses it with collages. At regular intervals it accesses news articles from a few internet sources, and takes the viewpoints of the authors by extracting most-content-indicative keywords (only nouns) from the articles. The keywords are used to retrieve digital images from a few online and local sources, including Corel, Flickr and Google Images. The retrieved images, in their whole or segments, are assembled according to one of the grid-based templates, which is then rendered with pencils, pastels or paints. In the example presented in the paper, ten nouns are extracted in order to cover all the central subjects of an article. The use of collage makes it easy to present the multiple facets of an event. In contrast, Vismantic relates only two objects and intends more specific messages. Moreover, it combines images in two additional ways, i.e. replacement and fusion.

Another computer visual artist, DARCI (Digital Artist Communicating Intent) (Norton, Heath, and Ventura 2010; 2011), renders a given image in order to represent a list of adjectives. It learns, from human-annotated images, the mappings between adjective synsets and low-level image features, including color, light, texture and shape. The mapping for each synset is encoded in a series of artificial neural networks (ANNs). To render an image, DARCI selects a set of image filters through an evolutionary mechanism, where the ANNs are used in each generation to assess how well a rendering reflects the specified adjectives. Unlike DARCI, which focuses on the overall impression of images and the meaning-carrying capacity of low-level image features, Vismantic primarily uses objects and their relations to convey meanings.

In addition, there is work on suggesting objects (in the form of concepts) for images to be generated. Xiao and Blat (2013) were interested in the use of pictorial metaphors in advertisements and created a program proposing metaphor vehicles, to which a product (metaphor tenor) and a few attributes with different levels of prominence are given as an input. The program first searches in several commonsense knowledge bases for concepts that have the main attribute as one of their stereotypical properties. Then, it evaluates the aptness of the concepts found as metaphor vehicles, in regard to imageability, affect polarity, attribute salience, secondary attributes and similarity with tenor. Another work is a software called Perception (De Smedt et al. 2013), which assists the brainstorming of artists in general. It is backed by a semantic network of concepts and their adjective properties. By concept clustering and graph path finding, Perception is able to find instances of novel concepts such as ‘creepy animals’, and make analogies, e.g., proposing a toad as a symbol of Brussels. Both works are made for creative visual tasks, and both touch only the conceptual aspect. They are relevant in augmenting the conceptual creativity of Vismantic.

Outside the Computational Creativity community, a relevant field is Content-aware Image Synthesis (Diakopoulos, Essa, and Jain 2004; Lalonde et al. 2007; Chen et al. 2009), which deals with composing scenes (images) using pictorial elements taken from photos. Its center of investigation is how to make a composition look as realistic as possible, considering that photos normally vary in camera pose, lighting, scale, resolution, etc. There is overlap in the image processing techniques used in this field and by Vismantic. The difference is that, in Content-aware Image Synthesis, it is the user who dictates the composite objects of an image, not the computer.

Vismantic Workflow

Vismantic takes as an input a subject word and a message word. To generate visual ideas, it follows three major steps:

I Find representative photos of the subject and message, respectively;

II Preprocess photos found;

III Apply visual operations (juxtaposition, replacement and fusion).

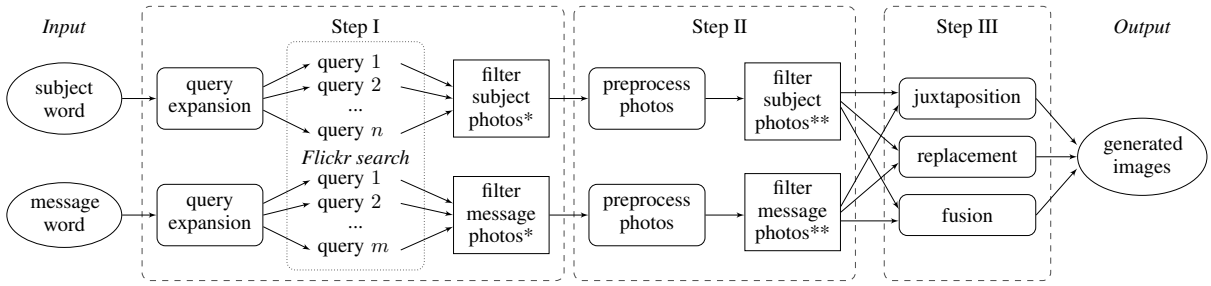


Figure 2: Vismantic workflow. * and **: two filterings have different content (see the text for details).

