The Workshops of the Thirty-Second
AAAI Conference on Artificial Intelligence

Visual Listening In: Extracting Brand Image Portrayed on Social Media

Liu Liu,[1] Daria Dzyabura,[1] Natalie Mizik,[2]
[1.] NYU Stern School of Business, New York, NY 10012
[2.] UW Foster School of Business, Seattle, WA 98195
{lliu, ddzyabur@stern.nyu.edu}, {nmizik@uw.edu}

Abstract

Marketing academics and practitioners recognize the importance of monitoring consumer online conversations about brands. The focus so far has been on text content. However, images are on their way to surpassing text as the medium of choice for social conversations. In these images, consumers often tag brands. We propose a "visual listening in" approach to measuring how brands are portrayed on social media (Instagram) by mining visual content posted by users, and show what insights brand managers can gather from social media by using this approach. We first use two supervised machine learning methods, traditional support vector machine classifiers and deep convolutional neural networks, to measure brand attributes (glamorous, rugged, healthy, fun) from images. We then apply the classifiers to brand-related images posted on social media. We study 56 brands in the apparel and beverages categories, and compare their portrayal in consumer-created images with images on the firm's official Instagram account, as well as with consumer brand perceptions measured in a national brand survey. Although the three measures exhibit convergent validity, we find key differences between how consumers and firms portray the brands on visual social media, and how the average consumer perceives the brands.

Introduction

Brand managers have long recognized the importance of creating, managing, and measuring brand image. With the rise of social media platforms, a profound shift has occurred not only in how individuals consume information, but also in the very origins of the information itself. Much brand-related content is now created and spread through Twitter postings, user discussion forums, social networking sites, and blogs. With the wider, more egalitarian distribution model, monitoring how a brand is portrayed on social media is essential to effective brand management.

Our focus is on consumer-created visual content, which is on the rise. With the proliferation of camera phones, cheap data plans, and image-based social media platforms, photo taking and sharing has become an important part of consumers social lives. Images are becoming an increasingly prevalent form of online conversations. In these shared photos, consumers often tag brands, resulting in a large volume of photos depicting a brand. For example, a search of hashtag #nike in Instagram returns over 52 million photos tagged with Nike.

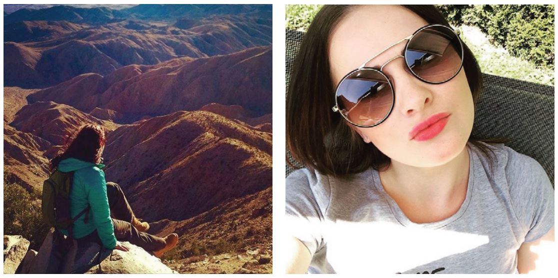

(a) #eddiebauer (b) #prada

Figure 1: Sample images from Instagram hashtagged with brands

By tagging brands in their social media posts, consumers communicate with each other about brands, by linking the brand with context, feelings, and consumption experiences. For example, consider the two instagram posts in Figure 1. The first image is tagged with *eddiebauer* and the second one is tagged with *prada*. The first image shows us the consumer wearing Eddie Bauer on a hike, sitting on a rock, facing mountains in the distance. The second image shows the user taking a picture of herself wearing Prada sunglasses and bright red lipstick. The two images differ in terms of their content (mountainous landscape vs. head shot) but also in their visual properties (color palate, contrast, amount and direction of edges, etc). These examples suggest that computer vision tools may be able to identify patterns in content linked to various brands. They also highlight the value of mining visual user generated content: photographs capture rich information about the consumption experience that can be harnessed to get a more complete understanding of consumer online brand communications.

