A Picture Paints a Thousand Words but Can it Paint Just One?

David A. Robb[1] **Stefano Padilla**[1] **Thomas S. Methven**[1] **Britta Kalkreuter**[2] **Mike J. Chantler**[1]
[1]School of Mathematical and Computer Sciences, [2]School of Textiles and Design,
Heriot-Watt University, Edinburgh, UK. EH14 4AS
{d.a.robb, s.padilla, t.methven, b.kalkreuter, m.j.chantler}@hw.ac.uk

ABSTRACT
Imagery and language are often seen as serving different aspects of cognition, with cognitive styles theories proposing that people can be visual or verbal thinkers. Most feedback systems, however, only cater to verbal thinkers. To help rectify this, we have developed a novel method of crowd communication which appeals to those more visual people. Designers can ask a crowd to feedback on their designs using specially constructed image banks to discover the perceptual and emotional theme perceived by possible future customers. A major component of the method is a summarization process in which the crowd's feedback, consisting of a mass of images, is presented to the designer as a digest of representative images. In this paper we describe an experiment showing that these image summaries are as effective as the full image selections at communicating terms. This means that designers can consume the new feedback confident that it represents a fair representation of the total image feedback from the crowd.

Author Keywords
Design feedback; image browsing; similarity; clustering; semiotics; image summarization; visual communication; perception; emotion; mood; crowdsourcing.

ACM Classification Keywords
H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces.

INTRODUCTION
The experiment described in this paper was motivated by previous work on the development of an image based crowd feedback method created for designers (initially fashion and interior designers) to get emotional feedback on their designs [31, 32]. The method (see Figure 1) was developed with the aim of redressing the asymmetry between the largely visual output of fashion and interior designers and the conventional text feedback they receive when seeking crowd feedback through services such as *Dribbble* [8] which is usually textual. The visual design feedback method allows members of a crowd (*crowd users*) to respond to some presentation of a design, such as a fashion garment or a room interior, by choosing images from an intuitive image browser which represent their emotional reaction. Commonly in other systems, such crowd users would be asked to describe their emotional response using text instead. Aside from the novel image browsers which enable image selection by the crowd, the other key component of this visual feedback method is the summarization of the responses. A single design would get a large number of image selections in feedback from the crowd, and the system compiles these into a montage, or digest, of a smaller number of representative images. The designers then view the image feedback summary montages and take inspiration from them.

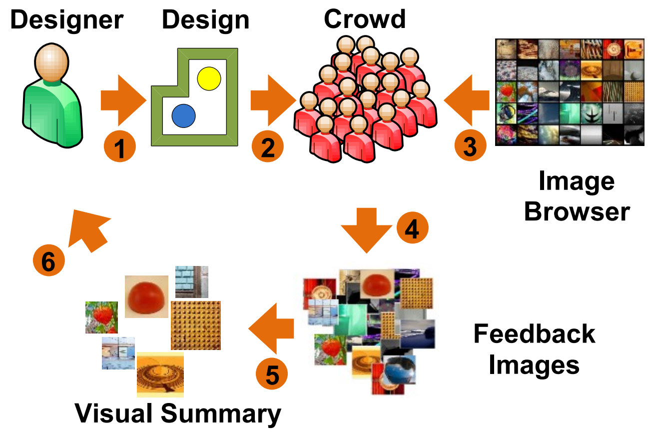

Figure 1. The crowdsourced visual feedback method (CVFM) allows designers to get image-based feedback from a crowd. After viewing a design the crowd choose images from a browser to represent how the design made them feel. The feedback is then shown to the designer as a concise, algorithmically generated, summary.

Contribution
Robb et al.'s [31, 32] evaluation of this visual feedback method revealed that designers and a section of crowd participants viewed the method as useful in communicating about the emotional reaction of the crowd participants to designs. However, given that the method is for communication, while that evaluation gave a picture of the views of the designer and crowd participants, it did not provide evidence that any of the message being intended by the crowd participants in their image choices was actually

being read from the visual feedback summaries by designer participants. Therefore the following two questions remain:

1) Can communication actually be achieved using abstract images chosen from an image browser?

2) Do the image summary montages constructed by the algorithm used by Robb et al. [32] preserve the meaning of the image selections that they summarize?

If the first question could be answered positively and communication were proved to take place using the method, then this would encourage uptake, further study and development of the new visual feedback method or indeed of other forms of image-based feedback. If the second question could be answered positively then designers would be secure in the knowledge that the inspiring, algorithmically generated, visual summaries represent a fair view of the totality of the image feedback, while avoiding being overwhelmed by a mass of images from a potentially large crowd. They could then have confidence in using the method to engage large numbers of feedback-givers in reacting visually to their designs thus contributing to an expansion of feedback-giving beyond those already involved through text-based methods.

In this paper we address those two questions. We describe an experiment which shows that a signal can be sent and received using the abstract image browser, and that the image summaries preserve the meaning from their constituent images.

In the rest of this paper, in *Background*, we describe in more detail what was found out about the experience of designer and crowd participants in the previously published work on the crowdsourced visual feedback method (CVFM). Then we discuss work on semiotics and communication relating it to the CVFM establishing the multi-faceted nature of communication. In *Experiment,* after separating out a single aspect of communication on which to focus, and stating our research questions, we set out the experiment to address those questions. In *Results* we present the analysis of the data from the experiment. Finally, in *Discussion and Conclusions* we discuss how the results have answered our research questions and describe how they might influence implementations of the CVFM.

BACKGROUND

In this section we describe the findings in the previous work on the CVFM and then discuss work on semiotics, communication and how certain types of pictures and images are used to convey meaning.

The Experience of CVFM Users

Robb et al. [32], while evaluating the CVFM, focused on the designers' point of view when comparing image based feedback with text. Designers put forward designs for feedback about how they made people feel. These were viewed by participants acting as the crowd. The feedback was in three formats: lists of text comments and two styles of images (distilled into two image summaries). One summary was from an emotion image set (people showing expression, peaceful landscapes etc.) and the other was of abstract images (rich in colors and unusual forms). Interviews with the designers showed that designer participants took inspiration from the image feedback. The new image based feedback method was well received by them and eleven of twelve designer participants wished to use a service offering the crowdsourced image feedback. Importantly, the interviews also revealed that designers thought the abstract image summaries showed the mood perceived in their designs by the crowd participants.

Further to that, an overview of the crowd participant experience was given separately with a description of software components used in the evaluation [31]. Crowd participants rated the feedback formats for engagement and utility during the feedback task and then completed a questionnaire. Analysis showed a portion of them, termed *image-likers*, found the *abstract* images more engaging than text while considering them to be just as useful as text for describing how designs made them feel. (The emotion image set was less popular than the abstract image set among both crowd and designer participants).

Taken together, these results from both designer and crowd participant sides of the evaluation of the CVFM showed that both the designer participants and a section of the crowd (the "image-likers") thought they could communicate on a perceptual or emotional level using the CVFM with the abstract image browser and summarization. The conclusion drawn from those results, that some people think they can communicate using these tools, however, does not prove that communication does actually take place. Also, if communication does take place using the tools, the summarization process which produces the image summaries viewed by designers might have an attenuating effect on the communication.

