



Shop by image: characterizing visual search in e-commerce

Arnon Dagan¹ · Ido Guy² · Slava Novgorodov³

Received: 21 November 2021 / Accepted: 28 September 2022 / Published online: 3 March 2023
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

Abstract

Visual search has become more 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. Despite the increasing popularity of visual search in e-commerce, no comprehensive study has inspected its characteristics compared to traditional search using a text query. 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 segmentation by user types. Additionally, we examine, for the first time, a wide variety of visual pre- and post-retrieval query performance predictors, several of which showing strong results. Our study 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.

Keywords E-commerce search · Product search · Query log analysis · Query performance prediction · Search by image · Visual search

An earlier partial version of the manuscript was reported in Dagan et al. (021).

✉ Slava Novgorodov
slavanov@post.tau.ac.il

Arnon Dagan
ardagan@ebay.com

Ido Guy
idoguy@acm.org

¹ eBay Research, Netanya, Israel

² Ben-Gurion University of the Negev, Netanya, Israel

³ Tel Aviv University, Tel Aviv, Israel

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 (Bitirim et al., 2020; Hu et al., 2018). Neural network techniques for image recognition support effective feature representation, classification, segmentation, and detection, and enable efficient retrieval of relevant results given an image query over huge corpora (Wan et al., 2014; Kim et al., 2016; Shiau et al., 2020). As Web content becomes ever more visual (Shiau et al., 2020), with the explosive growth in the number of online photos in social media and other websites (Zhai et al., 2017; Zhang et al., 2018), allowing users to express their information needs through an image becomes imperative. And in a culture already dominated by the visual, it is only natural that an image would be used to start a search.

In recent years, visual search has been implemented and studied in a variety of domains, such as travel (Parikh et al., 2018), news (Saez-Trumper, 2014; Elkasrawi et al., 2016), healthcare (Hegde et al., 2019; Gandomkar & Mello-Thoms, 2019), education (Shapovalov et al., 2018), and food (Lien et al., 2020; Zhu et al., 2019). Notably, one of the most popular visual search domains is electronic commerce. Sometimes referred to as *visual shopping* (Togashi & Sakai, 2020), 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 (Li et al., 2020a). 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 (Zhang et al., 2018; Bhardwaj et al., 2013). For instance, a customer may take a photo of a hat, either of another person, or in a store, and would like to instantly look up prices or stock availability from online stores (Bhardwaj et al., 2013). Alternatively, they may run into an image of an item they like on their social media feed and would like to quickly search for it (Goel, 2017).

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 (Li et al., 2020a). For example, some users might be interested in “jeans with holes”, but the relevant products are described as “distressed jeans” (Laenen et al., 2018). 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 (Shiau et al., 2020; Bhardwaj et al., 2013). 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 (Jiang et al., 2019). Even after filtering by a large number of attributes, users may be confronted with hundreds of items that differ by Style (Jiang et al., 2019). 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 (Bell et al., 2020). On the other hand, it should be noted that other aspects, such as size or brand names, can be more easily expressed by text.

Integrating visual search capabilities can enhance customer experience and increase engagement. 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.¹ Photo sharing service Pinterest reported that among its 350M monthly users, many have expressed a desire for visual shopping (Shiau et al., 2020). A study from The Intent Lab found that 85% of the young respondents put more importance on visual information than textual information.²

In recent years, many Web and e-commerce sites have introduced visual search functionality into their commercial applications (Jing et al., 2015; Yang et al., 2017; Zhang et al., 2018; Li et al., 2018; Hu et al., 2018; Bitirim et al., 2020). E-commerce platform Alibaba reported that their “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 (Zhang et al., 2018). Pinterest integrated “shoppable” pins into its visual search, making it easier for users to purchase products they have taken photos of (Shiau et al., 2020). 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 (Yang et al., 2017; Jing et al., 2015; Hu et al., 2018; Li et al., 2018; Zhang et al., 2018; Lin et al., 2019), demonstrating the complexity of real-world visual search systems (Jing et al., 2015), with challenges such as large gaps in image quality between the query and results; indexing of dynamic data; and training of large-scale ranking models (Zhang et al., 2018). Another focus of area has been the evaluation ranking models, including feature representation, retrieval models, and similarity calculation (Wan et al., 2014; Zhai et al., 2017; Hsiao et al., 2014; Misraa et al., 2020; Li et al., 2020a; Zhai et al., 2019; Zhang et al., 2019).

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, users, 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, users, 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.

¹ <https://www.visenze.com/blog/how-visual-search-has-transformed-the-modern-shopping-experience>.

² <https://www.businesswire.com/news/home/20190204005613/en/Visual-Search-Wins-Text-Consumers%E2%80%99-Trusted-Information>.

- We evaluate a set of query performance predictors for visual search and compare them with classic predictors in textual search. Our evaluation reveals a variety of strong predictors for the performance of visual queries.

This work extends the findings previously reported in Dagan et al. (2021). The main new contributions include the following:

- Additional analysis of result page characteristics, considering the title length and price (Sect. 6.2).
- Introduction of the notion of “intent sessions”, which captures a more focused sequence of interactions with the search engine (Sect. 7).
- Analysis of price, condition, and other “global” filters used for search refinement, alongside category-specific attributes (Sect. 8).
- Exploration of bias towards results with a similar image, showing an effect on visual e-commerce search click model (Sect. 9.2).
- A new user-based analysis, exploring in depth whether the observed differences between visual and textual search are due to the different types of users or persist across the exact same set of users (Sect. 10).
- Extension of our experiments with visual query performance prediction, introducing new pre-retrieval predictors and an additional evaluation metric (Sect. 11).

Overall, 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 and future research directions.

2 Related work

In this section, we cover related work, starting from search log analysis, through broad visual search, to visual search in e-commerce. For the latter, we also elaborate on work in the Fashion domain and a line of studies performed by the Pinterest image sharing service. Finally, we discuss the connection to work on visual e-commerce recommendation and image search.

2.1 Search log analysis

Studies of search log analysis, inspecting the queries submitted by users to search engines, and often times the returned results and the user’s interaction with them, have been published since the early days of the Web (Broder, 2002; Jansen, 2006). The emergence of mobile technologies led to a variety of studies inspecting queries submitted from mobile devices. For example, Kamvar and Baluja (2006) performed a large-scale analysis of Google mobile search, differentiating it from desktop search. Baeza-Yates et al. (2007) compared mobile and desktop search queries and found that mobile queries included more queries in the Business category and fewer in Art. Song et al. (2013) recommended, using log analysis of the Bing search engine, that ranking methods for smartphones, tablets, and desktop devices adapt to specific user behavior patterns in each of these platforms. Guy (2016) reported an in-depth analysis comparing spoken queries to typed-in queries on

mobile web search, indicating that the language used in voice search is closer to a natural language. In this work, we perform a log analysis of mobile e-commerce search on eBay, one of the world's largest marketplaces, in order to characterize visual search and compare it to textual search.

Log analysis of e-commerce search has been studied both as a prominent vertical of Web search (Jansen & Spink, 2006) and by inspecting the search logs of e-commerce platforms. Su et al. (2018) combined search log analysis and user survey to define three main intent categories for shoppers: target finding, decision making, and exploration. Sondhi et al. (2018) proposed a refined version with five categories based on analysis of e-commerce search logs, associating each of the categories with a distinct user search behaviour. Hirsch et al. (2020) provided a large scale and in-depth study of users' query reformulations in e-commerce search and showed that well over 50% of the queries take part in a reformulation session. In this work, we characterize visual search and compare it to textual search on e-commerce using a log analysis of one of the world's largest marketplaces.

2.2 Visual search

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 (Datta et al., 2005; Bhardwaj et al., 2013) to deep learning and semantic representation methods (Liao et al., 2018; Lin et al., 2019; Li et al., 2020a; Misraa et al., 2020). 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. (2018) 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. (2019) 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". While in classic image search, the query is textual and the results are images, in reverse image search, the query is an image and the results are (typically) textual documents. Bitirim et al. (2020) performed an evaluation of Google's reverse image search performance, in terms of average precision at varying sizes of result sets. Reilly and Thompson (2017) conducted a user study of reverse image search over a digital library.

2.3 Visual e-commerce search

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. (2018) 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; dealing with large-scale indexing of dynamic data; training

deep models for effective feature representation without massive human annotation; and improving user-based metrics by considering image quality for result re-ranking. Their mobile application, named “Pailitao”, which means shopping through the camera, enabled “search by image” using the visual search service. In a followup work (Zhang et al., 2019), the authors proposed learning image relationships based on co-click embedding, to guide category prediction and feature learning. They showed that the richness of click data enables to better reflect users’ interests and and improve visual search relevance. Li et al. (2018) 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. (2017) 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. They applied a supervised approach with multiple deep learning models to retrieve and rank listings from eBay’s huge inventory and showed, using an ImageNet benchmark, that their approach is faster and more accurate than several unsupervised baselines. Among the key challenges, they enumerated the dynamic nature of eBay’s inventory and its scale, the diverse image quality contributed from a variety of sellers; and the diverse quality of image queries, often captured by the shopper’s mobile device camera. 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. In an earlier study (Bhardwaj et al., 2013), eBay researchers presented a simple and fast search algorithm that uses color as a key feature for building visual search. They showed that low-level cues such as color can be used to quantify image similarity and distinguish between products with different visual appearance, to support fast and effective visual search.

2.4 Visual search for fashion

Much of the work on visual search in e-commerce has focused on the Fashion category. Many of the common attributes in this category, such as style (Kim et al., 2016; Kang et al., 2019; McAuley et al., 2015) or texture (Wróblewska & Rączkowski, 2016), are difficult to specify in words and easier to describe using an image. In addition, the notion of relevance in these categories is often mostly visual (Bhardwaj et al., 2013). Laenen et al. (2018) focused on the Dresses category and introduced a search interface and ranking model that supports the submission of an image query and the refinement of the image’s attributes using a text query. On eBay and other e-commerce platforms, such a refinement is enabled by facets (Tunkelang, 2009) presented according to the query’s category. Kim et al. (2016) described their visual Fashion search system on Korean e-commerce site SK Planet. Liao et al. (2018) incorporated the category tree hierarchy into a deep learning model to improve capturing user intent and reasoning of results in visual Fashion search. Togashi and Sakai (2020) examined the relationship between “visual intents” in terms of color, texture/material, design and user feedback in the form of clicks, likes, and purchases. They found that visual relevance positively correlates and can actually cause user feedback. They also recommended to diversify search results since different combinations of visual intents may lie behind the same image query. Our work examines visual search on eBay, which spans a variety of shopping categories. Our findings indicate that Fashion is indeed

a popular category for visual search, but a few other categories are even more popular, such as Collectibles, Art, and Toys.

2.5 Visual search in pinterest

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 (Jing et al., 2015) described how Pinterest built a cost-effective large-scale visual search system and showed its positive effect on user engagement. A follow-up work focused on the notion of visual discovery, enabling users to select any object in an image as a visual query (Zhai et al., 2017). They presented an overview of the visual discovery engine and shared the rationales behind technical and product decisions, such as the use of binarized features, object detection, and interactive user interfaces. Another work (Zhai et al., 2019) described the image embedding process behind Pinterest's visual search, using a multi-task learning architecture capable of jointly optimizing multiple similarity metrics, such as browsing and searching relevance. They detailed how to jointly train for multiple product objectives and how to leverage both engagement data and human labeled data. The resulting unified embedding outperformed all specialized embedding trained with an individual task. Recently, Pinterest introduced "shop the look" (Shiau et al., 2020), with an explicit goal to fulfill the user's shopping intent. The service detects objects within billions of inspirational scenes, and finds matching products from a huge product corpus that are visually similar to the detected objects. A related study named "complete the look" focused on the task of "style compatibility" in order to recommend complementary products for an outfit (Kang et al., 2019; Li et al., 2020b).

