# 1 INTRODUCTION

Addressing problems through online human-computer interactions, referred to broadly as crowdsourcing, has led to emerging platforms for massive design creation such as the 3D Warehouse and the Build with Chrome project. From a design research perspective, crowd-sourced design creations and evaluations provide an opportunity to learn and model quantitatively how humans deal with design problems. This paper explores crowdsourcing to evaluate perceptual design attributes and to create new design concepts using such attributes.

Quantitative analysis of product preferences has been an active research area since the introduction of conjoint analysis in the 1970s (Netzer et al., 2008). Modeling preferences for product design attributes that can be quantified easily, such as price and fuel consumption of automobiles, is well developed. Perceptual attributes, i.e., attributes that depend on subjective individual human perception such as aesthetics, luxury or safety, are important elements in design decision making, but they are not commonly incorporated in preference modeling because they are difficult to quantify.

Research on modeling perceived design attributes can be traced back to Kansei Engineering (Nagamachi, 1998) and Interactive Genetic Algorithms (IGAs) (Takagi, 2001). In Kansei Engineering, the perception model is built through regression over a collection of human ratings across a set of designs. For example, one can use Kansei Engineering to explain how safe a car is based on its design features such as silhouette curvature and exterior color. In IGAs, designs evolve based on user-assigned fitness and converge toward some user-defined target. This method has been used successfully to create computer graphics (Sims, 1991), track fashions (Kim and Cho, 2000), and design car shapes (Kelly and Papalambros, 2007). However, the evolutionary nature of IGAs makes them more suitable for exploring new designs rather than modeling perception (Ren and Papalambros, 2011). Both Kansei Engineering and IGAs require user-input ratings, e.g., a 1 to 5 score on how safe a car looks, which is not natural for human evaluations.

Building preference models requires a preference elicitation process, such as surveys or interactive queries. This paper investigates whether crowdsourcing can be a solution to modeling perceptual design attributes. As a running example, we study how the perception of safety can be modeled as a function of the exterior design of a car and how "safe-looking" cars can be created using the perception model.

Specifically, we study human-computer interactions comprised of pairwise comparisons. Based on the comparison data, a learning algorithm called *ranking SVM* is implemented to model the perceived safety of car exterior designs. The ranking SVM algorithm, originally developed for refining search engines (Joachims, 2002), performs a nonlinear regression that maps three-dimensional (3D) geometries to the perceived measure of safety. In the running example, we describe the modeling and design creation process with anonymous human responses collected through Amazon's Mechanical Turk (MTurk). We highlight the issues encountered and provide guidance for further algorithm development and interaction design. As a side contribution, we also demonstrate the use of WebGL, a recently standardized browser-side rendering language, for real-time generation of parametric 3D models.

The paper is structured as follows: In Section 2 we review recent developments in preference learning and crowdsourcing, and present the rationale for the algorithm employed in this work. Section 3 elaborates on the technical details of perceptual attribute quantification using pairwise comparisons. We show, in simulation, how noisy human responses and divergent human opinions can hamper the performance of

the perceived safety model. In Section 4 we discuss the crowdsourcing studies using MTurk and the issues faced in applying the proposed learning algorithm with actual users. We show that, while the model for perceived safety can generate safe-looking car designs, it can perform poorly when discriminating two random designs in terms of safety. Section 5 concludes with a summary of findings and suggests future research directions.