Semiotics and Communication

Chandler defined semiotics as *"…the study… of anything which 'stands for' something else."*[3]. The CVFM (Figure 1) proposes that communication between crowd and designer using images is feasible. However, an important aspect of any such conversation is whether or not a designer can understand what a crowd has attempted to say in its image selections. Sausseure, in his theory of language, as described by Guiraud [6, 13], argued that, in language, signs are an arbitrary combination of signifier and signified; e.g. there is no natural reason for the word, dog, to signify what we recognize as the furry animal that barks. Therefore, if the images used for communicating within the CVFM were considered to be a totally new language, this raises the prospect of an involved and time-consuming language learning process to be gone through before the crowd and designers can communicate. However, we think that using images exploits established visual conventions already implicit in the experience of the crowd and the designers, whether or not they are explicitly aware of them, allowing communication to take place [17].

Jakobson's six "functions" of communication encompass the purposes of each aspect of conversation [15]. Only two of these aspects of communication are concerned directly with the specific detail of the message. These are the semiotics (or code) used, and the specific content of the message, i.e. what the message is actually about. The remaining aspects in Jakobson's analysis can, from a purely semiotic view, seem to be peripheral. These are: the emotional state or attitude of the sender of the communication; its effect on the receiver; the simple purpose of maintaining or continuing the conversation; and the inherent artistic value of the message itself. On the other hand, taking a holistic view, this deconstruction highlights that communication is complex and there is more to it than simply what is said or written. This is confirmed, for example, by experimental work on communication of attitude during face-to-face conversation [23]. It was found that only 7% of attitude communication depends on the words spoken. The rest depends on tone of voice and facial expression. The weightings were found to be verbal - 7%, vocal - 38% and facial - 55% [22].

Visual communication is already done with pictographic symbols and icons. Signs without words at airports and on our roads are evidence that symbolic visual communication works. Indeed, pictographic languages, such as Chinese, use characters originally derived from stylized drawings. Figurative images can be used to communicate specific concepts or objects, and work has been done to establish standardized line-drawn pictures of household objects and animals, for use in psychology experiments [38]. Images have also been used in psychology for the purpose of evoking emotions in experimental participants [19, 7]. The importance of images in establishing and developing a perceptual and emotional theme (or mood) for a design is recognized in the design practice of mood boards. These are a well-established way in which designers in domains such as fashion and interior design gain inspiration. Designers use them as a creative and analytical tool when developing a design idea [9]. To avoid specific figurative connections having an undue influence on viewers of a mood board, abstract images are often used [12].

These examples of work on communication show that it is a complex issue. Certainly the direct meaning within any episode of communication is important but that is just one of six aspects according to the Jakobson model. Visual communication can be explicitly symbolic or it can make use of visual conventions. The fact that some participants in the evaluation of the CVFM [32] thought they could communicate using images in an impressionistic way was perhaps not unexpected. It could be that artistic value aspect of a message suggested by Jakobson [15] was one component in the inspiration taken from visual feedback via the CVFM by designers. However, it is the complexity in communication that leads us to focus on a single aspect for the purposes of our experiment which we describe in the next section.

EXPERIMENT

As described in the *Introduction*, designer and crowd users, in an evaluation of an image based feedback system, thought that they could communicate successfully using an abstract image browsing interface and summaries. However questions remained about whether communication was actually taking place and whether the image summarization used to condense a crowd's many feedback image selections affected the communication. (The purpose of the summarization was to avoid designers being overwhelmed by the massed image selection of a potentially large crowd, and instead provide them with a digest of the feedback images). We designed an experiment to address these questions.

Below in *Aims*, we state the research questions. In *Method* the experiment is described by identifying a single aspect of communication on which to focus (the specific meaning in a message) and then setting out, diagrammatically, how the experiment addresses the two research questions. In *Task 1 –Terms to Images* the apparatus, interface and participants used for that task are described. *Producing the Image Summaries* describes the steps between the tasks in which the image selections output from Task 1 were summarized to become the stimuli for use in Task 2. Lastly, *Task 2 – Images to Terms* describes the apparatus, interface and participants for the second task.

Aims

The aims were to address two research questions:

RQ1

To what degree can meaning be communicated by a crowd's selections from a bank of abstract images?

RQ2

Are visual summaries of image selections, produced using the method described by Robb et al. [32], as effective at communicating meaning as the image selections which they summarize?

Method

We decided to focus on a single aspect of communication so as to minimize confounding factors from other aspects of communication. Taking Jakobson's [15] deconstruction we chose to focus on the specific meaning sent in a message basing the experiment on testing how much of that meaning is received using the medium of abstract images and automated visual summarization. Specific terms would be used to define the meanings in the messages being sent thus reducing as much as possible the amount of ambiguity in the minds of the participants 'composing' them. One group of participants would choose images to represent each of a set of terms and later a second group of participants would "read" or assess the image selections to decide the degree to which the meanings of the terms were present.

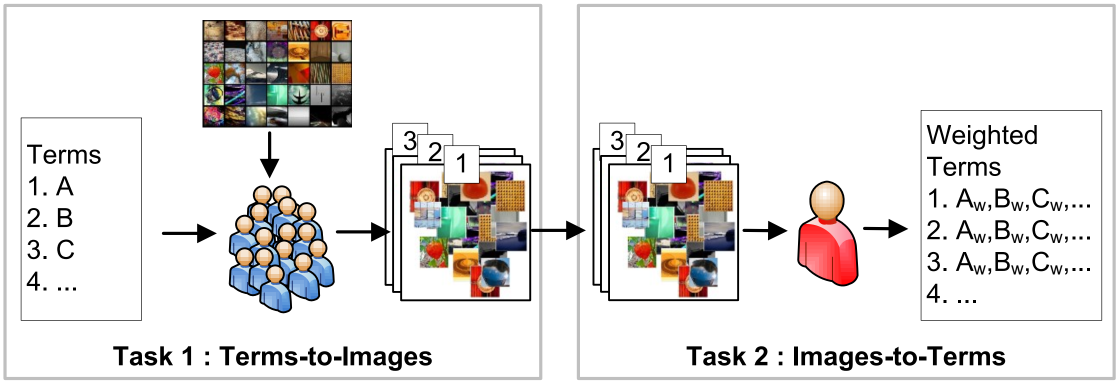

Figure 2. Addressing RQ1. Task 1: A group of participants selected images to represent each of a set of *n* terms producing *n* term image selection lists (TISLs). Task 2: A second group of participants viewed the *n* TISLs collated from Task 1 and, for each TISL, weighted the original set of terms to indicate the degree to which they could see the meaning of all *n* terms in that TISL.

Neither group would be aware of the purpose of the experiment other than to know that they were being asked either to choose images to represent given terms or judge the meaning of a given visual stimulus. Although the experiment as a whole can be viewed as being aimed at addressing RQ2 (the effectiveness of the image summarization), it is helpful to consider first how it addresses RQ1 (the degree of communication achievable with a bank of abstract images). For this reason the whole experiment is described with the aid of two diagrams introduced below.