2.6 Visual e-commerce recommendation

Similarly to Pinterest, several other studies focused on a recommendation scenario, where the image "query" is not necessarily input explicitly by the user. Hsiao et al. (2014) used visual similarity to refine personalized product recommendations. McAuley et al. (2015) proposed a deep learning method based on heterogeneous graph embedding to recommend which clothes and accessories will complement each other well based on their appearance. Wróblewska and Rączkowski (2016) described a content-based visual recommender on a Polish online marketplace based on color and texture characteristics. Parikh et al. (2018) used convolutional neural networks (CNNs) to recommend tourist attractions, restaurants, and hotels based on their visual characteristics. In this work, we focus on pure visual search, analyzing a large-scale log of searches performed using explicit image and text queries.

2.7 Image search

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

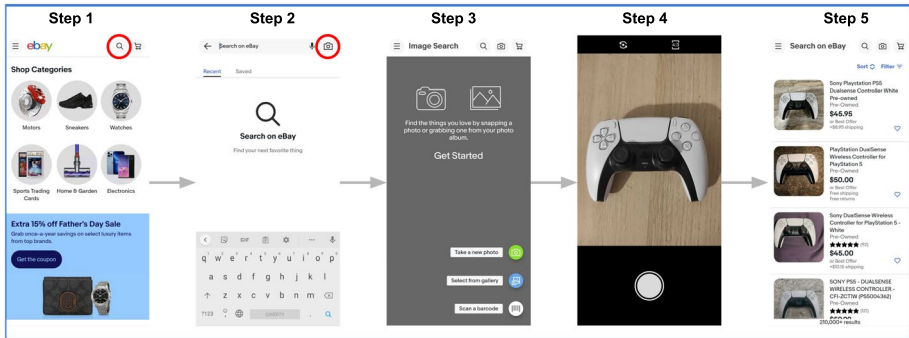


Fig. 1 Step-by-step demonstration of search by image experience using eBay's mobile app

consists of images, is a popular search vertical and has been extensively studied (e.g., (Datta et al., 2008; Goodrum & Spink, 2001)). 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, or CBIR (Datta et al., 2005; Wan et al., 2014). 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 setting

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, either by camera or by gallery, 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.

Figure 1 depicts a step-by-step demonstration of such process. After opening the mobile application, the user can choose to browse various categories or go to search mode by clicking the magnifying glass icon (Step 1). After entering search mode, the user can directly type the text query, use voice search by clicking the microphone icon, or use image search by clicking the camera icon (Step 2). Upon clicking the camera icon, the user enters the image search mode, where the three options include taking the photo instantly using the device's camera, uploading a photo from the device's gallery, or scanning a barcode (Step 3). If the user decides to take a new photo, the app opens a photo taking interface with an option to toggle between front and back camera and selecting of various aspect ratios (Step 4). Finally, after taking the photo (in this example a wireless controller for PlayStation 5), the search engine retrieves the most relevant results and present them to the user (Step 5).

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. Basic demographic data including gender, age, and location in terms of city and state was approved for analysis in aggregate across the entire image and text datasets. 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. The portion of unique queries out of all queries was similar between the image and the text samples: 69.5% and 73.5%, respectively.

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. In other words, the same product may appear multiple times on the SERP, with different sellers, prices, delivery options, and so forth. 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 and context. 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 inspects sessions characteristics. Section 8 examines the attributes used to refine image queries in comparison with text queries, while Sect. 9 inspects click characteristics, such as click-through rate and mean reciprocal rank. In Sect. 10, we examine user characteristics, focusing on those who use both visual and textual search. Finally, in Sect. 11, we describe our experimentation with a set of new pre- and post- retrieval performance predictors for visual search.

Table 1 summarizes key variable and measurements that were considered in this paper.

4 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

Table 1 Variable and measurements summary

Parameter	Value
# of image queries	1,635,632
# of unique users	250,000
# of meta categories	43
Time period	February 2nd, 2020– February 29th, 2020
Country	United States
Operating Systems	Android, iOS

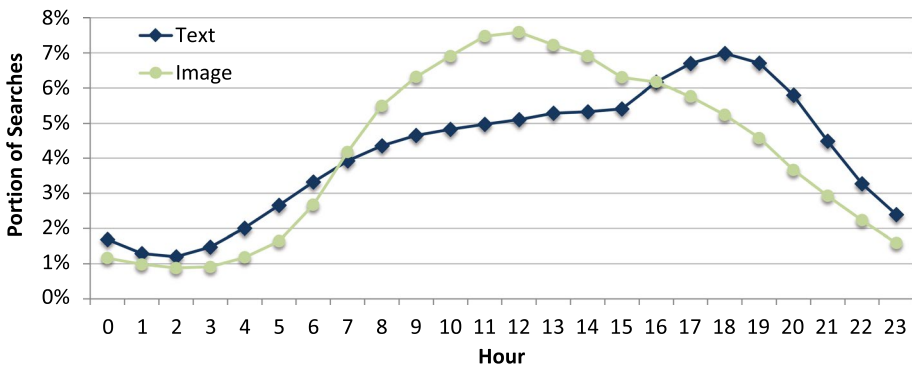


Fig. 2 Query distribution by hour of the day

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 mobile device’s operating system distribution was similar between iOS and Android, with a slightly higher portion of the image queries originating from Android compared to text queries (ratio between the two portions 1.06).

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 Fig. 2. 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. This trend was consistent across all seven days of the week.

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.



Fig. 3 Examples of image queries

5 Queries

As mentioned in Sect. 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 Fig. 3, 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 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. Zooming in on LCs, those related to cards, especially Sports cards, as well as vintage items, such as VHS tapes, have high portions of camera queries (over 90%), whereas fashion LCs such as dresses, earrings, heels, and women’s tops, boots, and swimwear have especially high portion of gallery photos. As mentioned in Sect. 4, gallery queries in general are especially popular with women. Overall, it can be observed that when searching for vintage and collectibles, as well as media items, users typically take their own picture to perform a visual search, while for “showy” items, such as fashion, jewelry, and beauty, they more often use a gallery image.

Throughout our analysis in the remainder of this paper, we compare key characteristics between the camera and gallery flow.

In the remainder of this section, we examine three image characteristics – orientation, brightness, and catalog quality – and compare them between the two flows. The differences in such query characteristics may influence the models used to retrieve results considering each of 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, and when it is exactly 1 the image is square. Table 2 summarizes the query image orientation. It can be seen that the portion of vertical images was substantially higher on camera queries compared to gallery queries. 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% (Tian et al., 2019) and 40.3% (Hadwiger & Riess, 2020). It can also be seen that square photos were very rarely used on camera queries, but were much more common on gallery queries.

Brightness Figure 4 depicts the brightness (Bezryadin et al., 2007) histogram (by buckets of brightness values) of camera and gallery queries. For reference, the figure also plots the brightness distribution of two publicly-available image datasets: a Flickr dataset (Young

Table 2 Distribution of image orientation by flow

Flow	Vertical	Horizontal	Square
Camera	92.72%	7.18%	0.10%
Gallery	72.04%	20.26%	7.70%
Total	88.60%	9.79%	1.61%

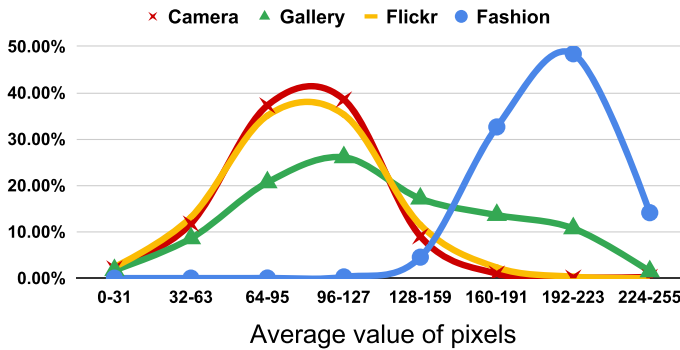


Fig. 4 Brightness of camera vs. gallery query images, compared with two public datasets: Flickr (Young et al., 2014) and Fasahion (2018). Lower values represent darker images

et al., 2014) that contains 30k pictures taken by Flickr users and a Fashion dataset (2018) 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 Fig. 3, examples 3 and 10 demonstrate uploaded user-generated screenshots, while 2 and 4 are uploaded stock photos. A similar comparable analysis of the luminance (Bezryadin et al., 2007) histograms showed the same trends and is excluded for clarity of presentation.

Image quality Online marketplaces often use models for image quality assessment in order to select the best images for their product catalogs (Chaudhuri et al., 2018). 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, additional elements besides the main product for sale, and package quantity. As could be 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. For reference, the average quality score for the Fashion dataset was 0.94 (std: 0.11, median: 0.98) and for Flickr 0.85 (std: 0.19, median: 0.95).

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 (Singh et al., 2012). 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 text null queries were typically longer at an average of 4.70 (std: 3.14, median: 4) compared to all text queries at 3.183 (std: 1.85, median: 3). For image, null queries had lower brightness (Bezryadin et al., 2007) (−11%) and aesthetic score (based on a publicly-available implementation of a aesthetic quality model (Talebi & Milanfar, 2018), somewhat resembles the image quality mentioned above) (−17%) compared to all other image queries. Figure 3 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, as indicated in our logs. 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 both types of search.

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, as noted in the previous paragraph.³

6.1 Categories

6.1.1 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 3 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: it almost always contains results of one MC and typically only one LC as well. These characteristics were similar for both camera and gallery queries. Figure 5 (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.

6.1.2 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

³ Throughout our analysis, we also calculated the statistics for the top 10 results and observed very similar trends.

Table 3 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%

top 40 results. In case of a tie, we considered the category 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 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 (Berger & Lafferty, 2017). Formally, KL divergence is defined as follows:

$$KL(P||Q) = \sum_{c=1}^M P_c \log \frac{P_c}{Q_c}$$

where P and Q are the two distributions (in our case of image and queries, respectively, across categories), and P_c and Q_c are the frequencies of a category c in each of the distributions. 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. In other words, we calculated the categories for which the formula in the summation above is the highest.

Table 4 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

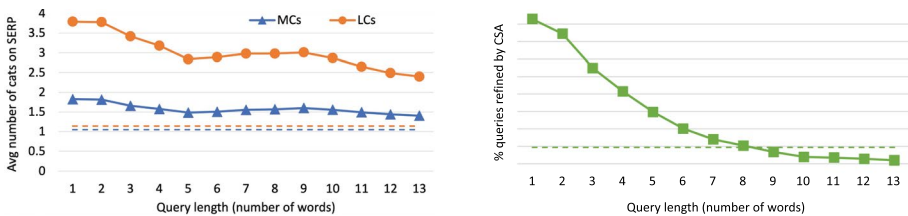


Fig. 5 Analysis of text queries by length: average number of MCs and LCs on the SERP (left plot) and percentage of queries refined by category-specific attributes (right plot). The dashed lines indicate the respective values for all image queries

Table 4 Most distinctive MCs and LCs in image queries relative to text queries according to KL divergence

	Meta category	Leaf category		Meta category	Leaf category
1	Collectibles	Mugs & cups Table lamps Collector plates	6	Coins & paper money	US coin errors Medals
2	Pottery & glass	Fenton art glass Crystal Pyrex	7	Stamps	Nepali paper money US collection and lots US postage
3	Antiques	Chinese figurines & statues Ceramic & porcelain vases Porcelain plates & chargers	8	Baby	US unused 1941-now Stroller parts Baby swings Baby monitors
4	Dolls & bears	Cloth dolls Dollhouse miniatures Cultures & ethnicities dolls	9	Crafts	Ready-to-paint pottery Wood items Acrylic paint
5	Toys & hobbies	TV & movie character toys TV, movie & video game action figures Contemporary manufacture traditional games	10	Jewelry & watches	Retro & vintage costume necklaces Retro & vintage costume pins & brooches Fashion bracelets

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. Other distinctive image query LCs not shown in Table 4, include glassware, drinkware, and candle holders under the Home & Garden MC; antiquarian, collectible, Fiction, and magazines under Books; cups and mugs memorabilia under Sports; and film stock under DVDs & Movies. Overall, we observe that while visual search was used across the board, it was especially popular for collectible and vintage products, art, and toys and babies. The Fashion category, which has been the subject of many previous studies on visual search (Liao et al., 2018; Bhardwaj et al., 2013; Bell et al., 2020; Shiao et al., 2020; Kang et al., 2019; Kim et al., 2016; Laenen et al., 2018), occurred in nearly 10% of the visual searches, but was not more popular than in textual searches. Distinctive image query LCs within the Fashion MC included outerwear, such as coats, jackets, and vests; bags, backpacks, and cases; and vintage/formal clothing such as suits, dresses, heels, and umbrellas.

