Boa, D., Ranscombe, C., & Hicks, B. (2015). Determining the similarity of products using pairwise comparisons and eye tracking. In C. Weber, S. Husung, G. Cascini, M. Cantamessa, D. Marjanovic, & F. Rotini (Eds.), DS 80-5 Proceedings of the 20th International Conference on Engineering Design (ICED 15) Vol 5: Design Methods and Tools - Part 1, Milan, Italy, 27-30.07.15. (Vol. 5, pp. 225-234). (ICED). The Design Society.

1 INTRODUCTION

The design process is complex and products must maintain a balance between a wide range of criteria such as cost, manufacturability, performance, and appearance. It is acknowledged that within this process judgements of appearance are subjective and can be hard to justify across all stakeholders/departments in the design process (engineering, marketing and design).
Product appearance plays a critical role in the success of products, as it is the first form of contact and the basis of initial product appeal. Styling is a term used to refer to the form giving process undertaken by industrial/product designers to establish the appearance of products.
A major motivation in styling is to generate brand identity (similarity to existing product family) or to differentiate from competitor's products. High-level guidelines exist that identify common features across product families and aid the styling of products to differentiate them against the competition (Person et al., 2007). While this information is valuable to design teams, it remains subjective and context/product specific guides for designers do not exist
The aim of this exploratory research is to develop a method for creating objective style guides based on how individuals engage with products visually. Eye tracking can be used to supplement existing understanding of how products and features are attended, inspected, and engaged with and will be used to describe aspects of product perception. Designers and other stakeholders may then use understanding from eye tracking to justify styling decisions or the relative importance of product features.
The paper begins by with an overview of visual behaviour and preference judgments in design as well as a discussion on approaches for identifying key features for a product's aesthetics. A short exploratory experiment conducted at Design 2014 is presented that demonstrates how visual behaviour and preference judgments can be used to identify key features.

2 BACKGROUND

In this section a review of styling and aesthetic feature management in the design process is discussed as well as the use of eye tracking for investigating product perception phenomenon.

2.1 Eye tracking and design

Eye trackers record a person's gaze relative to a scene allowing inference as to where an individual is directing their attention. Characterising a person's gaze can be used to detect tasks, various states of cognition and an individual's perceptive process (Duchowski, 2007). Two key gaze measures are fixations, where an individual's gaze stabilises over a region of an image for a period of time, and saccades, the rapid eye movement between fixations.
Kukkonen (2005) researches the use of eye tracking in design as part of a larger research project, Perception of Design. Eye tracking was employed to research the relationship between gaze, product attitude, and consumer preference of products. A pair of mobile phones was presented simultaneously (a pairwise comparison) to participants who were then asked to rate appearance, apparent usability and favourite of all the phones shown. A correlation between preference for products and fixation duration during comparisons was established but was inconclusive in relating gaze metrics and specific features and their role in establishing preference for appearance.
Similarly Carbon et al. (2006) explore the use of eye tracking to understand how consumers rated the attractiveness of innovative (more novel) appearance of car interiors. Data from eye tracking successfully demonstrates different characteristics in viewing behaviour when assessing more innovative designs. In particular eye tracking showed that participants tended to distribute their gaze in a more balanced way over key features of a design when it is more innovative.
Reid et al. (2012) use eye tracking to explore the visual behaviour associated with customer preference for different product representations. Participants compared different styles of representation (silhouette, outline, greyscale render) in a pairwise manner to investigate visual behaviour changes for these representations. This research offers some insights into the ideal mode of representation for different product types, however there are no significant outcomes with respect to product features limiting usefulness in justifying styling decisions.

Köhler et al. (2014) apply eye tracking as a component of Kansei engineering to better understand latent perceptual behaviour (eye movements) in comparison with emotional evaluations of products. Eye tracking results demonstrated that in the product stimuli used (wristwatches) there are critical features that are focused on for the majority of observation. Thus eye-tracking gives a clear basis to define components most relevant in consumer perception. Eye tracking was particularly useful in characterising and quantifying feature engagement that the individual participants were unable to directly recall or were unconscious of having viewed.

Perceptive processes are commonly conducted at a speed and level of cognition which makes them challenging for individuals to accurately and completely describe them. Eye tracking is a valuable tool for investigating these processes and is well suited for doing so. It has however not been used to its full potential for identifying key features within products when determining aesthetic qualities.

