HeadCrowd: visual feedback for design

Britta Kalkreuter, David Robb
School of Textiles and Design, Heriot-Watt University
B.Kalkreuter@hw.ac.uk, darl14@hw.ac.uk

Abstract
HeadCrowd is a collaboration between the School of Mathematical and Computer Sciences and the School of Textiles and Design at Heriot-Watt University. It investigates how rich web and mobile applications may be employed to provide designers with near instantaneous and highly visual feedback from thousands of potential customers, or crowds. We are exploring the use of state of the art rich media applications to add quantity, speed and statistical accuracy to the study of semiotics, and the use of visuals in fashion as communication.

The project seeks to add to participatory design and market intelligence processes by enabling rapid and iterated co-design cycles between crowds and designers based on visual forms of communication so as to mirror the highly visual nature of fashion design inspiration. Such a scheme shows applications for sustainability in fashion if it can give crowds a concrete sense of ownership of the design process and provide enthusiastic target markets, thereby offering potential to significantly reduce the risks of producing unwanted product.

The paper provides an analysis of prior knowledge before describing the first two stages of the project, in which a pilot browser has been constructed that allows observers to navigate a vocabulary of 500 images which have been ordered into 48 similarity stacks using a mixture of human and crowd sourced sorting techniques. A first test involved the presentation of 20 terms to observers and asking them to choose 3 images from the browser to represent each term. Analysis of the resulting pilot data has given insights into the communicative certainty that a selection of 3 images from a vocabulary of 500 can provide for certain types of terms, and amongst certain groups of testers. It has also prompted deeper analysis of the pilot browser.

To put the communicative value of visual feedback to the test, the current research phase is preparing the reverse experiment of asking a fresh cohort of participants to associate images back to the original terms, and various interfaces are currently being constructed to facilitate the presentation of visual choices from phase 1. The similarity relationship between test images is investigated and visualized before a case is made for comparative experiments of raw selection data and versions of visual summaries in this second research phase, in order to test which way of data presentations best convey the intended visual communication.

Keywords: visual feedback, rich web applications, participatory design, sustainability
HeadCrowd: visual feedback for design

HeadCrowd is a collaboration between the School of Mathematical and Computer Sciences and the School of Textiles and Design at Heriot-Watt University. It will employ web and mobile applications to provide designers with near instantaneous and highly visual feedback from thousands of potential customers and collaborators, instead of using the long development processes that have added value to them, thus contributing to sustainable design solutions.

Visual prompts for designers are routinely used in commercial forecasting, and individual ambitions in fashion design are often served by a growing interest in crowd-sourcing co-design, and while web 2.0 and similar platforms have opened these new possibilities of participation to designers, and while web 2.0 and similar platforms have opened these new possibilities of participation to designers, there are still many issues of design as a way to ensure less fast and more sustainable fashion. It is hoped our proposed system will be taken up by retailers and used to improve on-line promotion of designers and products.

An example of an existing business gathering the opinions of an internet crowd on clothing designs prior to manufacture is the online fashion retailer ModCloth.com. On “Be the Buyer” (Modcloth.com 2012) garments not yet in production are shown and members of the site, who need not necessarily be expert buyers, are asked to discuss the suitability of the designs. If they are, in fact, produced, the opinion can be in the form of positive or negative votes and a comment. The comments are in free text but guided towards several specific areas such as colour and “wearability”, with a majority of comments focusing on the former.

It is hoped our proposed system will be taken up by retailers and improved on-line promotion of designers and products, generating measurable gains in the marketplace in the short term. However, Sanders & Simons (2009) assert that additional benefits may accrue from co-creation in the form of long term benefits in brands and events including improved quality of life and improved ecological sustainability.

A major factor in co-design is the context in which the co-design activity takes place, or the design space. Sanders & Westerlund (2011) describe this as encompassing three aspects: the actual design problem; the physical space in which the activity happens, the experience and the design space. Sanders & Westerlund point out that the first two of these aspects are clearly important in the eventual outcome. In the case of our project is that the crowd will fragment into followings for the different designers. If that occurs, and such an individual following begins to form a group mentality, then the “wisdom” of the crowd could be compromised. If this does happen, it is hoped that the value added (and purchases made) within the designer’s following will outweigh any loss of global value in the finished design due to any degradation of the global’s judgement. Indeed it may be sensible for the feedback from the designer’s following to be collected and analysed separately from the global. With this in mind, the users will receive two separate streams of feedback and make design decisions accordingly.