We introduce a "visual listening in" approach for monitoring visual brand content created and shared by consumers on social media. We create metrics, derived from these images, that allow firms to compare how their brand is portrayed on social media relative to competitors. Specifically, we map the images from each brand onto specific brand per-

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ceptual attributes, and compare the brands along these attributes. We focus on intangible brand attributes, which go beyond the functional attributes of the product ((Park, Jaworski, and Maclnnis 1986)). Positioning brands along intangible attributes allows firms to differentiate themselves from one another in categories with functionally similar products. When choosing in categories such as beverages or apparel (our focus categories in this paper), which contain many brands offering products with very similar functionality, what often makes a bigger difference is the feelings consumers have about the brand. For example, in the apparel category, we see Old Navy appears to be positioned as fun, whereas Levis's is positioned as rugged, allowing the brands to target different groups of consumers. Without these intangible brand attributes, the two firms' products would likely be "just a pair of jeans." For consumers, these intangible attributes of brands provide the benefit of allowing consumers to choose the brand that seems most appropriate, and to express themselves through the brand.

The rest of the paper is organized as follows. First, we introduce our visual listening in approach for measuring brand attributes expressed in consumer-created images. Then, we demonstrate the use of this approach on brand-related images posted on Instagram and discuss the results of empirical studies. We conclude with a discussion of future directions.

Methodology

We adopt a two-stage approach. In the first stage, we build and examine image classifiers to predict whether a particular brand attribute is expressed in a given image, for example, whether an image looks rugged. This step requires a large set of labeled training data. For classification, we adopt both a traditional machine learning approach, using an SVM with human-defined image features, and a state-of-the-art deep learning approach, learning deep ConvNets for each perceptual brand attribute. The deep learning approach achieves better out-of-sample prediction accuracy, but SVM provides more interpretable insights on the relationships between image features and perceptual attributes. In the second stage, we apply the image classifiers from the first stage to consumer-created images to measure how brands are portrayed in these images.

Data

To train image classifiers of brand attributes, we need to collect an annotated data set, consisting of images labeled with respect to whether or not they express each attribute. To our knowledge, no existing data set is annotated with brand attributes, so we create one.

We gather an annotated training set from Flickr, an online photo-sharing website. Flickr lends itself well to gathering a training set, because, unlike Instagram, it provides a search engine that returns the most relevant photos for a keyword, based on text labels provided by users, image content, and clickstream data.[1]. Flickr has been used as a data source in previous visual and social network research (e.g., (Zhang et al. 2012; Dhar, Ordonez, and Berg 2011; McAuley and Leskovec 2012)).

Table 1: List of Features by Feature Type

Feature Type	Feature
Color	RGB color histogram HSV color histogram L*a*b color histogram
Shape	Line: number of straight lines Line: percentage of parallel lines Line: histogram of line orientations & distances Line: histogram of line orientations Corner: percentage of global corners Corner: percentage of local corners Edge Orientation Histogram Histogram of Oriented Gradients (HOG)
Texture	Local Binary Pattern (LBP) Gabor

We explore four brand attributes that are particularly relevant to the brand categories we study in next section: glamorous, rugged, healthy, and fun. For each attribute, we query the attribute word on Flickr's search engine and collect images in the top 200 result pages returned by Flickr, which is about 2,000 photographs. We use these images as our positive-labeled data. We also need negative-labeled data for each attribute, comprising images that do not express the perceptual attribute. We query the antonym of the perceptual attribute in Flickr's search engine and again collect the images in the top 200 result pages returned. For example, for the perceptual attribute rugged, we use "rugged" as the query to collect positive instances and "gentle" as the query to collect negative instances. The entire training set for all four perceptual attributes we study contains 16,368 photographs.

SVM with Human-Defined Image Features

We first use a more traditional machine learning approach: train SVM classifiers with a set of predefined image features. We extracted 13 types of features from each image, relating to the color, shape, and texture of the image. Many of these features are widely used in the computer vision literature, and are known to work well for object recognition and detection tasks. Color, shape, and texture are also among the fundamental visual design elements in design literature. Table 1 provides a list of all the features we use for classification.

Once the features are extracted, we trained a Support Vector Machine (SVM) with linear kernel. For each perceptual attribute, we train the SVM classifier with features from just one type of features, as well as combination of features across different types. We compared their performance out-of-sample.

Deep Learning with Transfer Learning

Second, we adopted a deep learning approach to classify images into brand perceptual attributes. We use a type of neural network called deep convolutional neural networks (ConvNets). This type of networks are widely used to process image data.

