A Video-Based Automated Recommender (VAR) System for Garments

Shasha Lu, Li Xiao, Min Ding*

Feburary 4, 2016

*Shasha Lu is Lecturer in Marketing at Cambridge Judge Business School, University of Cambridge, Cambridge, UK, CB2 1AG (email: s.lu@jbs.cam.ac.uk); Li Xiao is Assistant Professor of Marketing at School of Management, Fudan University, Shanghai, PRC 200433 (email: lixiao@fudan.edu.cn); Min Ding is the Bard Professor of Marketing, Smeal College of Business, Pennsylvania State University, University Park, PA 16802-3007 and Advisory Professor of Marketing, School of Management, Fudan University, Shanghai, PRC 200433 (email: minding@psu.edu). The authors thank the participants in presentations given by the authors in College of Business at City Univeristy of HongKong and Cambridge Judge Business School for their feedback, as well as the Editor, the Area Editor, and two anonymous *Marketing Science* reviewers for their insightful comments. This research was supported by two National Natural Science Foundation of China Fund (Grants 71232008 & 71502039), and the Institute for Sustainable Innovation and Growth (iSIG) at School of Management, Fudan University.

A Video-Based Automated Recommender (VAR) System for Garments

Abstract

In this paper, we propose an automated and scalable garment recommender system using real time in-store videos that can improve the experiences of garment shoppers and increase product sales. The video-based automated recommender (VAR) system is based on observations that garment shoppers tend to try on garments and evaluate themselves in front of store mirrors. Combining state-of-the-art computer vision techniques with marketing models of consumer preferences, the system automatically identifies shoppers' preferences based on their reactions and uses that information to make meaningful personalized recommendations. First, the system uses a camera to capture a shopper's behavior in front of the mirror to make inferences about her preferences based on her facial expressions and the part of the garment she is examining at each time point. Second, the system identifies shoppers with preferences similar to the focal customer from a database of shoppers whose preferences, purchasing and/or consideration decisions are known. Finally, recommendations are made to the focal customer based on the preferences, purchasing and/or consideration decisions of these like-minded shoppers. Each of the three steps can be implemented with several variations, and a retailing chain can choose the specific configuration that best serves its purpose. In this paper, we present an empirical test that compares one specific type of VAR system implementation against two alternative, *non-automated* personal recommender systems: self-explicated conjoint (SEC) and self-evaluation after try-on (SET). The results show that VAR consistently outperforms SEC and SET. A second empirical study demonstrates the feasibility of VAR in real time applications. Participants in the second study enjoyed the VAR experience, and almost all of them tried on the recommended garments. VAR should prove to be a valuable tool for both garment retailers and shoppers.

Keywords: Retailing, Video Analysis, Collaborative Filtering

1. Introduction

Clothing and accessories retailer Zara launches 11,000 designs on store shelves every year. Although a retail chain like Zara usually maintains a wide variety of products at its warehouses, each store displays only a much smaller set of items due to space constraints. Even so, evaluating all of the items in a single store would require significant time and cognitive effort from customers. Since they only evaluate a subset of the available inventory, shoppers may leave the store without finding anything they like, even when the retailer may carry products that appeal to their preferences—at that store, a different store, or the warehouse.

To address this problem in retail stores, it is crucial for retailers to understand customer preferences and recommend products accordingly. To this end, retailers usually adopt two strategies: (a) train salespeople to discern customer preferences based on solicited and/or unsolicited feedback from individuals and provide appropriate recommendations; and (b) use marketing research tools such as conjoint analysis to understand customer preferences for a variety of product features and make recommendations accordingly.

However, these two strategies have several limitations. The first strategy is expensive because an experienced salesperson must be paid well and more employees must be hired in order for the strategy to be effective. In addition, salespeople have a limited ability to remember every product in a company's available inventory and the quality of recommendations from salespeople may vary widely. Furthermore, some customers may not like being observed by salespeople, and may even be suspicious of their recommendations. The second strategy is rarely used to provide personalized customer recommendations in the real world, as the costs associated with asking every customer in a store to complete a conjoint task during the shopping process would be significant. The recommendations are thus based on average preferences; this is clearly a suboptimal strategy, since individuals are very likely to have heterogeneous preferences.

To address these problems in retail stores, we have created an automated and scalable garment recommender system using real time in-store videos that can improve the experiences of garment shoppers and increase product sales while requiring minimal extra effort from customers. The video-based automated recommender (VAR) system is based on observations that garment shoppers tend to try on garments and evaluate themselves in front of store mirrors. Combining state-of-the-art computer vision techniques and marketing models of consumer preferences, the system automatically identifies shoppers' preferences based on

their reactions and uses that information to make meaningful personalized recommendations. First, the system uses a camera to capture a shopper's behavior in front of the mirror to make inferences about her preferences based on her facial expressions and the part of the garment she is examining (region of interest) at each time point. Second, the system matches this shopper with a database of shoppers whose preferences, purchasing and/or consideration decisions are known, and identifies a small set of shoppers with similar preferences to the focal customer. Finally, recommendations are made to the focal customer based on the preferences, purchasing and/or consideration decisions of these like-minded shoppers (referred to as *nearest neighbors* in this study). Each of the three steps can be implemented with several variations, and a retailing chain can choose the specific configuration that best serves its purpose.

While the system can be used mainly for making real time individualized recommendations, it also provides useful self-evaluation tools for customers (e.g., allowing prior try-on videos and/or frames to be played and compared later). The information extracted from try-on videos across customers and garments may also provide valuable insights in the big data framework. To the best of our knowledge, this is one of the first attempts to integrate video analysis (real time facial expression recognition and hand detection) at the individual customer level with extant marketing research methods to create useful managerial tools in the retail context. Our aim is to demonstrate a proof of concept that the information inferred from video (automated detection of customers' facial expressions and regions of interest, in our case) has value in predicting customers' individual preferences toward different garments with minimal extra effort on their part. While the potential benefits of our model to retailers and customers are promising, we believe that many possible variations could be implemented. We hope that the present study will serve as a meaningful starting point for future work on the automated extraction of information from video to help both consumers and retailers.

The rest of this article is organized as follows. First, we review literature on video data in marketing and relevant practices in the retailing context. Second, we describe the system, its general structure and variations, and how each step is implemented. We then describe an empirical study (Study 1) designed to test the validity of the system by comparing one specific implementation of the system with two non-automated individual level recommender systems. The feasibility and value of the system in the real world is demonstrated in a second empirical study (Study 2). This is followed by a section on building a scalable VAR system for retail chains, and

a section on mining valuable information from large scale try-on video data. We conclude with a general discussion on the use of video analysis in marketing.

2. Literature Review

In this section, we begin by describing challenges associated with inferring customer preferences in retail stores. Then, we provide a brief review of extant video-based research in marketing, as well as potential uses of video data. Although we do not review the vast literature on video analysis from the computer science discipline in this section, we do discuss relevant literature in our methods and implementation sections.

2.1. Challenges Associated with Determining Customer Preferences in Retail

A typical customer has neither the time nor the mental capacity to evaluate even a small subset of store displays (Lee and Lee 2004; Scheibehenne et al. 2010). To reduce customer effort and increase sales, one possible approach is to provide individualized recommendations (Aljukhadar et al. 2012) based on customer preferences. Yet, identifying customer preferences in retail stores is challenging. Customers not only have different preferences, they also generally do not voluntarily provide structured preference information to retailers. Moreover, customers only evaluate a small subset of displayed products and partial product features while shopping, so preferences must be inferred based on just a few products.

To address these challenges, retailers typically use two strategies. First, retailers hire salespeople to communicate with customers about their product preferences and assist them while they shop. Salespeople usually make recommendations based on subjective judgments about customer preferences, the effectiveness of which depends greatly on their skills and experiences, and the quality of their interactions with customers (Franke and Park 2006; Weitz et al. 1986). It is also very difficult for a salesperson to memorize information about every item in stock in order to make informative recommendations (Weitz et al. 1986). Customers are at risk of experiencing undesirable outcomes if a salesperson is unable to provide useful information or is not motivated to protect a customer's interests (Swan et al. 1999). Second, practitioners use preference measurement models such as conjoint analysis to infer customer preferences about product features (Green and Srinivasan 1990). The typical practice in business is to estimate conjoint models at the aggregate level, given individual level part-worth estimates (DeSarbo et al. 1995). In this way, retailers can recommend popular items with the

features customers prefer most and display them in prominent places. However, this is a one-size-fits-all strategy; the same products are always recommended, regardless of a customer's preferences. Although providing individualized recommendations is much more preferable (Ariely et al. 2004) customers would need to indicate their preferences as they shop, which is rarely done in practice due to the extra effort involved.

2.2. Using Video Data in Marketing

Video can be used to record and analyze behavior in situational contexts, which means such data can potentially reveal both qualitative and quantitative insights (Basil 2011; Belk and Kozinets 2005). As video data and inexpensive video editing hardware and software become more prevalent, marketing researchers and practitioners are harnessing the potential of video data to generate insights for business practice (Belk and Kozinets 2005).

With the rapid increase in computation power, we are now able to capture, store and analyze video data for various purposes, such as face recognition (Zhao et al. 2003), facial expression recognition (Fasel and Luettin 2003; Russell and Fernandez-Dols 1997; Shergill et al. 2008) and gesture detection (Kuch and Huang 1995; Yoruk et al. 2006), among others. Video analysis has been applied in various disciplines such as artificial intelligence, human-computer interaction, biometrics, and marketing. Commercial tools[1] also have been developed using video analysis to help marketers identify customer demographics.

In marketing, video data provide rich information that can be useful to both consumers and managers if properly used (Belk and Kozinets 2005; Lee and Broderick 2007). Shergill et al. (2008) proposed a framework for using video data to allocate salespeople to customers, and Belk (2011) proposed using documentary videos to examine consumer behavior. Given its potential to yield rich insights into consumer behavior, video analysis is being increasingly used in retail contexts (e.g., Hui et al. 2009a, 2009b, 2013; Valizade-Funder et al. 2012; Zhang et al. 2012). Hui et al. (2013) first used in-store video tracking to collect data about customers' in-store shopping paths, shedding light on customers' product consideration processes.

Video data can provide information on the temporal, spatial and social dimensions of objects, as well as psychological information about customers (Kozinets and Belk 2006) such as emotional reactions, associated stimulating factors and other simultaneous behavioral responses without interrupting the normal shopping

[1] See http://www.videomining.com for an example.

process. Affective responses play an important role in information processing and product evaluation (Pham 1998; see also Schwarz and Clore 1983, 1988; Wyer and Carlston 1979), thereby driving customer behavior (Roseman et al. 1996; Zeelenberg and Pieters 2004). Positive affective responses result in positive evaluations of the focal product, whereas negative affective responses result in negative evaluations (Bloch 1995).The emotion theory literature (Frijda 1986; Frijda and Zeelenberg 2001; Roseman, Antoniou and Jose 1996) shows that different emotions can lead to "different behavioral tendencies (action tendencies or patterns of action readiness) and behavioral consequences" (Zeelenberg and Pieters 2004, p. 446).