6.2 Other characteristics

6.2.1 Title length

The average title length (in words) for image SERP results was similar for image compared to text searches at 11.06 (std: 2.01 median: 10.74) versus 11.56 (std: 1.93, median: 11.51). This was also reflected in a similar number of characters per title at an average of 70.33 (std: 19.18, median: 65.82) versus 70.30 (std: 15.49, median: 69.64).

6.2.2 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 0.87 (std: 0.27, median: 0.99) compared to 0.80 (std: 0.33, median: 0.98), respectively. There was also no significant difference from image queries in the gallery flow (avg: 0.83, std: 0.30, median: 0.98), even though as shown in Sect. 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. It may indicate that the search engine copes well with the relatively low quality of the image queries, and is not biased towards either low or high quality listing images.

6.2.3 Price

The average price for image SERP results was lower by a factor of 0.45 than for text SERP results. This ratio varied across MCs, but the general trend of image results being less expensive than text results on average remained: only 3 out of the 43 MCs had an average price higher for image results than for text results: Health & Beauty (average price ratio 1.19), Antiques (1.11), and Travel (1.10). Others had average price higher on text queries, e.g., Fashion (0.74), Pottery & Glass (0.73), Cellphones & accessories (0.42), Collectibles

(0.39), and, most sharply Toys & Hobbies (0.31), Jewelry & Watches (0.25), Books (0.16), and Music (0.12).

We also inspected image searches with an average SERP price at the top quartile compared to the bottom quartile (quartile calculation was performed by considering all visual searches). We observed that image searches at the top price quartile included a considerably higher portion of gallery queries (26.19%) compared to searches at the bottom price quartile (12.76%). This suggests that gallery photos are used more often for searching more expensive items, while the camera flow is especially common for low-cost goods. In terms of image quality, we observed no difference between the quality of the query image as well as the quality of the SERP images, between the top and bottom price quartiles.

7 Sessions

The query logs (both image and text) are partitioned into *eBay 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 (Hirsch et al., 2020; Jones & Klinkner, 2008). In this case, we refer to a query as any new combination of image or text with filters, as described in Sect. 3. The analysis in this section considers all queries - text and image - submitted throughout the experimental period, without any sub-sampling. This allows us to track complete sessions over the experimental period (February 2nd-29th, 2020).

Out of the image *eBay sessions*, 62.94% start with an image query, while the rest start with a text query, with the image query occurring later in the session. In addition, since the above definition of sessions often captures long sequences that involve more than one “quest” on the part of the user, we additionally define an *intent session* as a sub-sequence of an *eBay session* in which the dominant MC for all queries is identical (see the definition of a dominant category in Sect. 6.1).⁴ Using this definition, we aim to capture shorter sessions that focus on the same intent.⁵ We refer to an image session (either *eBay session* or *intent session*) as any session that contains at least one image query. All the other sessions are considered as text sessions.

Table 5 presents session statistics. For *eBay sessions*, it can be seen that image sessions tend to be longer than text sessions, 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. As for idle time between queries in a session, it is also longer for image sessions, even when considering specific transitions, such as from the first query to the second, or from the second to third (generally, idle times between queries tend to grow along the session). Inspecting *intent sessions*, the trends change, and session length, duration, and idle times become similar for image and text queries. It appears that while image queries are more commonly part of a longer sequences of activities, perhaps implying higher engagement levels, when moving to *intent sessions*, which capture a

⁴ We consider the longest possible *intent session* according to this definition, i.e., an *intent session* cannot be a sub-sequence of another longer *intent session*.

⁵ We experimented with other variants of this definition, e.g., based on dominant LC or on the top 10 results rather than 40; we converged to the above definition as we found it to best capture maximum-length sequences of similar intent.

Table 5 Session characteristics: length (number of queries) properties; duration properties; and idle time between consecutive queries in the session

	eBay sessions		Intent sessions	
	Text	Image	Text	Image
Avg (std) number of queries	2.99 (4.22)	7.83 (12.9)	2.00 (2.42)	2.14 (3.10)
Median number of queries	2	4	1	1
% 1-query sessions	44.00%	21.63%	61.47%	64.65%
Avg (std) duration in minutes	18.47 (29.81)	38.81 (57.87)	14.77 (26.67)	13.84 (29.97)
Median duration in minutes	7.65	18.13	4.82	3.74
Median duration of 2-query sessions	5.85	6.43	9.57	8.73
Median duration of 3-query sessions	9.55	10.55	13.00	12.27
Median duration of 4-query sessions	17.23	20.15	16.28	15.42
Avg (std) idle in minutes	3.85 (7.20)	4.31 (6.49)	3.19 (6.57)	3.28 (6.21)
Median idle in minutes	1.48	2.33	1.13	1.30
Avg (std) idle 1st query to 2nd query in minutes	3.91 (9.31)	4.60 (10.38)	3.19 (7.96)	3.32 (7.78)
Median idle 1st query to 2nd query in minutes	0.87	1.03	0.77	0.92
Avg (std) idle 2nd query to 3rd query in minutes	3.86 (8.95)	4.37 (9.77)	3.31 (7.92)	3.23 (7.47)
Median idle 2nd query to 3rd query in minutes	0.93	1.02	0.87	0.88

focused intent, the session characteristics of image and text sessions demonstrate comparable characteristics.

The eBay session characteristics for image queries are different for the two flows: the sessions that include gallery queries are shorter in terms of the number of queries (avg: 4.9, std: 6.98, median: 3) and duration (avg: 30.00 minutes, std: 40.00, median: 12.25). The idle times are nearly identical for camera and gallery queries. The length and duration differences disappear when inspecting intent sessions.

8 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 (Hirsch et al., 2020; Sondhi et al., 2018; Tunkelang, 2009). Upon submitting a query, the user is displayed with different refinement options, often defined based on the category(ies) of the returned results, and can narrow down the list of retrieved results based on specific values, such as a color, size, or condition. 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. We distinguish between general (global) filters (*GFs*), used across all categories, such as price and condition, and category-specific attributes (*CSAs*), such as color, material, or brand, which are defined based on the category(ies) of the returned results.

The analysis in this section is based on the original datasets of over 1.6 million image and text queries (Sect. 3). Generally, the use of GFs was much less frequent on image search compared to text search, with an image to text ratio of 0.23. Since most global filters focus on cost perspectives, from price through shipping to format (buy it now, auction, accepts offers), return policy, and savings and deals, this may imply that image searches are less focused on finding a good deal and, alternatively, on discovery and recollection of a specific item (Togashi & Sakai, 2020; Liao et al., 2018).

The use of CSAs to refine the search results was even rarer on visual compared to textual search, with an image to text ratio of 0.17. Indeed, CSAs can often be captured by an image, while global filters are often not characterizing the item's visuals (except for, arguably, the condition filter). Finally, queries that were filtered by both a GF and a CSA were less frequent on visual compared to textual search, by a ratio 0.27.

The sharp difference in CSA refinement gives another indication that image queries often reflect narrower information needs with richer sets of attributes than text queries, and therefore does not require further refinement. We further explore this by analyzing CSA refinement in text queries according to their number of terms. Figure 5 (right plot) shows a clear trend: the use of refinement by CSA 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 (Hirsch et al., 2020), 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.

The use of query refinement varies substantially across MCs. Collectibles, Toys & Hobbies, Stamps, Antiques, Pottery & Glass, and Entertainment memorabilia – all categories shown popular for visual search (Table 4) – have generally low refinement use on both text and image queries, with a relatively high ratio between image and text. For Fashion and Jewelry & Watches, the use of refinements is generally high, with an image -to-text ratio higher than the general. For “technological” MCs, such as Cell-phones & Accessories, Computers/Tablets & Networking, and Cameras & Photo, the ratio is especially low, with more frequent use on text and highly infrequent on image.

Table 6 presents the most common GFs and CSAs used for text and image queries. The data for each type of GF and CSA is based on at least a few thousands of data points. The upper section of Table 6 shows the most common GFs used in text and image queries. The four most popular GFs, which account for 75% of the image and 66.5% of the text total GF use were condition, format, location, and price. The rightmost column of the table shows the image -to-text ratio of relative use per each filter. It can be seen that the condition filter, one of the few GFs that does not directly relate to cost, is the most commonly used in both image and text queries, with comparable relative use on both. For image queries, the location filter comes at close second, whereas for text queries format is second, while much less popular on image. The filters for text queries include, further down the list (not shown in Table 6), a longer and tail of filters related to shipping (e.g., free or expedited), return policy, promotions, and seller authority.

The lower section of Table 6 shows the relative distribution of CSAs 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

Table 6 Most clicked GFs and CSAs for refining text and image queries

	Text	Image	Ratio
	GF	GF	
1	Condition	Condition	1.012
2	Format	Location	1.523
3	Location	Price	1.666
4	Price	Format	0.596
	CSA	CSA	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

The rightmost column shows the ratio between the percentage of the GF (CSA) use out of all GFs (CSAs) used in image queries and the same percentage in text queries

studied as a particularly popular attribute for visual search (Kim et al., 2016; Kang et al., 2019; McAuley et al., 2015), 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 3, 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 CSAs are style-related heel height, sleeve length, and dress length. With a particularly high ratio are pattern (also studied in various papers (Bhardwaj et al., 2013; Shiau et al., 2020; Togashi & Sakai, 2020; Yang et al., 2017; Laenen et al., 2018)), 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 CSAs , such as shoe size and size type, as well as technology-related CSAs such as network, storage capacity, screen size, and operating system.

Inspecting the values used for the popular GFs , we observed that when a price filter was used, the low bound (minimum price) was applied relatively more often in image queries (ratio 1.68), while the upper bound (max price) was used more often in text . The use of a low price bound is typically common in vintage categories. For condition, more image queries were refined to various forms of 'used' (ratio 1.18), likely due to the popularity of vintage categories, as shown in Table 4, while text queries

Table 7 Most common color and material values used for refinement of text and image queries, with respective image -to-text relative usage ratios

Color			Material		
Text	Image		Text	Image	
Value	Value	Ratio	Value	Value	Ratio
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

refined more frequently to ‘new’ items. For format, the proportion of filtering between ‘buy it now’ and ‘auction’ was identical between image and text queries.