# 2 RELATED WORK

## 2.1 Preference Elicitation

The task of quantifying a design concept presented in this paper can be considered as a special application of the more general research topic of design preference elicitation. Besides Kansei Engineering, other related work includes conjoint analysis from marketing science (Netzer et al., 2008) and preference learning from computer science (Joachims, 2002 and Herbrich et al., 1999). In general, these are all statistical methods to quantify subjective evaluations using observed "features", e.g., design variables and demographic data. Due to the large amount of work in these areas, here we summarize only some key ideas that lead to the choice of the interaction form and learning algorithm in the presented work. Many recent practices adopt pairwise comparisons instead of ratings or rankings; the rationale here is that, while comparison provides less information than the other two forms, it is the most natural form of human evaluation on products and can be observed freely from market data (Toubia et al., 2007) and online interactions (Joachims, 2002). Further, the introduction of a kernel machine like ranking SVM has shown great value to elicitation practices (Cui and Curry, 2005 and Evgeniou et al., 2007). This method enables use of nonlinear models without requiring a specified functional form of the model. Practice has shown that such nonlinear models have performance close to hierarchical Bayes models that are more computationally intensive (Evgeniou et al., 2007). Finally, with the growth of online interactions, more emphasis has been placed on how preference models can be built efficiently by adaptively presenting questions to users (Abernethy et al., 2008). Scalability issues have also been studied so that learning models can be developed efficiently when massive observations are present (Joachims, 1999 and Fan et al., 2005).

## 2.2 Crowdsourcing

With the advent of ubiquitous computing and social media, more humans are networked together than ever before. This allows previously unattainable access to large pools of potentially well-qualified designers, particularly for design tasks primarily based on human perception (Yuen et al., 2011 and Engel et al., 2012). As a result, much work with crowdsourcing comes from the machine learning literature, where a general strategy is to model humans as noisy labelers in human perception tasks such as annotating images and translating words (Raykar et al., 2010). A goal of this strategy is to estimate the true labels of a task when the labels are either latent or only a subset of them is known a priori. We adopt a common, though not necessarily justified, assumption that a true concept of safety exists that it is global across humans within the crowd, i.e., that there is homogeneity of safety perception (Viappiani et al., 2011). The corresponding set of pairwise comparisons elicited from the crowd then trains a relevance measure between perceived safety and the various designs (Alonso et al., 2008 and Tamuz et al., 2011). Arrow's impossibility theorem dictates caution in attempting to extend any results of such elicitations to broad social agreement, see e.g., Hazelrigg (1996) and Saari (2010).

# 3 DESIGN CONCEPT QUANTIFICATION

## 3.1 Interactive design platform

The idea for quantification of a perceptual attribute is to develop a model that achieves gradually refined accuracy (to "learn a model" in computer science jargon) using massive user responses from pairwise comparison tasks. For this purpose, we developed a web-based interaction, as illustrated in Figure 1.

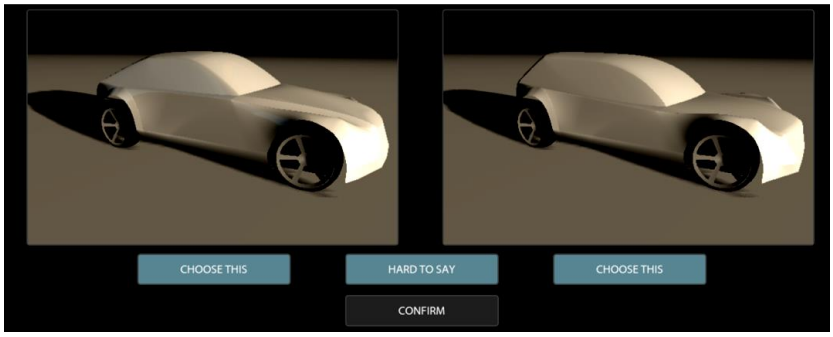



*Figure 1. The user interface for concept quantification (Ren, 2013)*

The 3D car shape models are rendered in real time using the WebGL language. Users are allowed to rotate the designs for comprehensive evaluation. Each car design is comprised of 18 pieces of Bezier surfaces defined by 52 control points. These control points are positioned by 19 continuous design variables in the range of 0 to 1. Manipulating these variables generates new car shapes.