Addressing RQ1

Figure 2 shows two tasks, each with a separate group of participants. In Task 1 participants viewed terms one at a time, and selected images to represent those terms. The image selections for each term, or term image selection list (TISL), were collected. In Task 2 the TISLs were shown as stimuli to a different participant group who were not informed that the stimuli they were to view had any intended meanings. For each TISL the participants output the full set of terms, assigning each term a weighting according to their judgment of the degree to which the meaning of the term was present in that TISL. The output weightings for each term were used as a metric for the effectiveness of communication of its intended term by each TISL; e.g. if the Task 2 participants, viewing the TISL which was intended to represent term *A*, allocated a high weighting for term *A* to that TISL relative to their weightings for the other terms, then this would be evidence that communication of term *A* had taken place. The success of the communication of each term relative to other terms was used to determine strengths and weaknesses of the abstract image browser for communication.

Addressing RQ2

Figure 3 shows where summarization was applied to the term image selection lists (TISLs) output from Task 1 to produce summaries as additional stimuli for Task 2. RQ2 was addressed by including these summaries along with the term image selection lists (TISLs) as the stimuli shown to the participants in Task 2. Thus, participants in Task 2 actually viewed two types of stimuli (TISLs and summaries). Therefore, because the output of from Task 2 for each stimulus consists of term weights, those term weights assigned to a TISL could be compared with those assigned to the corresponding summary thus allowing their communicative effectiveness to be compared.

Task 1- Terms to Images

The Terms

A set of terms was required for the experiment. It would need to be large enough to provide a range of different meanings, yet small enough for the experiment to be manageable in terms of participant cognitive load and fatigue. The factors weighed in choosing how many terms to use included the following: more terms would make it easier to discern a signal above possible random noise in weightings assigned by participants in task 2; fewer terms would decrease participant cognitive load and fatigue in both tasks (but particularly in task 2); fewer terms would present less difficulty in providing a suitable experiment presentation interface. With these considerations in mind, it was decided to seek a set of 20 terms. As the domain of fashion design was one of the original inspirations for the CVFM a sample of terms descriptive of material properties was sought as appropriate for that domain, and would serve as an abstraction for all material properties. The importance of emotions in design, decision making, and cognition is recognized in the literature [25, 20, 36, 41, 33]. Therefore to allow emotion communication to be assessed, a sample of emotion terms was sought. The emotion terms would serve as an abstraction of all emotion terms.

10 descriptive terms, e.g. “smooth”, were drawn from a study of consumer terms by Methven et al. which sourced 78 words used to describe fabrics from technical journals and from non-expert participants [24]. The perceived similarity between the terms was defined by having participants free group them based on their meanings. This similarity data was visualized using a dendrogram. We looked at Methven et al.’s data and exposed 11 clusters of terms by cutting their dendrogram at a particular height [10]. Two of the clusters contained terms such as “natural”

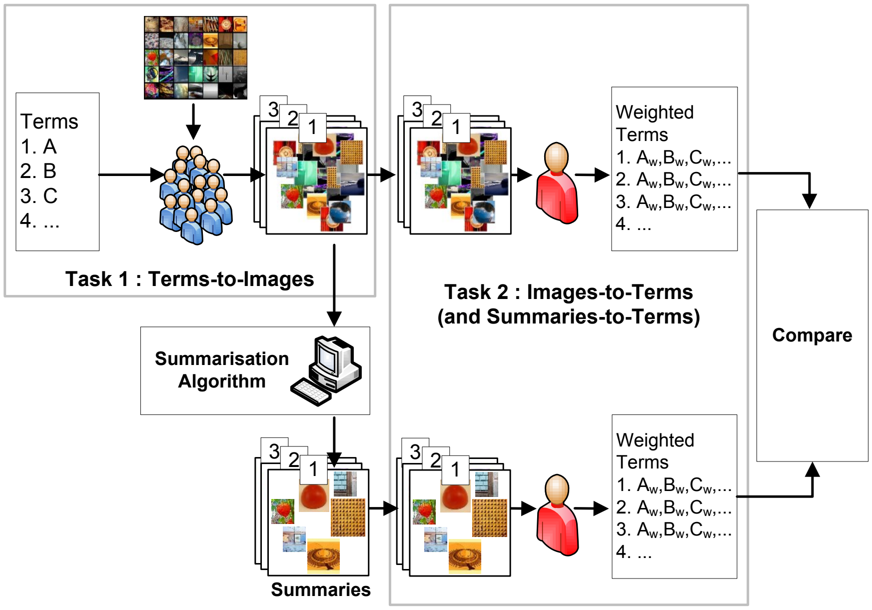

Figure 3. RQ2 was addressed by applying summarization to the image selections from Task 1 producing one summary for each term image selection list (TISL). This allowed comparison of the term weights for each TISL and its respective summary. Thus participants in Task 2 viewed two types of visual stimuli: term image selection lists (TISLs) and summaries. In this way both RQ1 and RQ2 were addressed in one experiment involving two separate participant groups.

and "even" and also "hot" and "cold", which were less relevant to fabric material than the other clusters. For our experiment we set aside these two clusters. We selected one term to represent each of the remaining nine clusters and one further term from the largest cluster to give 10 terms in total: Brittle; Coarse; Crumpling; Delicate; Flexible; Fuzzy; Smooth; Solid; Sticky; Textured.

10 emotion terms, e.g. "tenderness, love", were sourced from the Geneva emotion wheel model of emotions [35] which has been often used in research involving emotions [37, 28, 39]. Version 2 of the model [34] consists of 20 emotion terms arranged around the dimensions of valence (positive vs. negative or pleasant vs. unpleasant) and dominance (controlling vs. controlled or dominant vs. submissive) [29, 19]. Five terms from the negative valence and five terms from the positive valence regions of the wheel were chosen offering a balanced set of positive and negative emotion terms: "Astonishment,surprise"; "Disgust, repulsion"; "Embarrassment, shame"; "Enjoyment, pleasure"; "Involvement, interest"; "Irritation, anger"; "Sadness, despair"; "Tenderness, feeling love"; "Wonderment, feeling awe"; "Worry, fear".

Image Browser

Task 1 relied on participants being able to freely browse a collection of images while choosing some to represent the terms. We constructed a browser based on perceptual similarity as described by Padilla et al. [26, 27]. The source of the images was a random sample of Creative Commons licensed images from Flickr.com tagged with the word, '*abstract*'. We used abstract images so as to avoid an individual participant's own experiences or context biasing their choices in the experiment. Any images showing clear or conventional portrayals of objects, people, and writing, were discarded. The final set contained 500 images. Perceptual similarity data about the images gathered using human subjects was used to produce an intuitive organization for the browser. We decided not to depend on image similarity data based on computer vision features as these have been shown not to accurately reflect human perception [4]. Instead, we used similarity data gathered in perceptual grouping tasks with lab-based and crowdsourced participants to inform a rectangular self-organizing map (SOM) browser [16, 43]. (Figure 4.) A 6 x 8 configuration of image stacks was chosen as suitable for display on an iPad which was the platform we planned to use.