Table 7 shows the most common values used for refinement in two of the most common CSAs : color and material. We focused on these two attributes since their list of common values is rather small, as opposed to other common CSAs such as brand, size, and type. The analysis is therefore based on at least a few hundreds of data points (over 700) and in most cases a few thousands. 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 3 example 5, the color of the bottle is dark blue and the user explicitly used ‘blue’ as a color refinement. This also explains why color, despite being a highly visual attribute, is relatively high on image CSA use as shown in Table 6. We observed similar trends in color value distribution across different MCs , such as Home & Garden and Fashion. 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 compared to camera : a ratio of 2.83 for GFs and 3.51 for CSAs , but still not as frequent as in text queries. The distribution of the GFs was similar between the flows, with one noticeable difference: the price filter was used substantially more frequently in the gallery compared to the camera flow, by a ratio of 2.31. We conjecture that when customers search by a gallery image, they often use an image from another website and seek for comparing the price. The distribution of CSAs was also similar between the flows, with one difference: the size attribute was used substantially more frequently in the gallery compared to the camera flow, by a ratio 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 3, examples 5 and 6).

Table 8 Click-through rate (CTR), average number of click (AVC), mean reciprocal rank (MRR), and average click rank (ACR) ratios between image and text queries

	CTR	Clicked queries		
		AVC	MRR	ACR
eBay sessions	0.897	0.979	–	–
Intent sessions	0.553	0.777	–	–
Queries	0.481	0.739	1.113	1.150
Gallery flow queries	0.674	0.907	1.008	1.035
Home & garden queries	0.427	0.847	1.163	1.201
Collectibles queries	0.497	0.668	1.151	1.010
Toys & hobbies queries	0.521	0.677	1.211	0.865
Jewelry & watches queries	0.552	0.724	0.921	1.231
Fashion queries	0.595	0.786	0.957	1.202
Antiques queries	0.642	0.685	1.095	0.958
Pottery & glass queries	0.810	0.714	1.261	0.750

The leftmost column presents the subset of image queries or sessions considered for the ratio to all text queries or sessions, respectively

9 Clicks

9.1 Click-through rate and rank

The analysis in this section is based on the original datasets of over 1.6 million image and text queries (Sect. 3). Table 8 shows the ratio for various click characteristics between visual and textual search.⁶ 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), and average click rank (ACR). 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 5, the length of image sessions is substantially higher than text sessions. Indeed, inspecting intent sessions, whose length is similar between image and text (Table 5), the CTR and AVC are substantially lower for image. 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 (Guy, 2016), 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, despite the fact that users traverse a similar number of results, as note in Sect. 6. For example, a higher portion of the clicked image queries included a click on the first result compared to text queries, at a ratio of 1.29. On the other hand, the average click rank (ACR) was also higher for image queries, indicating that some clicks on image queries are also performed relatively lower on the result list. Indeed, the standard deviation of click ranks was considerably higher for image queries than for text queries (ratio 1.443). For voice queries, the MRR was reported to be similar to that of text queries, at a ratio of 0.97 (Guy, 2016). Overall, the lower CTR and AVC and high MRR and ACR imply

⁶ We cannot disclose actual values due to business sensitivity.

that visual search is often used for target finding (Su et al., 2018). 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 8: the CTR and AVC ratios were higher, while the MRR and ACR were almost identical to text queries.

In general, the CTR was quite diverse across different MCs (considering the dominant categories, as defined in Sect. 6): standard deviation was 36.5% and 27.6% of the mean CTR, for text and image queries, respectively. The lower section of Table 8 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 CTR ratio varied rather substantially across MCs: it was a low 0.43 for Home & Garden, while reaching as high as 0.81 for Pottery & Glass. The ratio does not reflect the frequency of the MC for image versus text queries as presented in Table 4: for instance, it is higher for Fashion (0.595), which is not on the list, than for Collectibles (0.497), which tops the list. 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 3 example 6 shows a query whose dominant category is Pottery & Glass). Typically, MCs with higher MRR also have lower ACR, indicating that clicks are performed on results ranked higher for such MCs.

9.2 Image similarity bias

Click models (Chuklin et al., 2015) have been extensively studied in textual search, aiming at modeling the user behavior when interacting with the retrieved results. In this section, we focus on one factor that may play a key role for click models in visual e-commerce search: the similarity between the image query and a listing's image. We hypothesize that a presentation bias may occur due to such similarity. To examine this, we define the *similarity* function 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 (Mikolov et al., 2013) trained over a corpus of 10 million titles sampled uniformly at random from the eBay inventory during the month of February 2020. 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 (Arora et al., 2017). For visual search, we used a ResNet-50 network (He et al., 2016) to learn image embeddings over more than 50 million listing images from the eBay site, across all major categories (Yang et al., 2017). We applied these embeddings to both the image query and the listing's main image.

Using this similarity definition, we performed the following analysis over image queries: for each image query q and rank $r \in 1, 2, \dots, 10$, we compared r to the image similarity rank r_{img} , which is calculated by ranking the images of the top 10 retrieved listings based on their similarity to the image query. In other words, r_{img} reflects an alternative ranking of the query's top 10 results according to image similarity only. In contrast, the original ranking as computed by the search engine considers additional factors, such as

Table 9 CTR ratio for ranks 2 to 10 between queries for which the image similarity rank is higher than the actual rank and queries for which it is equal to or lower than the actual rank for image, gallery, and text queries

	2	3	4	5	6	7	8	9	10
Image queries	1.21	1.39	1.17	1.41	1.50	1.29	1.27	1.56	1.37
Gallery flow queries	1.23	1.37	1.17	1.45	1.47	1.33	1.45	1.44	1.57
Text queries	1.15	1.06	1.00	1.00	0.85	1.19	1.04	1.25	1.01

other modalities of the listings (e.g., title or attributes) and user behavior signals (e.g., past clicks). We then compared, for each $r \in 2, \dots, 10$, its CTR when it is lower than r_{img} and its CTR when it is equal to or higher than r_{img} .⁷ Table 9 presents the CTR ratio between these two cases for $r \in [2, 10]$. It can be seen (according to the table's first row) that when the image similarity rank is higher than the actual rank, the CTR is considerably higher, by a factor ranging between 1.17 and 1.56, than when the image similarity is lower than or equal to the actual rank. This trend is consistent across all values of r from 1 to 10 and indicates that searchers who use an image as their query have a tendency to click on listing results with images similar to their query. The results for the gallery flow (second row of Table 9) showed similar trends to all image queries, as shown in the table. For comparison, we performed an analogous experiment for text queries, based on the similarity function described above between the textual query and the listing's title. As can be seen in third row of Table 9, we could not observe a similar bias for text queries. The CTR ratio ranges between 0.85 and 1.25 across ranks 2 to 10, and is close to 1 in many of the ranks, showing no clear bias towards results whose title is more similar to the query. For example 7 in Figure 3, all top 10 results included similar shoes, and the ones at rank 4,5, and 9 had the identical product. Yet, only the result ranked 9 was clicked, presumably since it had a very similar photo to the image query, while other higher ranked results had less similar photos. This demonstrates how image similarity comes to play when users interact with retrieved results in visual e-commerce search.

9.3 Examples

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 10 shows eleven examples, including the image and text queries, and the title and image of the purchased listing. In a few examples (4,6) the purchased item is prominently different than the image query, suggesting a decision making and exploration intent rather than target finding (Su et al., 2018).

10 Users

In this section, we focus on user characteristics. We first inspect the unique characteristics of relatively "heavy" (frequent) users of visual search. Then, we examine a variety of the characteristics analyzed in previous sections for text queries by users who also submitted

⁷ The CTR for a rank r and a subset of queries Q is the portion of Q for which the result at rank r was clicked.

Table 10 Example image and text queries that led to a purchase of the same item















#	Image Query	Text query	Purchased listing image and title
1		iphone 8 se screen protector	 3-Pack For iPhone 11 Pro 8 7 6s Plus X Xs Max XR Tempered GLASS Screen Protector
2		Sata 2.5" 8mb buffer 5400"	WL 500GB 5400RPM 8MB 2.5" SATA Notebook Hard Drive (PS3 Fat, PS3 Slim, PS4 HDD)
3		Carotone cream	DSP10 Black Spot Corrector Creme 1oz
4		Party rings t	Women Elegant 925 Silver Sapphire Amethyst Rings Wedding Engagement Jewelry Gift
5		Jadoo tv remote wireless	Universal Wireless Air Mouse Keyboard Remote Control For Mini PC Android TV Box
6		Bad boys chevelle	GREENLIGHT HOLLYWOOD SERIES 21 BAD BOYS 1968 CHEVROLET CHEVELLE SS
7		Wireless earbuds	Wireless Earbuds Bluetooth Headphones, in Ear Bluetooth 5.0 White-IPS03
8		Lucky step rainbow shoes 7.5	Womens shoes Rainbow Lace Up Sneakers Gym Sports Running Trainers Casual Shoe
9		Rocker baby	Fisher-Price Infant-to-Toddler Rocker - Pacific Pebble, Portable Baby Seat, Multi

Table 10 (continued)

#	Image Query	Text query	Purchased listing image and title
10		Griddle electric	 Eternal Deluxe 12" Electric Crepe Maker & Griddle Temperature Control PG93932
11		Klein ncvt-2	 Klein Tools NCVT-2P Dual-Range Non-Contact Voltage Tester - Brand New!!!

Examples 2,3,5,11 include camera queries and the rest include gallery queries

Table 11 The ratio between frequent image users and all image users w.r.t different characteristics: portion of dominant categories on the SERP (top 3 and others), use of query refinement filters, and clicks

Top Image MCs			Other MCs			Refinements			Clicks	
Collectibles	Pottery	Antiques	Cellphones	Healthcare	Sports	CSA	GF	CSA+GF	CTR	AVC
1.21	1.40	1.17	0.34	0.48	0.72	1.18	1.15	1.21	0.91	0.92

image queries. This allows us to control for user factors when comparing textual and visual search, inspecting both text and image queries originating from the same set of users. Throughout this section, we refer to *visual search users* or *image users* in short, as users who performed at least one visual search, i.e., submitted at least one image query, throughout our experimental period, as reported in Sect. 3. Image users account for 1.64% of all users who submitted any query (image or text) throughout the experimental period. Their overall number of queries (both text and image) amounts to 6.32% of all queries submitted during the experimental period. The analysis in this section considers all queries – text and image – submitted throughout the experimental period, without any sub-sampling of either users or queries. This allows us to track user behavior along our experimental period (February 2nd–29th, 2020), without discarding any of the user’s data due to sub-sampling.

10.1 Frequent visual search users

We define *frequent users* of visual search as the top decile of the group of visual search users during our experimental period, according to the total number of image queries they submitted throughout the period. The number of image queries submitted by the frequent users accounts for 53.97% of all image queries. Table 11 presents the ratio of occurrence of different characteristics between the frequent users and all users of visual search. In terms of query categories, determined by the dominant category on the SERP, as defined in Sect. 6.1, it can be seen that the top 3 most distinctive visual search MCs, as presented in Table 4, are even more common on visual searches from frequent users. For example, the Collectibles category’s portion of visual searches is higher by a factor of 1.21 for frequent users compared to all visual search users. The prevalence of these MCs comes at the expense of other MCs, which are anyhow less frequent for visual search, such as Cellphones, Healthcare, and Sports, for which, as can be seen on the table, the ratio is substantially lower than 1.

Table 11 also indicates that the use of refinements, both by category-specific attributes and global filters, is more common for image queries submitted by frequent users, by a factor of around 1.2. Finally, inspecting click behavior, the click-through rate and average number of clicks (for clicked queries) are both lower by a factor of about 0.9 for queries submitted by frequent visual search users. While the frequent users submit more image queries, they perform fewer clicks per query. This behavior is generally typical for frequent users, likely as they perform more exploration than target-oriented searches (Su et al., 2018). For instance, the CTR and AVC for frequent users of textual search in our experimental period (top decile of users by number of text queries submitted) is lower by a factor of 0.88 than for all textual search users.