Step I and II both involve user filtering. The above workflow is also illustrated in Fig. 2 for clarity. The details of the three steps are presented below. Because the user filterings in Step I and II are influenced by how the visual operations are implemented, Step III is introduced first, followed by Step I and II.

Implementation of Three Visual Operations

In this subsection, we introduce first the specifications we give to the visual operations and then how they are realized with several image processing techniques.

Juxtaposition means that two objects are shown side by side in an image. There is no restriction on whether the image of the subject or message should be on the left or right. Also, it does not rely on the context in the generated image to assist understanding.

Replacement means that an object takes the place of another object. The context of the replaced object has to be able to hint about it. Again, it is arbitrary whether the subject or the message object should be replaced.

Fusion means that an attribute of an object is fused with an attribute of another object, which creates a new object with mixed traits. The new object has to remind viewers of the original objects, which normally depends on the distinctiveness of the attributes.

The above visual operations suggest using pictures of objects in their natural surroundings. We chose Flickr[3] as image source, attempting to capitalize on its diverse content.

In order to implement juxtaposition, replacement and fusion, we have identified three image processing challenges. The first is discovering the most prominent object in an image. The second is removing an object from an image and filling the empty space left in order to make it a natural part of the background. For fusion, we currently use the texture of an object to blend with the object region (texture) in another image. This particular implementation does not require that the object region has a distinctive texture, except a good object extraction. Again, either the image of the subject or message can provide texture or object region. Hence, the third challenge is blending the texture of an object with another object so that traits of both objects are still recognizable.

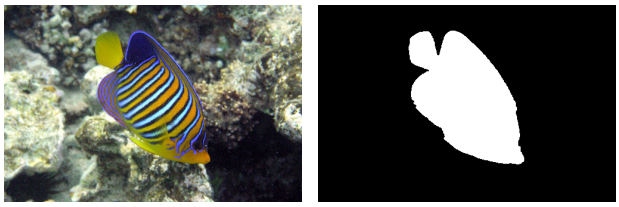
[3]www.flickr.com

The families of image processing techniques we have chosen for solving the above challenges are *saliency-based object extraction*, *inpainting* and *texture transfer*, correspondingly.

Object Extraction refers to finding the most prominent (salient) region in an image. The available algorithms usually provide floating point estimation of saliency for each pixel/segment, which is then binarized to obtain a mask of the most salient region (see Fig. 3b). This mask can then be used to extract the most prominent object (Fig. 3c). We use an algorithm created by Cheng et al. (2011,2015), which was concluded to be one of the better performing algorithms in a recent benchmark survey (Borji, Sihite, and Itti 2012). However, the robustness of object extraction algorithms is still far from perfection; the deficiencies include, e.g., partial object extraction and the object humans infer as the most prominent is not extracted. Furthermore, when there is no objectness estimation for the extraction results, regions in images with no clear separation of fore- and backgrounds, e.g., patterns, can be treated as objects.

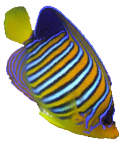
Inpainting techniques were originally created for restoring damaged images or concealing unwanted objects from images. Our intention is to remove objects from images by filling the saliency masks generated by the object extraction algorithm (see Fig. 3d where the object in Fig. 3c is removed using the saliency mask in Fig. 3b). Inpainting algorithms have to deal with textural and structural soundness; textural soundness means preserving the observed textures around the mask, and structural soundness means merging the continuing isophotes (contours of equal luminance) around the mask. As in object extraction, no existing inpainting algorithm gives decent results across the board. Especially, when the removed object is big and/or its surrounding area is diverse, it is difficult to make the inpainted region a natural part of the original image without manual processing. Typical defects are clear patch borders and blurred images. Moreover, in Vismantic, the defects in object extraction may propagate to inpainting.