Customers' behavioral responses also serve as important clues in evaluating their preferences. For example, touching (conscious or unconscious) plays a prominent role in garment evaluation (Grohmann et al. 2007; McCabe and Nowlis 2003). The first sense that humans develop, touching is a form of analytical or systematic (vs. relational) processing in which one feature is evaluated at a time (Yazdanparast and Spears 2012). People use their hands to acquire and process information about objects, sometimes simply for the sake of sensation (Klatzky et al. 1993; Peck and Childers 2003). Evidence shows that tactile cues are more influential than visual cues in customers' evaluations of clothing products, which have diverse material properties, such as texture and stretch (Grohmann et al. 2007; Holbrook 1983).

Very few marketing scholars have developed models to apply automatic video analysis to infer customer preferences in real business contexts. In the few existing empirical marketing studies, human judgment was used to analyze video or image data, which is labor intensive and time consuming for large datasets and cannot be scaled up or accomplished in real time. To the best of our knowledge, we are among the first to automatically infer individual preference information from video data in the retailing context and to develop managerial tools that combine video technology from computer science with current standard marketing research methods.

3. General Design of a Video-based Automated Recommender (VAR) System

In this section, we describe the general design of a video-based automated recommender (VAR) system. The system is designed to satisfy several key criteria that must be met in a retailing environment: (a) make recommendations for individuals based on their individualized preferences; (b) require minimal customer effort and not interfere with a customer's normal shopping experience; (c) address a customer's need for privacy; and

(d) be easy to implement in a retail store and scale well to large retail chains. We first describe the system setup and the initiation and termination of each recommendation. We then discuss each of the three analytic steps involved in making a recommendation in detail. The general structure of the system is illustrated in Figure 1.

Insert Figure 1 Here

3.1. System Setup and Recommendation Process

The system's hardware includes a central computing unit connected to many pairs of decentralized in-store user interface devices. Each pair is comprised of an input device (i.e., webcam) and an output device (i.e., display). The central computing unit can be either on-site, or virtual in the form of rented servers (including database servers) in the cloud computing environment. The system's software resides on the central computing unit, and includes codes for the three analytic steps as well as two databases: an inventory database and a customer database. The inventory database is updated dynamically as new items are added, and the customer database expands as more customers use the system. The inventory database may be comprised of an existing store database (e.g., item ID/barcode and a photo), or may include additional information on each garment (e.g., garment regions, feature descriptions). The customer database includes an organically growing set of information about which items a customer has tried and her associated reactions as captured via video analysis during her try-on experiences; these data are linked with eventual purchase decisions (if made) during a shopping trip. We call this a *try-on customer database*.

There are two caveats regarding the try-on customer database. First, depending on the opt-in level a customer selects, entries in the database may be: (a) independent (i.e., when a customer chooses not to be tracked across different try-ons during a shopping trip); (b) linked to the same customer during one shopping trip only (i.e., if a customer chooses to be tracked across different try-ons during a shopping trip but wants her identifier, such as face image, to be erased at the end of the business day); or (c) linked to the same customer over multiple shopping trips (possibly to different stores of the same chain, if she allows her identifier, either her face or an ID number, to be permanently stored). Independent entries are generally not useful in helping the system make recommendations to other customers. Second, the purchase decisions made by a customer during a given shopping trip may be connected to the try-on data in several different ways with potentially varied precision. For example, one way to link a try-on to a purchase is to assign each garment a unique ID, which is

recorded both while the customer is trying on the item and at purchase. Another way is to identify a customer's identity at the checkout counter using face recognition software and to match it against customers who tried garments on that particular business day in that store (assuming she has opted-in to tracking).

The VAR system is initiated once a customer opts-in by scanning the product barcode/ID on the system's barcode reader. This mechanism also enables the system to identify which product the customer is evaluating. By scanning the garment ID, the customer activates the camera to capture his or her reactions. As the customer evaluates the garment, the video camera records the process and sends information to the central computing unit for analysis. The system terminates the analysis when the customer leaves the evaluation area (i.e., when the frontal upper-body cannot be detected for a certain continuous period of time).

3.2. Step 1: Infer Preferences from Try-on Video Data

The purpose of the first analytic step is to use video data to infer some preference information about a focal customer based on her reactions toward a garment she is trying on. As described in detail in the literature review, two visual cues can be obtained from customers in a retailing context: affective responses and behavioral responses.

Facial expression is a good and natural indication of a customer's internal emotions and mental activities (Russell and Fernandez-Dols 1997) and thus is often used as a proxy for a person's affective state in both practice and research. For example, companies such as Procter & Gamble and Unilever collect high-frequency data on facial expressions to understand their influence on consumer behavior (Teixeira et al. 2012), and GfK tracks viewers' facial expressions for copy testing (Miller 2013). In research contexts facial expressions have been used to study viewers' preferences toward Internet video commercials (Teixeira et al. 2012, 2014), and how customers react to different online shopping contexts in virtual stores (Raouzaiou et al. 2002).

Behavioral responses (e.g., touching) also are crucial in the evaluation of products, especially apparel (Peck and Childers 2003; Peck and Wiggins 2006). It has been well documented in the literature that hand movements are associated with customers' exploratory and evaluative perceptions (Krishna 2009; Peck and Childers 2003); in fact, touching (haptic perception) is the dominant input for determining product quality, and can increase perceptions of ownership (Peck and Shu 2009). When a customer wants to assess a fashion element on garment, she may touch the corresponding area repeatedly. We call this a *region of interest* in the present paper.

The VAR system infers customers' preferences by simultaneously analyzing affective responses (facial expressions) and behavioral responses (region of interest being touched) captured on video as they evaluate garments in front of a mirror. Thanks to advanced computer vision techniques, a customer's facial expressions and the areas of a garment that a customer touches can both be automatically inferred from video data with reasonable accuracy (see a recent review by Xiao et al. 2013). We describe the analysis process below (see Figure 2).

Insert Figure 2 Here

3.2.1. Pre-processing

Once the process is initiated (i.e., a customer opts-in), the actual video analysis process begins when the system detects a person (more specifically, the frontal upper body) in the scene using the local context[2] detector (Kruppa et al. 2003). The detector uses the differences between the sum of the pixels within two rectangular regions (Haar-like features[3]) to encode the details of the head, neck and shoulder area, and Adaboost to select features and train the classifier. This upper body detector[4] is trained with images that contain a person's head, neck and shoulder area. After it detects the location of the human body, the system crops the image so it contains only the focal customer and the garment s/he is evaluating, for further analysis.

3.2.2. Face recognition and facial expression recognition[5]

Face recognition and facial expression recognition serve different purposes and are performed at different time points in the VAR system, but are achieved with similar techniques (Chavan and Kulkarni 2013; Fasel and Luettin 2003) and thus are discussed here in the same subsection. Face recognition allows the VAR system to link a customer's try-on experiences by matching faces in the video data (this only needs to be done at the beginning of each try-on video). Facial expression recognition is used to capture the affective state of a customer in each frame analyzed. Since video is actually a temporal sequence of still images (called frames) representing scenes in motion, facial expression recognition can be regarded as recognizing facial expression in each frame in

[2] Local context is defined as a local area surrounding the face, i.e., the head, neck and shoulder area.
[3] Haar-like features are weighted differences of integrals over rectangular sub-regions (see Viola and Jones, 2001).
[4] The upper-body detector is a modified version of the Viola-Jones detector (Viola and Jones 2001, 2004), and is available through the Open Computer Visions Library and MATLAB.
[5] Details on face and facial expression recognition can be found in online Appendix A.

the video.[6] The general process involves three steps: face detection, feature extraction, and face/expression classification (see Chavan and Kulkarni 2013; Fasel and Luettin 2003 for literature reviews on these steps).

Face detection. There are three main methods for automatic face detection—template-matching, feature-based (e.g., skin color), and image-based—that train machine systems on large numbers of samples (i.e., images labeled as face or non-face). Among these, image-based methods perform the best and achieve good accuracy in detecting faces in images (Rowley et al. 1998; Sung and Poggio 1998). The most popular computer vision software applications (e.g., MATLAB and OpenCV) include pre-trained face classifiers and toolboxes that can be incorporated into the VAR system. After the face is detected, the face image is typically normalized by size.

Feature extraction. The human face is complex, so decomposing it into an effective set of features is critical to the success of face/facial expression recognition. Feature extraction mainly involves three types of features: geometric-based, appearance-based, and a combination of both. Geometric-based features measure the displacements of certain face regions such as the eyebrows or corners of the mouth, while appearance-based features are concerned with face texture. Appearance-based features may be extracted either holistically or locally. Holistic features are determined by processing the face as a whole, for example, eigenface features (Abboud et al. 2004; Turk and Pentland 1991; Xiao and Ding 2014). Local features are specific facial features or areas that are prone to change with facial expressions, and are extracted by, for example, detecting local binary patterns (LBP, first introduced by Ojala et al. 1996) associated with the eye and mouth regions (Shan et al. 2009).

Due to illumination variations common in retail settings, techniques must be incorporated to reduce noise and make the information required for recognition more salient. Gamma correction, a nonlinear gray-level transformation, enhances the local dynamic range of the image in dark and shadowed regions while compressing the bright regions (Shan et al. 2003; Tan and Triggs 2010). Rotation invariant LBP (RI-LBP)[7] is then used to represent the various features. The RI-LBP technique is widely used in computer vision; it has been shown to be invariant to monotonic global illumination changes (Ojala et al. 1996; Shan et al. 2009; Tan and Triggs 2010),

[6] Facial expression can be recognized either by classifying facial expressions in each single frame or by tracking facial points of interest in a temporal sequence of images (Fasel and Luettin 2003). In the current paper, we use the former since we want to match the facial expression to the hand position in each frame.

[7] Detailed information on the RI-LBP operator can be found in online Appendix A.

and its computational simplicity makes it suitable for real-time applications. The RI-LBP features are then used in the classification task.