## 3.2 Ranking SVM

Let us assume that $m$ pairs of comparisons are presented to the user and responses are collected. We denote the set of pairs as $\Delta X = \{(x_1^{(1)}, x_1^{(2)}), \dots, (x_m^{(1)}, x_m^{(2)})\}$ where $(x^{(1)}, x^{(2)})$ is a pair and $x^{(1)}$ is the chosen design. We consider perceived safety as a utility function of design variables taking the form $f(x) = < w, \phi(x) >$, where $<\cdot,\cdot>$ is an inner product, $\phi(x)$ are *design features* and $w$ is an unknown parameter vector that can be estimated through the following a quadratic programming problem based on $\Delta X$:

$$\begin{aligned} &\min_w && w^T w \\ &\text{subject to:} && < w, \phi\left(x_i^{(2)}\right) > - < w, \phi\left(x_i^{(1)}\right) > \leq -1 \ \ \forall i = 1, \dots, m. \end{aligned} \tag{1}$$

The constraints here require the utility of $x_i^{(2)}$ to be less than that of $x_i^{(1)}$, while minimizing the $l_2$-norm of $w$ helps to avoid over-fitting from the responses. Problem (1) is similar to a regularized logistic regression that replaces the constraints with a log-likelihood penalty (Evgeniou et al., 2007). From a Bayesian perspective, Problem (1) finds $w$ that best interprets the responses based on the prior that $w$ follows a standard normal distribution.

In this study, we define the feature vector $\phi(x)$ of a design as a collection of distances among control points, which govern the surfaces of the car shape. The heuristic here is that these distances are better descriptors of the 3D car shape than the design variables $x$, and can be more useful at quantifying a perceptual attribute of a 3D car shape. A total of 276 distances are used based on 24 manually selected control points. As a preprocessing step of solving Problem (1), we compute features from all designs in

$\Delta X$, and obtain the normalized features $\hat{\phi}(x) = (\phi(x) - \mu)/\sigma$ where $\mu$ and $\sigma$ are vectors of the mean and standard deviation of each feature (distance). This normalization step is essential as it scales features to the same level and prevents an inferior model biased towards features with greater magnitudes.

The dual problem of Problem (1) can be derived as:

$$\begin{aligned} \min_{\beta} \quad & \frac{1}{2}\beta^T Q\beta - e^T\beta \\ \text{subject to:} \quad & \beta \geq 0. \end{aligned} \tag{2}$$

Here $\beta$ are Lagrangian multipliers of Problem (1), $e$ is a column vector of all ones, and $Q$ is an $m \times m$ matrix with $Q_{ij} = < \hat{\phi}\left(x_i^{(1)}\right) - \hat{\phi}\left(x_i^{(2)}\right), \hat{\phi}\left(x_j^{(1)}\right) - \hat{\phi}\left(x_j^{(2)}\right) >$. Throughout this paper, we use a Gaussian distance to define the inner product: $< x, y > \coloneqq \exp(-\gamma||x - y||^2)$, where the Gaussian parameter is set to $\gamma = 1/\#features$ according to the default value in LIBSVM (Chang and Lin, 2011). The computation cost of Problem (2) depends on the data size $m$ that can grow dramatically when the task is crowdsourced. In order to alleviate the computation burden, a scalable optimization algorithm is developed similar to the working-set algorithm of Fan et al. (2005). The details of this algorithm are omitted here for brevity.

Solving Problem (2), the safety measure of design $x$ can be updated by:

$$f(x) = \quad -\sum_{i=1}^{m} \beta_i < \hat{\phi}\left(x_i^{(1)}\right) - \hat{\phi}\left(x_i^{(2)}\right), \frac{\phi(x)-\mu}{\sigma} >. \tag{3}$$

Eq. (3) can be used to predict the choice on safety for any pair of designs. The accuracy of this model is then measured by the consistency between predicted choices and actual user choices.

