Font Specificity

Luther Power[1] and Manfred Lau[2]
[1]Lancaster University, UK [2]City University of Hong Kong

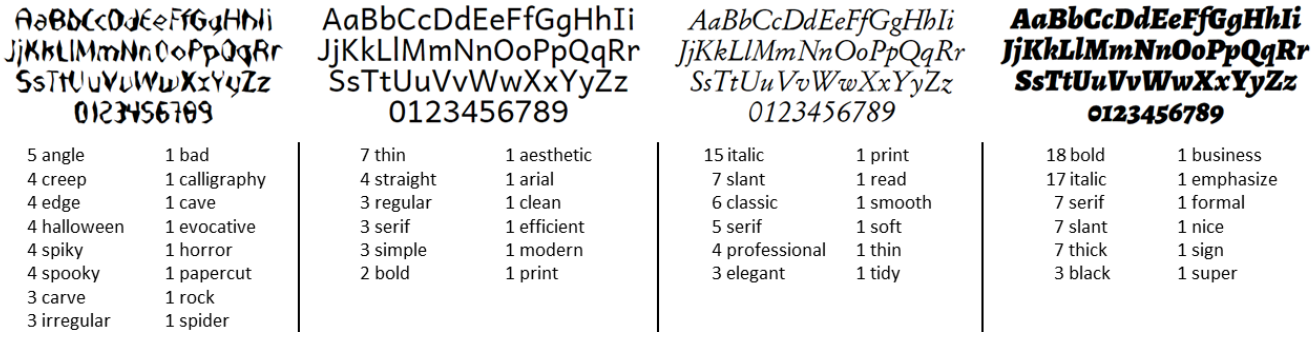

Figure 1: *Four fonts with increasing font specificity scores. The normalized specificity scores are 0.106, 0.359, 0.544, and 0.793 respectively. For each font, the top and bottom 15% of the number of unique words given by 24 participants to describe the font are shown (left and right columns and with corresponding number of participants). The notion of specificity can be seen in the distributions of words.*

Abstract

We explore the concept of "image specificity" for fonts and introduce the notion of "font specificity". The idea is that a font that elicits consistent descriptions from different people are more "specific". We collect specificity-based data for fonts where participants are given each font and asked to describe it with words. We then analyze the data and characterize the qualitative features that make a font "specific". Finally, we show that the notion of font specificity can be learned and demonstrate some specificity-guided applications.

CCS Concepts

• ***Computing methodologies*** → *Perception;* • ***Human-centered computing*** → *Human computer interaction (HCI);*

1. Introduction

The work of "image specificity" [JP15] introduces the concept that images that elicit consistent descriptions from different people are more "specific". In this paper, we explore this concept for fonts and introduce the notion of "font specificity". Analogous to image specificity, the idea of font specificity is that a font that elicits consistent descriptions from different people are more "specific". Fonts are different from images as an image has colors and usually consists of a larger scene, whereas a font focuses on text and the style that it is written in. In general, what makes a font specific (e.g. bold font) is different from what makes an image specific (e.g. image with people). We believe that this notion of font specificity can lead to new ways of understanding and thinking about fonts.

The contributions of this paper are:

- We introduce the concept of font specificity. The consistency of the human descriptions of fonts (i.e. the concept of specificity) can be used as a feature or descriptor of fonts, and the concept gives us new ways of thinking about fonts that have not been explored before.
- We collect specificity-based data for fonts.
- We explore the characteristics (i.e. qualitative features and quantitative image descriptors) that make a font more "specific".
- We show that the notion of font specificity can be learned and learn a function to predict specificity scores for new fonts.
- We demonstrate applications with the specificity-guided visualization and specificity-guided search of fonts.

2. Related Work

Image Captioning. There has been much work in computer vision in the problem of automatically generating captions or sentences to describe images [BCE*16, KFF17], and a detailed review of them is beyond the scope of this paper. The key difference is that they automatically generate sentences to describe images whereas we ask humans to provide words to describe fonts. More importantly, we care about the distribution of the provided words for the specificity concept and the actual words themselves are less significant.