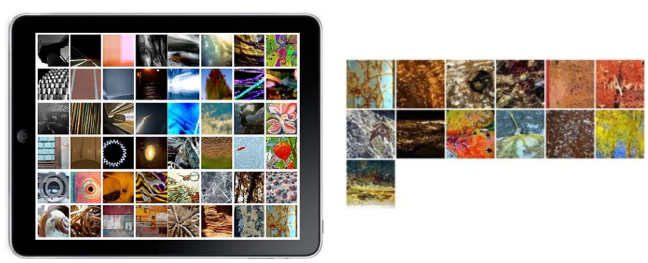

Figure 4. The abstract image browser showing the top level of 48 image stacks on an iPad (left) with an open stack (right).

This mobile platform would allow recruitment of participants from two university campuses, broaden the participant base (reducing any selection bias that might occur if relying on participants responding to email publicity and attending lab appointments), and increase the number of participants that could be on task at one time. It has been shown that, for example, usability testing can be successfully conducted and reliable quality data can be gathered away from the lab [1, 42]. The SOM browser is intuitive to use because image stacks open with a tap or click to reveal similar images. Adjacent stacks contain relatively similar images. Stacks far apart contain dissimilar images. Here "similar" is defined by the perceptual similarity data collected on the images.

Participants

Although the CVFM is intended to enable crowd feedback, because we were focusing on the utility of the browser and summarization, we decided campus sourced participants would be suitable to show that a sample of people can be stimulated to choose images from the browser for collection and summarization. 20 students (10 male, 10 female) were recruited from the two university campuses.

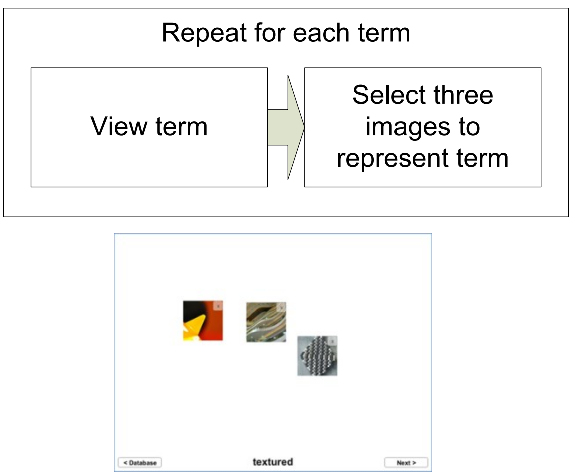

Figure 5. Top: Task 1 participant work flow. Bottom: Task 1 interface screen. For each term the screen started blank showing a term e.g. "textured" as the stimulus; participants tapped the *Database* button to access the image browser. After selecting three images, the participants used the "Next" button to get the next stimulus term.

The CVFM was intended for crowd feedback via the web and was hoped to transcend language barriers and perhaps cultural barriers too. Therefore, rather than restrict the sample to native English speakers, the minimum requirement was English as a foreign language (EFL). A sheet of dictionary definitions was provided in case any participant was in doubt about the meaning of any term. Each participant had to be confident of the meanings of the terms given the list of definitions. In Task 1, 7/20, 35% had EFL. The remaining 13 were native English speakers. There was a mixture of nationalities and cultures. 100g of chocolate was given as a reward for taking part.

Task 1 workflow and interface

Figure 5 illustrates the workflow for Task 1. Each participant viewed terms in a random order until they had selected 3 images for each of the 20 terms. We asked participants to choose 3 different images per term so that they would not be restricted to one single region of the image set. The data was recorded in a database. Mean time on task, excluding one outlier, was 25 minutes (median: 25; SD: 5; max.: 32; min.: 15). The outlier participant took 72 minutes. (She had spent the additional time browsing through the images, for fun as she found this enjoyable).

The Output from Task 1

Using database queries, the image selections were assembled into lists for each of the 20 terms forming the term image selection lists (TISLs). Each TISL contained 60 images. These TISLs became part of the input to Task 2 along with the summaries that would be produced from them.

Producing the Image Summaries

Each of the TISLs from Task 1 was processed following the algorithm described by Robb et al. [32] to produce a summary of 10 representative images (RIs). This method of summarization exploits the perceptual similarity data gathered for the image set and already used to organize the browser. The summarization is done through a combination of *k*-means clustering and multidimensional scaling (MDS) to produce a non-overlapping 2D montage of the RIs [5, 18]. This process is visualized in Figure 6.

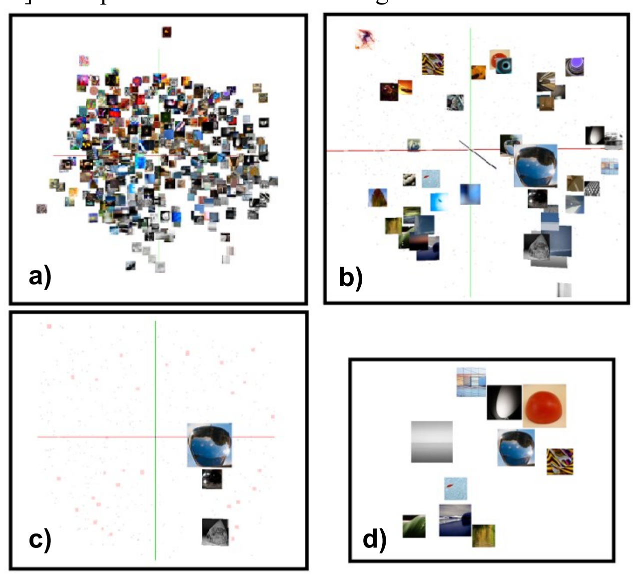

Figure 6. Summarization: a) The 500 abstract images arranged in a 3D MDS space according to their perceptual similarity. b) The TISL for the term "smooth" projected onto the 3D MDS space and sized by popularity. c) After clustering in multidimensional similarity space, one of 10 clusters isolated from the TISL is shown. (Pink dots mark the positions of images in other clusters). The image of the glass orb was calculated by the algorithm to be closest to the cluster centroid and became the representative image for that cluster. d) The summary for the "smooth" TISL. Image sizes are proportional to their popularity in the TISL.

Ten was chosen as the number of RIs for each summary as this would fit on an iPad screen while still allowing portrayal of a range of images within each 60-image TISL. The number of image repetitions within each TISL varied depending on the level of agreement on image choice among the Task 1 participants. Across the 20 TISLs the mean number of individual images was 46.2 (*s*:4.8; med.:46.5; max.:56; min.:36), with on average 13.8 images being repetitions of other images. Each representative image (RI) on the summary represented a cluster of images perceptually similar to each other. Repeated images in a cluster are more likely to become the RI for that cluster as the cluster centroid tends towards those images. The size (area) of each RI on a summary is proportional to the sum of the image selections in its cluster. The summary is a 2D projection of the RI's based on MDS. The proximity of one RI to another on the summary is related to the similarity of those two images as defined by the human-derived similarity data previously gathered about the image set.