### 3.3 Effect of noisy responses and divergent opinions

The above algorithm is developed under the assumptions that (i) human choices are noise free and (ii) users share the same safety model, referred to as homogeneity as described in Section 2.2. Since these do not hold in reality, we investigate below how noisy responses and divergent opinions, i.e., multiple models, will affect the prediction accuracy. The findings here will help to analyze the real-user experiment results in Section 4. The simulations are set in the design space $D = \{x|x_i \in [0,1], \forall i = 1, \dots, 20\}$. In each simulated experiment, 50 simulated users are set to take the comparison task in a sequence. The task contains 15 pairwise comparisons for each user, who makes a choice based on an artificial utility function:

$$f(x) = \exp(-||x - t||^2) + \theta U(0,1). \tag{4}$$

Here the parameter $\theta$ scales a uniformly distributed random variable $U(0,1)$ to represent the noise in human response and the target $t$ can be set different to represent different safety models. The prediction accuracy is recorded for each user interaction using the user's choices and the predicted choices prior to the interaction. Figure 2 shows the mean accuracy along the 50 simulated interactions under various noise levels in Fig. 2(a) and divergent opinions in Fig. 2(b). Each data point is averaged over 50 independent runs. The simulation result shows that the existence of noise or multiple utility models of perceived safety will largely undermine the prediction accuracy of the concept model learned using ranking SVM.

# 4 HUMAN-USER EXPERIMENTS AND ANALYSIS

In this section, we conduct user experiments using MTurk to quantify perceived safety of car designs. We evaluate the resulting models by the prediction accuracy throughout the experiments and the safest design suggested by the model. Note that predictions on skipped pairs are not taken into the calculation of the prediction accuracy.

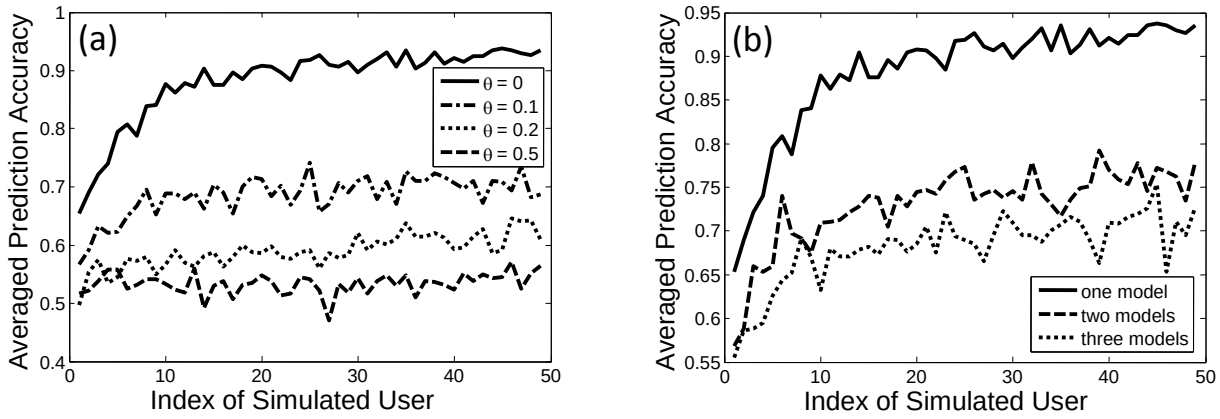



*Figure 2. Effect on model performance from (a) noisy responses and (b) heterogeneous utility models of perceived car safety*

## 4.1 Individual user experiment

An initial test was performed using the first author as an individual user experienced with car design. The user is presented with a sequence of pairs of car models and asked to make a pair-wise comparison of two car designs to assess which design has better perceived safety. It was observed that the user considers car models with larger frontal crush space and boxy shapes as safer ones. Figure 3**Error! Reference source not found.** (left) shows the recorded model accuracies through 12 consecutive tests on the user. The superior performance is partially due to the fact that the user skipped pairs whenever a choice is hard to make, reducing the chances of contradicting model predictions. Nonetheless, the result shows that the quantification process can efficiently capture user rules to measure a qualitative concept. To further support this argument, we optimized the resulting model to find the safest design, as shown in Figure 3 (right). Notice that the safety function derived from Eq. (3) is non-convex and the optimized design is obtained using a genetic algorithm.