10.2 Text queries by visual search users

Our analysis in previous sections revealed prominent differences between image and text queries across a variety of characteristics. However, all the analysis was performed over a random sample of text and image queries. Since visual search is still not mainstream, image queries originate from a small subset of users (1.64% of all users, as mentioned above), while the vast majority of text queries in the sample originate from users who have never used visual search. While, as mentioned in Sect. 4, we controlled for basic factors, such as gender distribution, it is still possible that the characteristics of the relatively small group of users who used visual search throughout the one-month experimental period carries unique characteristics that also influence the observed differences between visual and textual search. To further examine to what degree the reported characteristics were affected by user differences and to what degree they really stem from the query modality, we set out to explore the characteristics of text queries submitted by visual search users. We refer to these queries as *image user text (IUT)* queries. Overall, they account for 5.96% of the text queries in our dataset. In addition, for another level of comparison, we also examine text queries by frequent image users, referred to as *FIUT* queries. These account for 55.06% of all IUT queries, indicating that frequent image search users also tend to frequently submit text queries. As mentioned above, for the analysis we consider all queries throughout the experimental period, without sub-sampling.

Table 12 compares different characteristics of image queries (by all visual search users) in terms of their ratio to the same characteristics of text queries (by all users), IUT, and FIUT queries. This enables to examine to what degree the image-to-text characteristics, as reported in previous sections, can also be observed when considering the text queries originating from the same user group, as reflected in the image-to-IUT column. The image-to-FIUT column is included as a reference, to reflect the trends for text queries by frequent users of visual search (who are also, as mentioned, more frequent users of textual search). The left section of the table focuses on click and refinement characteristics, as discussed in Sects. 9 and 8, respectively, while the right section inspects dominant categories on the SERP, as discussed in Sect. 6.1.

10.2.1 Click characteristics

Inspecting click-through rate (CTR), it can be seen that the image -to-IUT ratio is 0.75, indicating that even for the same group of users, CTR is lower for image queries than for text queries. The ratio, however, is not as low as the general image -to-text ratio at 0.48, since the CTR for IUT queries is itself lower than the general CTR for text queries. We have already reported above that queries by more frequent users (either image or text) have lower CTR, and visual search users are in general more frequent search users (recall they account for 1.64% of all users, with 5.96% of all text queries). Overall, these results indicate that part of the CTR difference reported for image compared to text queries stems from the mere nature of the users, while another can be attributed to the modality difference between image and text queries. Inspecting FIUT queries, the ratio is even closer to 1, giving another indication that the CTR decreases for more frequent users (recall that they account for 10% of the image users, but over 55% of the text queries that originate from image users). Inspecting the average number of clicks (AVC) for clicked queries, we can observe similar trends to the CTR: the image -to-IUT ratio is lower than 1, yet not as low as the image -to-text ratio, indicating that some of the observed gap between image and

Table 12 Image-to-text, image-to-IUT, and image-to-FIUT ratios across a variety of characteristics: clicks and uses of query refinement (left) and portion of dominant categories on the SERP (right)

	Image-text	Image-IUT	Image-FIUT	Image-text	Image-IUT	Image-FIUT
CTR	0.48	0.73	0.93	3.00	2.38	2.17
AVC	0.74	0.86	0.90	8.04	3.18	2.86
CSA	0.17	0.26	0.27	7.10	4.88	4.96
GF	0.23	0.26	0.25	3.21	1.67	1.45
CSA+GF	0.27	0.40	0.38	1.57	1.30	1.48
Size	0.50	0.83	0.61	0.27	0.39	0.78
Color	1.75	1.33	1.37	0.45	0.54	0.49
Brand	1.60	1.25	1.32	0.48	0.83	0.92
Format	0.60	0.79	0.83	0.51	0.57	0.64
Price	1.67	1.34	1.38	0.59	0.70	0.67
				Collectibles		
				Pottery		
				Antiques		
				Dolls		
				Toys		
				Cellphones		
				Sporting goods		
				Musical instruments & gear		
				Consumer electronics		
				Cameras		

text queries can be attributed to the user group, while the other is purely due to the modality. The image -to-FIUT ratio is only slightly higher than for IUT queries for AVC .

10.2.2 Query refinement

Inspecting the use of query refinements, it can be seen that the IUT -to-image ratio is very low for CSAs and GFs , at 0.26 for both. While this is not as low as the image -to-text ratios, it still indicates that refinements are substantially less common on image queries than on text queries, even when considering the same set of users for the analysis. In this case, the ratios for FIUT queries are rather similar to IUT , i.e., there is no substantial difference in refinement use stemming from the frequency of image search usage, as opposed to clicks.

The bottom left section of Table 12 presents the ratio of a few specific CSAs and GFs for which there was a noticeable gap between image and text queries (i.e., a ratio far from 1), as reported in Sect. 8. It can be observed that the image -to-IUT (as well as the image -to-FIUT) ratios demonstrate similar trends to the image -to-text ratios, but with milder numbers (i.e., closer to 1), indicating that similarly to the general use of refinements, the main refinement type differences can also be attributed both to the different group of users and to the query modality.

10.2.3 Dominant category distribution

The right section of Table 12 presents the ratios for query category portions, calculated by the dominant category on the SERP, as described in Sect. 6.1. The upper section presents the top 5 most distinctive MCs for image compared to text queries, as reported in Table 4. The image -to-text ratios for each of these categories is therefore higher than 1 (this is the ratio between the portion of queries whose dominant category is the MC in question for image and text queries, respectively). For IUT (and FIUT) queries, it can be seen that the ratios are also all substantially higher than 1, indicating that these categories characterize visual search compared to textual search, rather than just the set of visual search users. The lower ratios compared to the image -to-text ratio indicate, however, that some of the difference in the latter ratio can be attributed to the set of users who use visual search rather than the modality of the query. The lower section of the table presents 5 categories that are more frequent on text queries (image -to-text ratio lower than 1), with similar trends, i.e., the image -to-IUT ratio is also lower than 1, even if not as low as image -to-text , suggesting that the less popular categories for visual search remain even if considering the same set of user.

10.2.4 Time-of-day

Finally, Figure 6 nicely demonstrates the trends observed thus far for another characteristic: time-of-day distribution, as reported in Sect. 4. Recall that we observed noticeable differences between image queries, more popular during day hours, and text queries, more popular in evening hours. The figure also includes the time-of-day distribution for IUT and FIUT users and shows that these “fall” in-between the image and text distributions, indicating that some of the difference between text and image can be attributed to the set of

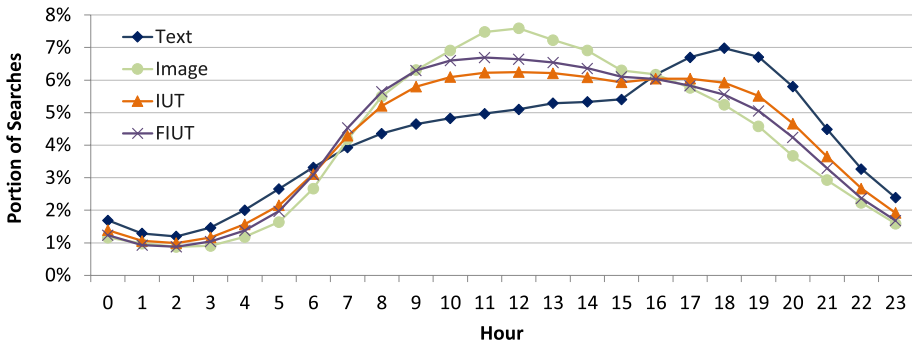


Fig. 6 Query distribution by hour of the day, including text queries by image users (IUT) and by frequent image users (FIUT)

users who use visual search, while the rest can be related directly to the modality difference between image and text queries.

Overall, the results in this section confirm our conjecture that some of the differences between visual and textual search stem from the different characteristics of the (small) user subset who is using visual search. On the other hand, the differences remain substantial even when controlling for this effect, and comparing text and image queries from the same set of users. The IUT -to-image ratios consistently show similar trends, even if somewhat milder (especially for CTR), to those reported throughout the sections of this paper.

11 Query performance predictors

The task of query performance prediction (QPP) (Cronen-Townsend et al., 2002; Carmel & Yom-Tov, 2010) 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 before the retrieval stage, based on the query itself and the corpus statistics (Hauff et al., 2008); and post-retrieval QPPs, which assess the query performance by considering the retrieved result list (Kurland et al., 2011). 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 (Li et al., 2012). In this section, we experiment with several pre- and post-retrieval QPPs for visual search. We evaluate their performance based on both traditional relevance judgements and direct query difficulty evaluation and compare them with classic QPPs applied to text queries.

11.1 Definitions

11.1.1 Textual QPPs

For text queries, we follow the list of QPPs described in a recent paper studying e-commerce textual search (Hirsch et al., 2020). For pre-retrieval, these include the query length (in words) (Carmel & Yom-Tov, 2010); the minimum, maximum, and sum of the IDF

values of the query terms (Ponte & Croft, 1998); and the minimum, maximum, and sum of the variance of TF.IDF values of the query terms across documents in the corpus (Zhao et al., 2008). These predictors have shown to be effective for document search in large-scale studies (Hauff et al., 2009; Shtok et al., 2012).

11.1.2 Visual pre-retrieval QPPs

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 and luminance, as described in Sect. 5; the portion of pixels with its 1, 2 and 3 most dominant colors (where colors are defined by clustering in the RGB space using k-means with $k=8$ (Burney & Tariq, 2014)); and the image quality, measured both using our own model for catalog quality estimation described in Sect. 5 and using a publicly-available implementation of a neural model for general image aesthetic quality assessment (Talebi & Milanfar, 2018). In addition, we considered the number of detected objects and the percentage of the area of the largest detected object as well as all objects in the image query as pre-retrieval QPPs, using the YOLO model (Redmon et al., 2016) for object detection. We also used the categorization of the image query for performance prediction, relying on an internal model (Yang et al., 2017). Given an image the model predicts the top 10 LCs, associating each LC with a probability. We used the sum of top k probabilities (for $k \in \{1, 3, 5\}$) and, inspired by Ozdemiray et al. (2014), the ratio between the probability of the of the top 2 categories as pre-retrieval QPPs.

For corpus-based predictors, 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 (Li et al., 2012). Both are based on concept extraction from images using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) over visual words, with visual words extracted using the SIFT algorithm for feature detection (Lowe, 2004): *q-INS* measures the query's information need specificity and *c-DSC* measures the discriminability of concepts across the corpus (Li et al., 2012). In addition, we experimented with minimum, maximum, and average IDF scores based on these visual words, analogously to the common text QPPs (these were not examined in the original paper (Li et al., 2012)).

11.1.3 Visual post-retrieval QPPs

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 (Mikolov et al., 2013) trained over a corpus of 10M titles sampled uniformly at random from the eBay inventory during the month of February 2020 (the same month as our dataset, as described in Sect. 3). 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 (Arora et al., 2017). For visual search, we used a ResNet-50 network (He et al., 2016) to learn image embeddings over more than 50M listing images from

the eBay site, spanning all major categories (Yang et al., 2017).⁸ We applied these embeddings to both the image query and the listing's main image.

We examine the following post-retrieval QPPs (Hirsch et al., 2020): (1) *Num results* - the total number of retrieved results (Carmel & Yom-Tov, 2010; 2) *STD*: the thresholded standard deviation (Cummins et al., 2011), with a 50% threshold; (3) *WIG*: the weighted information gain (Zhao & Croft, 2007) without corpus-based normalization; and (4) *SMV*: the score magnitude and variance (SMV) (Tao & Wu, 2014), 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.⁹ For textual search, these predictors were shown to be highly effective for document retrieval in various studies (Carmel & Yom-Tov, 2010; Shtok et al., 2012; Roitman et al., 2017). For visual search, to the best of our knowledge, we are the first to experiment with post-retrieval QPPs.