Task 2

Participants

We recruited 60 students (30 male, 30 female) from two university campuses. As with Task 1 the participants were provided with a sheet of dictionary definitions, required to have English as a foreign language (EFL) as a minimum and be confident of the meanings of the terms given the list of definitions. (19/60, 32% had EFL. The rest were native English speakers). For Task 2 we gave each a $15 gift voucher as compensation for their time. This was greater than for Task 1 as the task was longer and many more judgements were required of each participant [2].

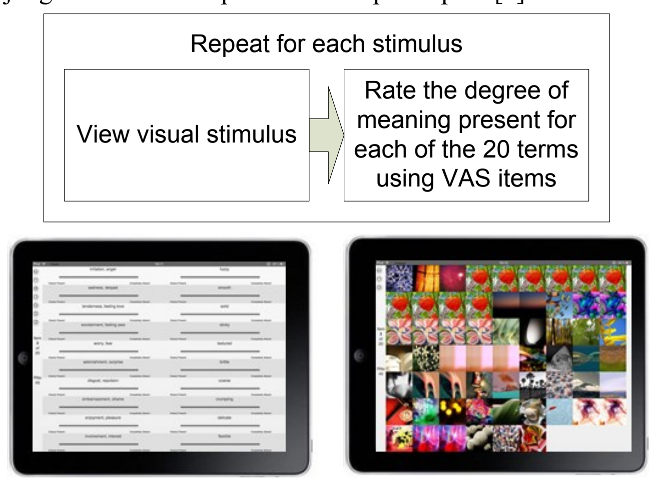

Figure 7. Top: The participant workflow for Task 2. Bottom: Master iPad (left) presenting VAS items, accepting input, and controlling progress; slave iPad (right) displaying stimuli (in this case one of the TISLs) and indicating progress.

Task 2 Workflow and Interface

Visual stimuli consisting of the TISLs from Task 1 and corresponding summaries were shown to the Task 2 participants. The stimuli were presented in random order and participants were required to rate each for the degree to which they could see the presence of the meaning of each of the 20 terms in the stimulus using visual analogue scale (VAS) items [14, 30]. We used VAS items as these offer much greater resolution than Likert items. There were 40 stimuli in all, 20 TISLs and 20 summaries. Figure 7 shows the workflow and interface. As requiring a participant to rate 20 meanings for all 40 stimuli would make the task too long, each participant was served a random selection of half the stimuli. The experiment application served the stimuli and recorded participant VAS item ratings in a database. The participants controlled progress by setting all the VAS items and then tapping a "continue" button on the master iPad. This triggered the web application to advance the slave iPad to the next stimulus. Each observation consisted of 20 interval scale measurements (Figure 8), representing the degree to which the participant could see the presence of the meaning of each of the 20 terms in a given visual stimulus. To avoid experimental bias due to VAS item positioning within the master display and scale anchor position (left or right), these were randomized for each participant. Mean time on task was 33 minutes (median: 30; SD: 10; max.: 61; min.: 16).

Figure 8. One of the 20 VAS items set for each stimulus. In addition participants viewed (and could recall at any time) a dialogue containing this question: "Is the meaning of the word or phrase present in the pictures?" The first tap on a VAS item scale caused a draggable cross to appear.

RESULTS

Analysis Methods

Two forms of analysis were used. First, we analyzed the frequency with which participants allocated their top ranking VAS score for a given stimulus to its intended meaning and used this to assess how well the intended meaning was being read from the various stimuli by the participants in Task 2. This allowed the performance of TISLs and summaries to be compared. Then, to show that any communication of meaning was not due to some bias in the image database, we analyzed the score distributions.

Frequency Analysis

Each TISL produced during Task 1 (and its summary) had an intended meaning (the term which the images were chosen to represent). The frequency with which participants in Task 2 allocated their top ranking VAS item score, among all 20 terms, for a given stimulus to its intended meaning (*f*1st) was analyzed. Standard competition ranking was used; i.e. a score's rank is always one plus the number of greater scores. This means a score which ties for first place counts as first rank. To establish the expected level due to random chance, 500 simulated studies each of 1000 random observations were generated. Sampling in this way was used to establish this random chance level as the VAS item scores ranged from 0 to 319. (The item scale length was based on its length in pixels as presented [31]). With 20 scores per observation the number of score combinations possible for one observation would be 319^{20}.

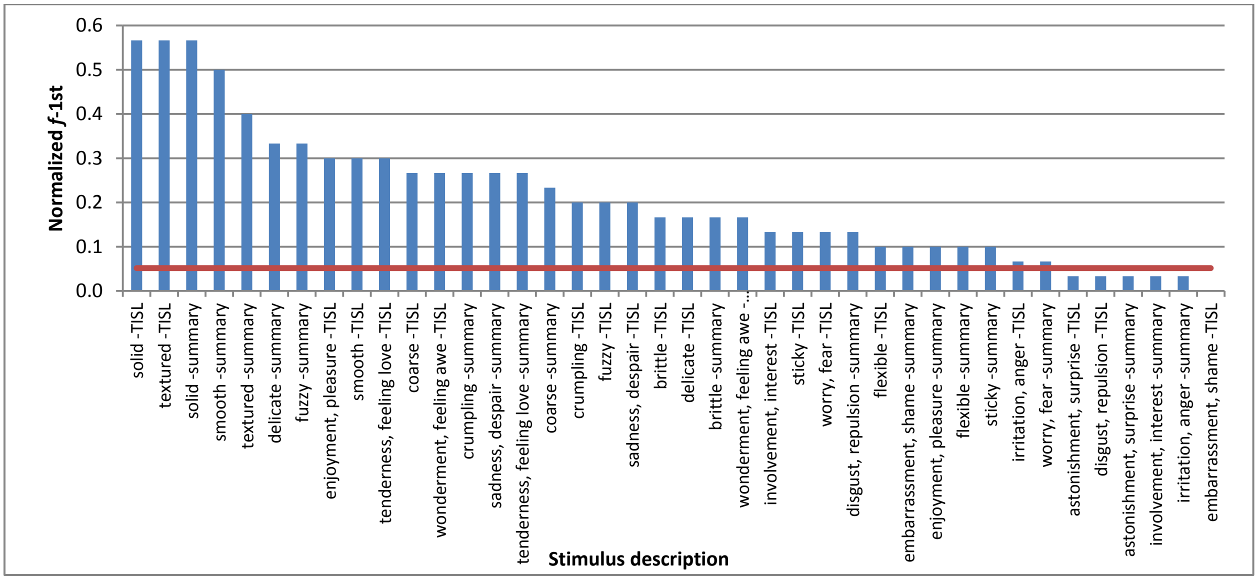

Figure 9. Chart showing the normalized *f*-1st (frequency of first rank score for intended meaning) for each stimulus, sorted high to low. The horizontal line shows the level that would be expected had participants set the VAS items at random.