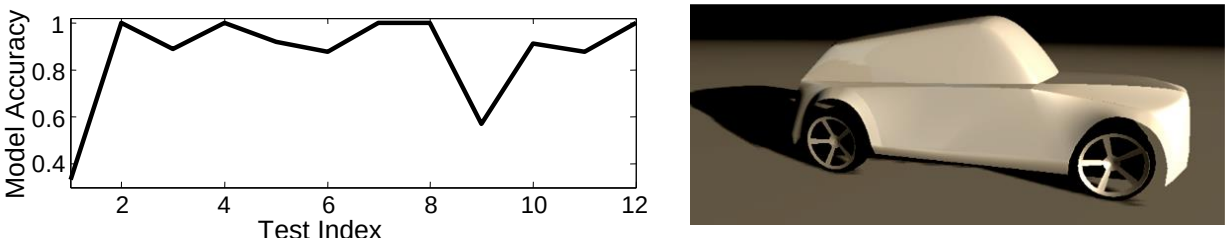



*Figure 3. Model accuracy (left) and "safest" design (right) from the individual experiment*

## 4.2 MTurk crowdsourcing experiment

The same comparison task on safety was then broadcasted on MTurk to MTurk "workers" with overall approval rates over 95% and located in the U.S. Two experiments were conducted in a sequence. We elaborate on the setup and results of these experiments, highlighting the observed issues.

### *4.2.1 Pilot experiment and lessons learned*

In the pilot experiment, each user was assigned 16 comparison tasks. To prevent random responses that could undermine model accuracy, two filtering rules were set up: For each user, we generated 14 random pairs and the first pair was then reproduced at the end of the task; we also inserted a predefined pair to check whether the user understood the task correctly or not. This predefined pair is shown in Figure 4. Prior to the experiment, it was believed that the "boxy" design on the right would be chosen as "safer" than the "sporty" design on the left, which later turned out to be an incorrect assumption.

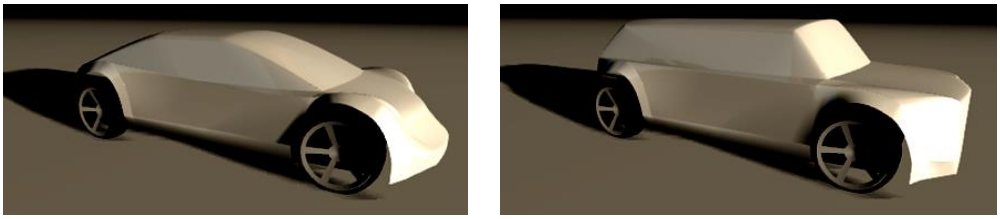


*Figure 4. A predefined "sporty" and "boxy" pair for sanity check: The "boxy" design on the right is believed to be "safer", and as such is used as a filter for human users.*

User responses were ignored either when the user answered differently in the repeated pair or chose the design on the left from the predefined pair. A total of 66 users completed the survey. Among all submissions, 10 users had inconsistent responses, and 30 users chose the design on the left. Before the users entered the comparison task, they were asked to provide written answers to a survey question. The question read "Which aspect of the car design affects your perception of "safety" the most: Frontal crush space, size, weight, handling or others?" From the 47 people who entered valid answers, there were 15 votes for "frontal crash space", 25 for "size or weight" and 15 for "handling." It is possible that users considered the predefined "sporty" design to be more maneuverable and therefore to have better handling or "active safety" performance than the alternative. However, the different numbers of people who chose this "sporty" design and those who voted for "handling" still leave doubt whether the MTurk responses are reliable in our experimental setup.

Due to the difference in user opinion for perceived safety and potential noisy responses, the resulting accuracy is unconvincing, as is shown in Figure 5 (left). Interestingly, the safest design suggested by the crowdsourced model in Figure 5 (right) captures features such as an overall boxy shape, sharp edges and a long frontal compartment.