2.2 Aesthetic/Styling feature management

Ranscombe et al. (2012) present a survey of the relationship between vehicle features and brand recognition. In this study aesthetic features of vehicles were isolated in varying combinations to ascertain which were most significant in communicating product brand. The study showed that certain features are far more critical than others, specifically vehicle grille and headlights.

In a similar manner Ranawat and Hölttä-Otto (2009) present a means to assess the prevalence of features throughout product families and market segments. This combines and corroborates approaches by Warell (2004) (Design Format Analysis) and Buddas (2006) (4D framework). The resulting framework objectively identifies the presence of different feature types across a brand or product families/ranges. The prevalence of features in ranges or across all products on the market gives an indication of the importance of a feature in creating familiarity or distinctiveness.

Shape grammars are the application of geometric rules or transformations to achieve designed outcomes that satisfy predefined structural, architectural or styling constraints (Pugliese and Cagan, 2002). Shape grammars have been used in a strategic styling context to generate designs for motorcycles (Pugliese and Cagan, 2002), cars (McCormack et al., 2004) and branded packaging (Chau et al., 2006) that maintain brand identity. Distinct from the other approaches discussed in this section, shape grammars provide a means to generate designs that maintain features rather than understand or define which features are significant to brand/style.

3 AIMS AND OBJECTIVES

From review of existing literature two findings can be elicited: first, previous studies verify relationships between eye movements and subjective perceptions of products. Second, literature relating to aesthetic features demonstrates that certain features have more potency in determining styling objectives e.g. brand identification. There is however, no prior work that can be considered to bridge these two aspects.

The aim of this paper is to explore whether it is possible to establish a set of features that are more influential/critical when making relative judgements and how this can be measured. Using the tasks of rating products on their similarity or difference an experiment is conducted to investigate the aim. For the purpose of this experiment products are rated on similarity or difference rather than overall preference. This is for two reasons; first, similarity and difference can be considered to be on the same spectrum and associated to brand identity and brand differentiation; second, it is possible to generate objective measures of similarity based on the geometry, features or collection of features if required for further analysis.

The corresponding hypothesis is proposed:

> *H0: The determination of similarity between product pairs is not dependent on engagement with specific product features*
>
> *H1: The determination of similarity between product pairs is dependent on engagement with specific product features*

4 METHOD

It is not clear whether the related tasks of asking individuals to rate products on their similarity or difference can change how an individual perceives a product. An experiment has been devised to incorporate the possibility that it might into the primary investigation of identifying key features for determining product similarity. This section details the experimental procedure adopted.

4.1 Participants

Seventeen academics from a range of countries were approached during the DESIGN 2014 conference to participate in the study. Of these 13 were male and 4 were female with an average age of 37.6 (S.D. 11 years). All participants spoke fluent English and freely volunteered to take part in the experiment.

4.2 Stimuli

Six products from four product categories (cars, gin bottles, smartphones and wardrobes) were selected as stimuli. These products were selected as they were deemed to cover a range of product sectors and values/price. Furthermore all products were deemed to be mature and mass-market, and hence products where appearance plays a significant role in consumer preference (Warell, 2004). The number of product categories was limited to four for the purpose of survey duration.

The six products from each category were grouped into three sets of pairs for the participants to compare. A single, consistent elevation or viewpoint was used for each of the products (front view for gin bottles and smartphones, side view for cars, and three-quarter perspective for wardrobes). The elevations were selected for their ability to convey the most information about each product and being representative of purchasing situations.

Each of the products was represented in outline form. This is a technique used in Ranscombe et al. (2012) to negate the influence of colour and surface finish/material while still defining features. While this type of representation is an abstraction from the real product form, studies in perception have shown that it is adequate to "communicate/identify products (Biederman, 1987).

Each outlined image was resized to 1920 x 1080 pixels. These were shown side by side to create the comparison image and displayed at full resolution on a laptop screen 350mm by 195mm. For the experiment the products were scaled such that the major dimension of each product was the same. This was done so that each product would occupy approximately the same area in the comparison image. The full set of product comparison images displayed to participants is provided in Figure 1.

