

# An Image is Worth a Thousand Terms? Analysis of Visual E-Commerce Search

Arnon Dagan  
ardagan@ebay.com  
eBay Research  
Israel

Ido Guy  
idoguy@acm.org  
eBay Research  
Ben-Gurion University of the Negev  
Israel

Slava Novgorodov  
snovgorodov@ebay.com  
eBay Research  
Israel

## ABSTRACT

Visual search has become popular in recent years, allowing users to search by an image they are taking using their mobile device or uploading from their photo library. One domain in which visual search is especially valuable is electronic commerce, where users seek for items to purchase. In this work, we present an in-depth comprehensive study of visual e-commerce search. We perform query log analysis of one of the largest e-commerce platforms' mobile search application. We compare visual and textual search by a variety of characteristics, with special focus on the retrieved results and user interaction with them. We also examine image query characteristics, refinement by attributes, and performance prediction for visual search queries. Our analysis points out a variety of differences between visual and textual e-commerce search. We discuss the implications of these differences for the design of future e-commerce search systems.

## CCS CONCEPTS

• **Information systems** → **Online shopping; Query log analysis; Web log analysis; Search interfaces.**

## KEYWORDS

E-commerce search; product search; query log analysis; query performance prediction; search by image; visual search.

### ACM Reference Format:

Arnon Dagan, Ido Guy, and Slava Novgorodov. 2021. An Image is Worth a Thousand Terms? Analysis of Visual E-Commerce Search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3404835.3462950>

## 1 INTRODUCTION

The growing popularity of search from mobile devices equipped with a camera and the advancement in computer vision techniques have given rise to a new form of search: search by image, also commonly referred to as *visual search*. Visual search enables users to input an image as a query and retrieve a ranked list of results

based on their relevance to the input image. Major Web search engines, such as Google and Bing, have introduced visual search functionally, which allows querying for information that is hard to articulate by text [10, 36]. Neural network techniques for image recognition support effective feature representation, classification, segmentation, and detection, and enable efficient retrieval over huge corpora [41, 60, 72]. As Web content becomes ever more visual [60], with the explosive growth in the number of online photos in social media and other websites [76, 79], allowing users to express their information needs through an image becomes imperative.

In recent years, visual search has been implemented and studied in a variety of domains, such as travel [55], news [23, 58], health-care [25, 33], education [59], and food [49, 82]. Notably, one of the most popular visual search domains is electronic commerce. Sometimes referred to as *visual shopping* [68], visual search in e-commerce allows customers to search for listed items or catalog products using an image instead of the keywords normally used in e-commerce search [46]. This type of search naturally reflects offline shopping processes, which are often driven by visual inspection, and brings a sense of visual discovery to the online world [8, 79].

Search by image has a number of potential advantages over traditional text-based search. First, it can be fast and intuitive, as simple as uploading or taking a picture and triggering a search. Second, it is agnostic to language, which becomes increasingly important as online shopping becomes global. In addition, it does not require from customers to be acquainted with the terminology used by the e-commerce site for the type of merchandise they are seeking [46]. Some e-commerce categories, such as Fashion, Home Decor, or Art, are fundamentally defined by visual characteristics that are sometimes difficult if not impossible to articulate by text [8, 60]. For instance, on Etsy, an online marketplace for handmade and vintage goods, Style is particularly important as buyers often seek items that match their eclectic tastes [37]. In Fashion, customers often seek a new look, outfit, or theme; visual search technology helps express these aesthetic aspects in a way text has never been able to capture [5].

In a recent survey by visual content company ViSenze, 62% of Millennials and Gen Z consumers indicated they wish for visual search over any other new technology [2]. Photo sharing service Pinterest reported that among its 350M monthly users, many have expressed a desire for visual shopping [60]. A study from The Intent Lab found that 85% of the young respondents put more importance on visual information than textual information [3].

In recent years, many Web and e-commerce sites have introduced visual search functionality into their commercial applications [10, 36, 38, 47, 74, 79]. E-commerce platform Alibaba reported that their

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
*SIGIR '21, July 11–15, 2021, Virtual Event, Canada*

© 2021 Association for Computing Machinery.  
ACM ISBN 978-1-4503-8037-9/21/07...\$15.00  
<https://doi.org/10.1145/3404835.3462950>

“search by image” application triggered high attention and wide recognition, and has experienced swift growth with an average of over 17 million daily active users in 2017 [79]. However, despite the growing popularity of visual search, to the best of our knowledge no study has performed an in-depth analysis of visual search usage. The majority of the literature on visual search in recent years has focused on describing the end-to-end system architecture [36, 38, 47, 50, 74, 79] and evaluating ranking models [35, 46, 54, 72, 76–78].

In this work, we perform a search log analysis of over 1.5 million image queries, issued to the mobile application of eBay, one of the most widespread e-commerce platforms, over a period of four weeks. We compare the image queries with a sample of text queries of similar size, performed on the same mobile application during the same time period. Our comparison encompasses characteristics of context, sessions, retrieved results, attributes (facets) used for query refinement, and clicks. We also analyze the image searches according to several unique characteristics of images, comparing searches with images captured from the device’s camera to searching with gallery images. In the final part of our work, we experiment with query performance prediction for visual search, revealing several novel pre- and post-retrieval predictors that demonstrate significant performance.

Our key contributions can be summarized as follows:

- To the best of our knowledge, we present the first comprehensive in-depth analysis of visual e-commerce search log.
- We combine analysis of queries, sessions, retrieved results, refining attributes, and clicks to shed more light on the common and different between image and text queries.
- We provide empirical evidence that image queries are more specific than text queries.
- We evaluate a set of query performance predictors for visual search and compare them with classic textual search predictors.

Our findings suggest different ways for e-commerce search systems to enhance their support and take advantage of the unique characteristics of image queries. We conclude the paper by summarizing the key findings and discussing their implications.

## 2 RELATED WORK

The task of visual search, or search via an image query, has been extensively studied by the Computer Vision and Multimedia communities. Techniques have evolved from feature-based and bag-of-words approaches [8, 21] to deep learning and semantic representation methods [46, 48, 50, 54]. With the growing popularity of mobile devices that made camera use ubiquitous, and the advancement in deep learning techniques for computer vision and particularly for visual search, more studies started to emerge introducing visual search systems. These studies focus on the end-to-end architecture and, in some cases, evaluation of the retrieval model, rather than on query log analysis and behavioral characteristics, as explored in our work. Hu et al. [36] provided an overview of the visual search system in Microsoft Bing. They described the methods used to address relevance (using a learning-to-rank approach with visual features), latency, and storage scalability and provided an evaluation of these three dimensions. Bhattacharya et al. [9] presented a multimodal dialog system to help online customers visually browse through large image catalogs, using both visual and textual queries. Web

search using images has also been referred to as “reverse image search”. Bitirim et al. [10] performed an evaluation of Google’s reverse image search performance, in terms of average precision at varying sizes of result sets.

Largely, the most popular domain of visual search research has been electronic commerce. In recent years, a variety of studies have been published describing the architectures of a “search by image” functionality introduced by multiple e-commerce platforms and evaluating different search algorithms to enable effective and efficient visual search. Zhang et al. [79] introduced the large-scale visual search algorithm and system infrastructure at Alibaba. They discussed challenges such as bridging the gap between real-shot images from user queries and stock images and dealing with large-scale indexing of dynamic data. In a followup work [78], the authors proposed learning image relationships based on co-click embedding, to guide category prediction and feature learning and improve visual search relevance. Li et al. [47] presented the design and implementation of a visual search system for real-time image retrieval on JD.com, one of China’s largest e-commerce sites. They demonstrated that their system can support real-time visual search with hundreds of millions of product images at sub-second timescales and handle frequent image updates through efficient indexing methods. Yang et al. [74] described the end-to-end approach for scalable visual search infrastructure at eBay, along with in-depth discussions of its basic components and optimizations, trading off search relevance and latency. To a large extent, our work takes advantage of the system described in that work to characterize visual search use on eBay and compare it with textual search.