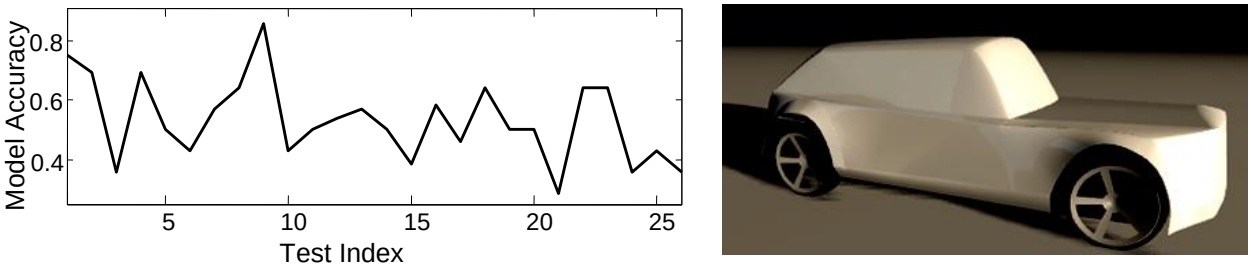



*Figure 5. Model accuracy (Left) and "safest" (Right) design from the pilot MTurk experiment*

#### *4.2.2 Follow-up experiment*

We attempted to address the observed issues above using a follow-up experiment. To deal with divergence in user opinion regarding the most important factor in their evaluation of perceived safety, we required users to choose from three options: (1) frontal crash space, (2) size/weight, and (3) handling. Three different learning models were then refined based on separately labeled user responses. The predefined pair was removed; instead, we placed another replicated pair from the randomized ones. This was used to strengthen the filtering of unreliable responses, and responses were considered valid only when choices were consistent on all of the repeated pairs.

The follow-up experiment received 117 valid responses out of 211 participants, with 39 for "frontal crash space", 34 for "size / weight" and 44 for "handling". Although a large amount of noisy responses were filtered out, the prediction accuracies were still less than satisfactory, as shown in Figure 6. We then generated the safest designs using the three perceived safety models, as shown in Figure 7. Consistent with the pilot test, these designs capture features that make them look safe from their own perspective.

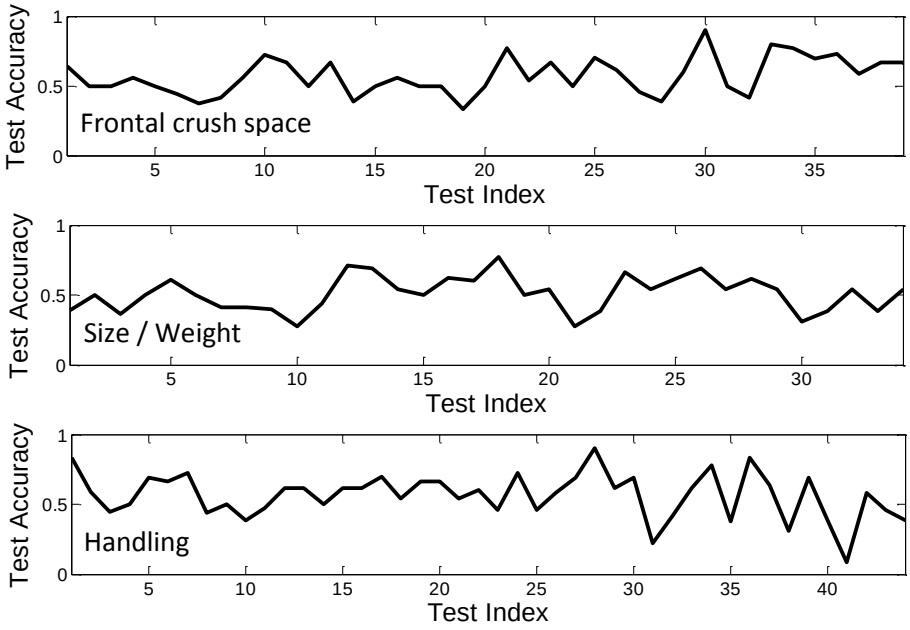
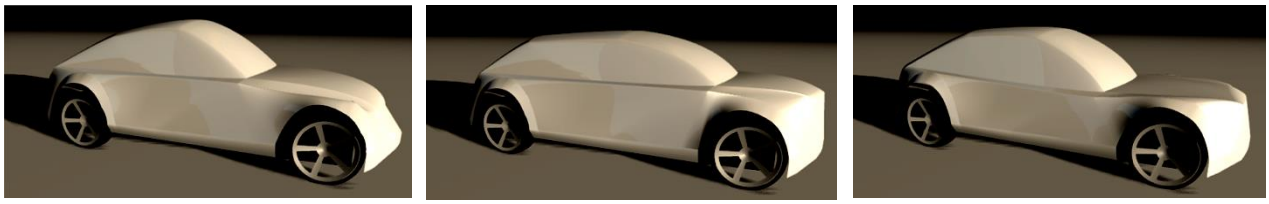