Therefore, an exhaustive comparison of all possible scores would have been computationally impractical. The probability of a given term being awarded first rank in the simulated random studies was 5.15%. The chart in Figure 9 shows how *f* 1st varied across the different stimuli in the actual experiment. The chart also shows the random chance level. The *f* 1st was normalized by dividing it by the number of times a given stimulus was presented. The highest value, 0.57, for the solid TISL stimulus which was presented 30 times, equates to 17 out of 30 participants giving the intended meaning, *solid*, the top ranking score among all 20 terms.

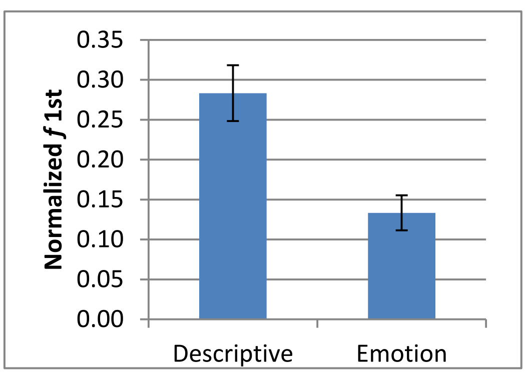

Figure 10. Comparison of mean *f* 1st, for stimuli representing descriptive terms vs. those representing emotion terms. (Error bars show 95% confidence intervals.)

The mean *f* 1st across the 20 stimuli (10 TISLs and 10 summaries) representing the descriptive terms was compared to the 20 for emotion terms. (Figure 10). The two means were compared using an independent *t*-test, the hypothesis being that the means were different. We found the mean *f* 1st for descriptive terms was significantly greater (M=0.283, SE=0.036) than for emotion terms (M=0.133, SE=0.023), $t(38)=3.543$, $p=0.001$, significant at the 95% confidence level. $r=0.498$ (a large effect) [11]. From that *t*-test we concluded our abstract image browser was significantly more effective at representing descriptive terms compared to emotion terms. We also compared the mean *f* 1st for stimuli consisting of each term's image selection list (TISL) with its corresponding summary. (Figure 11). We used a repeated measures *t*-test, the hypothesis being that the two means are different. We found the mean *f* 1st for the TISL stimuli was not significantly different (M=0.207 SE=0.034) than for summary stimuli (M=0.210, SE=0.035), $t(19) = 0.141$, $p=0.887$ which is greater than 0.05, not significant at the 95% confidence level. This represents a very small effect, $r=0.033$. Therefore, we can infer that, overall the summaries are equally effective at conveying their intended meanings as the TISLs which they summarize [11].

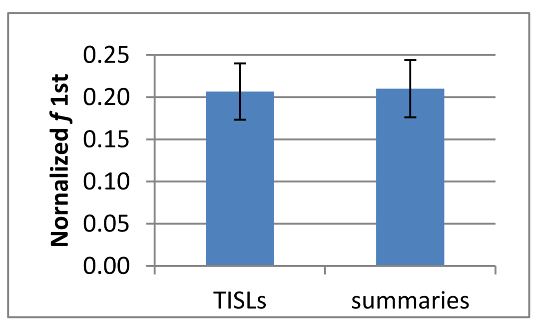

Figure 11. Comparison of mean *f* 1st for TISLs vs. summaries. (Error bars show 95% confidence intervals.)

Further confirmation of the similar performance of the summaries compared to their corresponding TISLs is shown by a correlation analysis. Figure 12 shows the *f* 1st for each TISL plotted against the *f* 1st for its corresponding summary. A Major Axis regression line is also shown. This is a line-of-best-fit for both x and y coordinates [40]. The regression line, $y=1.021x - 0.002$, crosses the axes close to the origin and has a gradient close to 1. A Pearson Correlation Coefficient (PCC) calculation shows that the PCC for summary *f* 1st vs. TISL *f* 1st is 0.77. This is evidence of a strong correlation [11].

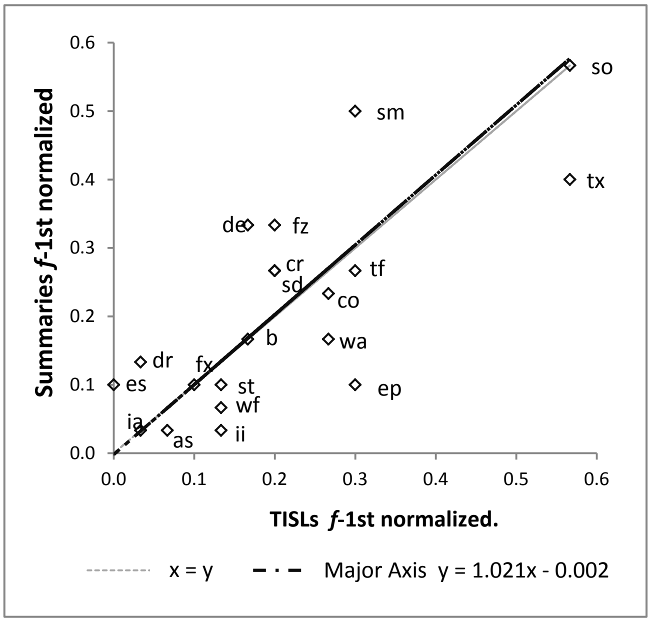

Figure 12. Scatter plot of the *f* 1st performance of TISLs vs. their corresponding summaries illustrating correlation. Also plotted is the Major Axis regression line lying almost on top of *x*=*y* plotted for comparison. Note: cr and sd are coincident points. Data point key: *as-astonishment, surprise; b-brittle; co-coarse; cr-crumpling; de-delicate; dr-disgust, repulsion; es-embarrassment, shame; en-enjoyment, pleasure; fx-flexible; fz-fuzzy; ii-involvement, interest; ia-irritation, anger; sd-sadness, despair; sm-smooth; so-solid; st-sticky; tf-tenderness, feeling love; tx-textured; wa-wonderment, feeling awe; wf-worry, fear.*

From these results comparing the performance of the TISLs with their summaries (by comparison of means, correlation and major axis regression) we conclude that the algorithm for creating summaries of image selections chosen from our abstract image browser is a valid and effective way of summarizing such image selections.

Score Distributions Analysis

To show that any communication of meaning taking place was not due to some bias in our image database, we analyzed the VAS score distributions. The score distributions of the intended term for a given stimulus were compared to the distribution of that term's scores over the other stimuli. This would confirm the presence of a signal and show that an effect apparent in the transmission of meanings was not simply due to a bias in the collection of 500 images arranged in the SOM browser. The Mann-Whitney-Wilcoxon (MWW) test allowed comparison of the term VAS item score distributions based on the relative rankings of the individual scores [11]. We divided the stimuli two ways. We separated the term image selection lists (TISLs) from the summaries and those representing descriptive terms from those representing emotion terms. This gave four groups: descriptive TISLs, descriptive summaries, emotion TISLs and emotion summaries.