Classification. Face/facial expression recognition is essentially a classification problem in supervised learning that involves assigning face/expression labels to focal face images. A large variety of algorithms are available for this step. For example, support vector machines (SVM), hidden Markov models (HMM), neural networks, and k-nearest neighbors (KNN) have been demonstrated to perform well in the literature (Ma and Khorasani 2004; Oliver et al. 2000). To achieve facial expression recognition, a set of training images for facial expressions needs to be collected and incorporated into the VAR system, generally using either a seven-expression scheme (i.e., neutral, happiness, sadness, fear, disgust, surprise and anger, Ekman 1994; Ekman and Friesen 1971) or a three-expression scheme (i.e., positive, negative, and neutral; Zeelenberg and Pieters 2004). The first few frontal face images detected from each try-on video can be used to train the system for face recognition in the future if matches are not found in the existing database.

3.2.3. Region of interest detection

The region-of-interest provides an estimate of a customer's focus of attention when evaluating a garment. The system first detects the hand position in the video frame, then relates it to a specific region on the garment (and the corresponding feature if a region-to-feature relationship has been pre-coded for that garment in the inventory database).

Hand detection. A key step in gesture recognition, hand tracking, and human-computer interaction applications, hand detection is an active research area in the computer vision field (Mitra and Acharya 2007). Approaches to hand detection include color-based detection (using local skin color), appearance-based detection (using pre-defined geometric hand templates), and motion-based detection (tracking hand movement by assuming different motion features for hand and the background regions). As a classical approach for hand detection, color-based detection has proven to be effective and robust (Li and Kitani 2013; Saxe and Foulds 1996) since skin color is fairly uniform and an individual's hands and face are typically the same color (Jones and Rehg 2002; Zhu et al. 2000).

Garment region map. After detecting the location of the garment in each frame, the system identifies the specific region of interest by determining where the hand is positioned on the garment. In Figure 3, we present a

typical 3×5 garment region map to specify the locations of design features on a garment. This follows common practice in garment design (see Cordier et al. 2003; Liu et al. 2010).

Insert Figure 3 Here

There are various ways to construct a garment region map for the VAR system. At one extreme, a retail store can code each garment to best capture its design features. Such an item-level garment region map enables more conceptually meaningful matching later, since each region corresponds to one particular feature; however, this approach requires extra work to code every new garment added to the inventory database. At the other extreme, a VAR system may use a generic garment region map (see Figure 3) for all garments. Somewhere in-between, a VAR system may include several garment region map templates to reflect general design categories (e.g., t-shirt, pants), and each garment could be coded to a template. (This could be automatically determined if the inventory database included such general classification information.) A VAR system based on collaborative filtering (CF) does not require a 1:1 correspondence between regions and features.

3.2.4. Preference inference

Preferences are inferred by matching the facial expression to the region-of-interest in each frame and aggregating similar facial expressions for the same region of interest among all frames analyzed throughout the try-on process. (Typically, the system will not analyze every frame in the video to reduce computational burden; for example, it may analyze 10 frames from each second in a 30 frames-per-second (fps) recording, i.e., it will skip every two frames.) Assuming a three-expression (positive, negative, neutral) scheme, a preference score for a given region (or feature, if the garment region map is pre-coded with features) is calculated as the total number of detected positive expressions minus the total number of detected negative expressions when a customer evaluates a particular region/feature of the garment throughout the try-on video. The total number of three expressions (positive, negative, neutral) detected respectively throughout the try-on video could be used as proxies for the overall garment preference. To help reduce computational burden, we used one summary element for each region/feature and three separate elements for overall item-level preference in our calculation. With F representing the number of regions/features on a garment, we used an $(F + 3) \times 1$ vector, denoted as $P_{u,g}$, to represent customer u's preference toward garment g, where the first F elements represent customer u's preferences(the total number of detected positive expressions minus the total number of detected negative

expressions when the focal customer evaluated the particular region/feature throughout the try-on video) toward *F* regions/features of garment *g*, and the last three elements represent customer *u*'s overall preferences (the total number of positive, negative and neutral expressions, respectively, detected throughout the try-on video) toward garment *g*.

3.3. Step 2: Identify Neighborhood for Focal Customer

Once the VAR system obtains the information about the focal customer's reactions to various regions/features of a particular garment, the next step is to match her to a set of customers with similar preferences from the try-on customer database. The widely used matching and recommendation method is collaborative filtering (CF). In general, CF is the process of filtering items using the collaborative opinions/preferences of other people (see Goldberg et al. 1992, 2001 for some applications in marketing). The fundamental assumption of the CF approach is that if Users A and B behave similarly when evaluating certain products, they are more likely to behave similarly toward other products. The CF approach has been widely deployed by firms such as Amazon.com and MovieLens to make recommendations to customers because it is highly effective and easy to implement (Su and Khoshgoftaar 2009). The basic idea is: Even when limited information exists about a focal customer, useful recommendations can be made based on what is known about other customers with similar preferences (e.g., what other items they have tried on or purchased). Matching generally can be accomplished by evaluating the similarities between the observed focal customer's preferences and those in the database, and selecting those whose similarities with the focal customer exceed a certain threshold.

In the garment retailing context, each try-on incident (one customer trying on one garment in front of the mirror) represents a customer-garment pair. For each try-on incident, the system first finds a set of customers in the database who have tried (i.e., evaluated) the same garment. These customers are called *candidate neighbors* for the target customer-garment pair. The system then selects a subset of these candidate neighbors who have most similar preferences on that garment with the focal customer. These people, called *nearest neighbors*, form a *neighborhood* for the target customer-garment pair.

In the garment retailing context, the CF algorithm must satisfy three criteria. First, it must be able to make recommendations to first-time users.[8] Second, it must be able to deal with highly sparse data since customers typically only evaluate (try on) a small subset of the garments in database, and purchase even fewer. Third, it must scale well with an increasing number of customers and items. We discuss our choice and implementation of the CF algorithm based on these three criteria.

CF can be implemented using one of several approaches: content-based, neighborhood-based, or model-based. We chose to use the neighborhood-based CF algorithm because it is easy to implement, is widely used in practice, and new data (about customers as well as garments) can be added easily and incrementally (Su and Khoshgoftaar 2009), which is crucial for scalability in the garment retailing context.

Two types of response data can be potentially used for the CF-based recommender system in the garment retailing context: customers' implicit/explicit responses to items considered (try-on data) and items purchased (choice data). We recommend the use of try-on data in the CF algorithm in VAR system for two reasons: (a) there are a lot more try-on data than choice data, which better addresses the data sparsity issue, one of the most important factors affecting the performance of CF-based recommender systems (Su and Khoshgoftaar 2009); and (b) since the purpose of the VAR is to suggest items for customers to consider/try on, it makes more sense to use the same type of data in the recommendation process. In addition, choice data might be influenced by factors other than customers' item preferences (e.g., prices), which may not be relevant to other customers.

The neighborhood selection and similarity weighting mechanisms are key components of a neighborhood-based recommendation method. The algorithm employed to construct a neighborhood for recommendation purposes depends on the ratio between the number of customers and the number of items (Su and Khoshgoftaar 2009). If the number of customers is much larger than the number of items, it is likely that an item would be rated by a large number of customers. In this case, it is optimal to use a mechanism to select a subset of high-confidence neighbors to form the neighborhood (Desrosiers and Karypis 2011), such as a correlation-based similarity weighting mechanism (Herlocker et al. 2004). Correlation-based similarity measures have been widely used in commercial recommender systems by firms such as MovieLens (Herlocker et al. 2004; Resnick et al. 1994; Resnick and Varian 1997). However, if the number of items is much larger than the number

[8] Here, a first-time user is a person who has no past try-on data stored in the database because the customer either is using the system for the first time, or has chosen not to be tracked across different try-on incidents.

of customers, an algorithm that utilizes information from all people who have rated the same item is more preferable, as there will only be a few such neighbors for each focal customer (Desrosiers and Karypis 2011). In garment retailing contexts, since there are typically a lot more customers than items, correlation-based similarity weighting is more suitable for selecting the neighborhood.

It is worth noting that although the traditional CF approach uses ratings with only one component (i.e., the overall item-level rating), there has been growing interest in using customers' responses to multiple aspects of items to generate more accurate recommendations (Adomavicius and Kwon 2007; Adomavicius et al. 2011; Sahoo et al. 2006). A CF method based on multiple-component responses can potentially solve the first-time user problem by learning customers' attribute-level preferences (Rashid et al. 2002). The multiple-component CF method has proven to be effective and outperform traditional single-rating CF methods in many applications where ratings on individual features carry meaningful information (Adomavicius and Kwon 2007; Manouselis and Costopoulou 2007). We implemented a multiple-component CF algorithm in this study since the VAR system can infer customers' responses to multiple features/regions of the garment.

The neighborhood-based CF method used in the VAR system operates as follows. First, given a focal customer's current try-on incident, all customers in the try-on customer database who have tried on the same garment g as the focal customer u are identified. These candidate neighbors are represented as $M_{u,g}$. Second, a similarity score is calculated between the focal customer and each candidate neighbor in $M_{u,g}$. This similarity score is used to form a proximity-based neighborhood between the focal customer and her like-minded neighbors. The purpose here is to rank order the candidate neighbors based on how similar they are to the focal customer. Similarity can be calculated based on correlations, distance and cosine, among others (Su and Khoshgoftaar 2009). As explained previously, we use a correlation-based similarity weighting mechanism here.

We can calculate the similarity score between two customers who have both tried on the same garment g using their video-inferred preferences about g. As described in Section 3.2, we use $P_{u,g}$, an $(F+3)\times 1$ vector, to represent the focal customer u's preferences (i.e., the feature-level and item-level preferences inferred from the try-on video) toward garment g, and $P_{v,g}$, another $(F+3)\times 1$ vector to represent a candidate neighbor v's preferences toward the same garment g. The similarity score between the preferences of focal customer u and her candidate neighbor v based on their reactions to trying on garment g is given by:

$$sim_g(u,v) = \frac{\left(P_{u,g} - \overline{P_{u,g}}\right)\left(P_{v,g} - \overline{P_{v,g}}\right)'}{\sqrt{\left(P_{u,g} - \overline{P_{u,g}}\right)\left(P_{u,g} - \overline{P_{u.g}}\right)'}\sqrt{\left(\left(P_{v,g} - \overline{P_{v,g}}\right)\left(\left(P_{v,g} - \overline{P_{v,g}}\right)\right)'}} \quad (1)$$

where $\overline{P_{u,g}}$ is an (F+3) ×1 vector, calculated as $\overline{P_{u,g}} = \begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix} \times \frac{1}{(F+3)} \sum P_{u,g}$; and $\overline{P_{v,g}}$ is also an (F + 3) ×1 vector, calculated as $\overline{P_{v,g}} = \begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix} \times \frac{1}{(F+3)} \sum P_{v,g}$ (Su and Khoshgoftaar 2009).