11.2 Evaluation

For evaluating the QPPs, we sampled uniformly at random 1000 text queries and 1000 image queries from our logs described in Sect. 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). In addition, we asked the annotators to provide a direct judgement for the query's difficulty. The guidelines explained what a query difficulty means (to what extent the intent of the query can be clearly and accurately understood), and annotators were asked to evaluate the query on a scale of 1 (very difficult) to 5 (very easy). Multiple examples of difficult and easy image and text queries were provided to the annotators. For example, an image with a clear object versus an image with a blurred object or an image where the object in question is not clear due to appearance of multiple non-related objects. Or, a clear wording (e.g., "nike red shirt" or "3ft round area rug") versus vague or ambiguous text ("2003 complete set" or "preowned case").

The Fleiss Kappa (Fleiss, 1971) among the three annotators was 0.91 and 0.82 for relevance judgements of text and image queries, respectively, and 0.86 and 0.84 for direct difficulty judgement of image and text queries, respectively. We used the relevance judgement 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) Kendall rank correlation ($K\tau$). Both of these metrics are commonly used for measuring the effectiveness of QPPs (Carmel & Yom-Tov, 2010; Kurland et al., 2011; Zhao et al., 2008). We use the two metrics for measuring the prediction quality w.r.t both the $AP@10$ values induced by the relevance judgements and w.r.t the direct query difficulty judgements.

11.2.1 Pre-retrieval QPPs

Table 13 presents the performance results for pre-retrieval QPPs for text queries, while Table 14 presents the same results for image queries. For text queries, all predictors

⁸ We also experimented with visual words that were extracted using the above-mentioned SIFT algorithm (Lowe, 2004) as alternative for the ResNet-50 based embedding; however, the achieved results were substantially lower and/or not significantly different.

⁹ For textual search, we also experimented with Okapi-BM25 (Robertson et al., 1994) 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 13 Pre-retrieval QPP performance results for text queries w.r.t relevance (AP@10) and direct judgements

QPP	AP@10		Direct judgement	
	<i>r</i> (<i>p</i>)	<i>K</i> − <i>τ</i> (<i>p</i>)	<i>r</i> (<i>p</i>)	<i>K</i> − <i>τ</i> (<i>p</i>)
Length	.084 (.422)	−.015 (.865)	.116 (.226)	−.013 (.886)
Min IDF	.118 (.257)	.082 (.323)	.005 (.965)	.012 (.884)
Max IDF	.095 (.361)	.023 (.783)	−.014 (.893)	−.091 (.256)
Avg IDF	.103 (.322)	.054 (.509)	−.031 (.768)	−.056 (.479)
Min TF.IDF Var	.112 (.284)	.139 (.093)	.198 (.056)	.234 (.004)
Max TF.IDF Var	.148 (.155)	.014 (.862)	.187 (.071)	.132 (.098)
Avg TF.IDF Var	.170 (.102)	.048 (.557)	.24 (.02)	.189 (.018)

Boldfaced correlations (Pearson's *r* and Kendall's *τ*) are statistically significant for $p < .01$

(as also described in Hirsch et al. 2020) did not show statistically significant correlations with either the relevance or the direct judgements. For image , it can be seen that a variety of predictors demonstrated statistically significant performance. One such query-based predictor 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 (Talebi & Milanfar, 2018), however, did not demonstrate a similar performance. In Figure 3, examples 7 and 8 show image queries with a high catalog quality, but low aesthetic quality score (a funkopop collectible toy and a handbag, respectively), while example 9 shows an image query with a high aesthetic quality and low catalog quality score (only a small part of the product is shown). The brightness QPP showed statistically significant correlations, but only with the direct judgement evaluation, perhaps as it is a straightforwardly visual characteristic.

The highest correlations across image pre-retrieval QPPs was attained by the top *k* category predictors. Recall we considered the probability sum of the top *k* predicted LCs as well as the ratio between the top LC to the second most probable. It can be seen that all category-related predictors yielded statistically significant correlations for both AP@10 and direct judgments. The highest correlations were achieved by merely considering the probability of the top LC or its ratio to the probability of the second most probable, indicating that image queries whose category can be clearly identified yield good retrieval performance. The QPPs based on object detection did not yield any significant results, suggesting no obvious correlation between the number of objects or their area in the image to its performance as a query.

The corpus-based predictors that consider similarity to the inventory's collection of images yielded significant performance. The maximum similarity with the corpus yielded significant performance w.r.t both relevance and direct judgments, while minimum and average similarity yielded significant results only w.r.t relevance and direct judgements, respectively. The two previously-proposed QPPs based on visual words (Li et al., 2012) 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 Sect. 5 and demonstrated in Figure 3, which are focused on objects for purchase rather

Table 14 Pre-retrieval QPP performance results for image queries w.r.t relevance (AP@10) and direct judgements

QPP	AP@10		Direct judgement	
	<i>r</i> (<i>p</i>)	<i>K</i> − <i>τ</i> (<i>p</i>)	<i>r</i> (<i>p</i>)	<i>K</i> − <i>τ</i> (<i>p</i>)
Size	.052 (.464)	.002 (.972)	.085 (.233)	.113 (.067)
Brightness	.062 (.383)	.074 (.195)	.091 (.006)	.165 (.002)
Luminance	.068 (.341)	.060 (.291)	.142 (.044)	.114 (.033)
Catalog Quality	.233 (.001)	.175 (.002)	.018 (.011)	.196 (< .001)
Aesthetics	−.082 (.252)	−.081 (.155)	−.050 (.478)	−.055 (.298)
1-Color	−.008 (0.912)	−.042 (.455)	.1 (.158)	.028 (.602)
2-Color	.021 (.766)	.002 (.969)	.084 (.238)	.04 (.459)
3-Color	.029 (.687)	.023 (.685)	.089 (.211)	.061 (.252)
Top-1 categories	.342 (.004)	.322 (.006)	.402 (.006)	.336 (.009)
Top-3 categories	.173 (.006)	.211 (.005)	.194 (.009)	.239 (.002)
Top-5 categories	.163 (.005)	.201 (.003)	.164 (.008)	.232 (.002)
1-to-2 categories ratio	.241 (< .001)	.319 (< .001)	.417 (< .001)	.385 (< .001)
# of objects	.106 (.264)	.114 (.145)	.133 (.314)	.102 (.499)
% of largest object' area	.163 (.543)	.162 (.269)	.143 (.244)	.086 (.272)
% of all objects' area	.217 (.427)	.121 (.183)	.052 (.394)	.092 (.312)
Min Corpus Sim	−.217 (.002)	−.163 (.004)	−.052 (.468)	−.059 (.267)
Max Corpus Sim	.273 (< .001)	.198 (< .001)	.313 (< .001)	.242 (< .001)
Avg Corpus Sim	.074 (.301)	.038 (.509)	.300 (< .001)	.164 (.002)
q-INS	.123 (.194)	.114 (.105)	.102 (.118)	.081 (.106)
c-DCS	.166 (.104)	.121 (.088)	.305 (.118)	.175 (.093)
Min IDF (visual words)	.108 (.252)	.124 (.183)	.032 (.168)	.094 (.298)
Max IDF (visual words)	.136 (.284)	.106 (.334)	.096 (.193)	.087 (.315)
Avg IDF (visual words)	.098 (.239)	.143 (.271)	.115 (.225)	.103 (.263)

Boldfaced correlations (Pearson's *r* and Kendall's *τ*) are statistically significant for *p*<.01

than scenes. Predictors based on visual words IDF showed insignificant correlations, similarly to their counterparts on text queries, as reported in Table 13.

Overall, we identified both query-based and corpus-based pre-retrieval QPPs that demonstrate high performance for image queries.

11.2.2 Post-retrieval QPPs

Table 15 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 (Carmel & Yom-Tov, 2010; Hirsch et al., 2020). For image queries, STD results were insignificant, while SMV was only significant by the *K* − *τ* metric. The WIG predictor, on the other hand, demonstrated significant results and yielded the best prediction by both the Pearson's *r* and Kendall's *τ* metrics with the relevance judgements 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. In summary, our experimentation with QPP for image queries showed various pre- and post-retrieval QPPs

Table 15 Post-retrieval QPP performance results w.r.t relevance (AP@10) and direct judgements

QPP	Text				Image			
	AP@10		Direct judgement		AP@10		Direct judgement	
	<i>r</i> (<i>p</i>)	<i>K</i> - τ (<i>p</i>)	<i>r</i> (<i>p</i>)	<i>K</i> - τ (<i>p</i>)	<i>r</i> (<i>p</i>)	<i>K</i> - τ (<i>p</i>)	<i>r</i> (<i>p</i>)	<i>K</i> - τ (<i>p</i>)
# Results	-.097 (.256)	.095 (.003)	-.004 (.964)	.234 (.001)	-.076 (.302)	.014 (.809)	.043 (.561)	.026 (.636)
STD	-.210 (.013)	-.286 (< .001)	-.221 (.014)	-.099 (.134)	.038 (.605)	-.064 (.273)	-.018 (.803)	-.035 (.528)
WIG	.206 (.015)	.144 (.027)	.296 (< .001)	.071 (.283)	.562 (< .001)	.449 (< .001)	.147 (.043)	.173 (< .001)
SMV	-.236 (.005)	-.292 (< .001)	-.254 (.003)	-.126 (.057)	-.156 (.033)	-.188 (.001)	-.086 (.242)	-.064 (.243)

Boldfaced correlations (Pearson's *r* and Kendall's τ) are statistically significant for $p < .01$

with significant prediction performance, leaving room for further development of effective QPPs for visual search. The broad set of effective pre-retrieval QPPs, including catalog quality, category association, brightness, and corpus similarity, indicate that image queries contain rich information that enable better performance prediction than for text queries, even without inspecting the retrieved results.

12 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 (Liao et al., 2018; Bhardwaj et al., 2013; Bell et al., 2020; Shiao et al., 2020; Kang et al., 2019; Kim et al., 2016; Laenen et al., 2018), which exhibits many visual characteristics. 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. On the other hand, categories that require many textual specification, such as Electronics or areas in Fashion that include characteristics that cannot be easily captured visually (e.g. size in sneakers) are less popular on visual search. 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. For example, more attributes related to vintage products, such as year of manufacture, may be presented for image queries.

Search broadness Previous work (Bhardwaj et al., 2013; Zhang et al., 2018) 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 (Laenen et al., 2018; Zhang et al., 2018; Wróblewska & Rączkowski, 2016) 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 (Hsiao et al., 2014). 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 (Laenen et al., 2018) or using an image to refine a textual search. This can help customers articulate their needs more easily in the multi-modal e-commerce domain, since some aspects are easy to express by image, while others are easier to articulate using text.

Sessions Search sessions capture user behavior often focused on one search goal, spanning multiple queries, SERP traversals, and interactions with the search results. Our results indicate that image sessions tend to be longer than text sessions, and more rarely consist of a single query solely. The length of image sessions is reflected, beyond the higher number of queries per session, in the duration of the session and the idle time between queries in the session. Overall, a visual search session is an experience that lasts more time, either because it is more engaging or because it requires more effort on the part of the user. In any of these cases, visual search systems should improve the tools they provide for users to navigate along their search sessions, allowing to more easily refine, expand, and edit their queries along the session.

Visual similarity bias User interaction with search results introduces a variety of biases, such as position bias and trust bias (Wang et al., 2016, 2018). This bias influences the way user interaction with the results can be interpreted in user behavior models, such as click models (Chuklin et al., 2015). Our experiments suggest a new type of bias introduced to visual search, towards results with particularly similar images to the query. By contrast, visual e-commerce search is not purely based on image similarity, and includes a variety of other features. We demonstrated that this type of bias is unique to visual search, and consistent in both camera and gallery flow. In contrast, it is not observed in a similar setup of textual search, for results with similar titles. While previous work has identified an “examine bias” for multimedia results when appearing in mix with textual results (Wang et al., 2013), the bias we observe occurs across results of similar nature, retrieved for a visual e-commerce query. Future work focusing on user behavioral models for visual search should account for this bias and provides means model it when comparing different retrieval models.