*Figure 6. Model accuracies from the follow-up experiment*



*(a) Frontal crush space* *(b) Size / Weight* *(c) Handling*

*Figure 7. "Safest" designs suggested by the three models*

One reason for the consistently low model accuracy is that users tended to make a choice instead of skipping the comparison when the two designs were hard to discriminate. In fact, the records show that MTurk users skipped only 11% and 8% of the comparisons in the pilot and the follow-up experiments, respectively, while this ratio was 44% in the individual experiment where the user was instructed to skip a comparison that was hard to make. Since most randomly generated pairs were hard to discriminate, and

user responses are likely to be noisy in such situations, making a hard choice instead of skipping the choice altogether leads to poor predictions in the presence of such pairs. On the other hand, the observation that the model can locate designs with high safety successfully indicates that the majority of users correctly identified the perceived safer design when they were able to discriminate between two designs.

## 5 FUTURE WORK AND CONCLUSION

From the experimental results obtained to date, we hypothesize that, when users face a pair of designs that are hard to discriminate, they tend to make a choice rather than skip the comparison even when such option is provided. This behavior leads to noisy responses and may partially account for the low model accuracy in the crowdsourcing experiments, namely, inadequacy at ordering perceived safety of two random designs. Nevertheless, the results showed that the learning models were able to capture important design features that compose a safe-looking design, and the suggested safest designs appear intuitively acceptable. Resolving this issue requires further investigation. Another area for future research is to relax the assumption of user homogeneity of perceived safety and instead "learn" clusters of human users in the crowd by their perception choices. The work to date is clearly in its early stages and more effort must be expended in designing the user interactions to draw more conclusive results.

From an algorithmic viewpoint, two future directions should be explored based on our findings to date. First, a kernel logistic regression method (Zhu and Hastie, 2005) compatible with pairwise comparison data should be developed in order to quantify the design concept robustly with noisy responses. Second, it would be interesting to extract design features and measure their relevance to a specific perceptual attribute. In the car example, in would be interesting to measure how much car body curvatures contribute to the perception of safety or sportiness. Computer science research in feature extraction may provide useful ideas for studying this problem.

## REFERENCES


Abernethy, J., Evgeniou, T., Toubia, O. and Vert, J.P. (2008) Eliciting consumer preferences using robust adaptive choice questionnaires, *IEEE Transactions on Knowledge and Data Engineering*, 20-2, pp:145-155.

Alonso, O., Rose, D.E. and Stewart, B. (2008) Crowdsourcing for relevance evaluation, *ACM SIGIR Forum*, 42-2, pp: 9.

*Amazon Mechanical Turk* [online], http://mturk.com (Jan. 03, 2013).

*Build with Chrome* [online], http://www.buildwithchrome.com (Jan. 03, 2013).

Chang, C.C. and Lin, C.J. (2011) LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, 2-3, pp:27.

Cui, D. and Curry, D. (2005) Prediction in marketing using the support vector machine, *Marketing Science*, pp:595-615.

Engel, D., Kottler, V. and Malisi, C. (2012) Crowd-sourcing design: Sketch minimization using crowds for feedback, *Workshops at the AAAI*.