Third, nearest neighbors are selected to form a neighborhood for the target customer-garment pair. One method used to select (or filter) candidates to form the neighborhood is threshold filtering (Desrosiers and Karypis 2011; Schafer et al. 2007). When a correlation-based similarity mechanism is used, a common practice is to use threshold filtering with 0 as a threshold (Herlocker et al. 2004; Schafer et al. 2007), since negative correlations "are generally believed to not be valuable in increasing prediction accuracy" (Schafer et al. 2007, p. 302). However, the number of candidate neighbors who share a non-negative similarity score with a focal customer *u* on garment *g* could be huge for large datasets. In such cases, top-K filtering can be further applied, where only the top *K* candidate neighbors with the highest similarity scores with focal customer *u* on garment *g* are selected. A general rule is to choose a *K* between 20 and 50, depending on the size and sparsity of the dataset (Desrosiers and Karypis 2011). These *nearest neighbors* are represented by $N_{u,g}$.

3.4. Step 3: Make Recommendations

Recommendations can then be made to the focal customer based on her nearest neighbors' preferences and/or behaviors. A try-on customer database has information on what other garments the nearest neighbors have tried on and their reactions (to each garment region/feature, and the overall garment) from the video analysis, as well as the garments they have purchased. The system can use a heuristic to rank order these garments and then recommend those ranked highest. While heuristic variations exist, they typically are based on some or all three types of information: how similar a nearest neighbor is to the focal customer, how often a particular garment is tried and/or bought by her nearest neighbors, and whether a garment is only tried, or tried and bought (and if one is only tried, the similarities between the focal customer's reactions to it and to another item that is tried and bought, which can be used as a proxy to represent how close this item is to being bought). A retailing chain, in practice, may need to test out various heuristics and find the one that works best.

One widely used heuristic in recommender systems is to compute predictions for a focal customer using nearest neighbors' responses (usually a rating-based response) to the items they have evaluated and their similarities to the focal customer by taking a weighted average of all responses from nearest neighbors (Adomavicius and Kwon 2007). The basic idea for the heuristic is that recommendations can be made by rank ordering all garments that were tried by at least one nearest neighbor (excluding the garments that the focal customer has already tried on), according to the weighted average of all nearest neighbors.

We implemented this heuristic in our VAR system by inferring responses to try-on items based on facial expressions and regions of interest. The try-on customer database contains information on which garments nearest neighbors have tried on and/or bought in the past. The system first estimates the purchase likelihood for the tried-but-not-purchased garments, assuming the customer's affective responses to the try-on garment signal how much she wants to buy that garment. The system can calculate the similarity between nearest neighbor *n*'s tried-but-not-purchased garments and purchased garments using the overall preferences (a three-dimensional vector, i.e., the total number of frames for three expressions—positive, negative and neutral, respectively—throughout the try-on video across all garment regions) inferred from her try-on videos. If nearest neighbor *n* has bought more than one garment, the system calculates the average similarity score for all purchased garments. The average similarity score for a tried-but-not-purchased item is treated as a proxy for nearest neighbor *n*'s likelihood to purchase that garment.

Using A_n to represent the set of garments purchased by nearest neighbor *n*, $a \in A_n$, where the total number of purchased garments is denoted as *TA*; and B_n to represent the set of garments that were tried but not purchased by *n*, $b \in B_n$; then $s_n(a, b)$ represents the similarity between *n*'s overall preference toward garments a and b, and is calculated as (Karypis 2001; Su and Khoshgoftaar 2009):

$$s_n(a,b) = \frac{R_{n,a}\, R_{n,b}{}'}{\sqrt{\left(R_{n,a}\, R_{n,a}{}'\right)\left(R_{n,b}\, R_{n,b}{}'\right)}}, \qquad a \in A_n; b \in B_n \qquad (2)$$

where $R_{n,a}$ is a 3×1 vector referring to the overall preferences of nearest neighbor *n* toward garment a, $R_{n,b}$ is a also a 3×1 vector referring to the overall preferences of nearest neighbor *n* toward garment b, and the purchase likelihood of nearest neighbor *n* toward garment *b* is given by:

$$PL_{n,b} = \frac{\sum_{A_n} S_n(a,b)}{TA} \qquad (3)$$

Once the purchase likelihoods for both considered and purchased items of all nearest neighbors have been calculated, the system determines which items to recommend to the focal customer. First, it calculates a weighted summation of all nearest neighbors who have considered a particular garment *i*; this is treated as the predicted purchase likelihood of the focal customer *u* toward garment *i*, given by:

$$Q_{u,i} = \sum_{n \in N_{u,g}} w_g(u,n) Q_{n,i} \qquad (4)$$

where $w_g(u,n)$ refers to the normalized similarity between focal customer *u* and nearest neighbor *n*, given by:

$$w_g(u,n) = \frac{sim_g(u,n)}{\sum_{n \in N_{u,g}} |sim_g(u,n)|} \qquad (5)$$

and $Q_{n,i}$ refers to the nearest neighbor *n*'s purchase likelihood toward garment *i*, given by:

$$Q_{n,i} = \begin{cases} PL_{n,i}, & if \ \ i \in B_n \\ 1, & if \ \ i \in A_n \end{cases} \qquad (6)$$

After rank-ordering all items according to the weighted summations (predicted purchase likelihood for the focal customer), the top-ranked items are recommended to the focal customer for consideration.

4. Empirical Study 1: Validation

As discussed in the previous section, different VAR system variations may be implemented based on a retailer's preferred tradeoff between accuracy and cost, as well as specific sets of garments and customers. In this section, we demonstrate the feasibility of one such implementation for exemplary purposes. Importantly, we do not claim that this is the best possible implementation of VAR, but a proof of concept. We also compare this particular VAR system with two state-of-the-art benchmark models: a self-explicated conjoint (SEC) model and a self-evaluation after try-on (SET) model. While we do not claim to have compared our system to all benchmarks, our objective is to demonstrate the relative usefulness of this particular VAR implementation. In this section, we describe the three models, explain the design and procedure of the empirical test, and present the analysis and results.

4.1. VAR System and Benchmark Models

The VAR model uses customers' affective and behavioral responses (i.e., facial expressions and regions-of-interest touched with their hands) to try-on items inferred from video clips in combination with the neighborhood-based CF algorithm described in Section 3. The VAR model relies on a *try-on customer database*, which contains customers' try-on and purchase data. In our VAR system implementation, we closely followed the analytic steps described in Section 3. Specifically, to infer preferences from the try-on video, we used RI-LBP to extract features and SVM as our classification tool during the face/facial expression recognition step. We used a three-expression scheme (positive, negative and neutral) for the training set of facial expressions. The garments in the inventory database were coded at the individual garment level to create a 1:1 correspondence between features and regions, which were then used to detect regions of interest.[9]

We implemented two individual level, non-automated recommendation models—self-explicated conjoint (SEC) and self-evaluation after try-on (SET)—as benchmark models. While both models make individual level recommendations, their applications require additional customer effort. We compare the three models in Table 1.

Insert Table 1 Here

The SEC model estimates each customer's preferences (i.e., part-worth data) for all garment features and then recommends the items ranked highest in utility. This model is widely used by marketers to understand customers' preferences and predict their purchases, especially when there are a large number of attributes with many variations (Green and Srinivasan 1990; Netzer et al. 2008), as is the case for garments. The implementation of the SEC model relies on customers' self-explicated conjoint preference data. The design of the SEC model in the study follows a procedure described in the literature (Srinivasan and Park 1997).

The SET model uses customers' feedback after evaluating the try-on garment in combination with the same neighborhood-based CF algorithm used in the VAR system. The implementation of the SET model relies on a customer-item database containing the overall ratings, and possibly feature-level ratings, for the items they have evaluated. This database is constructed by asking each customer to answer a mandatory question on her overall rating of each try-on garment (i.e., purchase likelihood on a scale of 0–100) and optional questions related to feature preferences using 7-point scales. This is then supplemented by the observed purchase decision for these

[9] We also tested an implementation using a generic garment region map shown in Figure 3 (see Footnote 19), where customers' preferences related to the 15 generic regions (instead of garment features) were detected from the try-on video.

items. The rest of the model (CF-based recommendation) is essentially the same as that described for the VAR system.

4.2. Study Design and Procedures

Following recent preference measurement literature in marketing (e.g., Kim et al. 2014; Narayan et al. 2011), we conducted our empirical study in two stages: a *calibration stage* followed by a *validation stage* about a week later. This two-stage design provided a good test and platform for comparing our model to the benchmark models. The delay is typically used to ensure that participants "forget" about the calibration tasks, so that the validation results are free of any potential interference due to the calibration. This design was particularly useful to us since we were constrained by a typical academic research budget and thus collected calibration data for all three methods from the same individuals. For the VAR and SET models, this also enabled us to use the calibration stage to construct the try-on customer database needed to make recommendations. In the second empirical study, we demonstrate and evaluate the real time recommendations of the VAR system using the existing try-on customer database obtained from the first empirical study and a different group of participants; thus, participants in the second study were not burdened with three different tasks.

In this first empirical study, we tested recommendations for women's casual tops. Based on literature related to apparel consumption behaviors (Abraham-Murali and Littrell 1995; Eckman et al. 1990; Fang et al. 2003; Zhang et al. 2002), interviews with female apparel consumers, and current online apparel retailers' business practices, we selected nine garment attributes that are most relevant for purchases of women's casual tops and constructed the garment database in a $2^2 \times 3^3 \times 4^2 \times 6^2$ attribute space (see Table 2). Using SAS fractional factorial design, we generated 160 garment profiles that closely resemble the product assortments offered in a small garment store, and purchased real garments corresponding to these profiles from almost 100 online stores (typically, each store only had one or two garments that fit our profiles). From the 160 garments, we randomly selected 140 garments for the calibration stage and used the remaining 20 garments for the validation stage. Due to budget constraints, all garments were size medium, a reasonable, but not ideal offering, given the physical characteristics of the participants. We recruited 127 female garment shoppers to participate in the study. Their average age was 21 (range: [18, 27]; SD = 2.2), average height was 163 cm (range: [154, 177]; SD = 4.8), average weight was 52 kg (range = [40, 67]; SD = 5.4), and average BMI was 19.53 (range = [15.06, 24.01]; SD = 1.78).

This indicates the participant sample reasonably represented garment shoppers in that age group in the real world, and medium was the preferred size in most cases.