We use fast spatial patch blending (Daisy, Tschumperlé, and Lézoray 2013) as the inpainting algorithm. It iteratively fits small areas (patches) surrounding the saliency mask into the masked area. Patch-based inpainting algorithms are a reasonably fast and convenient way of taking both of the textural and structural soundness into account. The characteristic of spatial patch blending is that it blends overlap-

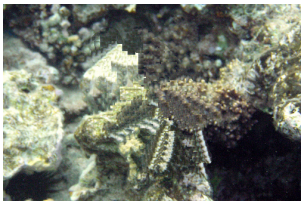


(a) Original

(b) Saliency mask



(c) Extracted object



(d) Inpainted

(e) Input texture

(f) Result of texture transfer

Figure 3: Results of image processing algorithms.

ping regions of adjacent patches making their seams less prominent. However, it has several parameters that should be tuned on an image-to-image basis in order to achieve satisfactory results.

Texture Transfer techniques take the texture of an image and apply it to another image so that the other image's characteristics are still recognizable. Comparing to more common texture synthesis methods, which only try to produce larger continuous texture based on a small sample image, texture transfer methods also take a map (usually a gray scale version of the other image or its segment) as an input, and generate texture to match the map's shape while trying to preserve the map's features. See Fig. 3f where the texture in Fig 3e has been transferred to the extracted object in Fig. 3c.

We use the texture transfer method by Harrison (2005), because we perceived it as more robust than other readily available methods in an informal evaluation. Unfortunately, it has the same shortcoming as the fast spatial patch blending – multiple input parameters need to be adjusted for each image in order to get the best results. Harrison's texture transfer method may produce inferior results on many occasions even with near optimal parameter settings. We noticed that, for our purposes, the best quality is obtained when the input texture and map exhibit similar features, but are still relatively different, e.g., the spatial variability of the texture and the map should be in the same order of magnitude.

Combining the three algorithms above, we can achieve our first implementation of the visual operations. Let I_S and I_M be the subject and message images, respectively; let I_i^s be the saliency mask obtained from the image I_i; let I_i^o be the object extracted from image I_i, given the saliency mask I_i^s; and, let I_i^p be the image, where the area of saliency mask I_i^s has been inpainted. With these notations, we can realize the visual operations as follows:

- Juxtaposition: Resize each of the extracted objects I_S^o and I_M^o to be within a bounding box of 240×240 pixels (refer to the resizing method below), and position the resized objects side by side on a blank 640×400 image, so that the centers of the objects' bounding boxes are vertically centered and at the $\frac{1}{4}$ and $\frac{3}{4}$ marks on the horizontal axis.
- Replacement: Resize I_S^o to be within the bounding box of I_M^o, and layer I_S^o to the same position as I_M^o in the inpainted image I_M^p, i.e. overlapping the centers of the two objects' bounding boxes.
- Fusion: Transfer the texture from I_M^o to I_S^o and overlay the resulting object upon the original subject image I_S.

Here, we have defined the operations only in one way, but as we pointed out earlier, subject and message images are interchangeable.

For resizing, let B_O^w, B_O^h be the width and height of the bounding box of the object to be resized, and B_T^w, B_T^h the width and height of the target bounding box (240×240 in juxtaposition and the bounding box of the object to be replaced in replacement). We formulate the resizing procedure as follows:

1. Calculate the width and height ratios between the bounding boxes: $r_w = B_T^w / B_O^w$ and $r_h = B_T^h / B_O^h$.
2. If $r_w \leq r_h$, then $r = r_w$, otherwise $r = r_h$.
3. Resize if $r < \frac{3}{4}$ or $r > \frac{4}{3}$ (this is for using the original image whenever we can, in order to avoid decreasing the image quality).