Evgeniou, T. Pontil, M. and Toubia, O. (2007) A convex optimization approach to modeling consumer heterogeneity in conjoint estimation, *Marketing Science*, 26-6, pp:805-818.

Fan, R.E., Chen, P.H. and Lin, C.J. (2005) Working set selection using second order information for training support vector machines, *The Journal of Machine Learning Research*, vol. 6, pp:1889-1918.

Hazelrigg, G.A. (1996) The implications of Arrow's impossibility theorem on approaches to optimal engineering design, *Journal of Mechanical Design*, 118-2, pp:161-164.

Herbrich, R., Graepel, T. and Obermayer, K. (1999) Support vector learning for ordinal regression, *9th International Conference on Artificial Neural Networks*, pp:97-102.

Joachims, T. (1999) Making large scale SVM learning practical, *Universität Dortmund*.
Joachims, T. (2002) Optimizing search engines using clickthrough data, *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp:133-142.
Kelly, J. and Papalambros, P.Y. (2007) Use of shape preference information in product design, *International Conference on Engineering Design*, Paris, France.
Kim, H.S. and Cho, S.B. (2000) Application of interactive genetic algorithm to fashion design, *Journal of Engineering Applications of Artificial Intelligence*, 13-6, pp:635-644.
Nagamachi, M. (1995) Kansei engineering: A new ergonomic consumer-oriented technology for product development, *International Journal of Industrial Ergonomics*, 15-1, pp:3-11.
Netzer, O., Toubia, O., Bradlow, E.T., Dahan, E., Evgeniou, T., Feinberg, F.M., Feit, E.M., Hui, S.K., Johnson, J. and Liechty, J.C. (2008) Beyond conjoint analysis: Advances in preference measurement, *Marketing Letters*, 19-3, pp:337-354.
Raykar, V., Yu, S. and Zhao, L. (2010) Learning from crowds, *Journal of Machine Learning Research*, vol. 11, pp:1297–1322.
Ren, Y. and Papalambros, P.Y. (2011) Design preference elicitation: Exploration and learning, *International Conference on Engineering Design*, Copenhagen, Denmark.
Ren, Y. (2013) *Experiment for ICED2013* [online], foriced2013.appspot.com/ICEDmturk.html (Jan. 11, 2013).
Saari, D.G. (2010) Aggregation and multilevel design for systems: Finding guidelines, *Journal of Mechanical Design*, 132-8, pp:081006.
Sims, K. (1991) Artificial evolution for computer graphics, *Journal of Computer Graphics*, 25-4, pp:319-328.
Takagi, H. (2001) Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation, *Proceedings of the IEEE*, 89-9, pp:1275-1296.
Tamuz, O., Liu, C. and Belongie, S. (2011) Adaptively learning the crowd kernel, *Proceedings of ICML*.
*Trimble 3D Warehouse* [online], http://sketchup.google.com/3dwarehouse (Jan. 03, 2013).
Toubia, O., Evgeniou, T. and Hauser, J. (2007) Optimization-based and machine-learning methods for conjoint analysis: Estimation and question design, *Conjoint Measurement: Methods and Applications*, pp:231.
Viappiani, P., Zilles, S., Hamilton, H.J. and Boutilier, C. (2011) A Bayesian concept learning approach to crowdsourcing, *AAAI Technical Report WS-11-13*, pp. 60–67.
Yuen, M., King, I. and Leung, K. (2011) A survey of crowdsourcing systems, *IEEE 3rd International Conference on Social Computing*, pp:766-773.
Zhu, J. and Hastie, T. (2005) Kernel logistic regression and the import vector machine, *Journal of Computational and Graphical Statistics*, 14-1, pp:185-205.

## ACKNOWLEDGMENTS

This research was partially supported by the Automotive Research Center, a US Army Center of Excellence in Modeling and Simulation of Ground Vehicle Systems headquartered at the University of Michigan. This support is gratefully acknowledged.