Semiology and Communication

It can be argued that the work of this project seeks to establish a visual language for fast intuitive visual feedback. In that case a broad examination of some principles of linguistics and semiotics is appropriate.

Semiology is usually taken to be the study of symbols and the various visual signs and signals. It is the study of the relationship between signs and what they represent. The study of semiotics is an important part of design, especially in design processes where there are many stakeholders with different interests, needs and expectations.

Semiology and Communication

For example, Saussure, in his theory of language...
Visuals to Words

Taking into account what we already know about visual feedback, HeadCrowd considers design feedback capabilities of images in their own right, but uses the written word in experiments to verify what feedback was intended and picked up by giver and recipient of that message respectively. To minimize the pre-determination of visual choice, the project uses a large image set, thus allowing for as free a selection as possible, and asks feedback givers to select three out of 500 visuals for each term described.

The first technical phase of the research concentrated on shaping prototype interfaces that allow fast and intuitive navigation of big numbers of visuals to facilitate selection by crowds; this essentially meant developing algorithms and protocols which automate the organization of large rich data sets. Easy navigation of these data sets was key to obtaining results that are relatively unaffected by fatigue, and our current work has made use of research on perceptually relevant image browsing by Halley et al. [2012]. It presents a “bootstrap” SOM of the 100 hand sorted image set was then enlarged to a visual “vocabulary” of the desired size of 500 by a crowd sourcing team that the number of Flickr accounts (or other appropriate sources) searched for images needed to be greater for future augmentation images. Each MTurk observer is shown 20 of the sorted images to observers sourced through Amazon Mechanical Turk (MTurk) who use this bootstrap SOM as being those most similar to the query image. The data from the observed likens between the 100 bootstrap images and the 400 augmentation images was used to calculate the augmented 500x500 similarity matrix. It is the larger similarity matrix on which the pilot study SOM browser is based. For the purpose of presenting this SOM on an iPad to the observers in the pilot study, an 8 x 6 grid was chosen. Fig 7 shows the full, 8 x 6 stack SOM browser.

Fig 3: A screen seen by an MTurk observer during the augmentation process. Fig 4 shows a screenshot seen by an MTurk observer during the augmentation process, where they are shown the query image (top left) and have to choose between 2 and 4 images from the bootstrap SOM as being those most similar to the query image. The data from the observed likens between the 100 bootstrap images and the 400 augmentation images was used to calculate the augmented 500x500 similarity matrix. It is the larger similarity matrix on which the pilot study SOM browser is based. For the purpose of presenting this SOM on an iPad to the observers in the pilot study, an 8 x 6 grid was chosen. Fig 7 shows the full, 8 x 6 stack SOM browser.

Fig 4: The full, 8 x 6 stack, SOM browser on iPad

Fig 2: Screenshot of image selection for “imagination”.

An example of near duplicates are the small and big glass apples on the screen in fig 1 which shows the full image selection by 20 observers for the term “smooth”; typical for the appearance of near duplicates in the pilot data, observers here clearly demonstrate their capability of identifying distinct images as similar despite their obvious and stark differences in scale, whereas an automated image sorter working on pixel levels could not. Since they are therefore treated as entirely unrelated images as far as electronic tagging is concerned, their obvious similarity will have to be considered as a limitation when evaluating statistics on frequency of identical image choices. This observation highlights the demand to use more sophisticated automated image sorting in future experiments.

Another realisation of the pilot data analysis was that a significant number of images made their way into the final 500-strong image set despite not actually following the project’s definition of abstract as they show people, full conventional depictions of objects or include writing or accepted symbols.

Limitions of Pilot Visual Data Set

A number of difficulties of this modus operandi were identified through analysis of the first pilot selection data: A small number of near duplicates were excluded by a process of screening the image pixel data, which meant sorting the images by average red, green and blue pixel level and then displaying the images in order side-by-side in screens of multiple images. One duplicate was found). Images which were very similar in composition to another given image, that specific pairing of images is given a high score. When communicating visually: Selections for the term “involvement, interest” by the 20 pilot testers (see fig 2) for example demonstrate that descriptive images rather than truly abstract ones, such as “smooth” (fig 1), where instances of semi-figurative images are notable lower. This trend holds firm for both emotional and 10 material terms in the visuals to word pilot.

Organising the Images into Manageable Interfaces

While HeadCrowd seeks to employ computer technology whenever appropriate, it recognizes earlier research that indicates the importance of a human basis for such an undertaking: The Self organizing Map (SOM) browser, as developed by Halley [2012], is effective because it is based on collected judgments of a small number of observers. For a small number of images this can be done by having observers freely group the images into however many similarity groups they wish, and 20 observers organized the first 100 images of this project in this way. Each time a given image is grouped with another given image, that specific pairing of images is given a similarity score. Before the score is added up and then divided by the number of opportunities there were for that pairing to have occurred. The resulting similarity matrix for the initial 100 is 100 rows by 100 columns and can be used to categorise the breadth of the available image “vocabulary”.