Insert Table 2 Here

In order to mimic the real-life experience of apparel shopping, we set up a room to resemble a small garment store (mock store), with real garments displayed on racks. Based on feedback from participants, the shopping process in this environment was similar to the experience in a real garment store. To eliminate the impact of brand and price, tags and brand labels were removed and participants were told that all garments were from the same brand and in the same price range. A fitting room was provided for participants to change into the garments they wanted to try on. A mirror was placed outside of the fitting room with a camera mounted on top (Logitech_C920, recording at 30 fps at a resolution of 1920*1080 pixels). The participants consented to being videotaped for research purposes, but did not know how the data would be analyzed (e.g., facial expressions, hand movements, etc.). To ensure that participants behaved as they would when shopping for clothes in real garment store, each participant received a particular garment based on her preferences revealed during the empirical study as compensation based on the procedures described in Dong et al. (2010).

At the calibration stage, we collected preference data using each of the three models. Participants came to the mock store one at a time. Each participant was asked to browse the 140 garments on the racks and choose any items that she would like to try on. A mock store assistant (research assistant) then put all of the chosen garments onto a separate rack outside the fitting room as in a real store. After putting on a garment in the private fitting room, the participant came out to evaluate herself in front of the mirror; this process was repeated for each garment selected. Each evaluation process was videotaped by the webcam mounted on top of the mirror. The video data were used for the VAR model.

After evaluating all chosen items, each participant was asked to provide feedback on a tablet for each item she had tried on. (The garments were still hanging on the rack outside the fitting room and participants were encouraged to respond to the questions while reevaluating each garment). Each participant was asked to answer a mandatory question on her item-level rating of each try-on garment (i.e., purchase likelihood on a scale of

0–100) and optional questions related to her feature-level preferences[10] using 7-point scales, followed by a question about her purchase decision. These data were used for the SET model.

Finally, participants were asked to complete the questionnaires for the SEC model. After excluding feature levels (see Table 2) that they would never accept (non-compensatory), they followed the standard two-step procedure of rating levels within attributes, and then assigned the relative importance across attributes. Recommendations were made by rank ordering the estimated utilities of all garments.

At the validation stage, participants came back after about a week to evaluate 20 different garments and make purchase decisions. Participants were free to evaluate and try on any of the 20 garments without any guidance. The data collected during the calibration stages were used to test the ability of the three models (VAR, SET and SEC) to predict each participant's choices/considerations from the 20 garments in the validation stage.

The calibration stage was comprised of 1,184 try-on incidents, with an average of 9 items per participant (ranging from 3 to 26). The validation stage was comprised of 450 try-on incidents, with an average of 4 items per participant (ranging from 1 to 9). On average, it took each participant around 20 seconds to evaluate a garment in front of the mirror, with a range from 3 seconds to 50 seconds, and a standard deviation of 12 seconds (excluding the time used to try on the garment in the fitting room). The average total evaluation time per customer was 3 minutes, varying from less than 1 minute to more than 10 minutes. The distributions of evaluation time are shown in Figures 4a and 4b.

Insert Figures 4a and 4b Here

4.3. Analysis and Results

We report the analysis and results of the VAR model following the three general analytic steps described in Section 3. We discuss each of the two benchmark models where relevant: SET in Sections 4.3.2 (neighborhood identification) and 4.3.3 (predictive performance) and SEC in Section 4.3.3 (predictive performance).

4.3.1. Infer preferences from video analysis (VAR model)

We analyzed each video clip by extracting frames at 10 fps. Following the procedures described in Section 3, for each frame extracted the model: (a) recognized the facial expression in the frame, (b) detected the region

[10] Features of each garment were determined by the garment design generated from the attribute space in Table 2.

of interest by identifying the customer's hand position on the garment region map, and (c) inferred the participant's preference for each region (or feature) and overall preference for each item.

Face detection. We used the Viola-Jones face detector[11] (Viola and Jones 2001, 2004), a face classifier that has already been trained and embedded into MATLAB, to detect the face in each frame. The Viola-Jones algorithm[12] makes it possible to process images rapidly while achieving high detection rates. It is a robust and rapid face classifier constructed by selecting a small number of important features using Adaboost.

It is worth noting that in natural settings, faces are usually viewed from various perspectives. The viewpoints (angle and position) of a face image have great impact on the accuracy of facial expression recognition (Wang and Ahuja 2003). To achieve reasonable accuracy, only the frames in which the frontal view of the face could be detected were considered valid and processed for further facial expression recognition. Among the 234,639 frames extracted from the 1,184 try-on videos in the calibration stage, 185,855 frames were valid (79%). Figure 5 shows the distribution of the percentage of valid frames for each try-on video clip.

Insert Figure 5 Here

Feature extraction and expression classification. Each detected face was then normalized in terms of size (Georghiades et al. 2001) and illumination (using gamma correction), and prepared for feature extraction. We extracted RI-LBP features from the detected face images. We then used the SVM classifier for the expression classification step.[13]

For expression classification, we constructed a training dataset with 540 facial images from 45 female participants following the three-expression scheme (Zeelenberg and Pieters 2004): positive (e.g., smile), negative (e.g., frown), and neutral (i.e., expressions that have no effect on determining whether the customer liked or disliked the garment). We adopted Zeenlenberg and Pieters' (2004) three-expression scheme rather than Ekman and Friesen's (1971) seven-expression scheme because several of Ekman and Friesen's primary expressions rarely occur in a garment shopping context (e.g., sadness and fear), and the latter system performs well only when the intensity of expressions is high.

[11] The MATLAB computer vision toolbox was used to detect faces in the video frames.
[12] We describe the procedure in more detail in online Appendix A (see also Viola and Jones 2001, 2004).
[13] See online Appendix A for technical details.

The facial expression training dataset was constructed following standard procedures in the computer science literature for collecting posed facial expression data (Cohen et al. 2003). The facial expression data were collected from 45 participants using a desktop computer with a camera mounted on top of the screen. The participants were asked to make each type of facial expression three times for the camera. First, they were asked to make a facial expression reflecting an emotion (e.g., happy, unhappy, neutral); then they were asked to read a sentence which was intentionally picked to trigger a certain emotion, causing them to make the facial expression naturally. Finally, the participants were asked to make the specific facial expression again. The expression sequences were randomized for each subject and the entire process was video recorded for each participant. After the training videos were collected, frames were extracted from each video. For each participant, four images were manually selected by three human judges from the video frames that best represented each of the three facial expressions (positive, negative and neutral) for each participant. Altogether, we collected 540 training images (4 facial images $\times 3$ facial expressions $\times 45$ participants) for the facial expression training dataset.

To test the performance of our model in correctly recognizing expressions from frames extracted from videos, we collected a test set of 100 facial expression videos from a different set of participants using the previously described procedure; each participant posed one of the three facial expressions (happy, unhappy, or neutral).[14] We randomly selected 1 frame from each of the 100 video clips and discarded 21 because they captured participants before or after the posed expression. The three human judges manually labeled each of the 79 frames as positive, negative or neutral, and their joint judgments were used as ground truth. We analyzed the 79 test frames and compared the facial expression recognition result with the ground truth. Table 3 shows the confusion matrix of our method for the 79 valid frames; the overall accuracy is 79%. The (correct) recognition rates for positive, neutral and negative expressions are 88%, 81%, and 68%, respectively.

Insert Table 3 Here

Region of interest detection. Region of interest detection involves detecting a hand in a frame and then matching it with the garment region map to identify which region is being touched. When the inventory database contains information that relates a region to a feature (as in this case), the feature of interest can be inferred.

[14] Using posed expressions to evaluate the performance of an expression recognition algorithm is a standard practice in the computer vision literature (Tian et al. 2011).

In hand detection, the hand is differentiated from other objects in the scene in order to determine its location. Because skin color is uniformly distributed in a small region of color space, it can serve as a strong cue for vision-based hand tracking (Jones and Rehg 2002; Yoruk et al. 2009; Zhu et al. 2000). This detection involves three steps.[15] First, the garment area is segmented from each frame using upper body detection (Kruppa et al. 2003), and a gray-world algorithm for color correction (Kovac et al. 2003) is applied to the detected garment area image in order to eliminate noise associated with different illuminant conditions. Second, color segmentation is employed to differentiate skin-color blobs from the background. Finally, assuming the colors of the hands and face are similar within the same image, the skin color of the face is used as a cue to find the hand blobs on garment region. Figure 6 shows an example of the detected skin blob using color clustering and the hand position detected.

Insert Figure 6 Here

Preference inference. As described in detail in Section 3.2.4, we used a simple heuristic to develop inferences of preference toward each garment region/feature and the overall preference toward the garment. Since we had a 1:1 correspondence between features and regions coded in the inventory database, a customer's preference could be inferred for each region (or corresponding feature, as shown in Table 2 in our implementation) of the try-on garment. The system uses the number of frames in which the focal customer displayed positive (coded as +1), negative (coded as -1) and neutral (coded as 0) expressions when touching a particular region/feature of a garment to infer her preferences toward that particular region/feature. The system then sums the scores for each type of expression to estimate her overall preference toward the garment in each specific video clip (try-on incident).

4.3.2. Identifying neighborhood (VAR and SET models)

In the VAR and SET models, a neighborhood is identified for the customer-garment pair in each customer's first try-on incident. Using only the first try-on incident data helped us better evaluate model performance, since the number of garments tried on by participants varied. Moreover, in real life, meaningful recommendations

[15] See online Appendix A for technical details.

may be based on only one try-on incident since customers are encouraged but not required to be tracked across different try-on incidents (see our discussion on first-time users in Section 3.3).[16]

In both the SET and VAR models, we implemented a neighborhood-based CF method using a correlation-based similarity weighting mechanism as described in Sections 3.3 and 4.1. In the SET model, participants' item-level and feature-level ratings were used to calculate the similarity between customers. Both models make recommendations based on nearest neighbors' response data for considered (try-on) garments. A leave-one-out cross validation approach is employed; that is, for each customer, we use the remaining 126 participants' try-on and choice data to make the CF-based recommendation.

4.3.3. Predictive performance (VAR, SEC, SET models)

We adopted metrics typically used in CF applications to compare the three models. Specifically, we used *recall* and *precision* as they are widely used for ranking-based systems in information retrieval research (Herlocker et al. 2004; Su and Khoshgoftaar 2009). If a model is allowed to recommend *t* items (*t* ranges from 1 to 4 in the present study),[17] *recall* (or *hit rate*) is the percentage of the considered (tried-on)/purchased items that are recommended by the model (a higher recall/hit rate indicates better model performance), and *precision* is the percentage of model-recommended items that are actually considered/bought by the customers (higher precision indicates better model performance).