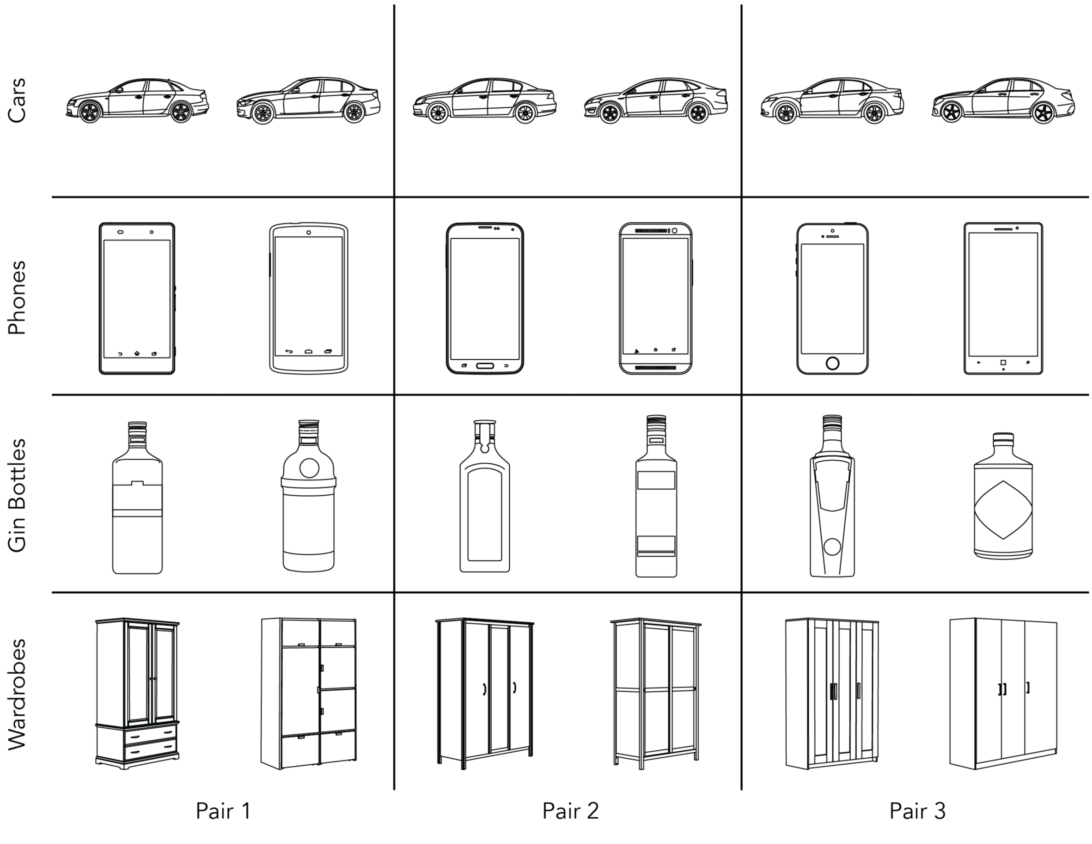

Figure 1 Pairs of product images shown to participants during the experiment. Line weight has been increased for clarity.

4.3 Equipment

A Tobii X2-30 eye tracker was used in conjunction with a Dell Precision M4800 Mobile Workstation running Tobii Studio 3.2.

4.4 Procedure

Testing took place in a quiet, but open access area of the conference. Participants faced a blank wall and were offered the use of noise cancelling headphones if they wished. A large natural light source introduced a degree of variability to the ambient lighting but the effect was deemed minimal on the tracker performance.

Each participant was calibrated for the laptop screen and eye tracker and asked to complete a basic participation questionnaire. Participants were instructed that they would be shown a pair of products for 15 seconds and then be asked to rate the two products on their degree of similarity. Twelve pairs were shown with the participants inputting their similarity rating on-screen immediately after viewing each pair. A distractor exercise lasting approximately 15 minutes immediately followed the first half of the experiment before participants completed the second half. In the second half participants were shown the same product pairs for 15 seconds but were asked to rate them on their degree of difference. By conducting the experiment a second time and asking participants to rate the product pairs on their degree of difference, it can be tested to see if the task has an affect on visual behaviour compared to rating on degree of similarity. All experimental instructions were delivered on screen in English.