Image sharing service Pinterest has been a source of a variety of studies describing its visual search and discovery system and some of the algorithms behind it. All applications were directed at online shopping, giving another indication of the relevance of visual search to e-commerce. The earliest work [38] described how Pinterest built a cost-effective large-scale visual search system and showed its positive effect on user engagement. Another work [77] described the image embedding process behind Pinterest’s visual search, using a multi-task learning architecture capable of jointly optimizing multiple similarity metrics. Additional studies focused on more specific use cases, such as selecting a detected object in an image as a visual query [60, 76] and recommending style-compatible complementary products for an outfit [40, 45].

Visual search should not be confused with the broad domain of image search, which refers to the results rather than the query: image search, i.e., search whose result set consists of images, is a popular search vertical and has been extensively studied (e.g., [20, 26]). Image search and visual search naturally integrate when both the query and returned results are images. This type of search is often referred to as content-based image retrieval [21, 72]. In our work, however, we explore visual search in another popular search vertical – shopping – with e-commerce listed items as returned results. To the best of our knowledge, no comprehensive log analysis of a visual e-commerce search engine has been reported.

## 3 RESEARCH SETTINGS

Our analysis is based on a random sample of 1,635,632 image queries from the eBay mobile search application, performed by over 250,000

unique users along a period of exactly four weeks (February 2nd-29th, 2020) in the United States. The eBay mobile search application allows searching with an image by clicking on a camera icon to the right of the textual search box. The user can then either instantly take a photo to be used as the query using the device’s camera or upload an image from the device’s photo gallery. After inputting an image, the eBay search engine retrieves a list of relevant results to the image query, presented to the user according to their relevance rank. The returned list of results can be traversed from top to bottom (and back) by scrolling. For comparison, we collected an identical number of queries performed using the “regular” textual search box of the same mobile application. We refer to the former set of queries as *image queries* and to the latter as *text queries*. The text queries were collected along the same period of four weeks for a similar number of users. Moreover, we sampled an identical number of image and text queries in each day of the experimental period. When inspecting day-of-week distribution and session statistics, we compared all queries from all users in our image sample with all queries from all users in our text sample, during four weeks of the experimental period, to allow suitable analysis.

Each query in the log, either image or text, included, in addition to the query itself, a timestamp (adapted to the timezone in which it was performed) and the list of retrieved results presented to the user on the search engine results page (SERP). Each returned result is a listed offer, or *listing* in short, by a specific seller. Our data included, for each result, its rank on the SERP (the top result is at rank 1) and a unique listing URL. In addition, for each query we had information about its associated clicks and purchases, if any were performed, including their ranks and corresponding listing URLs. After a query (image or text) is submitted and the results are presented, the user can refine the result list using attributes, such as color, brand, or size. Our log included the attributes used for refinement and their values or value ranges (e.g., color ‘blue’ or size over 40 inches).

eBay spans a variety of shopping domains. Each listing on eBay is associated with a *leaf category (LC)*, which is the most specific type of node in the eBay’s taxonomy. The taxonomy includes tens of thousands of LCs, such as Electric Table Lamps, Developmental Baby Toys, or Golf Clubs. Each listing is also associated with one out of 43 *meta-categories (MCs)*, such as Home & Garden, Toys & Hobbies, or Collectibles. For each result on the SERP, we had information about the LC and MC it belonged to.

Our analysis is organized as follows. Section 4 compares basic characteristics of image and text searches, including searcher’s demographics, context, and session characteristics. Section 5 examines the image query characteristics, including source (captured by camera or uploaded from gallery), orientation (vertical or horizontal), brightness, and catalog quality. Section 6 looks into the characteristics of retrieved results, including their category distribution and image quality. Section 7 examines the attributes used to refine image queries in comparison with text queries, while Section 8 inspects click characteristics, such as click-through rate and mean reciprocal rank. Finally, in Section 9, we describe our experimentation with a set of new pre- and post- retrieval performance predictors for visual search.

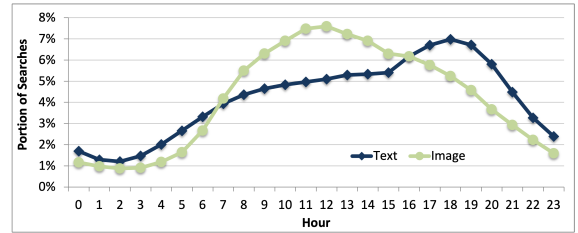


Figure 1: Query distribution by hour of the day.

## 4 BASIC CHARACTERISTICS

**Context and Demographics.** We found similar demographic characteristics for image and text queries in terms of searcher’s age and location (city and state). For gender, we observed a substantially higher portion of female searchers for visual search (ratio of queries performed by a female versus male up by a factor of 2.56 compared to textual search). This trend persisted across all MCs, such as Collectibles (ratio: 3.13), Home & Garden (1.97), and Fashion (ratio 1.61), and further intensified when inspecting only image queries performed using a gallery photo (ratio 3.67).

The distribution across day-of-week was similar for image and text queries: in both, there was a slight peak on weekends compared to weekdays. In contrast, there was a noticeable difference between image and text queries with regards to time-of-day, as depicted in Figure 1. Image queries were more frequent during day hours (from 6am to 4pm), with a peak at 12pm, while higher portions of the text queries (relative to image) were performed during late afternoon, evening, and night, peaking at 6pm.

In our analysis, we inspected the results while controlling for factors that were found to be different between image and text queries, including time-of-day and gender. When relevant, we report the influence of these factors on the results.

**Sessions.** The query logs (both image and text) are partitioned into sessions, based on a commonly used definition: a sequence of queries by the same user, without an idle time longer than 30 minutes between each pair of consecutive queries in the sequence [34, 39]. We refer to an image session as any session that contains at least one image query. All the other sessions are considered as text sessions. Table 1 presents session statistics. It can be seen that image sessions tend to be longer, with a substantially lower portion of 1-query sessions. As a result, their average and median duration is also substantially longer. Yet, even when controlling for the number of queries (2, 3, and 5 are presented in the table), the duration of image sessions is longer than text sessions. Inspecting idle time between queries in a session, it is also longer for image sessions, even when inspecting specific transitions, such as from the first query to the second, or from the second to third.

## 5 QUERIES

As mentioned in Section 3, visual search can be used by two *flows*: using the device’s camera to instantly take a photo and using the camera roll, or photo gallery, to upload one. We refer to these two flows as the *camera flow* and *gallery flow*, respectively. In Figure 2, examples 1,5,6,7,8 demonstrate image queries using the camera flow, while 2,3,4,9,10 demonstrate image queries using the gallery flow. In our sample, **80.07%** of the image queries were performed using the

**Table 1: Session characteristics.**

	Text	Image
Avg (std) number of queries	2.99 (4.22)	7.83 (12.9)
Median number of queries	2	4
% 1-query sessions	44.00%	21.63%
Avg (std) duration in minutes	18.47 (29.81)	38.81 (57.87)
Median duration in minutes	7.65	18.13
Median duration length=2, 3, 5	5.85, 9.55, 17.23	6.43, 10.55, 20.15
Avg (std) idle in minutes	3.85 (7.20)	4.31 (6.49)
Median idle in minutes	1.48	2.33
Median idle 1st-2nd, 2nd-3rd in minutes	0.87, 0.93	1.03, 1.02

camera flow and **19.93%** using the gallery flow. These portions vary substantially across categories: MCs with high portion of camera queries (over 85%) include media (Books, Music, Video Games), Collectibles, Antiques, and Art. On the other hand, MCs with high portion of gallery queries (over 30%) include Fashion (with nearly half of the queries), Jewelry & Watches, Cellphones & Accessories, and Health & Beauty. In the next section, we explain in more detail how we associate a query with an MC.

In the remainder of this section, we examine three image characteristics – orientation, brightness, and catalog quality – and compare them between the two flows.