In addition to these two metrics, we used Kullback–Leibler (K-L) divergence—a relative entropy that "measures the expected divergence in Shannon's information measure between the validation data and a model's predictions" (Ding et al. 2011, p. 121)—to investigate the model performance, since K-L measures discriminate among models even when hit rates are the same (Ding et al. 2011; Dzyabura and Hauser 2011; Hauser et al. 2010). Initially, we calculated divergence from perfect prediction, in which a smaller K-L indicates better performance. In order to better interpret the K-L natural bits, we calculated a K-L metric (termed K-L percentage) relative to the K-L divergence of the null model (random recommendation).[18] The K-L percentage

[16] In online Appendix B we describe an alternative VAR implementation; the VAR model performance is quite consistent across a customer's different try-on incidents (and comparable to combining inferences from multiple try-on incidents).

[17] We examined a recommendation set including up to 4 items because the average consideration set observed in the validation stage was approximately 4 items (3.5). Furthermore, it would be difficult to present a larger set to a customer on a computer screen in real life. We also chose to recommend 4 items each time in the second empirical study.

[18] A random recommendation model is used as the baseline in the calculation of K-L percentage, where recommendations are made by randomly selecting from the subset of garments that have been tried by at least one participant. In our case, all

is 0% for the null model and 100% for perfect prediction. Hence, a larger K-L percentage indicates better performance.

We evaluated the three models using these four metrics. The results are reported in Table 4. Table 4a summarizes the ability of each model to predict considerations (i.e., try-ons) for the validation task, while Table 4b focuses on predicting choices (purchases). Comparing the VAR model to benchmark models, namely SEC and SET, VAR model-based predictions are the best on all measures, including recall, precision, K-L, and K-L percentage, across four recommendation set sizes.[19]

More specifically, when predicting considerations, the VAR model performs significantly better than the SEC model across all recommendation set sizes on recall/hit rate and precision measures with $p < 0.10$, and better than the SET model across all recommendation set sizes on recall/hit rate and precision measures, but not significant. On the K-L measure, the VAR model performs better than SEC and SET models, but not significant. On the K-L percentage measure, the VAR model performs significantly better than the SEC model when the recommendation set size is 2 or 3 with $p<0.05$, and better than the SET model, but not significant except when the recommendation set size is 2 ($p < 0.05$). The differences between the SEC and SET models are not significant across all recommendation set sizes on all four metrics.

When predicting choices, the VAR model performs better than the SEC model on recall/hit rate and precision measures, but not significant, except when the recommendation set size is 3 ($p < 0.10$). The VAR model performs better than the SET model on recall/hit rate and precision measures across all recommendation set sizes, but not significant. The VAR model performs better than the SEC and SET models on the K-L measure, but not significant. On the K-L percentage measure, the VAR model performs significantly better than the SEC model when the recommendation set size is greater than 2 with $p<0.10$. The differences between the SEC model and SET models are not significant across all recommendation set sizes on recall/hit rate, precision and K-L measures. In general, the VAR system appears to not predict choices as well as it predicts considerations, possibly because the VAR system relies on a try-on customer database and recommends items that are considered, but not necessarily purchased by customers.

Insert Table 4a and Table 4b Here

---

20 items in the validation task had been tried by at least one participant.

[19] We ran additional VAR models using a generic garment region map instead of a feature map. The results are similar.

These results seem to indicate that implementing and using the VAR is feasible, and that the system provides valuable recommendations. Moreover, the system appears to perform better than the two benchmark models. To investigate the value of the additional complexity associated with feature-level information in the VAR system, we also compared its performance against three simpler benchmark models (results available upon request): (a) a simple CF model based only on purchase history (which can be tracked through a loyalty card or credit card number),[20] (b) a SET model that relies on overall preference only, and (c) a VAR model that relies on overall preference only. The model comparison result again shows that the VAR model reported here is the best or not significantly different from the best on all measures, including recall, precision, K-L, and K-L percentage, across four recommendation set sizes. All of the evidence supports the use of a multiple-component approach that captures feature-level information in the neighborhood CF-based recommender system.

5. Empirical Study 2: A Real-time Implementation of a VAR System

In Study 1, we demonstrated that the VAR model performs better than the non-automated benchmark models. The predictions in Study 1, however, were not made in real time. We now describe a second study demonstrating a real-time implementation of the VAR system and investigate its feasibility and performance.

5.1. Study Setup

We used the mock store from the first empirical study for this study with some additional decorations consistent with a real store. Our inventory consisted of the same 160 garment styles used in the first empirical study (thus allowing the VAR system to use the try-on customer database comprised of data from the 127 participants from the first empirical study), with brand labels removed and all offered at the same "on-sale" price ($16, close to our procurement cost). When a garment style sold out in the store, it became a catalogue item. Catalogue items were recommended by the VAR system; although participants were unable to try them on, they indicated whether they liked the recommended catalogue items and would try them on if available.

Following the general structure in Figure 1, the VAR system used in Study 2 was comprised of one central computing unit and one pair of input/output devices. The central computing unit was a desktop computer (3.60

[20] In the simple CF model, we select customers from the purchase history database who have purchased the garment that the focal customer is currently trying on as the neighborhood of the focal customer, and then use neighbors' past purchases to make recommendations to the focal customer.

GHz CPU with 16.0 GB RAM) located offsite from the mock store, which contained the inventory database (of the 160 garments), the try-on customer database (with data from the 127 participants from the first empirical study), and the codes for the VAR model. The computer was used to process video captured and generate appropriate recommendations in real time. Video analysis began when the frontal upper body was detected in the frame, and ended when either the frontal upper body could not be detected consecutively for 10 frames or the evaluation time had exceeded 50 seconds. (Note that the average evaluation time was 20 seconds per try-on incident in the first empirical study). The average time it took the system to analyze a video and make recommendations was about 3 minutes. (We discuss how we can substantially reduce this processing time in Section 6).

The input/output devices were located in the mock store, and consisted of a webcam connected to a laptop.[21] The laptop transmitted the video captured by the webcam to the central computing unit via the Internet, and displayed recommendations (garment images and IDs) sent by the central computing unit. We did not have a barcode reader (which would be used to identify the try-on garment in real life) in our mock store, so each participant was asked to manually type in the garment ID before she began to evaluate it in front of the mirror (thus activating the VAR system).

5.2. Procedure and Results

We kept the mock store open for 3 days, Wednesday through Friday, 9:00 am to 5:30 pm. In order to attract participants, we offered them the equivalent of $8 each for stopping by, which they could use to purchase garments in the store. A total of 25 female participants visited the store; none of them had participated in the first empirical study and they were demographically similar to participants in the first study.

When a participant came in, the store assistant (research assistant) provided a brief explanation about how the VAR system works, and explained that she would be filling out a short survey at the end of the shopping experience. The participant then selected a garment to try on from the rack, and activated the VAR system. Once the offsite central computing unit identified a set of four garments to recommend, their images and ID numbers were sent via the Internet and displayed on the laptop screen in the mock store. The participant then viewed the recommended garments, indicated which of the four recommended garments she would like to try on, and

[21] The laptop would not be required if an IP address-enabled webcam and display screen were used.

selected one garment to try on in the next round. If she did not want to try any of the recommended garments, or if the garment she wanted to try on next from the recommendation set was out of stock, she went back to the rack and selected a different garment to try. The process was repeated until the participant did not want to try on more garments and made a purchase decision (or not, if she found nothing she liked).

The 25 participants tried 101 garments in total, an average of 4 garments per person (ranging from 3 to 6). The VAR system made 94 real-time recommendations (each time recommending a set of 4 garments), but failed to make recommendations for 7 try-on incidents because it either failed to detect a face in the video (e.g., hair covered the face, or the participant bowed her head too much) or no suitable nearest neighbors could be identified for the focal customer because she had negative similarity scores with all candidate neighbors. During each participant's last round of recommendations, shoppers were asked to indicate which of the recommended garments (or none) they would like to try if they had the time and energy to try more.

The results of the second empirical study provide further support for the usefulness of the VAR system in real time. Figure 7 shows the distribution of the number of would-have-tried garments[22] across all recommendation sets. In almost all cases, participants indicated that they would like to try 1 or 2 garments from the set of 4 recommended garments. In just five cases out of 94, participants indicated no interest in any of the 4 garments recommended. Since some items recommended and liked by participants were out-of-stock (catalogue items), 56 items recommended by the VAR model were actually tried on by participants.

A simple CF recommendation system based on purchase histories tracked via loyalty cards or credit card numbers seems like a natural first choice for garment retailers. Using actual try-ons (excluding the first garments selected by the participants) as the ground truth, we compared the performance of the VAR model against a simple CF model[23] based on purchase history. With a recommendation set size of 4, the VAR model predicts considerations significantly better than the simple CF model on all measures, namely recall/hit rate, precision, K-L, and K-L percentage with $p = 0.00$. At the end of their shopping trips, 64% of participants indicated that they liked the garments recommended by the VAR system.

Insert Figure 7 Here

[22] Would-have-tried items are the items that customers were interested in trying on, but could not because the recommended items were catalogue items.

[23] See Footnote 20 for the recommendation mechanism of simple CF model. Like the VAR model, the recommendation set size for the simple CF model was set as 4.

Seven of the 25 participants bought a total of 8 garments, and 4 (50%) of these garments were VAR recommendations. Four additional participants told us that they would have liked to have purchased the garments recommended by the VAR system, but were not able to because they were catalogue items (i.e., we did not have the actual garments in the mock store), and one other participant would have bought a garment recommended by the system if we had had a larger size. With a recommendation set size of 4, the VAR model predicts choices significantly better than the simple CF model based on K-L (p=0.08) and K-L percentage (p=0.08) measures with $p < 0.10$, and better based on recall/hit rate and precision measures, but not significant (p=0.34).

5.3. Feedback on the VAR system

Each participant filled out a short survey at the end of her shopping trip on the overall shopping experience and provided reactions to the real time VAR system she had just used. All participants indicated that their privacy had been well protected during the shopping experience. Most participants also thought that their experiences had been very similar (20%) or similar (44%) to their regular shopping experiences. Some participants said, "It looks quite like a real fashion store, very comfortable and relaxing;" and "I feel no difference from a regular shopping trip."

Regarding the VAR system in particular, 56% of participants thought the system was easy or very easy to use; only 16% thought the system was somewhat difficult to use, mostly because of the long wait time (3 minutes for each recommendation) in the current implementation; and none thought it was difficult or very difficult to use. Over 70% of participants indicated that they would like to continue using the recommender system the next time they go shopping, and they would recommend the system to their friends. If a similar store (with only an automated recommender system and no salesperson following participants around) existed in the market, 89% of participants indicated that they would like to shop there.