[1] http://blog.flickr.net/en/2015/05/07/flickr-unified-search/

A challenge with using deep learning is that it requires very large sets of training data. We don't have a data set of sufficient size to design and train a network. A common approach is to "fine-tune" models that have been trained previously on a very large data set in a related domain. This type of approach is an example of transfer learning: using knowledge from one domain to help prediction in another domain. It is similar to using as a prior in Bayesian estimation parameters that are estimated on a different data set.

For each perceptual attribute, we fine-tuned two well-trained ConvNets on our data. The first is a Caffe reference model similar to the "AlexNet" model ((Krizhevsky, Sutskever, and Hinton 2012)). The ConvNet is trained on an ImageNet data set ((Deng et al. 2009)) with 1.2 million images of 1,000 different object categories. We call the resulting fine-tuned model ConvNet$_{ImageNet}$. The second is a ConvNet for Flickr-style recognition, which is fine-tuned from the Caffe reference model ((Krizhevsky, Sutskever, and Hinton 2012)) on 80,000 Flickr style images by (Karayev et al. 2013). It has the same network architecture but different parameter values. We call the resulting model ConvNet$_{FlickrStyle}$. We expect the ConvNet$_{FlickrStyle}$ model will perform better, because the Flickr-style images are more similar to our data. Also, the style recognition task is more similar to classifying perceptual attributes than is object detection.

We adapted the architecture of each pre-trained model but changed the number of neurons in the last layer to two, because we are doing binary classification for each perceptual attribute. When training, we initialize the weights of earlier network layers with those learned from the pre-trained model and fine-tune the weights by continuing back propagation on our data. We initialize the weights of the last layer (classification layer) with random values and train it from scratch on our data set. We use a high learning rate for the last layer so that the parameters in the last layer can change quickly with our data. However, we use a small learning rate for earlier layers where weights are fine-tuned, to preserve the parameters learned from the pre-trained model and to transfer that knowledge to our task.

We use the Caffe deep learning framework (see (Jia et al. 2014)) to fine-tune the ConvNets. Because our data set is relatively small (about 4,000 images for each classification problem), we run each model for 5,000 iterations. It converges quickly. Figure 2 shows the learning curve of the two types of fine-tuned models for each perceptual attribute. We choose the model snapshot when the training loss decrease in the training set between adjacent 100 iteration windows is smaller than 0.001.[2] It takes about eight hours in a single K80 GPU node in the university's high performance cluster.

Image Classification Performance

We report out-of-sample performance for both the SVM classifier and the two ConvNets. Recall the SVM uses three types of features: related to color, shape, and texture of images. To gain insight into what properties of an image are related to a brand attribute, we train the SVM classifier with features of only one type (e.g., only color features), as well as a combination of features of all types.

Table 2: Out-of-Sample Performance by Image Classifier

	SVM$_B$	SVM$_C$	SVM$_S$	SVM$_T$	ConvNet$_I$	ConvNet$_F$
glamorous	74.1%	69.5%	70.0%	70.9%	81.5%	**84.9%**
rugged	73.3%	65.6%	70.0%	67.2%	76.6%	**80.7%**
healthy	63.4%	63.4%	56.0%	51.4%	57.1%	**70.6%**
fun	65.3%	60.4%	57.3%	55.6%	74.5%	**81.5%**
Mean	69.0%	64.7%	63.3%	61.3%	72.4%	**79.4%**

Table 2 shows the performance of each of the perceptual attribute classifiers. The mean classification accuracy ranges from 61.3% with SVM with just texture features to 79.4% with ConvNets fine-tuned from the Flickr-style model. Recall the training set was balanced with 50% positive and 50% negative instances, so 50% accuracy is what we would get by guessing randomly. All classifiers outperform this benchmark. Among our classifiers, ConvNets fine-tuned from the Flickr style model perform the best across all perceptual attributes, as well as on average. The accuracy of this classifier is high for this type of prediction task. For comparison, photograph-style recognition tasks (binary classification of styles such as Vintage or Minimal) achieve an average per-class accuracy of 78% ((Karayev et al. 2013)). As we expected, the ConvNet fine-tuned from the Flickr-style model performs better than the one fine-tuned from the ImageNet object detection model. This difference is likely because the ImageNet ConvNet was optimized for object detection, whereas the Flickr ConvNet was optimized for image style classification. Our task, classification of brand attributes, is more similar to image style classification than to object detection; therefore, the optimal features for classifying image styles are better for classifying perceptual attributes than are features optimized for object detection.