**Orientation.** The aspect ratio of an image is the ratio of its width to its height. When it is higher than 1 the image has a horizontal orientation and when it is lower than 1 the image is vertical. The portion of vertical images was substantially higher on camera queries at 92.72% compared to gallery queries at 72.04%. This may stem from the fact that users shoot their camera queries while holding the phone in the more natural and common vertical orientation and do not bother to change to horizontal for querying. It should be noted that the portion of vertical images is relatively high even in gallery photos. Recent datasets of mobile photos include a much lower percentage of vertical photos, e.g., 44.5% [67] and 40.3% [29].

**Brightness.** Figure 3 depicts the brightness [7] histogram (by buckets) of camera and gallery queries. For reference, the figure also plots the brightness distribution of two publicly-available image datasets: a Flickr dataset [75] that contains 30k pictures taken by Flickr users and a Fashion dataset [1] that contains 44K images of professional stock photos of fashion products. It can be seen that gallery queries are generally brighter than camera queries. The camera query brightness histogram is almost identical to the Flickr dataset, with user-generated photos. The gallery histogram, on the other hand, spans almost the entire range and overlaps with both the Flickr and Fashion datasets. This implies that gallery queries include both user-generated photos and professional studio photos. In Figure 2, examples 3 and 10 demonstrate uploaded user-generated screenshots, while 2 and 4 are uploaded stock photos.

**Image quality.** Online marketplaces often use models for image quality assessment in order to select the best images for their product catalogs [17]. We used an in-house tool that assigns a quality score to an image, based on a supervised model trained over a large collection of images uploaded by sellers as part of their listing process. Quality scoring considers factors such as size, cropping, angle view, blur, background, frame, watermarks, inclusion of human body parts, and additional elements besides the main product for sale [71]. As expected, gallery queries had a substantially higher quality than camera queries, with an average score of 0.90 (std: 0.21, median: 0.97) versus 0.81 (std: 0.28, median: 0.94), respectively.

**Table 2: Distribution of the number of MCs and LCs among the top 40 retrieved results for image vs. text queries.**

# of categories	Meta Categories (MCs)		Leaf Categories (LCs)	
	Text	Image	Text	Image
1	70.30%	94.75%	40.6%	87.96%
2	15.19%	5.06%	18.71%	10.88%
3	5.81%	0.17%	11.13%	1.00%
4	3.08%	0.01%	7.33%	0.12%
5	1.94%	0.01%	5.00%	0.02%
6+	3.68%	0.00%	17.17%	0.02%

## 6 SERP

In this section, we inspect various characteristics of the retrieved results for image queries, presented to users in the search engine results page (SERP), in comparison with text queries. Our analysis in this section and those that follow excludes null queries, i.e., queries for which no results were returned [62]. The portion of null queries in our data was slightly higher for image queries than for text queries: 1.31% versus 0.80%, respectively. We observed that image null queries had lower brightness [7] (−11%) and aesthetic score [65] (−17%) compared to all other image queries. Figure 2 example 1 demonstrates a null image query.

Overall, the number of retrieved results was lower for visual search than for textual search, with a ratio of 0.35 between the two averages (std ratio: 0.28, median ratio: 0.57). We refer to the *last result viewed* (LRV) by the user as the result with the lowest rank that the user scrolled down to. The average LRV for image queries was 61.15 (std: 69.43, median: 40), comparable to text queries at 59.15 (std: 68.25, median: 46), indicating users traverse a similar number of results in search types. For our SERP analysis, unless otherwise stated, we considered the top 40 results, as this was the median number of results traversed by a user in visual search.<sup>1</sup>

**Number of categories.** A prominent characteristic of the SERP is the distribution of results across e-commerce categories. The average number of MCs on the SERP was 1.05 (std: 0.24) for image queries, compared to 1.67 (std: 1.47) for text queries. The average number of LCs on the SERP was 1.14 (std: 0.41) for image queries, compared to 3.46 (std: 3.94) for text. Table 2 shows a detailed distribution of the number of MCs and LCs on the SERP. It can be seen that while over 17% of the text SERPs span six LCs or more, virtually no image SERPs (0.02%) do. Overall, we observe that the SERP for image queries is considerably more focused on specific categories. These characteristics were similar for both camera and gallery queries. Figure 4 (left plot) demonstrates that for text queries, the number of MCs and LCs on the SERP decreases as the query length increases, however even for very long queries (10 terms or more), it is higher than for image queries.

**Category distribution.** For our next analysis, we assign each query to one MC and one LC according to its SERP. We define the *dominant category* (MC or LC) as the most common category among the top 40 results. In case more than one category is the most common on the SERP, we considered the one with the higher ranked top result as the dominant category. The use of such a tie breaker was infrequent: only 0.75% (2.41%) of the text queries and 0.001% (0.03%) of the image queries for MCs (LCs). During our experimental period, visual search was used across all categories at

<sup>1</sup>Throughout our analysis, we also calculated the statistics for the top 10 results and observed very similar trends.



Figure 2: Examples of image queries.

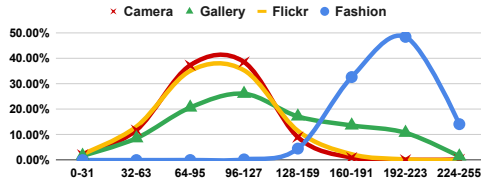


Figure 3: Brightness of camera vs. gallery query images compared with two public datasets. Lower values represent darker images.

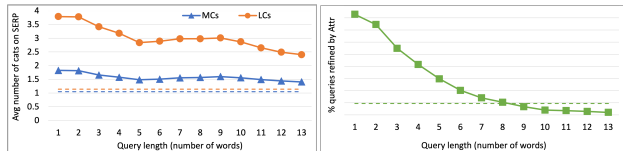


Figure 4: Analysis of text queries by length: average number of MCs and LCs on the SERP (left plot) and percentage of queries refined by attributes (right plot). The dashed lines indicate the respective values for all image queries.

eBay, spanning all 43 MCs and thousands of LCs. Yet, the distribution across categories was different for image queries than for text queries. To understand the most common distinctive categories in image queries relative to text queries, we used Kullback-Leibler (KL) divergence, which is a non-symmetric distance measure between two given distributions [6, 15]. Specifically, we calculated the categories (MCs and LCs, respectively) that contribute the most to the KL divergence between the distribution of image queries across MCs (LCs) and the distribution of text queries across MCs (LCs), which was 0.39 (1.11) in total.

Table 3 presents the 10 most distinctive MCs for image queries compared to text queries. For each such MC, the 3 most distinctive LCs that belong to it are presented, to demonstrate a finer-grained granularity of categories that are especially popular for visual search. The list of distinctive MCs is topped by Collectibles, such as mugs, lamps, and plates, with more related MCs further down the list, such as dolls, coins, and stamps. Pottery & Glass is the second most distinctive MC, especially glassware, as can be observed in the list of top-related LCs. Other MCs that relate to art can be observed further down the table and include crafts and jewelry. Antiques, the third most distinctive image MC combines characteristics of both collectibles and art. The fifth most distinctive MC is Toys & Hobbies, with character and action figures among the top LCs, and some vintage games such as traditional board games and puzzles further down the list (not presented in the table). The Baby category is also high on the MC list, with LCs related to strollers, swings, and monitors. Overall, we observe that while

Table 3: Most distinctive MCs and LCs in image queries relative to text queries according to KL divergence.

Meta Category (MC)	Leaf Category (LC)
Collectibles	Mugs & Cups
	Table Lamps
	Collector Plates
Pottery & Glass	Fenton Art Glass
	Crystal
	Pyrex
Antiques	Chinese Figurines & Statues
	Ceramic & Porcelain Vases
	Porcelain Plates & Chargers
Dolls & Bears	Cloth Dolls
	Dollhouse Miniatures
	Cultures & Ethnicities Dolls
Toys & Hobbies	TV & Movie Character Toys
	TV, Movie & Video Game Action Figures
	Contemporary Manufacture Board & Traditional Games
Coins & Paper Money	US Coin Errors
	Medals
	Nepali Paper Money
Stamps	US Collection and Lots
	US Postage
	US Unused 1941-Now
Baby	Stroller Parts
	Baby Swings
	Baby Monitors
Crafts	Ready-to-Paint Pottery
	Wood Items
	Acrylic Paint
Jewelry & Watches	Retro & Vintage Costume Necklaces & Pendants
	Retro & Vintage Costume Pins & Brooches
	Fashion Bracelets

visual search was used across the board, it was especially popular for collectible and vintage products, art, and toys & babies. The Fashion category, which has been the subject of many previous studies on visual search [5, 8, 40, 41, 43, 48, 60], occurred in nearly 10% of the visual searches, but was not more popular than in textual searches.