Finding Representative Photos of Concepts

At the first step, Vismantic searches in Flickr for photos that can represent well the subject and the message, respectively. Other concerns are *diversity*, *photo quality* and *(image processing) algorithm-friendliness*. We also pay attention to the copyright of photos, only retrieving photos under Creative Commons license with modification permission.

Both the subject and message can be a physical or abstract concept. For physical concepts, such as an object, pictures of the object itself or something closely related, i.e. its internal components or other objects interacting with it, are used to represent it. On the other hand, abstract concepts are represented by pictures of entities through connotation.

When searching in Flickr, the subject and message words are used as free text search, sorted by relevance. Photos with more than 15 tags are rejected, considering that too many tags might imply photographers' intention of boosting the rankings of their photos in every query. We also avoid photos tagged with 'illustration', 'painting', 'graphics', 'infographic', 'text', 'collage', 'scrapbook', 'photoshop', etc. The photos downloaded are of medium size: at most 640 pixels on the longer side.

Additionally, we take advantage of the 'related-tags'[4] provided by Flickr and one of our previous works (Xiao and Blat 2012) in order to improve the diversity of the search results for abstract concepts. (Xiao and Blat 2012) finds physical concepts that have the intended abstract concept as one of their stereotypical properties. This is achieved by retrieving strong associations from four semantic knowledge bases and subsequently filtering associations that are low in concreteness and imageability. Take the task in Fig. 1 as an example, the concept 'electricity' is concatenated with each of 53 physical concepts to form Flickr queries, such as "electricity storm", "electricity pylon", "electricity bulb", "electricity windmill", "electricity plant", "electricity outlet", etc. The photos retrieved by multiple queries are organized in groups, one group per query.

User Filtering The photos retrieved from Flickr might not be sufficiently representative for the concept of interest. Also, the photo quality might be low, due to, e.g., under/over exposure, blur, highlight, low resolution, colorization, or using a fisheye lens. Besides, as mentioned in the previous subsection, the image processing techniques currently used by Vismantic only work well with certain images. Photos having a recognizable object that is neither obscured nor too small, and is situated in a simple context, are preferred. At present, Vismantic needs a user to choose quality photos that are representative and algorithm-friendly.

Preprocessing

Each of the three visual operations has distinct requirements for a pair of input images:

- Juxtaposition: *good object extraction* for both images.
- Replacement: *good object extraction* for one image, and having a *suggestive context* for the other.
- Fusion: *good object extraction* for one image, and having a *distinctive texture* for the other.

The above requirements have to be satisfied before applying visual operations, which can be computationally expensive. Furthermore, they prevent unpromising results early on, which drastically saves the effort needed for evaluating the final output, since the number of images in the final output without filtering is quadratic to the number of input images (each photo of a subject is paired with every photo of a message).

At this step, object extraction and inpainting are applied to all the photos retrieved from Flickr and selected by a user. Currently, Vismantic does not have automatic means to judge the quality of object detection, the indicative capacity of a context, or the distinctiveness of a texture.

User Filtering For each photo, the object/region extracted is shown to a user, who is asked to decide if it represents the corresponding subject or message and if it has a distinct texture which alone cues the concept. The inpainted image is also shown to the user, together with the question whether the image reminds him of the concept.

[4] www.flickr.com/services/api/flickr.tags.getRelated.html

Evaluation and Analysis

To get a first idea of what Vismantic generates, we put it in a test consisting of five typical visual communication tasks, where only the authors interacted with the system. In this section, we first present the output and curation coefficient at each step of the workflow, along with our analysis of the output. Next, we reveal the major factors that cause a generated image to be uninterpretable or end up with unintended meanings. The five tasks are the following (subject and message words in italic):

1. *Electricity* is *green* (sustainable).
2. *Music* is *powerful*.
3. *Lipstick* is associated with *love*.
4. Heating system makes *house warm*.
5. *Earplug* reduces *noise*.