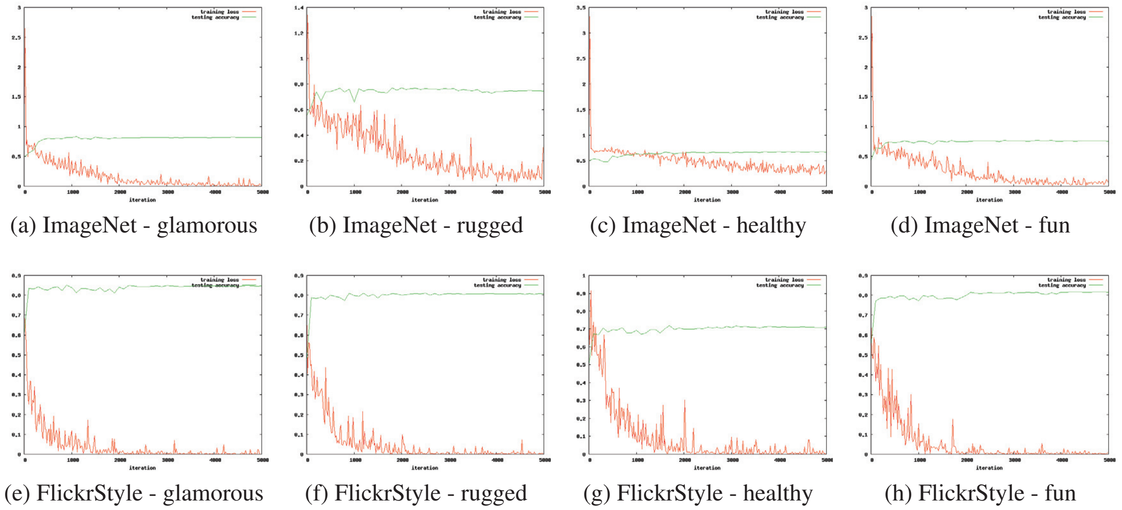
Image-Based Brand Metric

The ultimate goal of our research is to measure the perceptual attributes in consumer-created brand images in order to understand how their brand is portrayed on social media. The specific metric we use is the ratio of brand images classified as positive on each attribute.

Recall again the two motivating images in the introduction section, presented in Figure 1. The first image is hashtagged with *eddiebauer* and the second one is hashtagged with *prada*. The first image is classified as positive by the *rugged* classifier, and the second image is classified as positive by the *glamorous* classifier.

We are interested in the degree to which an attribute is expressed in consumer-created brand photographs on average. We therefore compute the proportion of images tagged with a given brand, that are classified as positive on an attribute. The higher this proportion, the more visual content portrays the brand as having the attribute. Coming back to our exam-

[2] ConvNets fine tuned from ImageNet model converged at iteration 3400, 3100, 600, and 1900 for glamorous, rugged, healthy, and fun, respectively. ConvNets fine tuned from Filckr Style model converged at iteration 2200, 4000, 2600, and 4100



(a) ImageNet - glamorous (b) ImageNet - rugged (c) ImageNet - healthy (d) ImageNet - fun

(e) FlickrStyle - glamorous (f) FlickrStyle - rugged (g) FlickrStyle - healthy (h) FlickrStyle - fun

Figure 2: Learning curve of fine tuned ConvNets (Axis x: training iteration. Axis y: red line for training loss, green line for testing accuracy)

Table 3: Prada vs. Eddie Bauer: t-test comparison between Prada and Eddie Bauer on mean of glamorous photos and mean of rugged photos

	Prada	EddieBauer	t-statistic	p-value
glamorous	60.0%	43.1%	8.61	$< 10^{-6}$
rugged	49.9%	63.2%	-6.75	$< 10^{-6}$

ple of Prada versus Eddie Bauer, we can see in Table 3[3] that, 60.0% of Prada images tagged by consumers are glamorous, whereas only 43.1% of Eddie Bauer are glamorous, which is significantly lower ($p < 10^{-6}$). On the ruggedness attribute, 63.2% of Eddie Bauer images are classified as rugged, which is significantly higher than Prada ($p < 10^{-6}$). These results match our intuition, because Prada is a glamorous brand and Eddie Bauer is a rugged brand. In the next section, we apply the method to a large set of brands on Instagram and show insights generated from social media.