**Image quality.** The average image quality on the SERP showed no significant difference between image and text queries ( $p > .05$ , two-tailed unpaired t-test) at 0.866 (std: 0.265, median: 0.985) compared to 0.796 (std: 0.333, median: 0.979), respectively. There was also no significant difference from image queries in the gallery flow (avg: 0.825, std: 0.304, median: 0.982), even though as shown in Section 5, there was a significant difference in terms of the image query quality. Overall, we see that the query’s modality does not have a significant impact on the image quality of the retrieved listings.

## 7 QUERY REFINEMENT BY ATTRIBUTES

Due to the structured nature of search results in e-commerce, refinement using a variety of attributes is a common feature of e-commerce search [34, 63, 70]. Upon issuing a query, the user is displayed with different attributes, typically defined based on the category(ies) of the returned results, and can narrow down the

list of retrieved results based on specific attribute values, such as a color, size, or brand. Despite the fact that such refinements are actually used in a rather small portion of the queries, they allow to gain understanding about specific information needs for visual versus textual search based on user interaction.

Generally, the use of attributes to refine the search results was considerably less frequent on image search compared to text search, with an image-to-text ratio of 0.17. This sharp difference gives another indication that image queries often reflect narrower information needs with richer sets of attributes than text queries. We further explore this by analyzing refinement use in text queries according to their number of terms. Figure 4 (right plot) shows a clear trend: the use of refinement by attribute decreases as the length of the text query increases. The portion of refinements in image queries is slightly lower than the portion of refinement for 8-word queries (which account for 0.58% of all text queries), suggesting that according to this signal, an image query is “worth” at least eight terms. This rough extrapolation likely reflects a lower bound, since as shown in previous work, users also refine their text queries by adding terms to the query itself [34], an option that does not currently exist in visual search. It should also be noted that users rarely input queries of more than 8 terms and these account for only 1.21% of all text queries.

Table 4 shows the relative distribution of attributes used for refining image and text queries (and the image-to-text ratio). For image queries, the list is topped by brand and color, with material and style also having particularly high ratio. While brand and material are indeed challenging to detect based on image, and style has been previously studied as a particularly popular attribute for visual search [40, 41, 52], the relative popularity of the color attribute is rather surprising and we therefore further explored it, as will be detailed later in this section. In Figure 2, example 3 was refined by a color (green) and example 4 was refined by both a material (leather) and size (women’s 8). Further down the list of common image attributes are style-related heel height, sleeve length, and dress length. With a particularly high ratio are pattern (also studied in various papers [8, 43, 60, 68, 74]), team (relevant to sports merchandise), and, most extremely, original/reproduction, which again indicates the prevalence of vintage and collectible items on visual search. The text list is topped by the size attribute, which has a double relative frequency compared to image queries. The list includes many other size-related attributes, such as shoe size and size type, as well as technology-related attributes such as network, storage capacity, screen size, and operating system.

Table 5 shows the most common values used for refinement in two of the most common attributes: color and material. The colors that are more popular on image search are blue, green, white, and clear (x7 more popular than for text). We conjecture that users need to distinguish colors that are hard to detect on image, prominently clear items, but also white (from other bright colors) and (dark) green and blue (from black). In Figure 2 example 5, the user explicitly used ‘blue’ as a color refinement. For material, the lists of image and text values are more disparate: the text list is dominated by types of fabric, while the image list is more diverse with a variety of materials, including ‘fabric’ itself.

We finally inspect image query refinement by flow. Generally, the use of refinements was more frequent on gallery image queries

**Table 4: Most clicked attributes for refining text and image queries. The rightmost column shows the ratio between the percentage of the attribute use out of all attributes used in image queries and the same percentage in text queries.**

	Text		Image	
	Attribute		Attribute	Ratio
1	Size		Brand	1.60
2	US shoe size (men’s)		Color	1.75
3	Brand		Size	0.50
4	Color		Material	2.88
5	US shoe size (women’s)		US shoe size (women’s)	0.95
6	Size type		Size type	1.07
7	Type		Style	2.07
8	Network		Type	1.43
9	Style		Sleeve length	1.49
10	Material		Heel height	2.66
11	Storage capacity		Dress length	2.33
12	Sleeve length		US shoe size (men’s)	0.17
13	Model		Original/reproduction	24.13
14	Screen size		Pattern	4.05
15	Operating system		Team	7.41

**Table 5: Most common color and material values used for refinement of text and image queries, with respective image-to-text relative usage ratios.**

Text	Color		Material		
	Image		Text	Image	
	Value	Value	Ratio	Value	Value
Black	Black	0.82	Leather	Wood	2.46
Blue	Blue	1.19	Cotton	Glass	5.77
White	White	1.21	Silk	Cotton	0.60
Gray	Green	1.39	Polyester	Leather	0.32
Brown	Red	1.01	Linen	Ceramic	5.91
Red	Pink	1.09	Wool	Fabric	5.30
Beige	Clear	7.00	Nylon	Mirror	17.63

compared to camera, by a factor of 3.51, but still not as frequent as in text queries. The distribution of attributes was similar between the flows, with one noticeable difference: the size attribute was used substantially more frequently in the gallery compared to the camera flow, by a factor of 2.52. We conjecture that size is easier to capture when taking a photo by the device’s camera and using one’s hand or another object of known size for reference (e.g., see Figure 2, examples 5 and 6).

## 8 CLICKS

Table 6 shows the ratio for various click characteristics between visual and textual search.<sup>2</sup> These include the click-through rate (CTR; the portion of queries for which at least one result was clicked), and, for clicked queries only, the average number of clicks (AVC) and mean reciprocal rank (MRR). At the session level, it can be seen that the CTR is only slightly lower for image sessions than text sessions and the AVC is almost identical. Yet, as shown in Table 1, the length of image sessions is substantially higher than text sessions. Indeed, moving to the query level, the CTR ratio between image and text queries is as low as 0.485. The AVC ratio is also below 1, indicating that even for clicked queries, fewer results are clicked when the query is an image. Lower CTR and AVC were also reported for voice queries [27], which represent another newly-introduced beyond-text query modality. Despite the fewer clicks, the MRR was higher for image queries than for text queries, indicating that clicks are more frequently performed on top results. For voice queries,

<sup>2</sup>We cannot disclose actual values due to business sensitivity.



**Table 6: Click-through rate (CTR), average number of clicks (AVC), and mean reciprocal rank (MRR) ratios between image and text queries.**

	CTR	Clicked queries	
		AVC	MRR
Sessions	0.897	0.979	-
Queries	0.481	0.739	1.113
Gallery flow queries	0.674	0.907	1.008
Home & Garden queries	0.427	0.847	1.163
Collectibles queries	0.497	0.668	1.151
Toys & Hobbies queries	0.521	0.677	1.211
Jewelry & Watches queries	0.552	0.724	0.921
Fashion queries	0.595	0.786	0.957
Antiques queries	0.642	0.685	1.095
Pottery & Glass queries	0.810	0.714	1.261

the MRR was reported to be similar to that of text queries, at a ratio of 0.97 [27]. Overall, the lower CTR and AVC and high MRR imply that visual search is often used for target finding [64]. These results also suggest there is more room for improvement in the ranking algorithms and user experience for visual search, as it is still in its infancy.