	Term	TISLs		Summaries	
		p	***r***	***p***	***r***
Emotion	astonishment surprise	0.034	0.18	0.107	0.13
	disgust, repulsion	0.000	0.34	0.000	0.35
	embarrassment, shame	0.097	0.13	0.365	0.08
	enjoyment, pleasure	0.000	0.34	0.000	0.33
	involvement, interest	0.084	0.15	0.396	0.07
	irritation, anger	0.013	0.22	0.001	0.30
	sadness, despair	0.000	0.38	0.000	0.43
	tenderness, feeling love	0.000	0.43	0.000	0.57
	wonderment, feeling awe	0.004	0.26	0.000	0.34
	worry, fear	0.001	0.31	0.012	0.23
Descriptive	brittle	0.002	0.29	0.001	0.30
	coarse	0.000	0.38	0.000	0.41
	crumpling	0.000	0.48	0.000	0.37
	delicate	0.000	0.33	0.000	0.39
	flexible	0.021	0.20	0.035	0.18
	fuzzy	0.000	0.44	0.000	0.54
	smooth	0.000	0.58	0.000	0.68
	solid	0.000	0.49	0.000	0.57
	sticky	0.000	0.46	0.007	0.25
	textured	0.020	0.21	0.009	0.24

Table 1. *p* and *r* values from the Mann-Whitney-Wilcoxon tests of the score distributions for stimuli. Underlined *p* values are the five non-significant results. The remaining 35 results are all significant at the 0.05 probability threshold.

We compared the set of scores for a given stimulus, e.g. the *brittle* TISL, on its matching term VAS item, with the *brittle* VAS item scores for all the other stimuli in their group (descriptive TISLs). The *brittle* scores on the other stimuli in that group would represent any bias in the image database towards the term, *brittle*, if bias existed. Table 1 shows these results. For the 20 descriptive stimuli the *p*-values were all significant at the 0.05 probability threshold. Effect sizes varied from 0.18 (the "flexible" summary) up to 0.68 (the smooth TISL). For the emotion stimuli the *p*-values for 15 of the 20 stimuli were significant at the 0.05 probability threshold. Effect sizes varied from 0.07 (the "involvement, interest" summary) up to 0.57 (the "tenderness, feeling love" summary). As we did 40 comparisons at the 0.05 probability threshold, it might be predicted that two (0.05 x 40) of these comparisons could be a false positive due to type 1 errors [11]. 35 out of the 40 were positive. Looking at Table 1 we can see that 17 of the 20 terms are associated with positive MWW test results, but, statistically, two could be false positives. Taking an overview and not being concerned with exactly which of these comparisons might be affected, these results confirm that for 15 of the 20 terms (75%) the MWW test detected a significant signal i.e. communication of that term was shown to have taken place using the abstract images.

DISCUSSION AND CONCLUSIONS

First we address the two research questions and then discuss further work and implications for implementing the crowdsourced visual design feedback method (CVFM).

RQ 1

To what degree can meaning be communicated by a crowd's selections from a bank of abstract images?

The score distribution analysis results (Mann-Whitney-Wilcoxon tests) show that two groups of people (the participants in Tasks 1 and 2) were able, to some degree, to communicate the majority (75%) of our descriptive and emotion terms through the medium of image selections from the abstract image browser. While showing that communication of terms can be achieved, the results also show that the relative effectiveness of the communication varied depending on the term. The analysis of the frequency of first rank for intended meaning (f 1st) bears this out with a t-test showing that descriptive terms were communicated better than emotion terms. While the implications of this varying effectiveness are discussed further in *Design Implications*, what these results mean, in terms of the crowd sourced visual feedback method, is discussed below.

The overall effectiveness of the communication of terms using the abstract image browser and summarization shows that one cannot expect to reliably send a literal message this way. To send a literal message one would, naturally, use written or spoken language. However, in attempting communication of terms using an abstract image collection as a "vocabulary", we found that there was a detectable signal amongst the noise. The presence of this detectable signal goes some way to explain how it is that Robb et al. found that a segment of a crowd and designer participants thought that they can communicate about the mood of a design using the abstract images and summaries [31, 32]. Those participants in those studies were not simply *imagining* that communication was possible. Our results here show that, on an impressionistic level, it *is* possible.

RQ 2

Are the summaries as effective at communicating meaning as the image selections which they summarize?

The analysis of the f 1st (frequency of first rank for intended meaning), by comparing means and by correlation, showed that although the effectiveness of the communication varied, the summaries were of equal effectiveness at communicating their intended terms as the associated TISLs (term image selection lists).

As described earlier, the summarization method depends on human perceptual similarity data, clustering and MDS. This result gives us confidence that by summarizing a crowd's abstract image feedback in this way a designer using the crowdsourced visual feedback method and viewing a summary is not missing out on significant amounts of the meaning contained in the original feedback images. The summarization process is effective at preserving meaning in image selections and does successfully portray a meaningful overview of the images that it summarizes.

Further Work

It would be possible in the future, to conduct a similar experiment with another similarly constructed image collection, in order to assess the relative effectiveness of such image collections for specific communication of terms. The results could be used to decide the appropriateness of image sets for inclusion in a platform implementing the crowdsourced visual feedback method. Alternative browser organizations could be compared for effectiveness in this way. An experiment such as this could be used to help optimize the number of images to include on the summaries. Rather than comparing summaries to term image selection lists, summaries containing a different number of representative images could be compared. Additionally, although visual communication offers language independence it is not culturally independent. Visual conventions can vary with culture e.g. color [21, 17]. How cultural background affects communication with image feedback would be an interesting avenue to pursue.

Design Implications

Our results, showing that the abstract image browser is better for communicating descriptive rather than emotion terms, prompted us to develop an additional browser of emotion images to be deployed alongside the abstract image browser. The emotion images were deliberately selected and categorized based on their perceived emotional content. The intention was that the two browsers would provide complimentary image collections for use in a visual design feedback platform allowing designers to see the reactions of the crowd to their designs via images.

The effectiveness of the summarization means that visual feedback consisting of massed image selections drawn from perceptually organized image browsers (such as that used in our experiment) can be presented to designers as an easily digestible summary. While we envisage this form of visual feedback working for any aesthetic design e.g. product and graphic design, it could also be used as a medium for visual commentary, e.g. on video posting sites, to attract image-based reactions to be displayed as a visual summary. As a feedback medium, the controlled nature of the image sets used means that there would be no problem with unsuitable feedback posts as there is with text in comment forums. Thus these image feedback summaries can be counted on to be inoffensive as well as visually stimulating.

Our main finding, that the image summarization algorithm used in the crowdsourced visual feedback method is effective, means that designers can consume the image feedback in this summarized form confident that it represents a fair summary of the total image feedback.

ACKNOWLEDGMENTS

Funded by Heriot-Watt University's CDI theme. We thank Prof. Patrick R. Green for his advice. Creative Commons browser images are acknowledged here: bit.ly/tlab-ack.