Crowdsourcing. There exists previous work in collecting data through crowdsourcing and then learning from such data to solve various graphics problems. Examples include learning a similarity measure for 2D clip art [GAGH14] and for fonts [OLAH14].

DOI: 10.2312/egs.20191015

Crowdsourcing has also been used to extract depth layers and image normals from a photo [GSCO12] and to convert low-quality drawings into high-quality ones [GVGH12]. In this paper, we use crowdsourcing as a platform to collect data in the form of words to describe fonts, but crowdsourcing itself is not our focus.

Learning. There has been much recent work in using methods based on neural networks to solve image processing and computer graphics problems. For example, neural networks have been used to explore the problems of visual similarity for products or objects [BB15], image colorization [ISSI16], object cutout [XLC*16], and photo adjustment [YZW*16]. While we use neural networks for learning a function to predict specificity, the goal is to show that font specificity is learnable and the introduction of the concepts of font specificity is the focus of our paper.

3. Collecting Font Specificity Data

Inspired by the "image specificity" paper [JP15], we extend the notion of specificity to fonts and call this notion "font specificity". We collect data on the specificity of fonts by presenting each font to people and asking them to describe it. We then observe if the descriptions are consistent across different people. The more consistent these descriptions are, the more specific the font is.

The concept of font specificity is different from image specificity, even though the fonts are shown to users in the form of images. For the images used in the image specificity work, it is generally the content of the images that people describe. For the fonts in this paper, it is generally the style of the fonts that people describe. We did not intentionally ask for this, but this naturally occurs as the fonts all have the same content (i.e. the same letters and numbers). In addition, the characteristics that make an image specific (e.g. image with people) are different from those that make a font specific (e.g. bold and/or italic font). Hence font specificity is a fundamentally different concept with interesting aspects to be explored.

We emphasize that the purpose of the data collection is *not* to collect the actual words themselves. It is how the set of words may differ (or be similar) across different users that is the interesting part of this paper and that is the concept of specificity.

Crowdsourcing Data. The task here is to present fonts to humans and collect data by asking them to describe each font with text. We first collected 100 fonts from an online library (*fontlibrary.org*). To represent a font, we prepare an image showing the letters A-Z in uppercase and lowercase and the numbers 0-9. The positions of the letters and numbers are placed as consistently as possible across the fonts.

We use crowdsourcing as a platform to collect data and post the prepared fonts on Amazon Mechanical Turk. Each HIT (a set of questions on Mechanical Turk) starts with instructions for the participants. The participant is asked to give words that describe each font and using any words that come to mind. They can specify as many or as few words (but at least one) as they want. These words are typed into a box to the right of the font and participants are asked to separate them with spaces such that they are already tokenized. There are 20 fonts per HIT. If we have the same set of fonts in a HIT, the order is randomized each time. It takes approximately 3 to 10 minutes to complete each HIT and we pay $0.10 for each HIT. For each font, we collected data for 24 participants.

In contrast to the image specificity paper, we do not ask users for sentences to describe each font, as sentences would work better for an image containing usually at least a few objects and possibly a larger scene. In our case, we only have one font (per question) with the same letters and numbers, and it is better to ask for individual words to describe it since the words would be more likely to directly relate to the font.

Quality of Data. We set a constraint within the Amazon system that only allows participants to work on the questions if their approval rate for previous questions that they have done is higher than 80%. This can help to filter out users who may give poor quality data. We originally included a few questions with example answers as part of the instructions. However, the answers may bias the user responses and we find that the users did not need these example answers. Hence they were not used. In addition, we find that a qualification test given to participants which they must pass before working on the real questions that is typical for crowdsourcing experiments is not needed here. We believe that participants find it easy to think of words and it seemed to be an interesting task given the surprising variety of words that they came up with. As there is no "correct" answer to our questions and the variety of answers is important to what we eventually measure, we did not reject any user responses otherwise.

Processing of Rawdata. The words from the rawdata are passed through a stemming process [Bir06] such that words with the same root are considered to be the same. Words that are synonyms [Mil95] are also considered to be the same. We also removed one-letter words and common articles such as *a*, *an*, *of*, *the*, and *with*. Users were not supposed to type these articles, but a small number of them gave phrases that included these. When analyzing the data, we will use the term "word" to refer to the processed words after the data collection.