4.5 Areas of interest (AOIs)

Areas of interest (AOIs) are generated within the eye tracking software to facilitate analysis of the location of the participant's gaze. Creating the areas of interest required the identification of elements of appearance that were consistent across each product category (Wardrobes, Cars, Smartphones and Gin Bottles). To do so a number of factors had to be considered.

Firstly the presence of features in all examples of a product from a product category were reviewed. In many cases, not all examples of products (within a category) had all of the same features. A second consideration was the proximity of different features. Here the accuracy of the eye tracker and its ability to establish the location of gaze is a barrier in being able to discern which of two closely located features are being looked at. This is also highlighted as a potential issue in Kukkonen (2005).

As a result of these considerations, areas of interest where defined by the overall, shape or architecture of the products. While the presence and proximity of features was not consistent within product categories, the overall location of features in the product was. For example in vehicles the headlights and grille details are always at the front, tail lights at the very rear and character lines, side windows and door handles are always toward the centre of the side-view. The areas of interest for each product category are illustrated below using a product from each category as an example Figure 2.

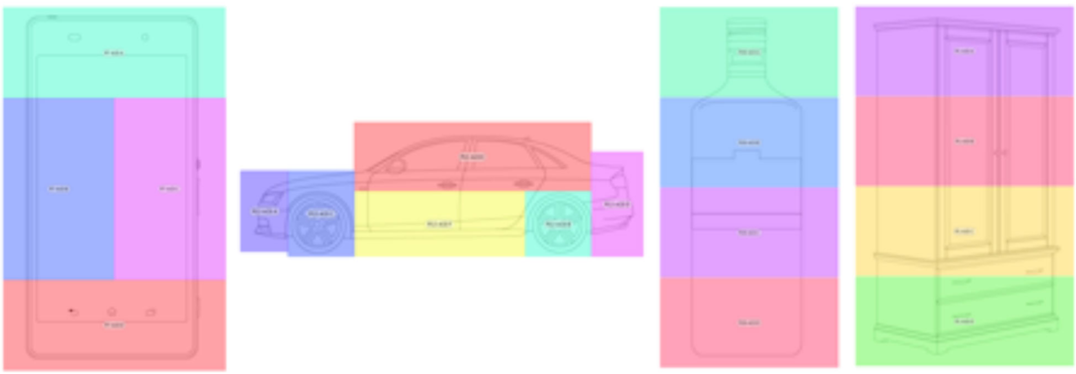

Figure 2. Example areas of interest for each product category. Each block of colour represents a distinct region (AOI) that is used by the eye tracker recording software for analysis.

5 RESULTS AND ANALYSIS

Modern eye trackers are robust and reliable in their data capture and are capable of dealing with head movement, corrected vision and variable light conditions. However, there is a limit and inevitably the tracker is unable to record 100% of an individual's gaze. The X2-30 Eye Tracker sample rate of 30 Hz limits the resilience of dealing with participant recording issues. Setting a recording sample rate threshold of 80%, 7 of the 17 tested individual's gaze had to be excluded from the analysis due to excessive missing data from testing.

Using the filtered results two research questions were explored. The first was used to investigate whether priming individuals to rate products on their similarity or their difference affected the way they perceived the products. The second question (corresponding to H0 and H1) aimed to determine whether the participant's similarity rating of a product corresponded to a specific engagement distribution. That is, do people rating the similarity of products in the same manner engage with the same features equally? All analysis was conducted using Excel and SPSS V.21.

5.1 Similarity rating task vs Difference rating task

Styling guidelines are often motivated to maintain brand identity or differentiate products from competitors. The terms similarity and difference are used somewhat interchangeably to achieve these goals. A foundational step in this research is to check whether priming individuals to determine similarity or difference results in a change in their perception of the product.

A commonly made distinction of different perceptive states is the ambient versus focal classification. Shorter fixations are often used to indicate an ambient state of perception, whilst longer fixations are used to indicate a focal state of perception (Unema et al. 2005). Thus, a meaningful change in perception and visual behaviour can be inferred by measuring the distribution of fixation durations for each rating task.

Simplistic analyses set a fixation duration threshold of around 200ms to distinguish between ambient (less than 200ms) and focal (more than 200ms) fixations. However, as there is a large degree of variation between individual's fixation durations this can be problematic, as the threshold for one person should not necessarily be equal to that of another person's.

