Towards Recognizing "Cool": Can End Users Help Computer Vision Recognize Subjective Attributes of Objects in Images?

William Curran[1], Travis Moore[1], Todd Kulesza[1], Weng-Keen Wong[1], Sinisa Todorovic[1], Simone Stumpf[2], Rachel White[1], and Margaret Burnett[1]

[1]Oregon State University, School of EECS
Corvallis, OR 97331, USA
{curranw, moortrav, kuleszto, wong, sinisa, white, burnett} @eecs.oregonstate.edu

[2]City University, London
London, UK
Simone.Stumpf.1@city.ac.uk

ABSTRACT

Recent computer vision approaches are aimed at richer image interpretations that extend the standard recognition of objects in images (e.g., cars) to also recognize object attributes (e.g., cylindrical, has-stripes, wet). However, the more idiosyncratic and abstract the notion of an object attribute (e.g., "cool" car), the more challenging the task of attribute recognition. This paper considers whether end users can help vision algorithms recognize highly idiosyncratic attributes, referred to here as *subjective attributes*. We empirically investigated how end users recognized three subjective attributes of cars—"cool", "cute", and "classic". Our results suggest the feasibility of vision algorithms recognizing subjective attributes of objects, but an interactive approach beyond standard supervised learning from labeled training examples is needed.

Author Keywords

Computer vision, interactive machine learning, classification, human factors.

ACM Classification Keywords

H.1.2 Models and Principles: User/Machine Systems; I.4.m Image Processing and Computer Vision: Miscellaneous; I.2.6 Learning

General Terms

Design, Human Factors, Experimentation

INTRODUCTION

Computer vision research on image interpretation has been primarily focused on *naming objects* occurring in an image. A common approach is to use machine learning techniques on features extracted from the image (e.g., textured patches, edges, or segments) to detect occurrences of an object class of interest (e.g., cars). Recently, computer vision has also developed methods to describe objects' *measurable attributes* (e.g., cylindrical, has-stripes, wet) that can be quantified directly from pixel values [3,4]. (Note that attributes of *objects* are different from features of *images*: attributes are descriptive characteristics of object appearance, such as has-stripes, whereas features are perceptually salient image parts, such as corners and edges.) Attribute recognition algorithms operate in a similar manner as object recognition algorithms -- by classifying a vector of image features. The classifier learns natural variations of object attributes from training examples with the attributes annotated.

This paper explores recognizing *subjective attributes*. For example, given an image of a car, can an algorithm recognize whether the car is "cool", "cute", or "classic"? Before addressing the feasibility of using computer vision for this task, however, we must first understand how *people* distinguish between such subjective attributes. This problem has not yet been addressed in computer vision or human-computer interaction research.

Why Subjective Attributes are Challenging

One possibility is that subjective attributes can be expressed in terms of concrete image features, and machine learning algorithms may be able to recognize such subjective attributes. In order to express these attributes as image features, we need to understand how people reason about subjective attributes like "cool". Then, in order for machine learning algorithms to recognize these attributes, we need to address at least three challenges to such recognition.

The first challenge is that subjective attributes may be vaguely defined in end users' minds. End users may not be able to communicate their personalized definition of an attribute in a "language" understood by a computer vision algorithm. Vision algorithms can operate only on observable, semantically low-level image features (e.g., two image regions share a boundary) but because human visual perception is largely an unconscious process, these low-level image features are often meaningless to end users.

Second, prior work [3,4] has shown that simply tagging an entire training image with an attribute name is insufficient for learning the attribute. Richer annotations, such as placing a bounding box around an example object and tagging the region with the attribute name, can improve attribute recognition but are extremely time-consuming for humans to provide.

IUI'12, February 14–17, 2012, Lisbon, Portugal.

Third, because a *specific* end user's own definition of the subjective attribute is the gold standard, the amount of training data available is inherently limited, especially when the algorithm is first deployed. The amount of training data cannot be increased by simply pooling together training examples of attributes labeled by different annotators. Even for non-subjective attributes, related work [3,6] has shown that this pooling decreases performance because annotators differ to a large degree in their visual perceptions of an attribute even though they may have a similar mental model of that attribute. These differences introduce labeling noise that we expect to be even more pronounced in the case of subjective attributes.

	Example from participant transcript	Code
Cool	"Mostly it's very aerodynamic, it's very cool in design"	Aerodynamic
Cute	"Cute ones are usually smaller."	Small
Classic	"I guess if they're not built in and they're really round shape, so I would actually label classic" [he circles headlights].	Headlights-Round, Headlights-External

Table 1: Three examples of how one participant's feedback was coded.

Together, these issues indicate that algorithms that recognize subjective attributes need to extend beyond standard supervised learning from labeled training examples. Our results suggest the need for a rich form of interaction wherein an end user can guide the learning algorithm to be able to find cool cars according to *that* user's definition.