User intent Image queries are used for two principal intents (Su et al., 2018; Togashi & Sakai, 2020): *target finding* desires to look up a specific item (Laenen et al., 2018), while *decision making* aims at discovery of visually-similar items (Zhai et al., 2017). 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 (Broder, 2002). 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 (Guy, 2016), 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 analysis indicates bias towards listings with a similar image to the image query, which may call for different click models and a more prominent presentation of images on the results page. 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. Our work expands the list of visual QPPs we experimented with, to also include predictors that consider the category of the query; the number of objects of the image query and their size; and richer surface characteristics, such as luminance and the portion of pixels with the top 1 and 2 most dominant colors. The evaluation of the QPP is also expanded to include direction judgement of the query difficulty, alongside the traditional correlation of the predictors with results' relevance judgements. Among the most effective pre-retrieval visual QPPs are those that consider the image's catalog quality, its similarity to the entire corpus, and its categories. The most effective post-retrieval visual QPPs include the score magnitude variance (SMV) and, most notably, the weighted information gain (WIG). The variety of effective pre-retrieval QPPs indicate that image queries can be used more effectively to predict the retrieval performance before applying it. 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 (even if the category themselves are different), the bias towards results with similar images, and performance prediction.

Item price Price plays a central role on e-commerce search, as buyers aim to achieve the best deal for the products they seek. Our results suggested that results on visual search are generally less expensive than for textual search, perhaps characterizing the early adoption stage, where users are more reluctant to use a new technology for a more significant transaction. Nonetheless, there was a noticeable difference between camera and gallery queries, with the latter yielding more expensive results. This implies that for more expensive goods, users would more likely make an effort to search for an appropriate gallery image, whereas camera images are more often used for casual search for low-cost items.

Visual search users As visual search use is still in its infancy, its group of users is a subset of e-commerce search users, representing early adopters of the technology. Our analysis delves into the behavioral characteristics of this unique group of users and indicates that some of the differences between visual and textual search stem from the unique characteristics of the (small) user subset who is using visual search. That said, the reported differences between visual and textual search remain considerable even after controlling for this difference, comparing visual and textual search across the same set of users. We refine our findings as reported here and in our previous work (Dagan et al., 2021) by distinguishing the contribution of the difference stemming from user groups and the contribution stemming from the actual search type (textual versus visual). Overall, a substantial portion of the difference persists even when considering the same set of users, and can thus be attributed to the actual difference between visual and textual search.

12.1 Limitations

The findings of this work are based on e-commerce search – visual and textual – at eBay. As one of the world's largest e-commerce platforms, we believe eBay well represents how search operates at scale over both textual and visual queries. Yet, the many implementation details and algorithmic choices made by the eBay's search engine can influence the results presented in this study. These include click-through rates, category distribution on the SERP, the use of attributes and global filters to refine searches, and session characteristics. We believe that the scale of this study, based on over 1.5 million visual and 1.5 million textual queries, provides a robust foundation for the analysis we presented. Future research should validate and contrast these results with findings over other e-commerce platforms, to see how they generalize beyond eBay. This work provides a significant first step towards deep understanding of visual e-commerce search and its differences from textual search.

Another caveat we wish to remind here relates to the relative sparsity of data used for the query refinement analysis. Since refinement is rather rare in itself and since different segmentation based on filter types and values render much smaller datasets, the results for this part of the analysis should be treated with extra care. While they mark clear trends, these are often based on only a few thousands of data points and may therefore be less robust than the rest of the findings presented in this work.

This work emphasizes the benefits of using images for visual search, but it should also be noted that some drawback exist. First, it requires either the use of a camera to capture a photo or the upload of a gallery image, both of which may often be less accessible than merely typing a query using a keyboard or a touch screen. Second, the processing of image queries requires more computational power from the search system. Finally, many e-commerce information needs are hard to express by image, from size and material aspects, through brand and model names, to technical specifications.

Finally, as visual search use is in early stages, its usage patterns may further evolve in the years to come. The results presented in this study represent early adoption by users whose characteristics are somewhat different than the general population using e-commerce search, as demonstrated in our user analysis section. Future research should further explore the characteristics and motivations of users who adopt visual search. We also hope to track the evolution of visual e-commerce search as it is destined to become more popular in the years to come.

12.2 Future directions

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 (Hirsch et al., 2020), 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. Editing and adding images can also help encourage query refinement, since the use of refining attributes is low with the current interfaces that are inherited “as is” from textual 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 (Shiau et al., 2020). 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.

One of the benefits of visual search, as mentioned in the introduction, is allowing users to search without familiarity of the domain's terminology (e.g., in special categories of Fashion or Collectibles). Once results are presented, users may learn the jargon by viewing the textual facets of the returned items, such as title, description, and attributes. Future research may further inspect user transition between visual and textual searches in specific domains. 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 (Carmel et al., 2020; Tsagkias et al., 2020). 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

- Arora, Sanjeev, Liang, Yingyu, & Ma, Tengyu. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of ICLR*.
- Baeza-Yates, Ricardo, Dupret, Georges, & Velasco, Javier. (2007). A study of mobile search queries in Japan. In *Query Log Analysis (WWW'07 workshop)*.
- Bell, Sean, Liu, Yiqun, Alsheikh, Sami, Tang, Yina, Pizzi, Edward, Henning, M., Singh, Karun, Parkhi, Omkar, & Borisjuk, Fedor. (2020). GrokNet: unified computer vision model trunk and embeddings for commerce. In *Proceedings of KDD*. 2608–2616.
- Berger, Adam, & Lafferty, John. (2017). Information retrieval as statistical translation. *SIGIR Forum*, 51(2), 219–226.
- Bezryadin, Sergey, Bourov, Pavel, & Ilinih, Dmitry. (2007). Brightness calculation in digital image processing. In *proceedings of TDPF*. 10–15.
- Bhardwaj, Anurag, Sarma, Atish Das, Di, Wei, Hamid, Raffay, Piramuthu, Robinson, & Sundaresan, Neel. (2013). Palette Power: Enabling Visual Search through Colors. In *Proceedings of KDD*. 1321–1329.
- Bhattacharya, Indrani, Chowdhury, Arkabandhu, & Raykar, Vikas C. (2019). Multimodal dialog for browsing large visual catalogs using exploration-exploitation paradigm in a joint embedding space. In *proceedings of ICMR*. 187–191.
- Bitirim, Yiltan, Bitirim, Selin, Ertugrul, Duygu Celik, & Toygar, Onsen. (2020). An Evaluation of reverse image search performance of Google. In *Proceedings of COMPSAC*. pp 1368–1372.
- Blei, David M., Ng, Andrew Y., & Jordan, Michael I. (2003). Latent dirichlet allocation. *JMLR*, 3(2003), 993–1022.
- Broder, Andrei. (2002). A taxonomy of web search. In *ACM Sigir forum*, 36, 3–10.
- Burney, SM Aqil., & Tariq, Humera. (2014). K-means cluster analysis for image segmentation. *International Journal of Computer Applications*, 96, 4.
- Carmel, David, Haramaty, Elad, Lazerson, Arnon, Lewin-Eytan, Liane, & Maarek, Yoelle. (2020). Why do people buy seemingly irrelevant items in voice product search? On the relation between product relevance and customer satisfaction in ecommerce. In *Proceedings of WSDM*. pp. 79–87.
- Carmel, David, & Yom-Tov, Elad. (2010). *Estimating the query difficulty for information retrieval*. Morgan & Claypool Publishers.
- Chaudhuri, Abon, Messina, Paolo, Kokkula, Samrat, Subramanian, Aditya, Krishnan, Abhinandan, Gandhi, Shreyansh, Magnani, Alessandro, & Kandaswamy, Venkatesh. (2018). A smart system for selection of optimal product images in e-commerce. In *Proceedings of Big Data*. pp. 1728–1736.
- Chuklin, Aleksandr, Markov, Ilya, & de Rijke, Maarten. (2015). Click models for web search. *Synthesis lectures on information concepts, retrieval, and services*, 7(3), 1–115.
- Cronen-Townsend, Steve, Zhou, Yun, & Croft, W Bruce. (2002). Predicting query performance. In *Proceedings of SIGIR*. pp 299–306.
- Cummins, Ronan, Jose, Joemon M., & O'Riordan, Colm. (2011). Improved query performance prediction using standard deviation. In *Proceedings of SIGIR*. pp 1089–1090.
- Dagan, Arnon, Guy, Ido, & Novgorodov, Slava. (2021). An image is worth a thousand terms? Analysis of Visual E-Commerce Search. In *Proceedings of SIGIR*. pp 102–112.

- Datta, Ritendra, Joshi, Dhiraj, Li, Jia, & Wang, James Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2), 1–60.
- Datta, Ritendra, Li, Jia, & Wang, James Z. (2005). Content-Based Image Retrieval: Approaches and trends of the new age. In *Proceedings of MIR*.pp 253–262.
- Robertson, Stephen E., Walker, Steve, Jones, Susan, Hancock-Beaulieu, Micheline, & Gatford, Mike. (1994). Okapi at TREC-3. In *Proceedings of TREC-3*.
- Elkasrawi, Sarah, Dengel, Andreas, Abdelsamad, Ahmed, & Bukhari, Syed Saqib. (2016). What you see is what you get? Automatic Image Verification for Online News Content. In *Proceedings of DAS*. pp 114–119.
- Fleiss, Joseph L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5), 378–382.
- Gandomkar, Ziba, & Mello-Thoms, Claudia. (2019). Visual search in breast imaging. *The British journal of radiology*, 92(1102), 20190057.
- Goel, Nishant. (2017). Shopbot: An image based search application for e-commerce domain.
- Goodrum, Abby, & Spink, Amanda. (2001). Image searching on the Excite Web search engine. *Information Processing & Management*, 37(2), 295–311.
- Guy, Ido. (2016). Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of SIGIR*. pp 35–44.
- Hadwiger, Benjamin, & Riess, Christian. (2020). The Forchheim Image Database for Camera Identification in the Wild. *arXiv preprint arXiv:2011.02241*.
- Hauff, Claudia, Azzopardi, Leif, & Hiemstra, Djoerd. (2009). The combination and evaluation of query performance prediction methods. In *Proceedings of ECIR*. pp 301–312.
- Hauff, Claudia, Hiemstra, Djoerd, & de Jong, Franciska. (2008). A survey of pre-retrieval query performance predictors. In *Proceedings of CIKM*.pp 1419–1420.
- He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, & Sun, Jian. (2016). Deep residual learning for image recognition. In *Proceedings of CVPR*. pp. 770–778.
- Hegde, Narayan, Hipp, Jason D., Liu, Yun, Emmert-Buck, Michael, Reif, Emily, Smilkov, Daniel, Terry, Michael, Cai, Carrie J., Amin, Mahul B., Mermel, Craig H., et al. (2019). Similar image search for histopathology: SMILY. *NPJ digital medicine*, 2(1), 1–9.
- Hirsch, Sharon, Guy, Ido, Nus, Alexander, Dagan, Arnon, & Kurland, Oren. (2020). Query Reformulation in E-Commerce Search. In *Proceedings of SIGIR*. pp. 1319–1328.
- Hsiao, Jen-Hao, & Li, Li-Jia. (2014). On Visual Similarity based Interactive Product Recommendation for Online Shopping. In *Proceedings of ICIP*. pp. 3038–3041.
- Hu, Houdong, Wang, Yan, Yang, Linjun, Komlev, Pavel, Huang, Li, Chen, Xi (Stephen), Huang, Jiawei, Wu, Ye, Merchant, Meenaz, & Sacheti, Arun. (2018). Web-Scale Responsive Visual Search at Bing. In *Proceedings of KDD*. pp. 359–367.
- Jansen, Bernard J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, 28(3), 407–432.
- Jansen, Bernard J., & Spink, Amanda. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information processing & management*, 42(1), 248–263.
- Jiang, Hao, Sabharwal, Aakash, Henderson, Adam, Hu, Diane, & Hong, Liangjie. (2019). Understanding the role of style in e-commerce shopping. In *Proceedings of KDD*.pp. 3112–3120.
- Jing, Yushi, Liu, David, Kislyuk, Dmitry, Zhai, Andrew, Xu, Jiajing, Donahue, Jeff, & Tavel, Sarah. (2015). Visual Search at Pinterest. In *Proceedings of KDD*. pp. 1889–1898.
- Jones, Rosie, & Klinkner, Kristina Lisa. (2008). Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of CIKM*. pp. 699–708.
- Kamvar, Maryam, & Baluja, Shumeet. (2006). A large scale study of wireless search behavior: Google mobile search. In *Proceedings of CHI*. pp. 701–709.
- Kang, Wang-Cheng, Kim, Eric, Leskovec, Jure, Rosenberg, Charles, & McAuley, Julian. (2019). Complete the look: Scene-based complementary product recommendation. In *Proceedings of CVPR*. pp. 10532–10541.
- Kim, Taewan, Kim, Seyeong, Na, Sangil, Kim, Hayoon, Kim, Moonki, & Jeon, Byoung-Ki. (2016). Visual fashion-product search at sk planet. *arXiv preprint arXiv:1609.07859*.
- Kurland, Oren, Shtok, Anna, Carmel, David, & Hummel, Shay. (2011). A unified framework for post-retrieval query-performance prediction. In *Proceedings of ICTIR*. pp. 15–26.
- Laenen, Katrien, Zoghbi, Susana, & Moens, Marie-Francine. (2018). Web search of fashion items with multimodal querying. In *Proceedings of WSDM*. pp. 342–350.
- Li, Bing, Duan, Ling-Yu, Chen, Yiming, Ji, Rongrong, & Gao, Wen. (2012). Predicting the effectiveness of queries for visual search. In *Proceedings of ICASSP*. pp. 2361–2364.