In previous sections, we observed that gallery queries demonstrated more similar characteristics to text queries than the rest of the image queries. This was also reflected in click characteristics, as shown in Table 6: the CTR and AVC ratios were higher, while the MRR was almost identical to text queries.














In general, the CTR was quite diverse across different MCs (considering the dominant categories, as defined in Section 6): standard deviation was 36.5% and 27.6% of the mean CTR, for text and image queries, respectively. The lower section of Table 6 presents the click ratio characteristics for seven of the most common image MCs. Most of the CTR ratios for the specific MCs were higher than the general CTR ratio: this is because visual search is relatively more popular on categories with lower CTR, such as Collectibles, than categories with higher CTR, such as Fashion. The ratio varied rather substantially across MCs: it was a low 0.427 for Home & Garden, while reaching as high as 0.81 for Pottery & Glass. The MRR also varied to some extent across MCs, with Fashion and Jewelry & Watches having an image-to-text ratio lower than 1, and Pottery & Glass having the highest ratio at over 1.25 (Figure 2 example 6 shows a query whose dominant category is Pottery & Glass).

Thus far, we have seen many quantitative characteristics by which image queries differ from text queries. Next, we show a few examples of image and text queries which are likely to reflect the same shopping intent. To this end, we inspected image and text queries that led to the purchase of the same listing during our experimental period of four weeks. Table 7 shows seven examples, including the image and text queries, and the title and image of the purchased listing. In some examples (2,4) the purchased item is prominently different than the image query, suggesting a decision making and exploration intent rather than target finding [64].

## 9 QUERY PERFORMANCE PREDICTORS

The task of query performance prediction (QPP) [16, 18] aims at estimating the query difficulty as reflected by its retrieval effectiveness, in the absence of relevance judgments or user interaction signals. Two main types of QPPs have been studied in the literature: pre-retrieval QPPs, which estimate the query’s quality based on

**Table 7: Example image and text queries that led to a purchase of the same item. Examples 1,3,7 include camera queries and the rest include gallery queries.**

#	Image Query	Text Query	Purchased Listing Image and Title
1		carotone cream	 DSP10 Black Spot Corrector Creme 1oz
2		party rings t	 Women Elegant 925 Silver Sapphire Amethyst Rings Wedding Engagement Jewelry Gift
3		jadoo tv remote wireless	 Universal Wireless Air Mouse Keyboard Remote Control For Mini PC Android TV Box
4		bad boys chevelle	 GREENLIGHT HOLLYWOOD SERIES 21 BAD BOYS 1968 CHEVROLET CHEVELLE SS
5		lucky step rainbow shoes 7.5	 Womens shoes Rainbow Lace Up Sneakers Gym Sports Running Trainers Casual Shoe
6		rocker baby	 Fisher-Price Infant-to-Toddler Rocker - Pacific Pebble, Portable Baby Seat, Multi
7		klein ncvt-2	 Klein Tools NCVT-2P Dual-Range Non-Contact Voltage Tester - Brand New!!!

the query itself and the corpus statistics [31]; and post-retrieval QPPs, which assess the query performance by considering the retrieved result list [42]. While query performance prediction has been studied in depth for traditional textual search, it has not been extensively studied for visual search. The only study we are aware of is a short paper that proposed two pre-retrieval visual QPPs [44]. In this section, we experiment with several pre- and post-retrieval QPPs for visual search. We evaluate their performance and compare them with classic QPPs applied to text queries.

For text queries, we follow the list of QPPs described in a recent paper studying e-commerce textual search [34]. For pre-retrieval, these include the query length (in words) [16]; the minimum, maximum, and sum of the IDF values of the query terms [56]; and the minimum, maximum, and sum of the variance of TF.IDF values of the query terms across documents in the corpus [80]. These predictors have shown to be effective for document search in large-scale studies [30, 61]. For visual search, we harness classic visual characteristics, taking advantage of the fact that the query is an image, to define the following pre-retrieval QPPs: the image size (in pixels); the image brightness, as described in Section 5; the portion of pixels with its 3 most dominant colors (where colors are defined by clustering in the RGB space using k-means with  $k=8$  [13]).<sup>3</sup>; and the image

<sup>3</sup>We experimented with other numbers of dominant colors: 1,2,4, and 5, which all showed similar trends.

quality, measured both using our own model for catalog quality estimation described in Section 5 and using a publicly-available implementation of a neural model for general image aesthetic quality assessment [65]. For a corpus-based predictor, we considered the minimum, maximum, and average similarity between the image query and a large collection of one million images sampled uniformly at random from the entire eBay inventory. In addition, we examined the two pre-retrieval QPPs previously proposed for visual search [44]. Both are based on concept extraction from images using Latent Dirichlet Allocation (LDA) [11] over visual words, with visual words extracted using the SIFT algorithm for feature detection [51]: *q-INS* measures the query’s information need specificity and *c-DSC* measures the discriminability of concepts across the corpus [44].

For post-retrieval QPPs, we first define the *similarity* between a query and its retrieved result (e-commerce listing) as the cosine similarity between the latent vector representations (size 300) of the query and the listing. For textual search, we used Word2Vec [53] trained over a corpus of 10M titles sampled uniformly at random from the eBay inventory. For the query, we used the TF.IDF weighted average of the query term vectors, while for the listing, we used the TF.IDF-weighted average of the listing’s title word vectors [4]. For visual search, we used a ResNet-50 network [32] to learn image embeddings over more than 50M listing images from the eBay site, spanning all major categories [74]. We applied these embeddings to both the image query and the listing’s main image.

We examine the following post-retrieval QPPs [34]: (1) *Num results* - the total number of retrieved results [16]; (2) *STD*: the thresholded standard deviation [19], with a 50% threshold; (3) *WIG*: the weighted information gain [81] without corpus-based normalization; and (4) *SMV*: the score magnitude and variance (SMV) [66], which can be viewed as integrating STD and WIG, with the average retrieval score in the corpus as a normalizer. The last three predictors were computed for the embedding-based similarity described above.<sup>4</sup> For textual search, these predictors were shown to be highly effective for document retrieval in various studies [16, 57, 61]. For visual search, to the best of our knowledge, we are the first to experiment with post-retrieval QPP.

For evaluating the QPPs, we sampled uniformly at random 1000 text queries and 1000 image queries from our logs described in Section 3. We asked three in-house annotators who specialize in relevance judgements for e-commerce to provide these for the top 10 results for each query (binary label per result: relevant or not relevant). The Fleiss Kappa [24] among the three annotators was 0.91 and 0.82 for text and image queries, respectively. We then used the human annotations to calculate the average precision (AP) at  $k=10$  for each query. We evaluate the image and text QPPs using two metrics: Pearson correlation coefficient ( $r$ ) between the predictor values and the  $AP@10$  values and Kendall rank correlation ( $K\tau$ ) between the ranks induced by the predictor values and  $AP@10$  over the queries. Both of these metrics are commonly used for measuring the effectiveness of QPPs [16, 42, 80].

Table 8 presents the performance results for pre-retrieval QPPs. For text queries, all predictors (as described in [34]) did not show

<sup>4</sup>For textual search, we also experimented with Okapi-BM25 [22] scores computed for listing titles w.r.t the query text, which yielded very similar performance to the embedding-based similarity approach reported in detail.