Unema et al. (2005) measure the difference between ambient and focal states by grouping fixations into 500ms bins over the course of a scene. Fixations commencing within a single bin are averaged together to provide a measure of how fixation duration changes over the course of the stimulus display. By binning the fixations, segments of the total scene viewing time can be directly compared between individuals.

To test whether there is a measurable difference between average fixation duration over time for the two experimental conditions, Unema et al. perform a non-linear regression and compare the time taken to reach an asymptotic level in the average fixation duration. Achieving the asymptotic level in a shorter period of time is taken to be indicative of a greater proportion of the scene viewing time as being given over to a focal state. The gaze data has been processed to show how fixation duration changes over time and an example from one product pair for the similarity rating task is plotted in Figure 3.

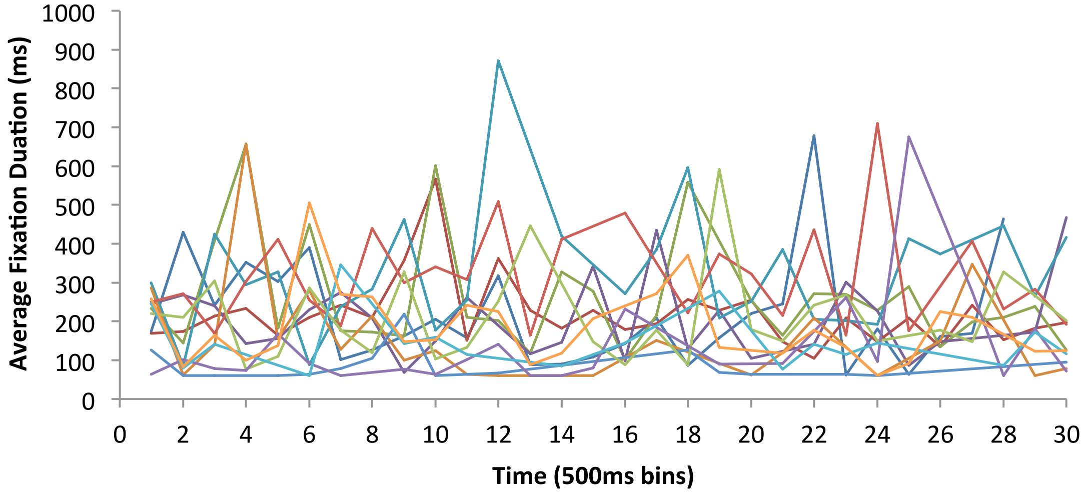

Figure 3. An indicative plot of time (in 500ms bins) versus average fixation duration for similarity rating task of one product pair. Each line represents the change in fixation duration over time for each participant.

The degree of variation within the population for the similarity versus difference prime test is considerably greater than Unema et al. (2005) (Figure 3). Unema et al. approximate their distribution by fitting a non-linear curve to their data, achieving a goodness of fit that explained around 78% of the variance. Approximation of the average fixation duration (AFD) over time for the similarity and difference conditions as a non-linear curve would result in a substantial margin of error (Figure 3).

A two-sample Kolmogorov-Smirnov (KS) test is performed to determine if the distribution of fixation durations of the two rating task conditions are from the same population. If they were from the same population it would suggest that there is no difference in the perceptive states for the two rating task conditions. Initially all 30 bins are tested to determine if the range of fixations durations are significantly different.

Out of the 12 product pairs two showed significance between the rating task conditions, Wardrobe pair 3 and Gin pair 2. Repeating the analysis for these two product pairs for 5 second sections (ten 500ms bins each) allows improved understanding about where in the total scene viewing time the difference in distribution occurs. In Table 1 it can be seen that significance occurs for both product pairs in the first two 5 second segments of the scene, but not the last 5 second segment.

Table 1. Two-Sample Kolmogorov-Smirnov test for 5-second scene segments to test for significance between experimental prime conditions of rating on similarity or difference.