Application

We have now trained image classifiers that can predict whether a given image represents the attributes, and defined a brand metric based on the classification results on brand images. In this section, we apply the classifiers to a large set of brand images on Instagram. We first describe this brand image data set. We then discuss three related but different brand metrics: two derived from brand images generated by consumers and firms on Instagram and one survey-based metric of brand perception. Finally, we present two empirical studies which compare these three types of brand metrics. We discuss the results and managerial implications of the empirical studies.

Instagram Brand Image Data Set

Instagram is an image-based social media platform that has quickly emerged as a popular communication medium. Since its launch in 2010, users have shared over 300 billions photos, and add an average of 70 millions photos daily ((Kane and Pear 2016)). Among these photos, users often hashtag brands, creating a collection of brand related images that may contain valuable insights for the firm. We collect data for two product categories for which consumers post a lot of photos, apparel and beverages, for a total of 56 large national brands.

Consumer-Created Brand Images We obtain consumer-created photographs by crawling Instagram for posts that are hashtagged with the name of the brand. When crawling, we filtered spam photos, resale photos, and photos that are posted by the official account of the brand. We collected about 2,000 photographs for each brand. The data set contains 114,367 photographs in total. All the data were collected between May and October 2016.

Firm-Created Brand Images We also obtain firm-created photographs from the brands' official Instagram account pages. There are 72,089 photos in total, and each brand's official account has 1,360 photos on average. Three beverage brands in our consumer data set do not have an official account on Instagram.

Brand Metrics

For each brand and each brand attribute, we applied the image classifiers trained in previous section to consumer-created brand images, and compute the ratio of the brand's

[3] The prediction is computed using ConvNet$_{FlickrStyle}$. The same comparison based on brand attribute prediction with the best SVM classifier and ConvNet$_{ImageNet}$ is consistent.

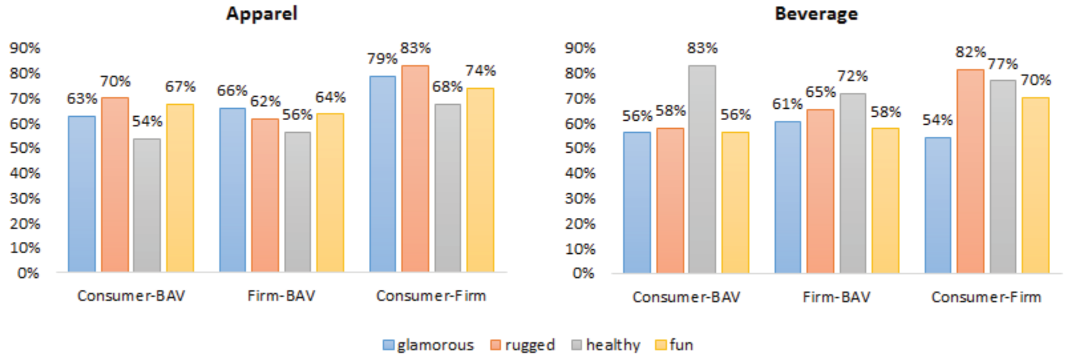


Figure 3: Percentage of Consistent Brand Pairs between Different Brand Metrics (BAV - Brand perception survey)

images that express the perceptual attribute. This brand attribute metric captures brand image portrayed by consumers on social media. It is closely related to usage context and consumption experience of brands. As in these photos, consumers put brands in a context and link brands with their lives.

