A Comparison of Crowd-Sourcing vs. Traditional Techniques for Deriving Consumer Terms

Thomas S Methven[1, 3] Pawel M Orzechowski[1] Douglas Atkinson[2] Sharon Baurley[2] Mike J Chantler[1]

[1]Texture Lab
Heriot-Watt University
Edinburgh, EH14 4AS, UK

[2]Brunel University
Uxbridge, UB8 3PH, UK
[3]Correspondence: tm112@hw.ac.uk

ABSTRACT

In this paper, we present a comparison between controlled laboratory experiments and crowd-sourcing techniques for gathering consumer terms. We discuss the difference between the data gathered from these techniques and how crowd-sourcing can be used to increase engagement and seek feedback from your customers via the internet or other digital methods. As a case study, we present a new set of unipolar scales for fabric description, which we believe are recognisable and understandable for non-experts in the field.

Categories and Subject Descriptors

H.5.m [**Miscellaneous**]: Information Interfaces and Presentation (e.g. HCI) Miscellaneous.

General Terms

Design, Experimentation, Human Factors, Languages

Keywords

Crowd-Sourcing, Design Language, Consumer Terms, Fabrics

1. INTRODUCTION

With the proliferation of online shops selling clothing and other such products which rely so heavily on a user's ability to accurately judge material characteristics, there is a growing need for an accurate perceptual language to describe them. While there has been some work done in defining the texture space [1, 2] and some looking at the perceptual dimensions fabrics, for example [3], most of these were done with traditional experimental techniques and in the case of [3], in French. In addition, many of these fabric descriptors are decided by expert or trained witnesses, rather than the average consumer.

Instead, we aim to describe a technique which we used to illicit a set of fabric descriptors via crowd-sourcing to increase the pool of data in a quick and scalable way, and which could easily be used to engage your customers in a simple and digital manner. We also believe that giving consumers the ability to talk about and share information about fabrics and clothing in a meaningful way will dramatically increase their engagement and interest.

Digital Engagement '11, November 15 – 17, 2011, Newcastle, UK.

2. EXPERIMENTS

2.1 Property Experiments

As our goal was to develop a language which could be used by those naive to fabric and fashion, our first requirement was to determine the amount of agreement between the words experts used and those non-experts understood. To this end, we decided to refine a list of words gathered from technical journals. Starting with a list of 69 words, we performed two experiments: one to remove words not understood by naive participants and one to capture any commonly used words not on the technical list.

For the first experiment, we presented a group of 30 participants with 11 fabrics out of a set of 20 (we balanced this so each 20 fabrics were presented the same number of times) and asked them to handle and look at them as they pleased. While they were handling the fabric for two minutes we asked them for any words which they believed described it and we recorded their responses. From this we gained 429 valid words. From this list, we then removed any hedonic, temporal or emotional words and those based on opinion such as 'ugly', 'old' and 'expensive'.

For the second experiment, we printed all of the original 69 words onto business cards and gave the same 30 different participants the stack in a randomised order. We then asked them to put the cards in three piles: Words they understood and used, words they understood but didn't use and words which they didn't know the meaning of. From this, we were able to give each of the 69 words a score of how well understood they were. Any scoring 0 or less was then removed as the majority of our participants didn't know their meaning.

Finally, we compiled a final set of words from our experiments. We removed 20 words from the initial set and replaced them with 29 words from naive participants. It is worth noting at this point that there were four words which were removed from the expert list, but then re-added from the word discovery experiment. We decided that the discovery experiment should take precedence as users offering a word themselves seemed a stronger indicator of use than rating a list of pre-defined words. Therefore, we were left with a final list of 78 words which non-experts could understand.

2.2 Grouping Experiments

Although we had a set of 78 words which we knew were well understood, this was clearly too many for use in simple experiments and online feedback. In addition, we suspected many of the words had similar meanings (such as 'Fluffy', 'Furry' and 'Fuzzy') and wanted to be able to use just the most representative.

As any work involving language like this is based so heavily on varying opinion and background and requires large numbers of

[1,2] Contributions of the first three authors: Methven and Orzechowski designed and performed the property and grouping experiments using a literature survey performed by Atkinson for the initial word set and an online sorting tool developed by Orzechowski.

multiple techniques to see if it would be more time efficient to use crowd-sourcing to gather participants in future experiments.

To better understand the set of words, we performed a series of 'free sorting' experiments where participants were asked to sort the words into groups based on their meanings. They were allowed to have any number of groups, but were asked to avoid singletons.

After the participant was finished grouping, they were asked to indicate one word in each group which best represented their opinion of the overall meaning of the group. This was so we could pick a representative word from each group to form a taxonomy.

Using the collected grouping data we created a similarity matrix. Therefore we have an average distance between every word and each other word in the set. This allows us to easily find which words are more similar to each other. For example, 'Fluffy and 'Fuzzy' appear much more similar than 'Fluffy' and 'Stiff'. This distance information also allows us to 'cluster' the data in as many groups as are required. When also using how often a word is indicated as being representative, a single word can then be chosen to name the group.