Product pair	Scene segment	Null hypothesis	P value	Result
Wardrobe 3	0 – 5 seconds	The distribution of AFD is the same across rating tasks	0.034	Reject the null hypothesis
	5 – 10 seconds	The distribution of AFD is the same across rating tasks	0.034	Reject the null hypothesis
	10 – 15 seconds	The distribution of AFD is the same across rating tasks	0.121	Retain the null hypothesis
Gin 2	0 – 5 seconds	The distribution of AFD is the same across rating tasks	0.015	Reject the null hypothesis
	5 – 10 seconds	The distribution of AFD is the same across rating tasks	0.015	Reject the null hypothesis
	10 – 15 seconds	The distribution of AFD is the same across rating tasks	0.125	Retain the null hypothesis

It can be concluded that it is unlikely that it matters whether individuals are asked to rate a product pair on its degree of similarity instead of on their degree of difference.

5.2 Product feature engagement for pairwise comparisons

To investigate participant's engagement with product features the distribution of gaze between areas of interest was analysed. Only fixation data gathered from the similarity rating condition part of the experiment was used.

In the analysis the distribution of engagement is measured within areas of interest (AOIs) that enclose key product features. Dwell time, the total time spent fixating within an AOI, is normalised for each AOI as a percentage of total scene viewing time to mitigate differences in data capture rates between individuals. Dwell time is used as a measure for the relative importance that an individual places on a product feature for determining the degree of similarity. Initial review of the dwell distribution (Figure 4) shows a number of AOIs being inspected more than others in each product category.

For the wardrobes this was the upper middle section including the door handles. For Phones it was the top section including the camera and speaker and for gin bottles it was the shoulder of the bottle/top of the label. These product features are clearly significant to individuals when performing pairwise comparisons of products and as such are likely to be important in selection activities, such as purchasing.

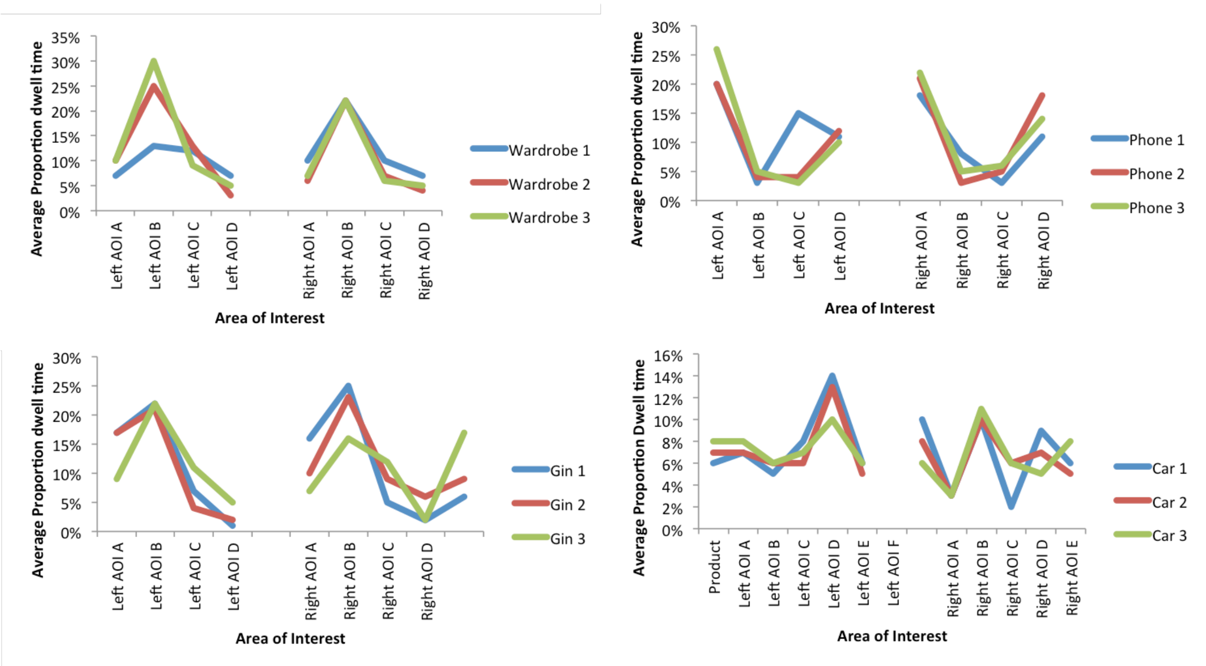

Figure 4 Average distribution of dwell time over AOIs for each product categories

To further interrogate the data a one-way ANOVA of the dwell time distribution over the product AOIs and individual's ratings of the degree of similarity between the two products was carried out. The effect of similarity rating on product AOI engagement, for significant AOIs only, is listed in Table 2.