Several participants described why they liked the VAR system: "The system is like magic! It made recommendations on some clothes that I would never pick myself, but when I follow the recommendations and try them on, they look amazingly good on me;" "There are too many clothes displayed in a store. I don't have the time or effort to carefully check every piece of clothing. The system saves me a lot of time and effort by finding what I like in several rounds;" "I trust the recommendations from a computer much more than a

salesperson, because the computer is more objective, scientific and systematic;" "I feel the system knows my preferences better than a regular salesperson;" "I don't trust a salesperson's recommendations since their recommendations are often very subjective, and they make recommendations based on what they would like to sell, rather than how I look."

Participants also identified a few aspects of the VAR system that should be improved. The biggest concern was the wait time. They indicated that the ideal wait time for recommendations should be 1 to 5 seconds; they would be willing to wait up to 30 seconds after the system had proven itself to be very helpful to them. Some participants stated that the quality of the recommendations after the first try-on would likely determine whether they would use the system in the future. They also suggested that recommendations could be presented better. Instead of the web-quality garment photo used in the second study, high-resolution photos, 3-D photos, and/or images of garments on models would be more desirable.

6. Implementation and Scaling Up in the Real World

In this section, we discuss a few key implementation concerns from the perspective of both customers and retailers, followed by detailed discussion on scaling up the VAR system to large garment retailing chains.

6.1. Implementation Issues

Some implementation challenges relate to customers' attitudes toward the VAR system. We surveyed 417 people about their attitudes and willingness to use the SET and VAR recommender systems based on descriptions of the purpose, procedure, and customer effort required for each. We refer to this survey as the *attitude survey* in the rest of this paper, and discuss some of the relevant survey results below where appropriate.

6.1.1. Opt-in and Linked Try-on Data

The VAR system does not need to store the raw video data for each try-on incident. It analyzes customers' try-on videos in real time, and then stores customers' reactions to various regions (features) of interest for each garment in the database. Each try-on experience is thus simply a numeric record in the database, with variables identifying the garment, customer and preferences inferred from each try-on video. Nonetheless, customers must be willing to make some small privacy sacrifices in order for the VAR system to perform well. Specifically: (a) customers must be willing to be videotaped while evaluating new garments in order to receive personalized

recommendations; and (b) some customers must be willing to be tracked across multiple try-on experiences in order to build the try-on customer database for CF-based recommendations. It is worth noting that once data across different try-on incidents are linked, the VAR system does not need to maintain customer identifiers such as face images (thus ensuring privacy) in order for CF to work.

The general environment nowadays has made it much easier for customers to accept being videotaped. Retailers, for example, have been using surveillance video to monitor customers' purchase processes for quite a while (Lyon 2001); thus, being videotaped during the shopping process is not novel to customers. In our studies, the participants seemed to be open to being videotaped during the evaluation process. Yet retailers must exercise caution and investigate their target customers' acceptance of the practice before widespread implementation.

The bigger challenge, we believe, is convincing customers to be tracked across multiple try-on experiences. Two possible approaches to help address this challenge are allowing multiple levels of opt-in and providing additional incentives for tracking. A retailer could offer multiple levels of opt-in such as: (a) no tracking; (b) tracking during a single shopping trip, but deleting all personal identifiers after a day; (c) tracking during all shopping trips at the store; and (d) tracking during all shopping trips at all stores in the chain. The customer can also be given the option to save certain try-on videos in the system for later viewing. Since almost all retail stores use surveillance cameras to record customers' activities and then erase the videos after a few days, we suspect many customers will be amenable to the idea of being tracked during a single shopping trip (i.e., having their identifiers such as face images erased at the end of the business day). All opt-in levels (except no tracking) will contribute useful records to the database.

There is an inherent opt-in rate, but it is also possible to entice customers to allow the system to keep their face images and track them across try-ons by promoting the benefits of doing so: customers can not only play back prior try-on videos, but also view extracted frames from the videos so as to easily compare how they look in different garments. In the attitude survey, we asked participants to allocate 100 points across eight potential benefits of a recommender system; the average weight that they assigned to playing back their prior try-on videos is 13.67 (SD = 8.76), and the average weight that they assigned to being able to compare frames from prior try-on videos is 19.36 (SD = 10.29). When asked to tell us their acceptance of "allow the automatic system to record my try-on history" on a scale from 1 (completely unacceptable) to 7 (completely acceptable), they

provided an average rating of 5.30 (SD = 1.47). These numbers seem to indicate that people are generally open to being tracked in this way. Furthermore, the additional benefits are important to participants, and may help convince more individuals to opt-in (at least during the same shopping trip).

On the technical side, some type of customer tracking technology must be implemented in the VAR system, such as face recognition (see Section 3). Although we did not use face recognition to track customers in our empirical studies, the system's face recognition function is very accurate.[24]

6.1.2. Customer effort

A customer who uses the VAR system must exert extra effort. Specifically, a customer must (a) step out of the fitting room to look in the mirror outside the fitting room; and (b) scan the garment's bar code to opt-in and activate the system. In the attitude survey, we asked participants to respond to the following statements using a 7-point scale (1 - completely disagree, 7 - completely agree): "It is not inconvenient to step out of the fitting room to use the video based recommender system" (Mean = 5.02; SD = 1.71); and "It is not inconvenient to step out of the fitting room to use the mirror" (Mean = 5.15; SD = 1.72). Their responses seem to indicate that while additional effort must be expended, on average, most people do not see it as a major hurdle.

Customers can use a handheld scanner to scan a garment ID and opt-in to use the system. Using the same 7-point scale, we asked customers whether they would accept scanning the garment ID themselves. Although responses were positive (Mean = 5.54, SD = 1.38), this may still be a challenge if the tag is on the back of the garment or in a hard-to-reach place. This may require extra work from the retailers' side to attach tags in locations that are easy for customers to scan while wearing a garment.

6.1.3. Addition of new garments or new garment features

The VAR system only works if a garment (identified in a try-on video or a possible candidate for recommendation) is in both the inventory and the try-on customer database. Adding new garments with potentially new features to an inventory database can be costly if the retailer wants to use garment-specific region maps, even if it only needs to do it once for all stores in the chain. To address this issue, a retailer may choose to use a generic garment region map when adding new garments to the inventory database, since such a map requires no additional coding effort and has been shown to generate consistently accurate recommendations

[24] We randomly selected 27 frames with frontal faces of 27 different participants described in Section 4, and the system was able to correctly match each with one of the 127 faces in the database.

similar to a system using item-level garment region maps (which enable a 1:1 correspondence between a region and a feature) in our first empirical study (see Footnote 19). As a middle-ground solution, several garment region map templates could be coded first, and a worker would only need to assign the appropriate one to a new garment. For the first few customers who try a recently-added garment, a VAR system based on a try-on customer database would be unable to make recommendations since no other try-on data for that garment would be available, making it impossible to find neighbors (referred to as *new item problem* in the CF literature) (Bobadilla et al. 2012). To fill this initial recommendation void, a common practice is to ask a set of motivated customers to rate new items (Bobadilla et al. 2012). The try-on customer database could also be seeded by recommending recently-added garments to randomly-selected customers (e.g., 1 out of 4 in each recommendation set is a new garment with minimal try-on data in the customer database).

6.2. Scalability of the VAR System

Large garment retailers usually have hundreds of retail stores (typically with multiple fitting rooms in each), and offer thousands of products. Several challenges are associated with implementing a VAR system in such contexts: (a) processing video as video length increases; (b) making CF-based recommendations as the numbers of features, customers, and garments increase; (c) running the system in a large number of stores where many customers may be using the system simultaneously; and (d) the costs of achieving these computation goals.

6.2.1. Video processing

Since the VAR system analyzes video information at the frame level, video processing time grows linearly with video length. For each frame, the VAR system performs two main tasks: facial expression recognition and hand detection. As a result, video processing computing time is approximately linear with the length of the try-on video. Note that with current video streaming technology, the latency of real-time video transmission is usually less than 150 ms and is independent of the length.

To investigate scalability regarding the video processing time, we explored whether we could reduce the 3-minute average processing time reported in Study 2 on a remotely-located desktop equipped with a 3.60 GHz CPU and 16.0 GB RAM. The video processing time can be improved through at least two approaches: code optimization, which requires no hardware upgrades or technology investments; and parallel computing, which

invokes some hardware requirements and potentially costs more. We investigated the potential to improve video processing time using these two strategies.

We performed an extensive analysis on the video processing time and found that the most time was spent reading unstructured video information (i.e., frames from the video file); this can be optimized by substantially reducing the I/O time required. We also investigated the benefit of using parallel computing; since video processing is done at the frame level, each frame of video can be analyzed on a different CPU core. In order to gauge the reduction in video processing time using the two strategies, we randomly selected two video clips for each of the five lengths (10s, 20s, 30s, 40s, 50s) and calculated the average processing time for each pair. We then compared the total video processing time of the three models: (a) the original code we used in Study 2 (*total-before*), (b) optimized code without parallel computing (*total-nopar*), and (c)optimized code with parallel computing (*total-par4*). We analyzed all videos on a desktop with four Intel 3.6 GHz i7-4790 CPU cores and 16.0 GB RAM. The processing was split across all four cores in the third model only. The results are reported in Figure 8. For a 20s try-on video from our empirical studies, processing time was reduced from 264s in the original model to about 40s when both code optimization and parallel processing were employed. In a real life implementation, one can rent even a 40 vCPU at a very affordable price from the Amazon cloud computing service, thus further decreasing the processing time substantially. These results provide good evidence that the video processing time can be shortened sufficiently to avoid unacceptable delays in real life applications.

Insert Figure 8 Here

6.2.2. Neighborhood-based CF algorithm

Two factors may affect the computing time of the CF recommendation algorithm: the number of customers and the number of garments. Computing time is only marginally related to the number of garment features involved, as the similarity score between two customers is calculated only once in our model, regardless of the number of features. The increase in computing time is due to the increase in the vector size involved in the calculation.

Although the computational complexity of neighborhood-based CF is not NP hard (Su and Khoshgoftaar 2009), scholars have shown that a naive neighborhood-based CF algorithm could have limited scalability for large datasets. With millions of customers and millions of items, even a CF algorithm with the complexity of

O(n)[25] could not react immediately to real-time requirements (Su and Khoshgoftaar 2009). Thus, like other CF-based recommender systems used in practice, two key modifications must be made to our model to ensure that the system can generate real time responses.

The first modification is to adopt top-K filtering and use only a small set of "best" neighbors when the database is large, instead of using all neighbors that satisfy a set of rules (the latter strategy tends to yield a huge number of nearest neighbors as the number of customers and garments increases). A typical number suggested in the literature is 20 to 50 (Bell and Koren 2007). Our simulations show that the computing time for a database with 500,000 customers, 3,000 garments, and try-on data for 30 garments for each customer can be reduced from over 2400s when using all non-negatively correlated candidate neighbors to about 30s when using the top 50 nearest neighbors on a 2.6 GHz PC with 16 GB RAM. According to our simulation results,[26] when the top-K filtering method is used for the recommender system, the computing time increases in an approximately linear fashion with both the number of customers and the number of garments.