**Table 8: Pre-retrieval QPP performance results w.r.t relevance judgements ( $AP@10$ ). Boldfaced QPPs have statistically significant Pearson’s  $r$  and Kendall’s  $\tau$  for  $p < .01$ .**

QPP	Text		Image		
	$r(p)$	$K-\tau(p)$	QPP	$r(p)$	$K-\tau(p)$
Length	.084 (.422)	-.015 (.865)	Size	.052 (.464)	.002 (.972)
Min IDF	.118 (.257)	.082 (.323)	Brightness	.062 (.383)	.074 (.195)
Max IDF	.095 (.361)	.023 (.783)	3-Color	.029 (.687)	.023 (.685)
Avg IDF	.103 (.322)	.054 (.509)	<b>Catalog Quality</b>	<b>.233 (.001)</b>	<b>.175 (.002)</b>
			Aesthetics	-.082 (.252)	-.081 (.155)
Min TF.IDF Var	.112 (.284)	.139 (.093)	<b>Min Corpus Sim</b>	<b>-.217 (.002)</b>	<b>-.163 (.004)</b>
Max TF.IDF Var	.148 (.155)	.014 (.862)	<b>Max Corpus Sim</b>	<b>.273 (&lt;.001)</b>	<b>.198 (&lt;.001)</b>
Avg TF.IDF Var	.170 (.102)	.048 (.557)	Avg Corpus Sim	.074 (.301)	.038 (.509)
			q-INS	.123 (.194)	.114 (.105)
			c-DSC	.166 (.104)	.121 (.088)

**Table 9: Post-retrieval QPP performance results w.r.t relevance judgements ( $AP@10$ ). Boldfaced correlations (Pearson’s  $r$  and Kendall’s  $\tau$ ) are statistically significant for  $p < .01$ .**

QPP	Text		Image	
	$r(p)$	$K-\tau(p)$	$r(p)$	$K-\tau(p)$
Num Results	-.097 (.256)	.095 (.003)	-.076 (.302)	.014 (.809)
STD	-.210 (.013)	<b>-.286 (&lt;.001)</b>	.038 (.605)	-.064 (.273)
WIG	.206 (.015)	.144 (.027)	<b>.562 (&lt;.001)</b>	<b>.449 (&lt;.001)</b>
SMV	<b>-.236 (.005)</b>	<b>-.292 (&lt;.001)</b>	-.156 (.0333)	<b>-.188 (.001)</b>

statistically significant correlation. For image, one query-based predictor showing statistically significant performance was the catalog quality score. While calculated using an internal model, this demonstrates that relying only on image characteristics, a significant performance prediction can be achieved. The external aesthetic quality score [65], however, did not demonstrate a similar performance. In Figure 2, examples 7 and 8 show image queries with a high catalog quality, but low aesthetic quality score, while example 9 shows an image query with a high aesthetic quality and low catalog quality score. The corpus-based predictors based on both the minimum and maximum similarity to the inventory’s collection of images yielded significant performance, while the average did not. The two previously-proposed LDA-based QPPs [44] yielded clear correlations, albeit not statistically significant, indicating they are not as effective for visual e-commerce search as reported for general visual search. An explanation to this may lie in the unique characteristics of e-commerce image queries, as described in Section 5 and demonstrated in Figure 2, which are focused on objects for purchase rather than scenes. Overall, we identified both query-based and corpus-based pre-retrieval QPPs that demonstrate high performance for image queries.

Table 9 shows the evaluation results for post-retrieval QPPs. For text queries, the STD and especially the SMV QPPs showed statically significant prediction performance, aligned with past work showing post-retrieval predictors are more powerful than pre-retrieval predictors [16, 34]. For image queries, STD results were insignificant, while SMV was only significant by the  $K-\tau$  metric. The WIG predictor, on the other hand, demonstrated significant results and yielded, by a large margin, the best prediction by both the Pearson’s  $r$  and Kendall’s  $\tau$  metrics out of all pre- and post-retrieval QPPs. Finally, we note that for all visual QPPs, the correlations showed very similar trends when inspecting camera and gallery queries separately.



## 10 DISCUSSION AND IMPLICATIONS

Our study disclosed various differences between visual and textual search. In this section, we summarize the key findings, discuss implications, and suggest directions for future work.

**Query Categories.** Much of the existing literature on visual e-commerce search focuses on the Fashion category [5, 8, 40, 41, 43, 48, 60]. Our analysis, however, shows that visual search is widespread across many e-commerce categories, and is especially popular in comparison with textual search for collectibles, vintage, art, toys, and baby products. These categories often share information need aspects that are harder to verbally express, but can be captured visually, such as style, type, and pattern. The substantial differences between image and text in query categories and their characteristics, as exhibited throughout our study, suggest that search tools that build on query classification, such as pre-retrieval category identification, sponsored or promoted results, query expansion, and even result ranking, may need to be adapted when used for visual search due to the different span of categories.

**Search broadness.** Previous work [8, 79] noted that visual search provides a superior entry to text for fine-grained item description, but provided no empirical evidence. Our analysis shows that image queries are indeed more specific than text queries. This is reflected in a lower number of retrieved results, narrower span of categories on the SERP, and a substantially sparser use of refinement by attributes. While the use of refinement by attributes decreases for text queries as they become longer, it only compares to the level of image queries for highly verbose queries (over 8 tokens), which are very rare. Using an image as a query allows users to convey more information about the desired item than with a textual query [43, 73, 79] and, as our analysis shows, influences the retrieval process and user interaction with the retrieved results. Image queries remove challenges related to name-entity disambiguation (e.g., is ‘orange’ a color or a brand?), but at the same time add new challenges, such as differentiating dark colors from black or distinguishing types of material. With the rapid development of e-commerce and explosive growth of online shopping markets, efficiently guiding users through a huge inventory has become essential [35]. For visual search, the choice of attributes presented for users to refine their query should be different than for textual search, and focus on aspects that are hard to articulate by image. In addition, search interfaces should evolve to provide support for easy and natural combination of image and text, such as expanding a visual search with keywords [43] or using an image to refine a textual search.

**User intent.** Image queries are used for two principal intents [64, 68]: *target finding* desires to look up a specific item [43], while *decision making* aims at discovery of visually-similar items [76]. The two use cases are different in nature and to some extent resemble the navigational versus informational intent classification suggested for Web search in its early days [12]. Our analysis and examples demonstrate the use of both types of intent, but suggest no obvious way to distinguish between them at retrieval time. Visual search interfaces may therefore consider to provide an explicit means for users to indicate if they are looking for an “identical item” or “similar look” when they input an image query, so the intent can be better captured and served.

**Query performance.** The click-through rate and average number of clicks are substantially lower for image queries than for text queries. While this is not uncommon for a new query modal [27, 28], it also implies there is more room for improvement in serving image queries as visual search is still in early stages. This is also reflected by the higher MRR and longer sessions that involve image search. These findings imply that users undergo a more disparate and less coherent experience when they search by an image, leaving room for improvement in ranking methods, retrieval models, and result presentation. Our experimentation with query performance prediction indicates it is applicable for visual search. We identified both pre- and post-retrieval QPPs that demonstrate high prediction performance, even in comparison with traditional QPPs used for textual search. Further research is required to enrich and expand the list of visual QPPs, explore their combinations, and apply them to improve the search experience at large.

**Camera vs. gallery queries.** Our analysis revealed a variety of fundamental differences between visual search performed using a photo captured at query time by the device’s camera and an image uploaded from the device’s gallery. Gallery image queries are rarer (20% of all queries), brighter and of higher catalog quality, and horizontal at higher portions. Camera queries are more common on vintage, collectibles, and media, whereas gallery queries are more frequent on fashion, jewelry, and watches. These differences are reflected in user interaction with the SERP: gallery images demonstrate higher click-through rate and more frequent use of refining attributes. These findings suggest that visual search engines may benefit from serving differently the two types of image queries. For example, the selection of similarity metrics, categorization model, and refining attributes can be adapted accordingly. Despite the different quality of camera and gallery queries, the images of retrieved listings were found to be of similar quality in both cases. Camera and gallery queries also share similar characteristics in terms of the number of categories on the SERP and performance prediction.

Additional future directions of visual e-commerce search research are abundant (and necessary). For example, query reformulation has been studied in textual e-commerce search [34], and can serve to track the evolution of image queries along a session and identify difficulties and gaps. Additional methods to reformulate an image query using visual means, such as by editing the image query or using multiple images as an input can help make reformulation more applicable in visual search. We inspected camera and gallery image queries as the two principal flows to prompt a visual search. Future research should explore the triggering of visual e-commerce search from an external context, e.g., by clicking an image in a news article or a social media feed [60]. This type of use case can play a central role as an entry gate to e-commerce; understanding how to make the context transition productive and engaging can yield substantial benefits. Finally, our analysis gave rise to some common characteristics between visual and voice search. The connection between the two should be further studied as both become more widespread [14, 69]. The integration of visual and voice search can also be explored as a means to provide a more complete experience of e-commerce search that does not require typing.