Table 2. ANOVA results for product similarity ratings that correlate to features that were fixated on a significantly different amount of time

Product	AOI	P Value	Product features within AOI
Wardrobes 2	Right product AOI C	0.027	Bottom middle quarter of right wardrobe
Gin Bottles 3	Left product AOI D	0.009	Bottle base of left gin bottle
Cars 1	Right product AOI B	0.036	Rear wheel of right car
Cars 3	Left product AOI B	0.033	Rear wheel of left car

From product pairs Wardrobe 2, Gin Bottles 3, Car 1, Car 3, only a single AOI showed significance between how a group of participants rated the similarity of those products. This means that across the participants, people that rated the similarity of the product pairs the same dwelled on a particular AOI a significantly different amount of time to groups that gave the product pairs a different rating for

similarity. This suggests that for those participants, certain features were critical in the determination of how similar the two products presented were. For the remaining product pairs a lack of significance suggests that no single product feature significantly affected how people rated the similarity of the product pairs.
Reviewing the features within the significant AOIs, the only commonality amongst the features is that in the car AOIs both focus on the wheel. In the Gin and Wardrobe AOIs it is both lower features being reviewed. Due to the low occurrence of AOIs exhibiting significance it was deemed unnecessary to perform statistical corrections for multiple testing.

6 DISCUSSION

Limited effect on the perceptual state has been found from asking participants to rate the product pairs on either their degree of similarity or degree of difference. The method for identification of product features important in the determination of product similarity shows potential and a discussion of how the procedure could be refined is stated in this section.

6.1 Similarity rating task versus Difference rating task, the effect on visual behaviour

Two of the twelve stimuli products showed a significant difference between the prime conditions. The finding that in these two cases, a significant difference in average fixation duration (AFD) was seen in the first two thirds of the scene suggests that this initial time frame should be the focus in future studies. It is expected that a difference in perceptual effect might be evident in more product pairs by correcting some experimental issues regarding how the rating tasks were implemented. Several participants reported during post-testing interviews that they weren't aware of a change in task between the two halves of the experiment. While it is imperative that participants are not over-primed, it is suggested that the change in task should be made more explicit in future.

6.2 Product feature engagement for pairwise comparisons

Eye tracking enables the engagement of specific features to be characterised for products. From review of Figure 4 it can be seen that there is a high degree of symmetry in the plots indicating that product features are directly and equally compared. Where an equal engagement for a corresponding feature does not occur, such as in Phone pair 1, Left product AOI C, there may be a number of reasons for asymmetric engagement. Across the product categories, not all products have the same features located in the corresponding AOI. For the phone pair 1, the two phones have their volume buttons on opposite sides. Categorising AOIs by the features contained within them would address this issue.
In pairs where features are consistently present in the same location (such as in all of the cars), differences in engagement for identical features may be attributable to the manner in how individuals determined how similar the two products are.
Identifying product features that are used to determine how similar two products are, a useful output for testing brand identification or brand differentiation, could be aided by modifying the experimental procedure. Conducting a Retrospective Think Aloud (RTA) or asking individuals more detailed questions relating to product features could provide rationale for why participants were engaging with features. Participants could also be asked to rate the degree of difference for the individual product features.

6.3 Future work

The results of the experiment show a large degree of variation between participants and a small effect (Table 2, Figure 4). In such situations a larger number of participants is required, or the effect that is measured needs to be amplified.
Firstly, it is contended that the positional accuracy of the eye tracker and small display size adds an element of variation. It is harder to distinguish whether a recorded fixation corresponds to a specific product feature when the positional accuracy of the tracker and the physical distance between the features are of the same magnitude. A larger display size would increase the physical distance between features and make attributing fixations directly to features simpler. Whilst enlarging the scale at which they are displayed would be an abstraction for some products, for the likes of cars and wardrobes it would increase their realism.