To eliminate the dangers of human fatigue while sorting, this nuclear, human sorted image set was then enlarged to a visual “vocabulary” of the desired size of 500 by a crowd sourcing modus operandi. HeadCrowd focused on the final stage of the process, where they are shown the query image (top left) and have to choose between 2 and 4 images from the bootstrap SOM as being those most similar to the query image. The data from the observed likens between the 100 bootstrap images and the 400 augmentation images was used to calculate the augmented 500x500 similarity matrix. It is the larger similarity matrix on which the pilot study SOM browser is based. For the purpose of presenting this SOM on an iPad to the observers in the pilot study, an 8 x 6 grid was chosen. Fig 7 shows the full, 8 x 6 stack SOM browser.

Visuals

In order to arrive at an appropriately large and varied image set, over 1800 images tagged as abstract were screen scraped from Flickr.com account holders who make their images available with a Creative Commons licence. Although only images tagged as abstract were used, a large number of miss-tagged images had to be discarded manually as they included images of people, full conventional depictions of objects and writing. Duplicates were excluded by a process of screening the image pixel data, which meant sorting the images by average red, green and blue pixel level and then displaying the images in order side-by-side in screens of multiple images. One duplicate was found). Images which were very similar in composition to another given image, that specific pairing of images is given a high score. When communicating visually: Selections for the term “involvement, interest” by the 20 pilot testers (see fig 2) for example demonstrate that descriptive images rather than truly abstract ones, such as “smooth” (fig 1), where instances of semi-figurative images are notable lower. This trend holds firm for both emotional and 10 material terms in the visuals to word pilot.

Again, this deviation from the originally intended selection criteria can partly be explained by human fatigue in the face of having to scrutinize such a large set of relatively small images, but it also calls for a tighter set of criteria for future image sets and further investigation of whether the margin for human error in privately tagged image collection makes them suitable for use in a controlled experiment.

However, while clearly a potential limitation of the current image set’s usefulness, we were able to use the existence of these non-abstract images in the pilot study to evaluate the varying efficacy of different image types for certain purposes, and certain groups with big numbers of non-abstract ones. Selections are asked to associate images with emotional terms, as opposed to choices for material terms such as “smooth” (fig 1), where instances of semi-figurative images are notable lower. This trend holds firm for both emotional and 10 material terms in the visuals to word pilot.

Despite their obvious and stark differences in scale, whereas an automated image sorter working on pixel levels could not. Since they are therefore treated as entirely unrelated images as far as electronic tagging is concerned, their obvious similarity will have to be considered as a limitation when evaluating statistics on frequency of identical image choices. This observation highlights the demand to use more sophisticated automated image sorting in future experiments.

Another realisation of the pilot data analysis was that a significant number of images made their way into the final 500-strong image set despite not actually following the project’s definition of abstract as they show people, full conventional depictions of objects or include writing or accepted symbols.

An example of near duplicates are the small and big glass apples on the screen in fig 1 which shows the full image selection by 20 observers for the term “smooth”; typical for the appearance of near duplicates in the pilot data, observers here clearly demonstrate their capability of identifying distinct images as similar despite their obvious and stark differences in scale, whereas an automated image sorter working on pixel levels could not. Since they are therefore treated as entirely unrelated images as far as electronic tagging is concerned, their obvious similarity will have to be considered as a limitation when evaluating statistics on frequency of identical image choices. This observation highlights the demand to use more sophisticated automated image sorting in future experiments.

Another realisation of the pilot data analysis was that a significant number of images made their way into the final 500-strong image set despite not actually following the project’s definition of abstract as they show people, full conventional depictions of objects or include writing or accepted symbols.

An example of near duplicates are the small and big glass apples on the screen in fig 1 which shows the full image selection by 20 observers for the term “smooth”; typical for the appearance of near duplicates in the pilot data, observers here clearly demonstrate their capability of identifying distinct images as similar despite their obvious and stark differences in scale, whereas an automated image sorter working on pixel levels could not. Since they are therefore treated as entirely unrelated images as far as electronic tagging is concerned, their obvious similarity will have to be considered as a limitation when evaluating statistics on frequency of identical image choices. This observation highlights the demand to use more sophisticated automated image sorting in future experiments.