Similar to consumer-created brand images, for each brand and brand attribute, we also compute the image-based brand metric from official photos. It can be considered a metric of the degree to which the firm portrays itself along the brand attribute. It captures part of firms' marketing effort to create their brand identities. Marketing and branding theory tells us that firms use brand elements, such as advertising, social media, and product packaging, to create a brand image in consumers' minds.

We also get a brand perception measure from a large national survey. The survey metric aims to capture the perception of a nationally representative sample of consumers. The large national brand perception survey we use is Young and Rubicam's Brand Asset Valuator (BAV) ((Lovett, Peres, and Shachar 2014)). In this survey, respondents are asked to indicate whether they perceive each brand as representing each attribute. The brand perception score is captured by the percentage of consumers who indicate the brand represents the attribute. Since its launch in 1993, BAV has become the largest database for brand perceptions. We use BAV data that were collected during the first quarter of 2016.

The three brand metrics capture related but different aspects of brand image. By comparing the three metrics, firms can either correct problems, or leverage and identify new opportunities.

Empirical Studies

We conducted two empirical studies to compare the brand attribute metrics extracted from consumer-created images with those extracted from firm-created images, and brand perception measure from a large national survey.

Product Category Level Consistency We first checked the consistency between the three brand metrics on the product category level. Brand attribute measures are usually used for brand comparison; for example, "Is my brand more or less healthy than my competitors?" Hence, we compare the three brand metrics by predicting the order of brand pairs: given a pair of brands, which one is more associated with a certain attribute, for example, which of a pair of brands is healthier?

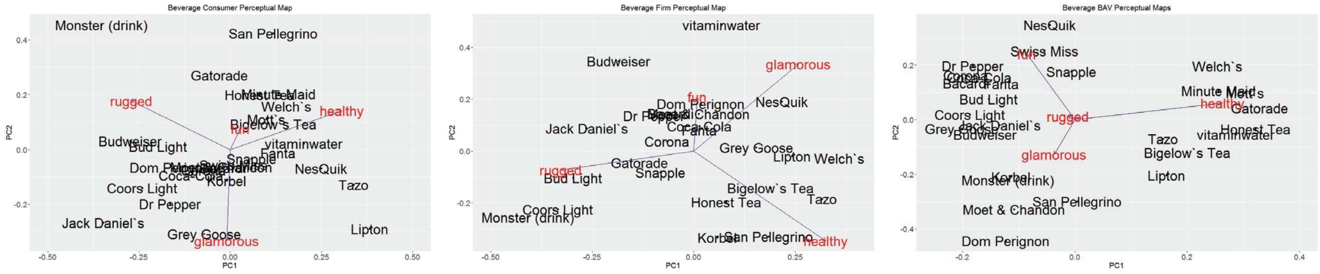
Figure 3 reports the percentage of consistent pairs between the brand metric computed from consumer-created brand images, firm-created brand images, and survey metrics, based on the classification results from ConvNet$_{FlickrStyle}$ in previous section. All are above 50%, which would have been obtained if guessing randomly, and most of them above 60%. The consistency indicates that brand image portrayed on social media reflects consumers' brand perception. It also provides convergent validity to the method.

The consistency of the three metrics on each attribute varies by product category. In the apparel category, we see high consistency on the attributes glamorous, rugged, and fun. These three attributes are key differentiators for apparel brands; for example, Victoria's Secret and Prada are perceived as much more glamorous than Eddie Bauer and LL Bean. In the beverages category, the perceptual attributes with more consistent brand pairs are healthy and rugged. Rugged is a relevant attribute in both categories: some brands are clearly positioned and perceived as rugged in apparel (e.g., Levi's, Eddie Bauer) and beverages (e.g., Jack Daniels, Gatorade).

Brand Maps The brand metric computed from consumer-created brand images on Instagram allows us to derive brand maps of the brands in each product category.

Figure 4 presents the perceptual maps of beverage brands based on brand metrics computed from consumer-created brand images, firm-created images, and the BAV survey. These maps provide important insights for managers with respect to where their products fall in the competitive landscape, the relevance of the attributes, and the discrepancy between how consumers and firms portray the market on social media, versus how the average population perceives it.