Research Questions

To investigate the issues we have raised in this section, we conducted an empirical study in which we asked end users to explain how to classify images of cars as "cool", "cute", or "classic". Our research questions were:

RQ1: What visual image properties do end users use to identify subjective attributes of objects?
RQ2: Can these visual image properties be mapped to low-level image features used by vision algorithms?
RQ3: How consistent are these visual image properties for subjective attributes across multiple users?
RQ4: How distinct are these visual image properties for different types of subjective attributes?

STUDY SET-UP

To investigate the viability of computer vision algorithms for recognizing subjective attributes, we conducted an empirical study in which participants explained how they reasoned about "cool", "cute", and "classic" subjective attributes of cars in images. We chose cars because they are widely used in computer vision and they present challenges pertinent to subjective attributes such as large variations in shape, arbitrary color patterns, and differences in sizes.

Participants and Procedures

We recruited 12 participants (7 males and 5 females) from the local community. These participants had little or no programming experience, no machine learning experience, and none were computer science majors. Participants were compensated $20 for their time.

We introduced participants to the idea of "thinking-aloud" by reasoning about cars. Participants were asked to describe prominent areas that stand out in a car by verbalizing their thoughts and marking up a printed image. Participants practiced this skill before the main study began.

The participants' task for the study was to describe which visual properties make (or do not make) a car "cool", "cute", or "classic" from a set of 15 car images. To avoid forced classifications, participants were asked to perform the task only on images that sparked their interest. This task lasted 20 minutes, which we observed to be suitable in our pilot runs. We video-recorded all sessions and transcribed the participants' verbalizations and image mark-ups for detailed analysis.

The Images

We obtained 67 images of cars from the PASCAL 2010 database [2], a well-known collection of images for object recognition. From these we selected a subset of images in which an entire car was in the center of the image and the image contained few background objects (e.g., pedestrians). Three researchers manually classified these as "cool", "cute", or "classic". We used the majority's decision to resolve any disagreements. For the study, we used 15 images, four for each subjective attribute, and three that did not clearly represent the subjective attributes.

RESULTS

Which Visual Image Properties Matter to End Users?

Subjective attributes rely on tacit knowledge, so it could be difficult for end users to precisely describe them as concrete visual properties [6]. We thus investigated how end users described the "cool", "cute", and "classic" subjective attributes, with particular attention to the concrete visual properties they discussed.

We used a fine-grained code set to characterize participants' feedback about "cool", "cute", and "classic" cars. The codes were words we extracted directly from the participant transcripts from part one of the study. Each code represents a visual property that the coders believed expressed a low-level image feature. We coded feedback as *part-property* when the focus was on a specific car part, or simply *property* when the participants' feedback was about the entire car. Table 1 shows an example for each code.

The primary purpose of this code set was to group synonyms together. Thus, if a participant described a car's headlights as "not built in" (Table 1), we applied the code "headlight-external". Two researchers iteratively built a list of such codes, refining the set to include each part and property participants discussed. The final code set had 37 codes for parts, 99 codes for properties of these parts, and 229 unique part-property combinations.