REFERENCES

1. Chris Andrzejczak & Dahai Liu. (2010). The effect of testing location on usability testing performance, participant stress levels, and subjective *testing* experience. *Journal of Systems and Software*, 83(7), 1258-1266
2. Colin F. Camerer, & Robin M. Hogarth. (1999). The Effects of Financial *Incentives* in Experiments: A Review and Capital-Labor-Production Framework. *Journal of Risk and Uncertainty*, 19(1-3), 7-42
3. Daniel Chandler. 2002. *Semiotics : the basics (2nd ed.)*, Routledge.
4. Alasdair D.F. Clarke, Fraser Halley, Andrew J. Newell, Lewis D. Griffin, & Mike J. Chantler. 2011. Perceptual similarity: a texture challenge. In *Proceedings of the 22nd British Machine Vision Conference* (BMVC'11), 120.
5. Trevor F. Cox, & Michael A. A. Cox. 2001. Multidimensional *Scaling (Second Edition ed.).* Chapman & Hall/CRC.
6. Jonathan Culler. 1976. *Saussure.* Fontana.
7. Elise S. Dan-Glauser & Klaus R. Scherer. 2011. The 8 Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*, 43(2), 468-477.
8. Dribbble 2015. https://dribbble.com/ (Last accessed January 10/2015)
9. Claudia Eckert & Martin Stacey. 2000. Sources of inspiration: a language of design. *Design Studies*, 21(5), 523-538.
10. Brian Everitt 1974. *Cluster Analysis*. Heinemann.
11. Andy Field 2009. *Discovering Statistics Using SPSS (3rd ed.)* Sage.
12. Steve Garner & Deana McDonagh-Philp. 2001. Problem interpretation and resolution via visual stimuli: the use of 'mood boards' in design education. *Journal of Art & Design Education*, 20(1), 57-64.
13. Pierre Guiraud. 1971. *Semiology,* Routledge.
14. Joeri Hofmans and Peter Theuns. 2008. On the linearity of predefined and self-anchoring Visual Analogue Scales. *British Journal of Mathematical and Statistical Psychology,* 61(Pt 2).
15. Roman Jakobson. 1960. Closing statement: Linguistics and poetics. *Style in language*, 350, 377.
16. Teuvo Kohonen. 1990. The self-organizing map. *Proceedings of the Institute of Electrical and Electronics Engineers*, 78(9), 1464-1480.
17. Gunther R Kress, & Theo Van Leeuwen. 1996. *Reading images: The grammar of visual design*, Psychology Press.
18. Joseph B. Kruskal 1964. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2), 115-129.
19. Peter J. Lang, Margaret M. Bradley, & Bruce N. Cuthbert. 2008. *International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8.*University of Florida.
20. Lerner, J.S., Small, D.A. and Loewenstein, G. 2004. Research Report Heart Strings and Purse Strings Carryover Effects of Emotions on Economic Decisions. *Psychological Science*, 15(5), APS, 337-341
21. David McCandless. (2009). *Information is beautiful*. Collins.
22. Albert Mehrabian & Susan R. Ferris. 1967. Inference of attitudes from nonverbal communication in two channels*. Journal of consulting psychology*, 31(3), 248.
23. Albert Mehrabian & Morton Wiener. 1967. Decoding of *inconsistent* communications. *Journal of personality and social psychology*, 6(1), 10.
24. Thomas S. Methven, Pawel M. Orzechowski, Mike J. Chantler, Sharon Baurley & Douglas Atkinson. 2011. Comparison of Crowd-Sourcing vs. Traditional Techniques for Deriving Consumer Terms. In *Digital Engagement '11*, http://de2011.computing.dundee.ac.uk/?page_id=211 (last accessed September 24th 2015).
25. Mitsuo Nagamachi. 1995. Kansei engineering: a new ergonomic consumer-oriented technology for product development, *International Journal of Industrial Ergonomics* 15(1), 3-11.
26. Stefano Padilla, Fraser Halley, David A. Robb, and Mike J. Chantler. 2013. Intuitive Large Image Database Browsing using Perceptual Similarity Enriched by Crowds. In *Proceedings of the 15th International Conference on Computer Analysis of Images and Patterns* (CAIP'13), Springer, 169–176.
27. Stefano Padilla, David A. Robb, Fraser Halley, & Mike J. Chantler. *2012*. Browsing Abstract Art by Appearance. In *Proceedings of the 3rd International Conference on Appearance: Predicting Perceptions*, Lulu Press, 100-103.
28. Sathish Pammi & Marc Schröder. 2009. Annotating meaning of listener vocalizations for speech synthesis. In *Proceedings of 3rd IEEE International Conference on Affective Computing and Intelligent Interaction* (ACII'09), 1-6.
29. Mick J. Power. 2006. The structure *of* emotion: An empirical comparison of six models. *Cognition & Emotion,* 20(5), 694-713
30. Ulf-Dietrich Reips and Frederik Funke. 2008. Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods*, 40(3), 699-704.

31. David A. Robb, Stefano Padilla, Britta Kalkreuter, and Mike J. Chantler .2015. Moodsource: Enabling Perceptual and Emotional Feedback from Crowds. In *Proceedings of the ACM Conference Companion on Computer Supported Cooperative Work & Social Computing* (CSCW'15), 21-24.
32. David A. Robb, Stefano Padilla, Britta Kalkreuter, and Mike J. Chantler. 2015. Crowdsourced Feedback With Imagery Rather Than Text: Would Designers Use It? In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI'15), 1355-1364.
33. Mark A. Runco. 2014. *Creativity: Theories and themes: Research, development, and practice*, Elsevier.
34. Vera Sacharin, Katja Schlegel and K. R. Scherer. 2012. *Geneva Emotion Wheel rating study (Report).* University of Geneva, Swiss Center for Affective Sciences.
35. Klaus R. Scherer. 2005. What are emotions? And how can they be measured? *Social Science Information*, 44(4), 695-729.
36. Norbert Schwarz, Herbert Bless, & Gerd Bohner. 1991. Mood and persuasion: Affective states influence the *processing* of persuasive communications. *Advances in experimental social psychology,* 24, 161-199.
37. Ingo Siegert, Böck Bock, Bogdan Vlasenko, David Philippou-Hubner & Andreas Wendemuth. 2011. Appropriate emotional labelling of non-acted speech using basic emotions, Geneva emotion wheel and self-assessment manikins. In *Proceedings of the IEEE International Conference on Multimedia and Expo* (ICME'11), 1-6.
38. Joan G Snodgrass, & Mary Vanderwart. 1980. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory*, 6(2), 174.
39. Mohammad Soleymani & Maja Pantic 2012. Human-centered implicit tagging: Overview and perspectives. *In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics* (SMC'12), 3304 – 3309.
40. Andrew R H Swan & Michael Sandilands. 1995. *Introduction to geological data analysis*, Blackwell. 174-177.
41. Larissa Z. Tiedens and Susan Linton. 2001. Judgment under emotional certainty and uncertainty: the *effects* of specific emotions on information processing. *Journal of personality and social psychology*, 81(6), APA, 973.
42. Tom Tullis, Stan Fleischman, Michelle McNulty, Carrie Cianchette & Margaret Bergel. 2002. An empirical comparison of lab and remote usability testing of web sites.
43. Johan H. J. Vesanto, Esa Alhoniemi and Juha Parhankangas. 1999. Self-organizing map in Matlab: The SOM Toolbox. In *Procedings of the Matlab DSP Conference*. 35-40.