## REFERENCES

- [1] 2018. Fashion Product Images Dataset. <https://www.kaggle.com/paramaggarwal/fashion-product-images-dataset/>, last accessed on 05/13/21.
- [2] 2019. How Visual Search has transformed the modern shopping experience. <https://www.visenze.com/blog/how-visual-search-has-transformed-the-modern-shopping-experience/>, last accessed on 05/13/21.
- [3] 2019. Visual Search Wins Over Text as Consumers' Most Trusted Information Source. <https://www.businesswire.com/news/home/20190204005613/en/Visual-Search-Wins-Text-Consumers%E2%80%99-Trusted-Information>, last accessed on 05/13/21.
- [4] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proc. of ICLR*.
- [5] Sean Bell, Yiqun Liu, Sami Alsheikh, Yina Tang, Edward Pizzi, M. Henning, Karun Singh, Omkar Parkhi, and Fedor Borisjuk. 2020. GrokNet: Unified Computer Vision Model Trunk and Embeddings For Commerce. In *Proc. of KDD*. 2608–2616.
- [6] Adam Berger and John Lafferty. 2017. Information Retrieval as Statistical Translation. *SIGIR Forum* 51, 2 (2017), 219–226.
- [7] Sergey Bezryadin, Pavel Bourov, and Dmitry Ilinih. 2007. Brightness calculation in digital image processing. In *Proc. of TDPF*. 10–15.
- [8] Anurag Bhardwaj, Atish Das Sarma, Wei Di, Raffay Hamid, Robinson Piramuthu, and Neel Sundaresan. 2013. Palette Power: Enabling Visual Search through Colors. In *Proc. of KDD*. 1321–1329.
- [9] Indrani Bhattacharya, Arkabandhu Chowdhury, and Vikas C. Raykar. 2019. Multimodal Dialog for Browsing Large Visual Catalogs Using Exploration-Exploitation Paradigm in a Joint Embedding Space. In *Proc. of ICMR*. 187–191.
- [10] Yiltan Bitirim, Selin Bitirim, Duygu Celik Ertugrul, and Onsen Toygar. 2020. An Evaluation of reverse Image search performance of Google. In *Proc. of COMPSAC*. 1368–1372.
- [11] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *JMLR* 3 (2003), 993–1022.
- [12] Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. 3–10.
- [13] SM Aqil Burney and Humera Tariq. 2014. K-means cluster analysis for image segmentation. *International Journal of Computer Applications* 96, 4 (2014).
- [14] David Carmel, Elad Haramaty, Arnon Lazerson, Liane Lewin-Eytan, and Yoelle Maarek. 2020. Why do people buy seemingly irrelevant items in voice product search? On the relation between product relevance and customer satisfaction in ecommerce. In *Proc. of WSDM*. 79–87.
- [15] David Carmel, Erel Uziel, Ido Guy, Yosi Mass, and Haggai Roitman. 2012. Folksonomy-Based Term Extraction for Word Cloud Generation. *ACM Trans. Intell. Syst. Technol.* 3, 4, Article 60 (2012), 20 pages.
- [16] David Carmel and Elad Yom-Tov. 2010. *Estimating the query difficulty for information retrieval*. Morgan & Claypool Publishers.
- [17] Abon Chaudhuri, Paolo Messina, Samrat Kokkula, Aditya Subramanian, Abhinandan Krishnan, Shreyansh Gandhi, Alessandro Magnani, and Venkatesh Kandaswamy. 2018. A smart system for selection of optimal product images in e-commerce. In *Proc. of Big Data*. 1728–1736.
- [18] Steve Cronen-Townsend, Yun Zhou, and W Bruce Croft. 2002. Predicting query performance. In *Proc. of SIGIR*. 299–306.
- [19] Ronan Cummins, Joemon M. Jose, and Colm O'Riordan. 2011. Improved query performance prediction using standard deviation. In *Proc. of SIGIR*. 1089–1090.
- [20] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2008. Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Comput. Surv.* 40, 2, Article 5 (2008), 60 pages.
- [21] Ritendra Datta, Jia Li, and James Z. Wang. 2005. Content-Based Image Retrieval: Approaches and Trends of the New Age. In *Proc. of MIR*. 253–262.
- [22] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proc. of TREC-3*.
- [23] Sarah Elkasrawi, Andreas Dengel, Ahmed Abdelsamad, and Syed Saqib Bukhari. 2016. What you see is what you get? Automatic Image Verification for Online News Content. In *Proc. of DAS*. 114–119.
- [24] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378–382.
- [25] Ziba Gandomkar and Claudia Mello-Thoms. 2019. Visual search in breast imaging. *The British journal of radiology* 92, 1102 (2019), 20190057.
- [26] Abby Goodrum and Amanda Spink. 2001. Image searching on the Excite Web search engine. *Information Processing & Management* 37, 2 (2001), 295–311.
- [27] Ido Guy. 2016. Searching by Talking: Analysis of Voice Queries on Mobile Web Search. In *Proc. of SIGIR*. 35–44.
- [28] Ido Guy. 2018. The Characteristics of Voice Search: Comparing Spoken with Typed-in Mobile Web Search Queries. *ACM Trans. Inf. Syst.* 36, 3, Article 30 (2018), 28 pages.
- [29] Benjamin Hadwiger and Christian Riess. 2020. The Forchheim Image Database for Camera Identification in the Wild. *arXiv preprint abs/2011.02241* (2020).
- [30] Claudia Hauff, Leif Azzopardi, and Djoerd Hiemstra. 2009. The combination and evaluation of query performance prediction methods. In *Proc. of ECIR*. 301–312.
- [31] Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. 2008. A survey of pre-retrieval query performance predictors. In *Proc. of CIKM*. 1419–1420.
- [32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. of CVPR*. 770–778.
- [33] Narayan Hegde, Jason D Hipp, Yun Liu, Michael Emmert-Buck, Emily Reif, Daniel Smilkov, Michael Terry, Carrie J Cai, Mahul B Amin, Craig H Mermel, et al. 2019. Similar image search for histopathology: SMILY. *NPJ digital medicine* 2, 1 (2019), 1–9.
- [34] Sharon Hirsch, Ido Guy, Alexander Nus, Arnon Dagan, and Oren Kurland. 2020. Query Reformulation in E-Commerce Search. In *Proc. of SIGIR*. 1319–1328.
- [35] Jen-Hao Hsiao and Li-Jia Li. 2014. On Visual Similarity based Interactive Product Recommendation for Online Shopping. In *Proc. of ICIP*. 3038–3041.
- [36] Houdong Hu, Yan Wang, Linjun Yang, Pavel Komlev, Li Huang, Xi (Stephen) Chen, Jiawei Huang, Ye Wu, Meenaz Merchant, and Arun Sacheti. 2018. Web-Scale Responsive Visual Search at Bing. In *Proc. of KDD*. 359–367.
- [37] Hao Jiang, Aakash Sabharwal, Adam Henderson, Diane Hu, and Liangjie Hong. 2019. Understanding the Role of Style in E-Commerce Shopping. In *Proc. of KDD*. 3112–3120.
- [38] Yushi Jing, David Liu, Dmitry Kislyuk, Andrew Zhai, Jiajing Xu, Jeff Donahue, and Sarah Tavel. 2015. Visual Search at Pinterest. In *Proc. of KDD*. 1889–1898.
- [39] Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proc. of CIKM*. 699–708.
- [40] Wang-Cheng Kang, Eric Kim, Jure Leskovec, Charles Rosenberg, and Julian McAuley. 2019. Complete the look: Scene-based complementary product recommendation. In *Proc. of CVPR*. 10532–10541.
- [41] Taewon Kim, Seyeong Kim, Sangil Na, Hayoon Kim, Moonki Kim, and Byoung-Ki Jeon. 2016. Visual Fashion-Product Search at SK Planet. *arXiv preprint abs/1609.07859* (2016).
- [42] Oren Kurland, Anna Shtok, David Carmel, and Shay Hummel. 2011. A unified framework for post-retrieval query-performance prediction. In *Proc. of ICTIR*. 15–26.
- [43] Katrien Laenen, Susana Zoghbi, and Marie-Francine Moens. 2018. Web Search of Fashion Items with Multimodal Querying. In *Proc. of WSDM*. 342–350.
- [44] Bing Li, Ling-Yu Duan, Yiming Chen, Rongrong Ji, and Wen Gao. 2012. Predicting the effectiveness of queries for visual search. In *Proc. of ICASSP*. 2361–2364.
- [45] Eileen Li, Eric Kim, Andrew Zhai, Josh Beal, and Kunlong Gu. 2020. Bootstrapping Complete The Look at Pinterest. In *Proc. of KDD*. 3299–3307.
- [46] Fengzi Li, Shashi Kant, Shunichi Araki, Sumer Bangera, and Swapna Samir Shukla. 2020. Neural Networks for Fashion Image Classification and Visual Search. *arXiv preprint abs/2005.08170* (2020).
- [47] Jie Li, Haifeng Liu, Chuanghua Gui, Jianyu Chen, Zhenyuan Ni, Ning Wang, and Yuan Chen. 2018. The Design and Implementation of a Real Time Visual Search System on JD E-Commerce Platform. In *Proc. of Middleware*. 9–16.
- [48] Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua. 2018. Interpretable Multimodal Retrieval for Fashion Products. In *Proc. of MM*. 1571–1579.
- [49] Yen-Chieh Lien, Hamed Zamani, and W. Bruce Croft. 2020. Recipe Retrieval with Visual Query of Ingredients. In *Proc. of SIGIR*. 1565–1568.
- [50] Kevin Lin, Fan Yang, Qiaosong Wang, and Robinson Piramuthu. 2019. Adversarial Learning for Fine-Grained Image Search. In *Proc. of ICME*. 490–495.
- [51] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.
- [52] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proc. of SIGIR*. 43–52.
- [53] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. of NIPS*. 3111–3119.
- [54] Aashish Kumar Misra, Ajinkya Kale, Pranav Aggarwal, and Ali Aminian. 2020. Multi-Modal Retrieval using Graph Neural Networks. *arXiv preprint abs/2010.01666* (2020).
- [55] Viken Parikh, Madhura Keskar, Dhwanil Dharia, and Pradnya Gotmare. 2018. A Tourist Place Recommendation and Recognition System. In *Proc. of ICICCT*. 218–222.
- [56] Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proc. of SIGIR*. 275–281.
- [57] Haggai Roitman, Shai Erera, Oren Sar Shalom, and Bar Weiner. 2017. Enhanced mean retrieval score estimation for query performance prediction. In *Proc. of ICTIR*. 35–42.
- [58] Diego Saez-Trumper. 2014. Fake Tweet Buster: A Webtool to Identify Users Promoting Fake News Ontwitter. In *Proc. of HT*. 316–317.
- [59] Yevhenii B. Shapovalov, Zhanna I. Bilyk, Artem I. Atamas, Viktor B. Shapovalov, and Aleksandr D. Uchitel. 2018. The Potential of Using Google Expeditions and Google Lens Tools under STEM-education in Ukraine. *arXiv preprint abs/1808.06465* (2018).
- [60] Raymond Shiau, Hao-Yu Wu, Eric Kim, Yue Li Du, Anqi Guo, Zhiyuan Zhang, Eileen Li, Kunlong Gu, Charles Rosenberg, and Andrew Zhai. 2020. Shop The