- Li, Eileen, Kim, Eric, Zhai, Andrew, Beal, Josh, & Gu, Kunlong. (2020b). Bootstrapping Complete The Look at Pinterest. In *Proceedings of KDD*. 3299–3307.
- Li, Fengzi, Kant, Shashi, Araki, Shunichi, Bangera, Sumer, & Shukla Swapna Samir. (2020a). Neural networks for fashion image classification and visual search. *arXiv preprint arXiv:2005.08170*.
- Li, Jie, Liu, Haifeng, Gui, Chuanghua, Chen, Jianyu, Ni, Zhenyuan, Wang, Ning, & Chen, Yuan. (2018). The Design and Implementation of a Real Time Visual Search System on JD E-Commerce Platform. In *Proceedings of Middleware*. pp. 9–16.
- Liao, Lizi, He, Xiangnan, Zhao, Bo, Ngo, Chong-Wah, & Chua, Tat-Seng. (2018). Interpretable multi-modal retrieval for fashion products. In *Proceedings of MM*. pp. 1571–1579.
- Lien, Yen-Chieh, Zamani, Hamed, & Croft, W. Bruce. (2020). Recipe Retrieval with Visual Query of Ingredients. In *Proceedings of SIGIR*. pp. 1565–1568.
- Lin, Kevin, Yang, Fan, Wang, Qiaosong, & Piramuthu, Robinson. (2019). Adversarial Learning for Fine-Grained Image Search. In *Proceedings of ICME*. pp. 490–495.
- Lowe, David G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91–110.
- McAuley, Julian, Targett, Christopher, Shi, Qinfeng, & van den Hengel, Anton. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of SIGIR*. pp. 43–52.
- Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S, & Dean, Jeff. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*. pp. 3111–3119.
- Misraa, Aashish Kumar, Kale, Ajinkya, Aggarwal, Pranav, & Aminian, Ali. (2020). Multi-modal retrieval using graph neural networks. *arXiv preprint arXiv:2010.01666*.
- Aggarwal, P. (2018). Fashion Dataset. <https://www.kaggle.com/paramaggarwal/fashion-product-images-dataset>.
- Ozdemiray, Ahmet Murat, & Altıngöve, İsmail Sengör. (2014). Query Performance Prediction for Aspect Weighting in Search Result Diversification. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*. pp. 1871–1874. <https://doi.org/10.1145/2661829.2661975>.
- Parikh, Viken, Keskar, Madhura, Dharia, Dhwanil, & Gotmare, Pradnya. (2018). A tourist place recommendation and recognition system. In *Proceedings of ICICCT*. pp. 218–222.
- Ponte, Jay M., & Croft, W. Bruce. (1998). A language modeling approach to information retrieval. In *Proceedings of SIGIR*. pp. 275–281.
- Redmon, Joseph, Divvala, Santosh, Girshick, Ross, & Farhadi, Ali. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 779–788.
- Reilly, Michele, & Thompson, Santi. (2017). Reverse image lookup: Assessing digital library users and reuses. *Journal of Web Librarianship*, 11(1), 56–68.
- Roitman, Haggai, Erera, Shai, Shalom, Oren Sar, & Weiner, Bar. (2017). Enhanced mean retrieval score estimation for query performance prediction. In *Proceedings of ICTIR*. pp. 35–42.
- Saez-Trumper, Diego. (2014). Fake tweet buster: a webtool to identify users promoting fake news on twitter. In *Proceedings of HT*. pp. 316–317.
- Shapovalov, Yevhenii B., Bilyk, Zhanna I., Atamas, Artem I., Shapovalov, Viktor B., & Uchitel, Aleksandr D. (2018). the potential of using google expeditions and google lens tools under stem-education in Ukraine. *arXiv preprint arXiv:1808.06465*.
- Shiau, Raymond, Wu, Hao-Yu, Kim, Eric, Du, Yue Li, Guo, Anqi, Zhang, Zhiyuan, Li, Eileen, Gu, Kunlong, Rosenberg, Charles, & Zhai, Andrew. (2020). Shop the look: Building a large scale visual shopping system at pinterest. In *Proceedings of KDD*. pp. 3203–3212.
- Shtok, Anna, Kurland, Oren, Carmel, David, Raiber, Fiana, & Markovits, Gad. (2012). Predicting query performance by query-drift estimation. *ACM TOIS*, 30(2), 11.
- Singh, Gyani, Parikh, Nish, & Sundaesan, Neel. (2012). Rewriting null e-commerce queries to recommend products. In *Proceedings of WWW Companion*. pp. 73–82.
- Sondhi, Parikshit, Sharma, Mohit, Kolari, Pranam, & Zhai, ChengXiang. (2018). A taxonomy of queries for e-commerce search. In *Proceedings of SIGIR*. pp. 1245–1248.
- Song, Yang, Ma, Hao, Wang, Hongning, & Wang, Kuansan. (2013). Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In *Proceedings of WWW*. pp. 1201–1212.
- Su, Ning, He, Jiyan, Liu, Yiqun, Zhang, Min, & Ma, Shaoping. (2018). User intent, behaviour, and perceived satisfaction in product search. In *Proceedings of WSDM*. pp. 547–555.
- Talebi, Hossein, & Milanfar, Peyman. (2018). NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8), 3998–4011.

- Tao, Yongquan, & Wu, Shengli. (2014). Query performance prediction by considering score magnitude and variance together. In *Proceedings of CIKM*. pp. 1891–1894.
- Tian, Huawei, Xiao, Yanhui, Cao, Gang, Zhang, Yongsheng, Zhiyin, Xu., & Zhao, Yao. (2019). Daxing smartphone identification dataset. *IEEE Access*, 7(2019), 101046–101053.
- Togashi, Riku, & Sakai, Tetsuya. (2020). Visual intents vs. clicks, likes, and purchases in e-commerce. In *Proceedings of SIGIR*. pp. 1869–1872.
- Tsakias, Manos, King, Tracy Holloway, Kallumadi, Surya, Murdock, Vanessa, & de Rijke, Maarten. (2020). Challenges and research opportunities in ecommerce search and recommendations. *SIGIR Forum*, 54, 1.
- Tunkelang, Daniel. (2009). Faceted search. *Synthesis lectures on information concepts, retrieval, and services*, 1(1), 1–80.
- Wan, Ji, Wang, Dayong, Hoi, Steven Chu Hong, Wu, Pengcheng, Zhu, Jianke, Zhang, Yongdong, & Li, Jintao. (2014). Deep Learning for Content-Based Image Retrieval: A comprehensive study. In *Proceedings of MM*. pp. 157–166.
- Wang, Chao, Liu, Yiqun, Zhang, Min, Ma, Shaoping, Zheng, Meihong, Qian, Jing, & Zhang, Kuo. (2013). Incorporating vertical results into search click models. In *Proceedings of SIGIR*. pp. 503–512.
- Wang, Xuanhui, Bendersky, Michael, Metzler, Donald, & Najork, Marc. (2016). Learning to rank with selection bias in personal search. In *Proceedings of SIGIR*. pp. 115–124.
- Wang, Xuanhui, Golbandi, Nadav, Bendersky, Michael, Metzler, Donald, & Najork, Marc. (2018). Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of WSDM*. pp. 610–618.
- Wróblewska, Anna, & Rączkowski, Łukasz. (2016). Visual Recommendation Use Case for an Online Marketplace Platform: Allegro.Pl. In *Proceedings of SIGIR*. pp. 591–594.
- Yang, Fan, Kale, Ajinkya, Bubnov, Yury, Stein, Leon, Wang, Qiaosong, Kiapour, Hadi, & Piramuthu, Robinson. (2017). Visual Search at EBay. In *Proceedings of KDD*. pp. 2101–2110.
- Young, Peter, Lai, Alice, Hodosh, Micah, & Hockenmaier, Julia. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *ACL*, 2(2014), 67–78.
- Zhai, Andrew, Kislyuk, Dmitry, Jing, Yushi, Feng, Michael, Tzeng, Eric, Donahue, Jeff, Du, Yue Li, & Darrell, Trevor. (2017). Visual Discovery at Pinterest. In *Proceedings of WWW Companion*. pp. 515–524.
- Zhai, Andrew, Wu, Hao-Yu, Tzeng, Eric, Park, Dong Huk, & Rosenberg, Charles. (2019). Learning a unified embedding for visual search at pinterest. In *Proceedings of KDD*. pp. 2412–2420.
- Zhang, Yanhao, Pan, Pan, Zheng, Yun, Zhao, Kang, Wu, Jianmin, Xu, Yinghui, & Jin, Rong. (2019). Virtual ID discovery from e-commerce media at alibaba: Exploiting richness of user click behavior for visual search relevance. In *Proceedings of CIKM*. pp. 2489–2497.
- Zhang, Yanhao, Pan, Pan, Zheng, Yun, Zhao, Kang, Zhang, Yingya, Ren, Xiaofeng, & Jin, Rong. (2018). Visual search at Alibaba. In *Proceedings of KDD*. pp. 993–1001.
- Zhao, Ying, Scholer, Falk, & Tsegay, Yohannes. (2008). Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Proceedings of ECIR*. pp. 52–64.
- Zhou, Yun, & Croft, W. Bruce. (2007). Query performance prediction in web search environments. In *Proceedings of SIGIR*. pp. 543–550.
- Zhu, Bin, Ngo, Chong-Wah, Chen, Jingjing, & Hao, Yanbin. (2019). R2GAN: Cross-modal recipe retrieval with generative adversarial network. In *Proceedings of CVPR*. pp. 11477–11486.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.