Secondly, adapting the experimental procedure to have participants compare three products simultaneously is proposed as a means of amplifying any observable effect. Out of the three products displayed, one product would be a 'datum' product and the participants would be then asked to select out of the two alternatives, which is the most similar or different to the datum. This would partly address issues with participants commenting on how they should determine similarity.

7 CONCLUSIONS

The nature of this study is exploratory in proposing and developing how eye tracking could be used to add objective input into styling decisions. Preliminary findings indicate that the observation of various features (Gin bottle shoulder and top label, Phone speaker and camera, Wardrobe handles, Car wheels) gives an indication of the contribution that can be generated. It is however, too early to relate findings to more generalised design and styling guidelines. With more robust data, eye-tracking findings could give clearer insight to designers about the perceptual impact of product features on styling without the burden of significant interference from protocol analyses.

There are some constraints in using similar approaches in industry. The requirement for more precise eye tracking equipment and need for fairly intensive coding and interpretation of data makes such a review process relatively resource intensive. As such it is likely that this type of research would be used as a benchmarking activity undertaken intermittently to understand key brand features.

An underlying pattern of feature engagement that might be influencing visual behaviour in comparing similarity and difference was not evident from the data. However, it is possible to draw conclusions on improvements for future studies into features and visual behaviour. Key features of a method adopted in future studies are the use of a datum product in three-way comparisons and retrospective-think-aloud (RTA) protocols to provide context to fixations.

REFERENCES

BIEDERMAN, I. 1987. Recognition-by-Components: A Theory of Human Image Understanding. Psychological review, 1000, 22.

BUDDAS, S. 2006. Quantifying Visual Product Attributes in Product Platform Design. Master's Thesis. Department of Mechanical Engineering. Helsinki University of Technology.

CARBON, C.-C., HUTZLER, F. & MINGE, M. 2006. Innovativeness in design investigated by eye movements and pupillometry. Psychology Science, 48, 173.

CHAU, M. C. A. H. H., MCKAY, A. & DE PENNINGTON, A. 2006. Combining evolutionary algorithms and shape grammars to generate branded product design. Design Computing and Cognition'06. Springer.

DUCHOWSKI, A. (2007). Eye Tracking Methodology Theory and Practice (2nd ed., p. 360). London: Springer London.

KÖHLER, M., FALK, B. & SCHMITT, R. 2014. Applying Eye-Tracking in Kansei Engineering Methodology for Design Evaluations in Product Development. INTERNATIONAL CONFERENCE ON KANSEI ENGINEERING AND EMOTION RESEARCH. LINKÖPING.

KUKKONEN, S. 2005. Exploring eye tracking in design evaluation. Joining Forces, 119-126.

MCCORMACK, J. P., CAGAN, J. & VOGEL, C. M. 2004. Speaking the Buick language: capturing, understanding, and exploring brand identity with shape grammars. Design Studies, 25, 1-29.

PERSON, O., SNELDERS, D., KARJALAINEN, T. M. & SCHOORMANS, J. 2007. Complementing intuition: insights on styling as a strategic tool. Journal of Marketing Management, 23, 901-916.

PUGLIESE, M. & CAGAN, J. 2002. Capturing a rebel: modeling the Harley-Davidson brand through a motorcycle shape grammar. Research in Engineering Design, 13, 139-156.

RANAWAT, A. & HÖLTTÄ-OTTO, K. Four dimensions of design similarity. ASME IDETC Design Theory and Methodology Conference, 2009.

RANSCOMBE, C., HICKS, B., MULLINEUX, G. & SINGH, B. 2012. Visually decomposing vehicle images: Exploring the influence of different aesthetic features on consumer perception of brand. Design Studies, 33, 319-341.

REID, T. N., MACDONALD, E. F. & DU, P. Impact of Product Design Representation on Customer Judgment with Associated Eye Gaze Patterns. 2012. American Society of Mechanical Engineers, 749-762.

UNEMA, P. J. A., PANNASCH, S., JOOS, M., & VELICHKOVSKY, B. M. (2005). Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. Visual Cognition, 12(3), 473–494.

WARELL, A. 2004. Towards a Theory-Based Method for Evaluation of Visual Form Syntactics. In: XIROUCHAKIS, H. (ed.) TMCE, Tools and Methods for Competitive Engineering. Lausanne: Millpress.