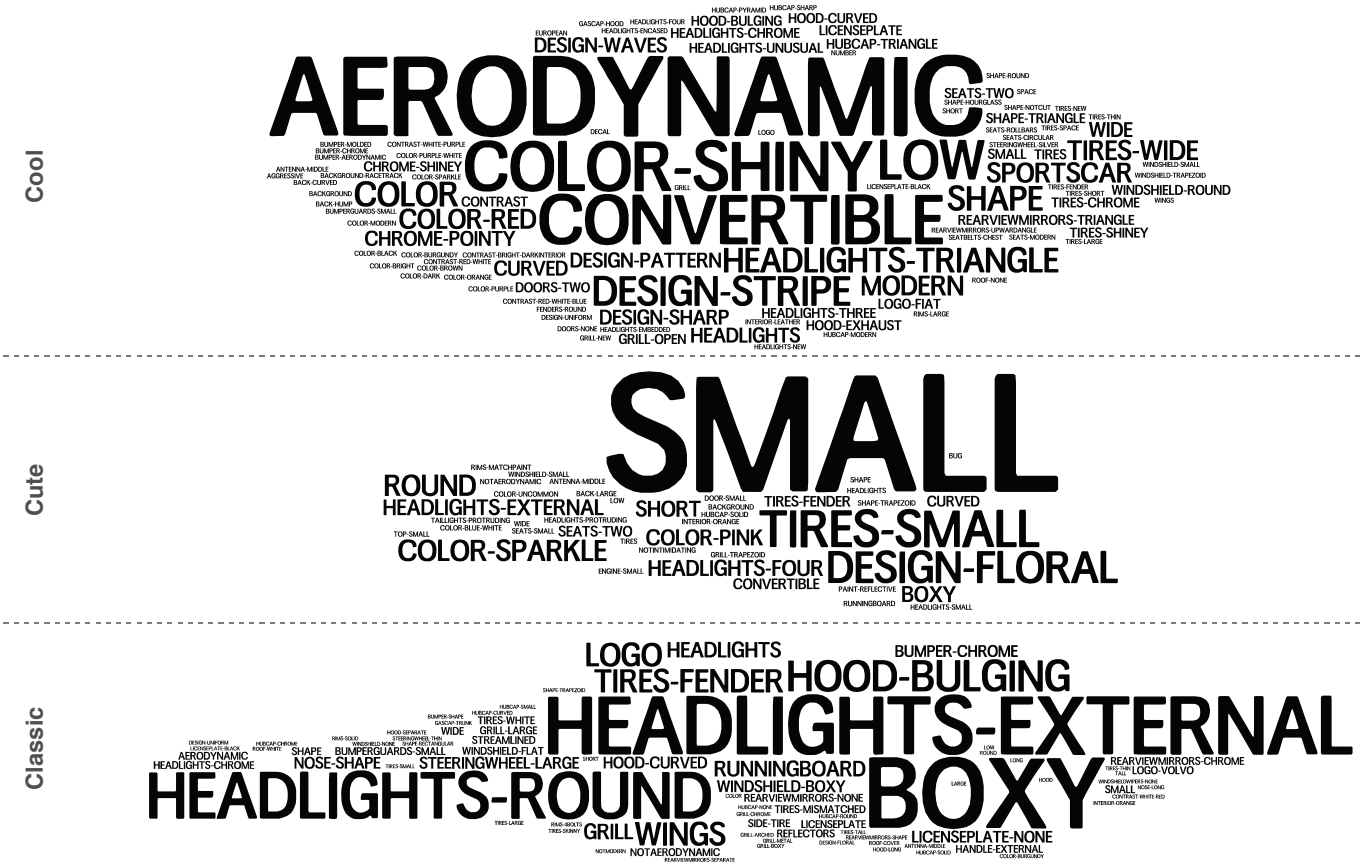

Figure 1: Tag clouds showing the frequency of participants' feedback for the subjective attributes "Cool,", "Cute," and "Classic."

To validate our code set's reliability, the two researchers independently coded 187 transcript samples (representing 40% of the total number of transcript samples). We computed reliability using the Jaccard index, where the number of agreeing codes (size of the intersection) was divided by the total number of applied codes (size of the union). Part codes and property codes were treated independently. For example, if Researcher 1 coded a segment as "headlight-round" while Researcher 2 coded the same segment as "headlight-external", agreement would be 1/3 because "headlight" agreed but "round" and "external" did not. The two researchers achieved a reliability of 80% over their 187 transcript samples. Despite the large number of codes, high reliability was achieved relatively easily as the code set functioned as a look-up dictionary for part and property synonyms. Given this acceptable level of reliability, one researcher independently coded the remaining transcripts.

Figure 1 shows how often each code occurred in participants' explanations about why a car was "cool", "cute", or "classic". The size of each code represents its popularity across all participants, i.e., between-participant consistency. (We did not analyze within-participant consistency because participants almost never explained the same things twice; for example, once they explained to us that aerodynamics was important to "coolness", they did not explain it again.) Participants consistently used certain visual properties of each subjective attribute. For example, in Figure 1, the property *Small* dominates the "cute" attribute. The most common descriptions applied to the entire car (e.g., *Aerodynamic* or *Boxy*), rather than part of a car, suggesting that participants focused on the whole *gestalt* before examining individual sub-objects.

Participants used a wide range of properties to explain subjective attributes but each attribute differed in the variety of properties (Figure 2). For instance, participants used nearly three times as many image properties to describe "cool" as they did for "cute". The degree of dominance also differed (e.g., *Small* was more dominant for "cute" than *Aerodynamic* was for "cool"). Most importantly, the participants' descriptions of properties had very little overlap between the attributes, as shown in Figure 2. In fact, only eight codes (4.5% of the total) were shared by all three attributes.

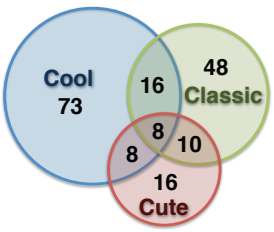

Figure 2: The number of codes for each attribute and their overlap.