Checking for Data Consistency. The consistency of the words used to describe the fonts tells us about the specificity. We are not referring to this consistency here. For this subsection, we wish to check whether the data collection process itself is consistent. We split the data into two even groups randomly for each font. The split is based on participants, so we do not split a set of words provided by the same participant. This simulates, for example, a data collection process where we collect two different groups of participants on different days, and we wish to check whether the two groups of data would be different. We then compare the two distributions by comparing the two bins of words. We perform a χ^2 (chi-squared) test or a G-test if the frequencies are too small. We find that the chi-square or G-test statistic does not exceed the critical value at the 95% significance level for all fonts. This provides evidence that the two groups in each case come from the same distributions. At a higher level, this provides evidence that repeating the data collection process will give similar data.

4. Results and Analysis

Words Given by Participants. Figure 1 shows some examples of words collected and their frequency counts for four fonts. We initially thought that there may be some (but not too many) words to describe each font. However, we find that even for a relatively mundane font, the number of words participants came up with is surprisingly large. There are some participants who tend to describe each font with one or two words, while there are others who tend

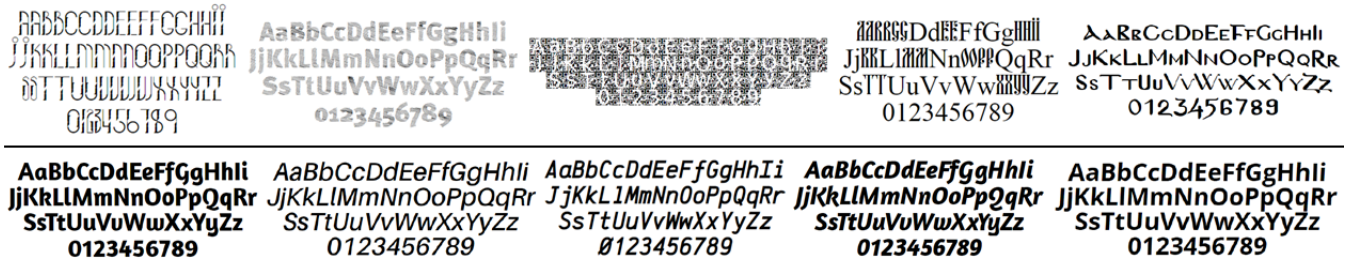

Figure 2: *Top row: The 5 fonts with the smallest specificity scores. Bottom row: The 5 fonts with the largest specificity scores.*

to be thoughtful and provided many words to describe each font (including words that indirectly relate to the font). For all fonts, the average number of total words per font is 75.4 and the average number of unique words (not counting repeats) is 44.4.

We take the top-50 most frequent words given by participants across all fonts and place them in a word frequency plot (Figure 3). The geometry-oriented words (e.g. bold, italic, thin, thick, straight) tend to have higher frequency or appear more often. As word frequency decreases, the words tend to relate more to subjectivity or emotion (e.g. nice, boring, pretty, fancy).

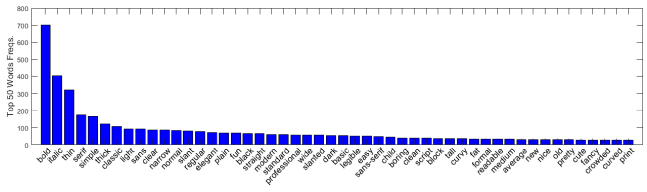

Figure 3: *The top-50 frequent words (rawdata before processing) given by participants across all fonts. Please zoom in to see the words more clearly.*

In addition, we consider the word types via their part-of-speech [TKMS03]. The words mostly belong to four categories: adjectives (63% of all words), singular noun (24%), plural noun (3.7%), and past-tense verb (2.7%).

Patterns Observed from Computed Font Specificity. To compute a font specificity score for each font given the collected data, we use an image specificity based method [JP15], except we consider the words provided by participants and not sentences. Since there is already a method to compute specificity, we focus on understanding this concept for fonts in this paper.