At the first step, the purpose is to find representative, diverse, high-quality and algorithm-friendly photos of concepts (subjects and messages). In the test, 50 photos were collected for each subject and message. For abstract concepts, which lead to multiple queries for searching in Flickr, the photos were collected by visiting the photo groups (one group per query) one by one and picking up the first unpicked photo (the photos in a group are sorted by relevance). The upper part of Table 1 shows the number of disqualified, qualified, selected and surprising photos for each concept. Averaging across all ten concepts, 46.4% of the photos retrieved from Flickr are qualified. The disqualified photos are divided into three categories, i.e. 'non-representative', 'non-algorithm-friendly' and 'low-quality'. Non-representativenessness, amounting to 35.2%, was the top reason for rejecting photos. Non-representative photos either lack relevance or represent a sense of a concept other than the one intended.

We selected photos from the qualified ones and only kept those that look quite different from each other. On average, around 9 photos were selected for each concept. We also noticed that there were novel representations of concepts among the retrieved photos (the row of 'surprising' in Table 1), which counts for 4.2%.

At the second step, preprocessing, object extraction and inpainting were applied to the photos selected in Step I, and the output is shown in the lower part of Table 1. Averagely speaking, good object extraction was found in 69.3% of the selected photos. The major types of incorrect object extraction include: only part of an object was extracted and the part was not recognizable; the object was not extracted at all but some other part of the photo instead, e.g., another object or part of the background; or the whole photo was extracted. Within the properly extracted objects, distinctive textures were not so common, counting for 20.5%. Some examples are green grassland, red lipstick, water, brick wall, flame and textile. Besides, only 22.7% of the selected photos had suggestive context around an object region. In many photos, the object was relatively big and the context was too small to be distinguishable; or the context could not hint at the object if it were removed. However, surprisingly, the

Table 1: Output of Step I Finding representative photos of concepts and Step II Preprocessing.

	electricity	green	music	powerful	lipstick	love	house	warm	earplug	noise	avg.	%avg.
photos retrieved	50	50	50	50	50	50	50	50	50	50		
non-representative	13	6	6	25	22	30	2	26	17	29	17.6	35.2
non-algorithm-friendly	6	2	5	1	9	4	10	4	21	2	6.4	12.8
low-quality	3	0	9	3	1	3	3	1	3	2	2.8	5.6
qualified	28	42	30	21	18	13	35	19	9	17	23.2	46.4
surprising	0	2	0	5	5	2	0	1	3	3	2.1	4.2
selected (at Step I)	10	7	9	9	13	8	8	9	6	9	8.8	17.6
good object extraction	6	5	5	5	10	5	8	6	5	6	6.1	69.3
distinct texture	0	5	0	3	3	1	2	4	0	0	1.8	20.5
has-context	0	4	4	4	0	2	0	3	2	1	2	22.7
suggestive context	3	5	6	6	2	4	0	3	5	5	3.9	44.3

Table 2: Output of Step III Applying visual operations. gene. = generated, expr. = expressive, supr. = surprise, % = ratio between the two numbers ahead.

	electricity-green			music-powerful			lipstick-love			house-warm			earplug-noise			avg.		
	genr.	expr.	%	genr.	expr.	%	genr.	expr.	%	genr.	expr.	%	genr.	expr.	%	genr.	expr.	%
juxtaposition	60	0	0	50	46	92	100	54	54	96	56	58.3	60	44	73.3	73.2	40	54.6
replacement	45	12	26.7	60	28	46.7	50	27	54	24	21	87.5	55	30	54.5	46.8	23.6	50.4
fusion	30	9	30	15	0	0	25	1	4	44	3	6.8	0	0	0	22.8	2.6	11.4
total	135	21	15.6	125	74	59.2	175	82	46.9	164	80	48.8	115	74	64.3	143	66.2	46.4
	expr.	supr.	%	expr.	supr.	%	expr.	supr.	%	expr.	supr.	%	expr.	supr.	%	expr.	supr.	%
surprise	21	21	100	74	30	40.5	82	40	48.8	80	80	100	74	32	43.2	66.2	40.6	61.3