Implications for Attribute Recognition Algorithms

As seen in Figure 1, the most common codes participants discussed, specifically 67% of them, involved shapes, sizes, and textures – all of which could be expressed

using the types of low-level image features commonly used in computer vision. For example, "boxy" can be captured by shape descriptors. Codes referring to car parts, such as "tires", can also be readily detected in images using off-the-shelf part detectors available in open-source computer vision libraries. A compound code such as "headlights-round" can be detected using a combination of part and shape detectors. An interesting code for "cool" pointed out by most participants was "aerodynamic", which can be interpreted as a simpler object attribute that defines the more complex attribute "cool". This hierarchical approach to attribute recognition, however, would require new theoretical formulations in computer vision.

The diffuse yet distinct nature of these subjective attributes has implications for attribute recognition algorithms. On one hand, the lack of overlap between the codes, as well as the presence of a core set of primary image properties, suggest that learning algorithms can learn the main concept defining the subjective attribute. On the other hand, while the primary properties can be identified as the amount of training data grows, identifying the subtler secondary image properties that define the attribute for a *specific* user is more challenging. The personalized definition of the subjective attribute for a specific user is much smaller than the diffuse set seen in Figure 1. Using the secondary properties collected from *all* the participants makes the concept noisier and harder to learn. A better alternative to supervised learning of the subjective attribute is to employ a richer form of interaction between the specific user and the learning algorithm, such as active learning [5], feature labeling [7] and interactive concept learning [1].

Altogether, our results strongly suggest that participants' descriptions of subjective attributes may be of real use to computer vision algorithms. Participants generally agreed on the core properties about *what* made a car "cool", "cute", or "classic", with little overlap between these attributes, and in terms that vision algorithms could readily leverage.

DISCUSSION

These results suggest two open issues for attribute recognition algorithms.

First, participants in our study viewed the shape of cars as the most prominent property defining "cool", "cute", and "classic". For example, Figure 1 shows that shape properties such as "round", "boxy", and "aerodynamic" were more relevant for recognition of the three car attributes than a car's material or color. This raises an open question regarding the relative importance of visual properties (shape versus color and texture)—a fundamental and as yet unanswered question for computer vision algorithms.

Second, participants were asked whether they thought properties for "cool", "cute", and "classic" cars could be applied to objects other than cars. For example, could the property "Small" apply equally to cute cars and cute cats? Most participants said the properties were independent of the object itself: "*I think so. I think the same criteria in general can be followed for bikes or clothes, tables, chairs, etc.*" These responses suggests new theoretical developments in computer vision, which currently treat attributes as tightly related to specific objects, and thus tie object recognition to attribute recognition. Our study suggests it might be possible to develop more general algorithms for recognizing attributes that transcend individual objects (e.g., an algorithm that recognizes "cuteness" regardless of whether the image shows cars or cats or cartoon characters).

CONCLUSION

This work is a first step toward expanding the scope of vision systems to include subjective, user-defined attributes—from simply "is there a car?" to "is this car *cool*?" Participants in our study consistently relied upon shapes, sizes, and textures to describe subjective attributes (RQ1). These image properties can be readily mapped to low-level image features commonly used by computer vision algorithms (RQ2). Participants agreed on a core set of primary image properties defining each subjective attribute but some attributes were more conceptually diffuse than others (RQ3). The properties participants identified for a specific subjective attribute did not substantially overlap, suggesting discriminative power for computer vision algorithms (RQ4). Overall, these results provide evidence that end users could interact with computer vision algorithms to recognize subjective attributes of objects in images.

ACKNOWLEDGMENTS

This work was supported in part by NSF grant IIS-0803487.

REFERENCES

1. Amershi, S., Fogarty, J., Kapoor, A. and Tan, D. Overview-based examples selection in mixed-initiative interactive concept learning. In *Proc. UIST*, ACM (2009).
2. Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A. The PASCAL Visual Object Classes Challenge 2010 Results. http://www.pascalnetwork.org/challenges/VOC/voc2010/workshop/index.html
3. Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. A. Describing objects by their attributes. In *Proc. CVPR*, IEEE (2009), 1778-1785.
4. Farhadi, A., Endres, I. and Hoiem, D. Attribute-centric recognition for cross-category generalization. In *Proc. CVPR*, IEEE (2010), 2352-2359.
5. Vijayanarasimhan, S. and Grauman, K. What's it going to cost you? Predicting effort vs informativeness for multi-label image annotations. In *Proc. CVPR*, IEEE (2009), 2262-2269.
6. Welinder, P., Branson, S., Belongie, S., and Perona, P. The multidimensional wisdom of crowds. *Advances in NIPS 23*, (2010), 2424-2432.
7. Wong, W-K., Oberst, I., Das, S., Moore, T., Stumpf, S., McIntosh, K., and Burnett, M. End-user feature labeling: A locally-weighted regression approach. In *Proc. IUI*, ACM (2011), 115-124.