We observe that *a font with lower font specificity tends to have a larger variety of words to describe it*. This is evident in the four fonts in Figure 1. Furthermore, *a font with a distribution of words that is more biased or less uniform tends to be more specific, as this means that different people are more likely to describe it with the same words*. In Figure 1, the rightmost font is the most specific (among the four) and has a more biased distribution of words as the words "bold" and "italic" are highly frequent.

Figure 2 shows the fonts with the smallest and largest font specificity scores. The fonts with the smallest specificity tend to be creative, unusual, and/or unique. The fonts with the largest specificity tend to be bold, italic, or both.

For all our fonts, the correlation between their specificity scores and corresponding numbers of unique words is -0.560. This correlation is significant as the p-value is much less than 0.05. It shows there is some negative correlation, which matches with the concept of specificity that we want to capture. The value is not high which is as expected, as we did not expect specificity to be simply predicted by the number of words alone, but it is a more complex concept.

Correlation with Qualitative Features. We wish to see whether some human understandable qualitative features can describe the concept of font specificity. We collect data to test whether these features are related to being "specific". We use the same setup on Amazon Mechanical Turk as described in Section 3, but with different participants. For each of ten features, we ask users to provide a Likert score on a 1-5 scale. We collected data for 15 users and paid $0.10 per HIT. Table 1 shows the results of correlating (across all fonts) between the mean score for each feature and the font specificity scores. There are some significant positive or negative correlations, but the correlation values are not large. Some of these features can partially explain the concept of font specificity, but no single feature can explain it well, providing evidence that the concept can be complex.

Between	Font Spec	Between	Font Spec
Stand-out	-0.038	Fun	-0.184
Unique	**-0.359**	Boring	**0.441**
Visually Appealing	**0.347**	Elegant	**0.226**
Legible	**0.468**	Modern	-0.196
Creative	**-0.517**	Normal	**0.365**

Table 1: *Correlations between each of ten qualitative features and font specificity scores. The values are Pearson correlation coefficients and bolded values indicate that the corresponding p-value is less than 0.05 which means the correlation is significant.*

5. Learning and Predicting Font Specificity Scores

We learn a function to predict font specificity scores for new fonts, such that we do not need to collect specificity-based data in general. The function is a multi-layer neural network, with a patch (size 200x200) of a font image as input and its font specificity score as output. The largest size of a font image is 800x250, and each patch still covers a significant portion of the image. We perform data augmentation with 100 randomly sampled patches per font, thereby giving 10,000 training samples. We take all samples and perform 10-fold cross-validation to predict a font specificity score for each sample. We then correlate between the predicted font specificity

Figure 4: *Specificity-Guided Font Search: Query font shown on left. In each row (on the right), the top-5 closest fonts are shown (only first 3 letters shown for clarity). The seven rows correspond to searching with: font specificity score, pixel-wise Euclidean distance, contour curvature, Sobel filter, SIFT, SURF, and HOG.*

scores and ground truth font specificity scores. The correlation coefficient is 0.710, providing evidence that the notion of font specificity can be learned.

Prediction with Image Descriptors. We also learn functions to take as input an image descriptor of the font and compute as output its font specificity score. The image descriptors are: Sobel filters, SIFT, SURF, and HOG. Each of these gives us an input vector of dimensions between 100 and 480. The function is a multi-layer fully-connected neural network. We perform the same data augmentation and 10-fold cross validation as described above. The correlation coefficients are 0.43 for Sobel filter, 0.25 for SIFT, 0.06 for SURF, and 0.52 for HOG. These results suggest that the image descriptors may partially explain the concept of font specificity, but they cannot be used to predict it well.

6. Applications

We demonstrate the potential uses of the concept of font specificity in some specificity-based applications.

Specificity-Guided Visualization. The specificity concept can be used to visualize a set of fonts. For example, the plots in Figures 1 and 2 and the results from Section 4 show that the fonts that tend to be more creative are grouped together since they have lower font specificity, while the fonts that are bold and/or italic are grouped together since they have higher font specificity.