errors in object extraction sometimes provided suggestive context. When only part of an object was extracted, the remaining part might be able to cue the object. For instance, in Fig. 1c, the smoke coming out of the power station was not extracted. Including these cases, suggestive context were found in 44.3% of the selected photos.

At the third step, applying visual operations, on average 143 images were generated for each task (Table 2). Most of them were juxtapositions and replacements, because distinctive textures were rare. Our primary evaluation criterion is whether a generated image expresses the meaning specified in a task. Averaging across all five tasks, 46.4% of the generated images were considered expressive.

Regarding if there is a general trend that one visual operation works better than another, there was no significant difference between juxtaposition and replacement. Both operations produced expressive images about half of the time, 54.6% and 50.4% respectively. The difference showed more in specific tasks. For instance, there was no expressive juxtaposition for Task 1. As seen in Table 2, it seems that fusion yields less than juxtaposition or replacement. However, a factor that has to be taken into account is that fusion can easily go wrong in the current implementation. In juxtaposition and replacement, the image processing involved are mainly resizing and positioning, while the quality of object extraction and the indicative capacity of context have been evaluated in the previous step, preprocessing. On the other hand, the parameters of the texture transfer technique used in fusion have to be fine tuned for each image for optimal performance, which is not yet available in Vismantic. In addition, the number of textures available for fusion was quite small. When there are more varieties of textures, one set of parameters that does not work well with one image may work for another, which could bring us more expressive images.

A few examples of the images generated for each task are shown in Fig. 4. More examples can be visited online[5]. Besides, we have a few interesting observations. Firstly, Vismantic sometimes generates "perfect images" (see Fig. 1a), when some visual features of two objects, such as size, shape, angle and lighting, match by coincidence. Secondly, fusion sometimes produces images of high artistic skill (see Fig. 1b for an example).

As in Step I, we counted the number of surprising ideas among the generated images, and found that on average 61.3% of the expressive images were surprising. The surprise came from novel representations of concepts and unexpected combinations of objects in terms of the concepts they denote/connote or the exact visual representations. Additionally, we call attention to Fig. 1c. The meaning of this image is not as straightforward as "the power station (covered by the leaves) is as green as the leaves", which is what we had envisioned. A plausible interpretation is "the leaves (or the concept of 'sustainability' accompanying it) replaces the traditional power station". Owing to the drastic contrast in size and solidity between leaf and power station in common sense, this image exemplifies immense boldness, which has not been explicitly modeled in Vismantic.

In the following subsection, we analyze why some of the generated images do not express the intended meanings.

Failure Analysis

We have observed that a generated idea may fail mostly in three aspects, namely *semantic interaction*, *visual operation implementation* and *object affordance and composition*.

Semantic Interaction For some generated images, there seems to be no plausible interpretation, divergent interpretations, or an interpretation either the one not intended or

[5]http://vismantic.hiit.fi/examples/

(a) (b) (c) (d)

Figure 4: Examples of generated visual ideas. 4a: for Task 2 “music is powerful”, a singer replaces part of waves; 4b: for Task 3 “lipstick is associated with love”, the heads of a kissing couple are fused with a red lipstick; 4c: for Task 4 “heating system makes house warm”, a house is fused with a pair of crochet mittens; 4d: for Task 5 “earplugs reduce noise”, a helicopter replaces part of a man’s head with fingers stuck in the ears.

the opposite. For instance, Fig. 5a is a juxtaposition generated for Task 4 “heating system makes house warm”. It seems rather difficult to get the feeling that the house is being warmed up. Nonetheless, this is well achieved by the fusion of the two objects (Fig. 4c). Another example is shown in Fig. 5b, a replacement (a power station (without smoke) replaces a line of trees) generated for Task 1 “electricity is green”. The image looks like a power station in its natural surroundings, which is unable to allure viewers into thinking of other connections between the two objects, such as grass gives energy to a power station (see Fig. 1a for a comparison). Moreover, semantic interaction may not happen as expected for other reasons, such as objects having opposing emotional valence.