This experiment was performed in three ways, one in a typical controlled laboratory setting using business cards with words printed on and two using an online tool which allowed users to group words on a computer screen. Of these two experiments, one was performed by getting participants in the local area, (via word of mouth, posters in local libraries, etc.) the other performed on Amazon's Mechanical Turk (M-Turk) crowd-sourcing framework. We discuss the merits of the different techniques in Section 4.

3. RESULTS

The final clustering data is available online [www.macs.hw.ac.uk/texturelab/resources/]. A dendrogram of these results has been provided in Figure 1, with a list of the groups and representative word scores provided in Figure 2.

For future experiments, we decided (for experiment length, and to avoid singletons or overly large groups) to cluster the results into 11 groups, and as such got 11 words which we could use for rating scales. We choose to present these as unipolar scales, as Picard et al. showed that often the most obvious bipolar scales can be incorrect [3].

The 11 unipolar scales are: Smooth, Synthetic, Natural, Delicate, Flexible, Irregular, Textured, Coarse, Warm, Fuzzy and Heavy.

4. TECHNIQUE COMPARISONS

As mentioned in Section 2.2, we performed the grouping experiment with three distinct groups. In addition to collecting data, we were therefore also able to come to some conclusions about the differing methods. The graph in Figure 3 shows the normalised scores of each participant as a distance from the average groupings. We used this to look at the difference in distributions between the different methods as each group had different numbers of participants.

To perform some simple analysis of the results, we decided to take the laboratory experiments as the ground truth and compare the other methods to see if they were statistically similar. It became readily apparent that both the laboratory experiments and those performed with word of mouth participants were not statistically different ($p = 0.53$), whereas the M-Turk results were ($p = 0.012$ and $p = 0.039$ vs. laboratory and word of mouth). This presents an interesting problem, because we were able to get vastly more participants using M-Turk than the other methods, but the data is much noisier.

One of the problems which we encountered with M-Turk was the motivation of the participants. In the laboratory experiments and with the participants sourced from the local area, they approached us and as such were eager to do the experiment. Therefore they were likely to listen to the instructions and perform the best they were able. Participants on M-Turk, however, are much more likely to be participating for the small financial reward.

Although necessary in many experiment designs, it therefore becomes much more important in M-Turk experiments to attempt to remove outliers and 'cheaters' in the process. This is a complex problem and requires many different techniques to solve. For example, we noticed a clear difference in results between those who took a very short time and those who took longer. We decided to remove any participant who took less than 5 minutes to complete the experiment, as well as those who were too far from the 'average' result of all the participants. Using these two metrics we found our M-Turk results were greatly improved and the average participant distance was no longer statistically different.

Another problem perhaps more prevalent for us was the background of the participants. Due to experiment being about language, we decided to limit our participants to those who were native English speakers. While M-Turk has options to limit experiments to those who fit your specific criteria, we found it insufficient. To try and combat this, we prefaced our experiment with a word meaning test to remove those who were unsuitable.

5. CONCLUSIONS

Our findings showed us that M-Turk is a viable way to get large quantities of data in a short amount of time. There were, however, a few important considerations when using it. First, of course, the experiment must be able to run on a wide variety of computers. Second, the data returned is much noisier than data collected from people who volunteer in the usual way and so, before you begin, it helps to have metrics to remove cheaters or outliers. Finally, if the experiment requires any specific weeding out of participants (such as native language or background) you can't rely on the inbuilt M-Turk tools and should test participants before the experiment.

Despite this, we were able to use a combination of both digital and traditional laboratory experiments to get a large quantity of participants in a short amount of time, which we believe led to a much stronger result and more universally acceptable and understandable unipolar scales for future use.

6. ACKNOWLEDGEMENTS

We wish to thank the EPSRC for funding this project through grants EP/F02553X/1 and EP/H007083/2.

7. REFERENCES

[1] Gurnsey, R. and Fleet, D. J. Texture space. *Vision Research*, 41, 6 2001), 745-757.

[2] Rao, A. R. and Lohse, G. L. Identifying high level features of texture perception. *CVGIP: Graphical Models and Image Processing*, 55, 3 1993), 218-233.

[3] Picard, D., Dacremont, C., Valentin, D. and Giboreau, A. Perceptual dimensions of tactile textures. *Acta Psychol*, 114, 2 (Oct 2003), 165-184.