Another realised of the pilot data analysis was that a significant number of images made their way into the final 500-strong image set despite not actually following the project’s definition of abstract as they show people, full conventional depictions of objects or include writing or accepted symbols.

An example of near duplicates are the small and big glass apples on the screen in fig 1 which shows the full image selection by 20 observers for the term “smooth”; typical for the appearance of near duplicates in the pilot data, observers here clearly demonstrate their capability of identifying distinct images as similar despite their obvious and stark differences in scale, whereas an automated image sorter working on pixel levels could not. Since they are therefore treated as entirely unrelated images as far as electronic tagging is concerned, their obvious similarity will have to be considered as a limitation when evaluating statistics on frequency of identical image choices. This observation highlights the demand to use more sophisticated automated image sorting in future experiments.

Another realisation of the pilot data analysis was that a significant number of images made their way into the final 500-strong image set despite not actually following the project’s definition of abstract as they show people, full conventional depictions of objects or include writing or accepted symbols.

Another realisation of the pilot data analysis was that a significant number of images made their way into the final 500-strong image set despite not actually following the project’s definition of abstract as they show people, full conventional depictions of objects or include writing or accepted symbols.

Another realisation of the pilot data analysis was that a significant number of images made their way into the final 500-strong image set despite not actually following the project’s definition of abstract as they show people, full conventional depictions of objects or include writing or accepted symbols.

Another realisation of the pilot data analysis was that a significant number of images made their way into the final 500-strong image set despite not actually following the project’s definition of abstract as they show people, full conventional depictions of objects or include writing or accepted symbols.

Another realisation of the pilot data analysis was that a significant number of images made their way into the final 500-strong image set despite not actually following the project’s definition of abstract as they show people, full conventional depictions of objects or include writing or accepted symbols. 
The image on the top of a particular stack in the SOM browser is the image in the stack which is nearest the centroid position of the perceptual similarity space represented by all the images in that stack. Once an observer clicks on one of the images on this top level, that centroid image, together with all images that were associated as similar (and have therefore been assigned to its stack) are shown on a screen for the observer to choose from. (See fig 5).

A detailed analysis of all pilot image choices will need to consider their position on the SOM browser as a possible determining factor for selection or non-selection by an observer. Already it emerges that although the 48 top level (6x8 iPad interface) images only make up less than 1/10 of the entire image set, their representation in final image choices is nearer 1/7, suggesting their prominent position has made them more likely to be chosen by observers. However, image position in the stacks and on the SOM browser interface is, of course, far from random as it correlates to the complex interrelationship of similarities that was assigned to its stack) are shown on a screen for the observer to choose from. (See fig 5).

Visualising similarities between images in stacks, and images in adjacent stacks through 3D MDS visualisation of the similarity space

A highly navigable way to observe how images had been assigned to their stack, that centroid image, together with all images that were associated as similar (and have therefore been assigned to its stack) are shown on a screen for the observer to choose from. (See fig 5).

The 3 dimensions depicted in the 3D MDS view describe over 80% of the variability in the similarity data and can also be viewed in clusters. (See fig 7).

Communicating Visually

With the aim of discovering whether meaning can be reliably communicated between participants of different backgrounds using a fixed set of images we are conducting a selection experiment in two phases:

**Phase 1 Visual to Word Pilot**

Phase 1 of the pilot has already been completed and involved the presentation of 20 terms to 20 observers, asking them to choose 3 images from the SOM browser to represent each term. The terms comprised of 10 descriptive words sourced from non-experts as being descriptive of fabrics or products [Methven et al 2011] and 10 emotional terms sourced from the “Geneva Emotional Wheel” [Scherer 2005]

20 Observers (10 male, 10 female, 10 designers, 10 non-designers from mixed (European and Asian) backgrounds in the 20-30 year age bracket) were shown the 10 descriptive terms followed by the 10 emotional terms in a random order and chose 3 different images from the SOM to represent each given term.

The experiment was presented to the observers on an iPad, and their responses were recorded on a central server, making data collection and analysis easy, for almost any size of future sample. Analysis concentrated on criteria and research questions based on somatic and sociological questions surrounding the image set and selection behaviour of the distinct observer groups in the pilot:

Quantitative evaluation of the pilot data, namely considering the frequency with which a particular visual was chosen by more than one observer to describe the same term, was identified as a key measure for assessing how strongly (and strongly) the communication value of an image is amongst respondents. Some interesting general trends emerged as well as some specific ones when filtering image choices by specific groups, and for specific terms:

When considering all of the 20 observers’ choices separately by type of term, it emerged that singular choices, seen above as an indicator for low communicative certainty of an image, accounted on average for 69.50% of selections for the emotional terms but for just 36.42% of images when observers had been asked to describe textural terms.