The maps created from the consumer- and firm-created images have healthy, rugged, and glamorous as three important factors, whereas fun is a less important factor. By



(a) Consumer (b) Firm (c) BAV

Figure 4: Beverage perceptual maps based on brand metrics from consumer-created brand images, firm-created brand images, and survey brand perception measures (BAV)

contrast, the map derived from the BAV survey results has healthy as the most important factor, fun and glamorous as less important, and rugged is not at all important. Therefore, observing the online conversations rather than asking directly about consumers' perceptions uncovers a new relevant attribute. The brands vary on ruggedness in the two image data sets, but not on the survey data set.

Based on the consumer images, a set of brands exist that are rugged and unhealthy, which consists primarily of alcohol (Jack Daniel's, Coors Light, Bud Light, Budweiser, Corona). Interestingly, Coca Cola and Dr. Pepper are in the same general region as these brands on the images. By contrast, waters, juices, teas, and sports drinks are identified as healthy based on consumer images. This separation provides face validity to our method: it is able to separate the sodas and alcohol from the healthy beverages based on simply the Instagram images (recall that we use no other information on the brands or products).

By comparing brand maps from difference sources, firms can also identity gaps in their position strategies. Take the brand Fanta for example. On the maps created based on firm photos and BAV metrics, Fanta falls into the same cluster of unhealthy brands such as Dr. Pepper and Coca Cola (Figure 4b and Figure 4c); on the consumer-created images map, however, Fanta falls closer to the juices. One possible reason for this discrepancy may be that consumers know that Fanta is a soda and therefore identify it as unhealthy when asked about it in a survey. However, the consumption experiences of Fanta are more similar to those of juice.

Conclusion

With the rapidly growing amount of visual brand-related content consumers create on social media, these images are a promising source for marketers and brand managers to track their brands‘ performance. This paper proposes an approach to leveraging these image data by extracting scores of brand perceptual attributes expressed in the images. We have demonstrated the resulting metrics are consistent across consumer- and firm-generated images, as well as large survey-based metrics of consumer perceptions. We also showed that brand managers can use this approach to identify relevant brand attributes and gaps in their positioning strategies.

Although text-mining approaches have gained popularity in leveraging user generated content for brand monitoring, image-mining approaches are still relatively new. This paper bridges the image-processing literature with the branding literature by proposing an approach to online brand monitoring and market intelligence through consumer-generated images. This approach enables managers to monitor how their brands are portrayed on image-based social platforms by mining consumer-created brand images.

The image-mining methods presented in this paper provide a first step in analyzing rich image data generated by consumers and firms. Future research can extend the application to analyzing how images affect consumer search behavior, learning consumer preference of product design, designing Ads targeting strategy based on consumer-posted images on social media, and so on. Given visual content is a ubiquitous part of modern life and affects consumers‘ decision making in multiple stages, being able to capture and incorporate visual content into marketing models is important.

References

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 248–255. IEEE.

Dhar, S.; Ordonez, V.; and Berg, T. L. 2011. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 1657–1664. IEEE.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, 675–678. ACM.

Kane, G. C., and Pear, A. 2016. The rise of visual content online. *Sloan Management Review*.

Karayev, S.; Trentacoste, M.; Han, H.; Agarwala, A.; Darrell, T.; Hertzmann, A.; and Winnemoeller, H. 2013. Recognizing image style. *arXiv preprint arXiv:1311.3715*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012.

Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Lovett, M.; Peres, R.; and Shachar, R. 2014. A data set of brands and their characteristics. *Marketing Science* 33(4):609–617.

McAuley, J., and Leskovec, J. 2012. Image labeling on a network: using social-network metadata for image classification. In *European Conference on Computer Vision*, 828–841. Springer.

Park, C. W.; Jaworski, B. J.; and Maclnnis, D. J. 1986. Strategic brand concept-image management. *The Journal of Marketing* 135–145.

Zhang, H.; Korayem, M.; Crandall, D. J.; and LeBuhn, G. 2012. Mining photo-sharing websites to study ecological phenomena. In *Proceedings of the 21st international conference on World Wide Web*, 749–758. ACM.