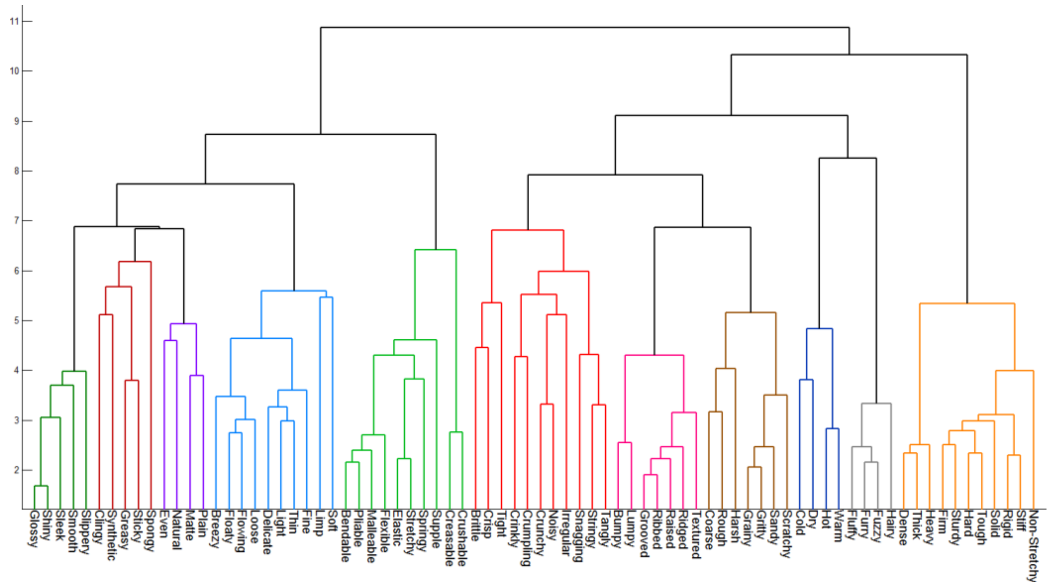

Figure 1 - Properties Dendrogram

Smooth 28	Glossy 24	Sleek 24	Shiny 21	Slippery 14						
Synthetic 37	Sticky 26	Clingy 25	Greasy 13	Spongy 10						
Natural 37	Plain 35	Even 14	Matte 14							
Delicate 38	Flowing 31	Soft 31	Loose 28	Fine 24	Light 20	Breezy 14	Floaty 13	Thin 11	Limp 5	
Flexible 34	Stretchy 30	Pliable 22	Elastic 21	Creasable 20	Crushable 18	Malleable 17	Supple 17	Springy 12	Bendable 5	
Irregular 35	Stringy 21	Tight 19	Crumpling 19	Tangly 19	Crisp 18	Noisy 15	Crinkly 14	Brittle 11	Snagging 10	Crunchy 9
Textured 64	Ribbed 14	Bumpy 13	Grooved 11	Raised 11	Ridged 11	Lumpy 6				
Coarse 34	Rough 20	Grainy 18	Harsh 13	Scratchy 12	Gritty 10	Sandy 10				
Warm 40	Hot 29	Dry 14	Cold 12							
Fuzzy 27	Fluffy 20	Furry 18	Hairy 10							
Heavy 29	Sturdy 25	Rigid 19	Thick 18	Solid 17	Dense 16	Non-Stretchy 16	Firm 14	Tough 13	Stiff 13	Hard 11

Figure 2 – Property Groups and Representative Word Scores (N.B. Colours correlate to Figure 1)

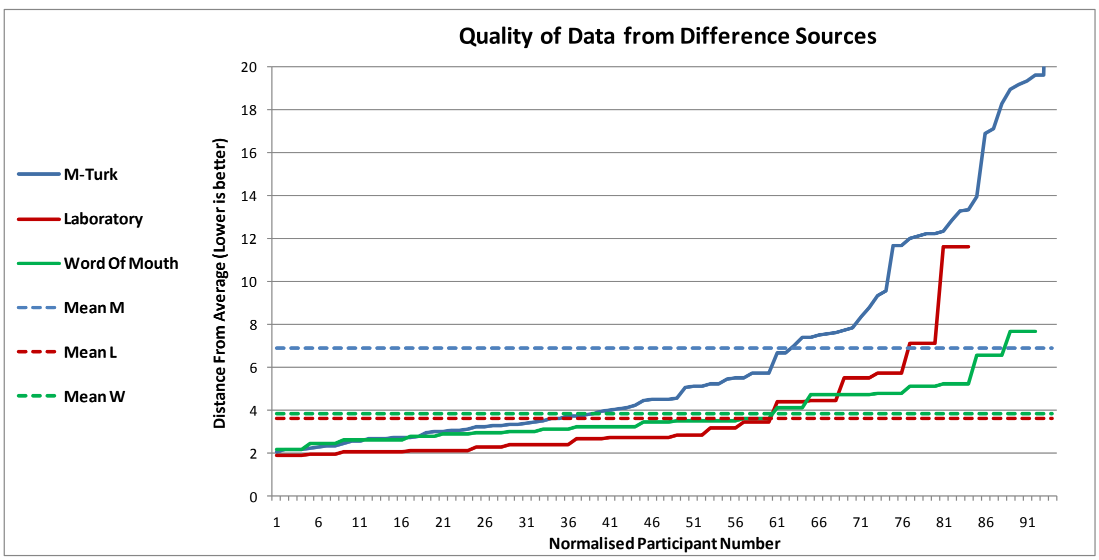

Figure 3 - Technique Results Comparison