- Look: Building a Large Scale Visual Shopping System at Pinterest. In *Proc. of KDD*. 3203–3212.
- [61] Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting query performance by query-drift estimation. *ACM TOIS* 30, 2 (2012), 11.
- [62] Gyanit Singh, Nish Parikh, and Neel Sundaresan. 2012. Rewriting Null E-Commerce Queries to Recommend Products. In *Proc. of WWW Companion*. 73–82.
- [63] Parikshit Sondhi, Mohit Sharma, Pranam Kolari, and ChengXiang Zhai. 2018. A Taxonomy of Queries for E-Commerce Search. In *Proc. of SIGIR*. 1245–1248.
- [64] Ning Su, Jiyin He, Yiqun Liu, Min Zhang, and Shaoping Ma. 2018. User Intent, Behaviour, and Perceived Satisfaction in Product Search. In *Proc. of WSDM*. 547–555.
- [65] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural image assessment. *IEEE Transactions on Image Processing* 27, 8 (2018), 3998–4011.
- [66] Yongquan Tao and Shengli Wu. 2014. Query performance prediction by considering score magnitude and variance together. In *Proc. of CIKM*. 1891–1894.
- [67] Huawei Tian, Yanhui Xiao, Gang Cao, Yongsheng Zhang, Zhiyin Xu, and Yao Zhao. 2019. Daxing Smartphone Identification Dataset. *IEEE Access* 7 (2019), 101046–101053.
- [68] Riku Togashi and Tetsuya Sakai. 2020. Visual Intent vs. Clicks, Likes, and Purchases in E-Commerce. In *Proc. of SIGIR*. 1869–1872.
- [69] Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten de Rijke. 2020. Challenges and research opportunities in ecommerce search and recommendations. *SIGIR Forum* 54, 1 (2020).
- [70] Daniel Tunkelang. 2009. Faceted search. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009), 1–80.
- [71] Hen Tzaban, Ido Guy, Asnat Greenstein-Messica, Arnon Dagan, Lior Rokach, and Bracha Shapira. 2020. Product Bundle Identification Using Semi-Supervised Learning. In *Proc. of SIGIR*. 791–800.
- [72] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. 2014. Deep Learning for Content-Based Image Retrieval: A Comprehensive Study. In *Proc. of MM*. 157–166.
- [73] Anna Wróblewska and Łukasz Rączkowski. 2016. Visual Recommendation Use Case for an Online Marketplace Platform: Allegro.PL. In *Proc. of SIGIR*. 591–594.
- [74] Fan Yang, Ajinkya Kale, Yury Bubnov, Leon Stein, Qiaosong Wang, Hadi Kiapour, and Robinson Piramuthu. 2017. Visual Search at eBay. In *Proc. of KDD*. 2101–2110.
- [75] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL* 2 (2014), 67–78.
- [76] Andrew Zhai, Dmitry Kislyuk, Yushi Jing, Michael Feng, Eric Tzeng, Jeff Donahue, Yue Li Du, and Trevor Darrell. 2017. Visual Discovery at Pinterest. In *Proc. of WWW Companion*. 515–524.
- [77] Andrew Zhai, Hao-Yu Wu, Eric Tzeng, Dong Huk Park, and Charles Rosenberg. 2019. Learning a Unified Embedding for Visual Search at Pinterest. In *Proc. of KDD*. 2412–2420.
- [78] Yanhao Zhang, Pan Pan, Yun Zheng, Kang Zhao, Jianmin Wu, Yinghui Xu, and Rong Jin. 2019. Virtual ID Discovery from E-Commerce Media at Alibaba: Exploiting Richness of User Click Behavior for Visual Search Relevance. In *Proc. of CIKM*. 2489–2497.
- [79] Yanhao Zhang, Pan Pan, Yun Zheng, Kang Zhao, Yingya Zhang, Xiaofeng Ren, and Rong Jin. 2018. Visual Search at Alibaba. In *Proc. of KDD*. 993–1001.
- [80] Ying Zhao, Falk Scholer, and Yohannes Tsegay. 2008. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Proc. of ECIR*. 52–64.
- [81] Yun Zhou and W. Bruce Croft. 2007. Query performance prediction in web search environments. In *Proc. of SIGIR*. 543–550.
- [82] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. 2019. R2GAN: Cross-modal recipe retrieval with generative adversarial network. In *Proc. of CVPR*. 11477–11486.