(a) (b)

Figure 5: Semantic interaction.

Visual Operation Implementation As explained earlier, fusion requires fine tuning the parameters of the inpainting algorithm, which is not available in Vismantic at present. The current resizing method used in replacement does not produce ideal results when the objects involved have quite different shapes. Besides, we noticed that additional constraints might be applied to visual operations, e.g., texture-based fusion should avoid objects with similar colors.

Object Affordance and Composition Fig. 6 shows two different light bulbs placed in the same context, both of which are replacements for Task 1 “electricity is green”. Fig. 6a works while Fig. 6b does not. The difference between the light bulbs is that one is for putting on a horizontal surface, such as the ground, and the other is for hanging vertically, such as from a ceiling. The context is a forest with the ground covered by grass and leaves. The bulb for the horizontal plane suits this context well, which suggests that it is the forest where the bulb gets energy. In contrast, the vertical light bulb can not connect to the forest in a similar way. This comparison reveals that two objects can only be connected meaningfully at certain parts, but not every part.

Besides, the left and right order (orientation) of two objects sometimes can not be arbitrary. Consider whether the idea is still effective if the singer in Fig. 4a turns his head to the opposite direction.

(a) (b)

Figure 6: Object affordance.

In Table 3, the numbers of different types of failure are presented. It shows that semantic interaction was a major cause of failure for all three visual operations. Failure of visual operation implementation occurred mainly in fusion and replacement, since juxtaposition has less constraints on resizing and positioning. Failure of object affordance and composition happened largely in replacement and juxtaposition, because the current implementation of fusion primarily replies on texture.

Table 3: Failure type. The ratio in parenthesis is against the number of disqualified images generated by each operation.

	juxtaposition	**replacement**	**fusion**
disqualified images	166	116	101
semantic interaction	158 (95.2%)	65 (56.0%)	50 (49.5%)
visual operation implementation	0 (0.0%)	29 (25.0%)	52 (50.5%)
object affordance & composition	8 (4.8%)	21 (19.0%)	0 (0.0%)

Conclusions and Future Work

This paper presents Vismantic, a semi-automatic system for generating visual ideas. The workflow it exemplifies has generality, in the sense that it starts from a conceptual task (described in text) and outputs visual compositions, which fit real-life practice. Vismantic takes advantage of both conceptual and visual creativity in its ideation. At present, with basic conceptual knowledge (semantic associations) and the first implementation of three visual operations (juxtaposition, replacement and fusion), it demonstrated the potential of producing images that are expressive, diverse and surprising.

Vismantic generates surprising ideas by using novel representations of concepts and unexpected combinations of objects in terms of the concepts they denote/connote or the exact visual representations. It also generates images with certain particular flavors, such as extreme boldness, though it is not supposed to have such sense. In the future, when deciding objects to be combined, additional effects, such as surprise, boldness and humor, can be considered.

For Vismantic to have a higher level of automation and generate more ideas that make sense, we have identified challenges in three areas:

- *Visual Resources*: sources of photos with high relevance and diversity, sources of distinctive textures and sources of indicative context;
- *Image Processing*: automatic means of selecting photos that are high-quality and algorithm-friendly, automatic means of tuning algorithm parameters, taking into account visual features (such as color, shape, orientation and camera angle) when applying the visual operations, and making use of more sophisticated image analysis to accurately locate objects in complex scenes;
- *Visual Semantics*: more visual knowldge, such as object affordance and the meanings of visual features (e.g., orientation, position and contrast), and the ability of interpreting images, i.e. simulating the interaction between all the meaning fragments generated by visual cues at various levels.