Specificity-Guided Font Search. The specificity concept can be used for search and retrieval applications of fonts. The idea is to use the font specificity score as a distance metric such that the distance between two fonts is the difference between their specificity scores. Figure 4 shows the advantage of searching with font specificity scores over various image descriptors. The first row shows the results for searching with font specificity scores, and these 5 fonts visually match with the query better than those in the other rows. The other rows are for searching with various image descriptors, and in each row there is at least one font that is somewhat strange or different from the query.

7. Discussion

We have introduced the notion of font specificity and studied various aspects of this problem. For future work, there are extensions that can be made to the major parts of the paper, for example in the data collection and applications. We believe that this short paper provides a good start to understanding font specificity and hope that our work will inspire further research and uses of this notion.

One limitation of our work is that the number of fonts can be larger and it would be useful to collect more data in the future. However, collecting the fonts data is difficult as there has to be multiple participants to label every font. The number of human labelers per font is therefore an important parameter that can be tested. In addition, future work can explore the fonts of different languages.

Furthermore, although the participants provide words for each font individually, the context of the other fonts shown to participants at the same time may affect their responses. Future work can consider this aspect when collecting and analyzing the fonts data.

References

[BB15] BELL S., BALA K.: Learning visual similarity for product design with convolutional neural networks. *ACM Trans. Graph. 34*, 4 (July 2015), 98:1–98:10. 2

[BCE*16] BERNARDI R., CAKICI R., ELLIOTT D., ERDEM A., ERDEM E., IKIZLER-CINBIS N., KELLER F., MUSCAT A., PLANK B.: Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research 55*, 1 (Jan. 2016), 409–442. 1

[Bir06] BIRD S.: Nltk: The natural language toolkit. *COLING/ACL on interactive presentation sessions* (2006), 69–72. 2

[GAGH14] GARCES E., AGARWALA A., GUTIERREZ D., HERTZMANN A.: A Similarity Measure for Illustration Style. *ACM Trans. Graph. 33*, 4 (July 2014), 93:1–93:9. 1

[GSCO12] GINGOLD Y., SHAMIR A., COHEN-OR D.: Micro Perceptual Human Computation for Visual Tasks. *ACM Trans. Graph. 31*, 5 (Sept. 2012), 119:1–119:12. 2

[GVGH12] GINGOLD Y., VOUGA E., GRINSPUN E., HIRSH H.: Diamonds from the Rough: Improving Drawing, Painting, and Singing via Crowdsourcing. *Proceedings of the AAAI Workshop on Human Computation (HCOMP)* (2012). 2

[ISSI16] IIZUKA S., SIMO-SERRA E., ISHIKAWA H.: Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Trans. Graph. 35*, 4 (July 2016), 110:1–110:11. 2

[JP15] JAS M., PARIKH D.: Image Specificity. *CVPR* (June 2015), 2727–2736. 1, 2, 3

[KFF17] KARPATHY A., FEI-FEI L.: Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell. 39*, 4 (Apr. 2017), 664–676. 1

[Mil95] MILLER G. A.: Wordnet: A lexical database for english. *Communications of the ACM 38*, 11 (Nov. 1995), 39–41. 2

[OLAH14] O'DONOVAN P., LIBEKS J., AGARWALA A., HERTZMANN A.: Exploratory Font Selection Using Crowdsourced Attributes. *ACM Trans. Graph. 33*, 4 (July 2014), 92:1–92:9. 1

[TKMS03] TOUTANOVA K., KLEIN D., MANNING C. D., SINGER Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. *North American Association for Computational Linguistics on Human Language Technology - Volume 1* (2003), 173–180. 3

[XLC*16] XU H., LI Y., CHEN W., LISCHINSKI D., COHEN-OR D., CHEN B.: A holistic approach for data-driven object cutout. *ACCV* (2016), 245–260. 2

[YZW*16] YAN Z., ZHANG H., WANG B., PARIS S., YU Y.: Automatic photo adjustment using deep neural networks. *ACM Trans. Graph. 35*, 2 (Feb. 2016), 11:1–11:15. 2