While fashion undoubtedly consists of a whole lot more than cloth stitched into shape, it means that the terms that are arguably more immediately meaningful to fashion design found more agreement across all 20 observers on which image might represent them. In fact it seems somewhat remarkable that almost half of the image choices for textural terms were agreed on by at least two of 20 individual observers, each faced with choosing just 3 per term from a bank of 500. Despite specific observers being less convinced on agreement on terms cannot be discarded in this study of visual feedback for design as they are frequently present in aforementioned forecasting and mood boards alike. The conclusion we must draw from this part of the pilot evaluation is however that particular care must be taken when eliciting and evaluating feedback on these more ephemeral aspects of design. The conclusion we must draw from this part of the pilot evaluation is however that particular care must be taken when eliciting and evaluating feedback on these more ephemeral aspects of design.

Even less agreement on the most appropriate visual signifier of a given term can be detected when looking at the pilot data from certain social groups (or distinct crowds’) point of view: sorting the choices by gender, males top the list of disagreeing by choosing on average 79.17% of images just once while

The Nordic Textile Journal
females display just 74.33% of singular choices. Looking at the pilot data according to professional background, designers and non-designers interestingly do not on average display a less or more unambiguous attitude to using visual signifiers (as might have been expected given different work practices as regards visuals) as both groups share an identical 77.17% of singular image choices. Looking at the pilot data in detail rather than by averages confirms some of the above described trends; e.g. when “solid” scores the lowest percentage for singular choices (i.e. the highest level of agreed communication) across all 20 observers with just 40%, while “involvement/interest” scores 100% of individual choices, meaning total disagreement of communication value. Detailed scrutiny of the pilot results also shows some startling deviations from the norm as females score just 50% of singular choices on “tenderness and love”, showing an unusually high level of choice overlap for an emotional term, while 93.33% of designers’ choices for “delicate” are not matched by any fellow professionals, though material terms generally generated more agreement. Minute scrutiny of the results will continue to identify all factors that may have had a significant influence on choices. Amongst these is the way in which the 500 images were presented on the iPad interface:

Is the number of images in the full set proportionate to the choices allowed each observer, i.e. can observers easily handle 500 images?

Do observers of a certain background (female, male, designer, non-designer) favour certain types of images to communicate specific things? A thorough analysis of mood boards and forecasts of choice for various types of images (as might be expected given different work practices as regards visuals) as both groups share an identical 77.17% of singular image choices.

As this semiotically motivated analysis of the pilot continues, HeadCrowd’s premise to use visuals as suggestive parts of a cumulative whole (the visual aggregation) is being researched in terms of a computer’s capability of calculating and representing what such a collage of image choices might look like, when informed by a set of collected human judgements about image similarity. A pilot aggregation of the 60 image selections per term according to a raft of popularity and similarity factors has now been designed to be used to facilitate the reverse experiment of eliciting term selections in response to images, or as Saussure would say, selecting a signified to the signifiers presented.

Visualisations of the Responses in Phase 1

The responses were visualised in two ways: All the chosen images for a given term were displayed simply in order of image ID number. Images that were chosen more than once were displayed multiple times to indicate this. (See Fig 1 and 2)

The images chosen for each term were displayed in 3D similarity space with the size (area) of each image being increased proportionately to signify its popularity. (See Fig 9).

Visual Summaries of the Image Selections

As the eventual volume of image feedback selections by an internet crowd is envisioned to be very large, unmanageable in fact for visual feedback, a way of summarising the selections was needed as an alternative to simply listing the images. To this end a cluster analysis was carried out on the chosen images based on their similarity data. The responses were divided into 10-clusters. The image nearest to a cluster’s centroid in the similarity space was chosen to represent that cluster in a summary collage which will then consist of just 10 images of size depending on each cluster’s popularity. The position of the images is determined by their position in the similarity space, projected into two dimensions. (See fig 10)

The next phase of the pilot will be a comparative study showing the raw (60 image) selections and their automated summary collages to a fresh set of observers in random order, asking them to indicate which of the 20 terms they think the given view represents. It will provide firmer answers as to how well meaning can be conveyed using mass selections from large image set. Showing raw selection data and its aggregated summary separately to the same observers will indicate whether summary collages are more or less effective than full sets of image selections at conveying meaning, with potentially wide ranging consequences for the usability of crowd-sourced visual feedback for design.
References