Last but not least, other visual operations can be added to Vismantic.

Acknowledgments

This work has been supported by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733 (ConCreTe). We would like to thank Hannu Toivonen for his valuable suggestion and Flickr users who grant their photos Creative Commons licenses.

References

Borji, A.; Sihite, D.; and Itti, L. 2012. Salient object detection: A benchmark. In Fitzgibbon, A.; Lazebnik, S.; Perona, P.; Sato, Y.; and Schmid, C., eds., *Computer Vision ECCV 2012*. Springer Berlin Heidelberg. 414–429.

Chen, T.; Cheng, M.-M.; Tan, P.; Shamir, A.; and Hu, S.-M. 2009. Sketch2photo: Internet image montage. In *Proceedings of the ACM SIGGRAPH Asia 2009*, SA '09, 124:1–124:10.

Cheng, M.-M.; Zhang, G.-X.; Mitra, N. J.; Huang, X.; and Hu, S.-M. 2011. Global contrast based salient region detection. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, 409–416.

Cheng, M.-M.; Mitra, N. J.; Huang, X.; Torr, P. H. S.; and Hu, S. 2015. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(3):569–582.

Daisy, M.; Tschumperlé, D.; and Lézoray, O. 2013. A fast spatial patch blending algorithm for artefact reduction in pattern-based image inpainting. In *SIGGRAPH Asia 2013 Technical Briefs*, SA '13, 8:1–8:4.

De Smedt, T.; De Bleser, F.; Van Asch, V.; Nijs, L.; and Daelemans, W. 2013. Gravital: Natural language processing for computer graphics. In Veale, T.; Kurt, F.; and Forceville, C., eds., *Creativity and the Agile Mind: A Multi-Disciplinary Study of a Multi-Faceted Phenomenon*. Berlin: Mouton. 81–98.

Diakopoulos, N.; Essa, I.; and Jain, R. 2004. Content based image synthesis. In Enser, P.; Kompatsiaris, Y.; O'Connor, N.; Smeaton, A.; and Smeulders, A., eds., *Image and Video Retrieval*. Springer Berlin Heidelberg. 299–307.

Harrison, P. 2005. *Image Texture Tools*. Ph.D. Dissertation, Monash University.

Krzeczkowska, A.; El-Hage, J.; Colton, S.; and Clark, S. 2010. Automated Collage Generation – With Intent. In *Proceedings of the International Conference on Computational Creativity*, ICCC '10, 36–40.

Lalonde, J.-F.; Hoiem, D.; Efros, A. A.; Rother, C.; Winn, J.; and Criminisi, A. 2007. Photo clip art. In *Proceedings of the ACM SIGGRAPH 2007*, SIGGRAPH '07.

Norton, D.; Heath, D.; and Ventura, D. 2010. Establishing Appreciation in a Creative System. In *Proceedings of the International Conference on Computational Creativity*, ICCC '10, 26–35.

Norton, D.; Heath, D.; and Ventura, D. 2011. Autonomously creating quality images. In *Proceedings of the 2nd International Conference on Computational Creativity*, ICCC '11, 10–15.

Phillips, B. J., and McQuarrie, E. F. 2004. Beyond visual metaphor: A new typology of visual rhetoric in advertising. *Marketing Theory* 4(1-2):113–136.

Xiao, P., and Blat, J. 2012. Image the imageless: Harvesting connotation knowledge for visual expression. In *Proceedings of the 7th International Conference on Design Principles and Practices*.

Xiao, P., and Blat, J. 2013. Generating apt metaphor ideas for pictorial advertisements. In *Proceedings of the 4th International Conference on Computational Creativity*, ICCC '13, 8–15.