The second modification is to implement the recommender system in a computing environment via a distributed data processing framework, which one can rent from Amazon (see example in Section 6.2.4). Most of the computing time is spent searching for nearest neighbors and garments tried for a specific user, which are processes that can be parallelized. Thus, according to Amdahl's law, using parallel computing—specifically, a cloud computing platform—would help to reduce the computing time required to make one recommendation to an acceptable level, even when huge datasets are used. For example, large web companies such as Twitter use clusters of machines to scale recommendations for their millions of users, and most computations are performed on machines with large amounts of memory (e.g., 1 TB RAM) (Gupta et al. 2013).

6.2.3. Concurrent recommendations

The ability to provide concurrent recommendations also needs to scale up easily as the number of stores (i.e., the number of end-user devices) increases, since more customers will be using the system simultaneously. Fortunately, this is a general problem that has already been addressed by major commercial firms, typically through the use of cloud computing and database services for data distribution and replication, enabling the

[25] O(n) is a general notation representing the time complexity of an algorithm. For an algorithm with a time complexity of O(n), as input size increases to infinity, the computing time increases linearly with the size of the input.
[26] Simulation results are available upon request.

demand for computing power to be scaled up and down in real time. For example, Cassandra (NoSQL Database) is used by firms such as Netflix, Twitter, and Cisco to achieve this objective (Klein et al. 2015; Klems et al. 2012). Note that normally a retail chain would rent such services from firms such as Amazon (see Section 6.2.4), instead of building and owning them in-house. The former is a lot more cost effective and creates much more flexibility to respond to demand fluctuations (scale up and down). To ensure fast recommendation, the retailing chain should reserve sufficient instances from service providers like Amazon in accordance to predicted number of concurrent use of the VAR system at different time period.

6.2.4. Cost of achieving desired computation capacity

System costs are comprised mainly of hardware costs (i.e., user interfaces and central computing unit devices) and human costs. The cost of a user interface mainly amounts to the cost of webcam and an LCD screen. Assuming the average cost for such a system is $400 (based on current prices listed on the Internet), the total cost of the user interface would be $400 * X * Y, where X is the number of stores and Y is the average number of fitting room interfaces per store. A chain with 1,000 stores and 10 fitting rooms per store would spend $4 million on user interfaces. The cost of central computing unit depends on whether a company decides to buy (establish internal computing capability) or rent (use computing capability from a provider). In the rent model which currently is very popular, a firm would rent a cloud computing engine from a provider such as Amazon(e.g., Amazon EMR (Elastic MapReduce) service combined with its Amazon Elastic Compute Cloud (Amazon EC2).) , cloud database engine and storage space (e.g., Amazon Relational Database Service (Amazon RDS) for Amazon Aurora). Based on a back-of-the-envelope calculation,[27] the variable cost of generating one recommendation would likely be a few cents. There would be also be fixed costs associated with renting cloud database storage space (10 TB of data storage costs about $1,000/month at Amazon). Finally, experienced engineers must be employed to develop, deploy and maintain the central computing unit, and people must be hired to set up and maintain the user interfaces.

These costs are manageable, but not trivial. It is thus important for firms to perform careful cost benefit analyses for their specific situations before deciding whether or not to deploy a VAR system. It is also advisable that firms perform initial small-scale tests of the system to gauge their own specific benefits and costs.

[27] Details available upon request.

7. Mining Information from Try-on Video Data

In addition to providing individualized recommendations, try-on videos enable firms to mine valuable information when pooled across individuals and garments. Video data from the retail industry are a main source of (unstructured) big data, as in-store data are rich in volume, variety and velocity (i.e., changing in real-time). At least two types of valuable insights can be mined from large scale try-on video data, which cannot be obtained from other data sources: (a) haptic evaluation patterns (i.e., customers' hand movements when evaluating a particular garment or garment type); and (b) reasonably accurate demographic (e.g., age) and physical attributes (e.g., height, body shape) of customers who have tried each garment. Useful information can be mined from a large number of try-on videos with minimal privacy sacrifices from customers.

In the rest of this section, we illustrate how haptic pattern data can be mined from try-on videos and discuss how they might be used. For ease of interpretation, we present haptic pattern data as superimposed images that we call *hand heat maps*. A hand heat map is customer normalized and centered, with dots representing the hand positions in all frames examined, and different dot colors representing the density of dots in a particular position. Figure 9 shows a simple example with four hand heat maps (2 customers $\times$ 2 garments). The first column presents a garment image (ID 5), followed by the hand heat maps of two different customers from their try-on videos for this garment. The second column presents a second garment image (ID 76), followed by two similarly-obtained hand heat maps. Each dot indicates the hand position in a frame. We use three colors (red for one observation, purple for multiple observations, and blue for the highest concentration of observations) to represent different dot densities, but a continuous scale can be employed if needed.

Insert Figure 9 Here

A comparison of the hand heat maps reveals several useful pieces of information. First, the density measurements highlight regions of a garment that seem to attract the most attention. Second, a particular customer's evaluation patterns can potentially be discerned by comparing her hand heat maps for different garments. For example, comparing the two hand heat maps of the first customer across two different garments (row 2), we notice that she touched the hemline quite often, indicating that this region of the garment was important to her evaluation. Third, we might be able to infer the important regions of a particular garment for a group of customers by comparing hand heat maps of multiple people (possibly from the same target segment)

for the same garment. Comparing the two hand heat maps of the second garment across two different customers (column 2), we notice that the first customer evaluated how the garment looked on her when she stretched out completely (hand positions scattered around), and the second customer was concerned about how the garment flowed around her shoulder (hand positions concentrated in the shoulder region). In addition, it appears that both customers cared about how the outer part of the hemline looked on them.

In real life implementations, these individual level (customer × garment) hand heat maps can be pooled for further data mining, for example, across all garments tried by a single customer or a segment of customers, or across all customers who have tried a particular garment or garment type, to provide more robust insights about a customer or customer segment, a garment or garment type, and various combinations of customers and garments. Such insights mined from large-scale try-on videos could be quite useful for designers and retailers. For example, when hand heat maps are combined with the respective purchase data, designers and retailers may be able to identify additional underlying reasons for the (un)popularity of items, and adjust inventories, displays, promotions and future designs accordingly.

8. Discussion

In this paper, we proposed a video-based automated recommendation system for use in garment retail stores, with potential benefits to both customers and retailers. The VAR system may be implemented with several variations. We demonstrated one such implementation and compared its usefulness to two state-of-the-art (individual level but not automated) benchmark product recommendation models in an empirical study. We also demonstrated the system's real-life feasibility in a second empirical study in which we provided real-time recommendations to shoppers.

We contribute to the marketing literature in two ways. First, to the best of our knowledge, we are among the first to create a system based on video data (as opposed to survey and scanner data) to estimate customer preferences in a retail context. Second, unlike existing marketing studies in which commercial software packages were used for data analysis, we are among the first to write and test algorithms for video analysis in marketing, thus opening the black box of existing commercial software. This should help pave the way for new modeling innovations in marketing related to video data.

Through two empirical studies, we demonstrated a proof of concept for the VAR system in garment retailing contexts. We did not exhaustively compare our model to all possible benchmark models, nor do we claim that our implementations of benchmark models are the best ways to implement. It is worth to further test other types and implementations of benchmarks using larger datasets collected from real-life retailing contexts. While potential benefits of the VAR system include generating individualized recommendations, decision aids for consumers and insights from rich and big data, a firm should evaluate the tradeoffs between specific benefits and costs when comparing the VAR to other recommendation systems. It will be interesting to see what type of stores and contexts are more suitable for the implementation of VAR system.

Since this is the first paper in this domain, much work needs to be done to further improve the model prior to widespread implementation and many promising extensions can be made in future research. First, it is worthwhile to explore and test alternative implementations of the VAR system. Model performance also can be improved by identifying and removing irrelevant information from video data, for example, habitual or random facial expressions not related to the garment itself, or behaviors (facial expressions and hand positions) related to interactions with other people (e.g., friends). Furthermore, customer preferences could be inferred more accurately by utilizing other information in the videos, such as an individual's weight, height, age, skin tone, etc. This information can be incorporated into the VAR system to help improve the system's performance. It may also be worthwhile to explore other methods of detecting regions of interest, such as eye tracking (Hui et al. 2013). Moreover, researchers can explore whether a VAR system can be used in other retailing contexts.

Video data are a new source of information in the retailing context. Applications of video analysis in marketing may include detecting gender, skin color and eye gaze; recognizing facial expressions and body or hand gestures; tracking trajectories; and counting people (see Xiao et al. 2013). As a first attempt to use video analysis to infer customers' individual preferences, the VAR system may yield benefits such as reducing customer searching effort, increasing retail sales by recommending garments that customers are likely to purchase, and helping companies adjust designs or inventory to match customer preferences. As discussed in Section 6, the costs associated with implementing a large-scale VAR system are nontrivial. A firm should carefully weigh the expected benefits and costs before deploying such a system. We hope that this new method can become a valuable tool to retailers and create benefits for both customers and retail stores.

References

Abboud B, Davoine F, Dang M (2004) Facial expression recognition and synthesis based on an appearance model. *Signal Processing: Image Communication* 19(8): 723–740.

Abraham-Murali L, Littrell MA (1995) Consumers' conceptualization of apparel attributes. *Clothing and Textiles Research Journal* 13(2): 65–74.

Adomavicius G, Kwon Y (2007) New recommendation techniques for multicriteria rating systems. *Intelligent Systems, IEEE* 22(3), 48–55.

Adomavicius G, Manouselis N, Kwon Y (2011) Multi-criteria recommender systems. Ricci R, Rokach L, Shapira B, Kantor PB, eds. *Recommender Systems Handbook* (Springer, New York), 769–803.

Aljukhadar M, Senecal S, Daoust CE (2012). Using recommendation agents to cope with information overload. *International Journal of Electronic Commerce* 17(2): 41–70.

Ariely D, Lynch Jr JG, Aparicio IV M (2004). Learning by collaborative and individual-based recommendation agents. *Journal of Consumer Psychology* 14(1): 81–95.

Basil M (2011) Use of photography and video in observational research. *Qualitative Market Research: An International Journal* 14(3): 246–257.

Belk R (2011) Examining markets, marketing, consumers, and society through documentary films. *Journal of Macromarketing* 31(4): 403–409.

Belk RW, Kozinets RV (2005) Videography in marketing and consumer research. *Qualitative Market Research: An International Journal* 8(2): 128–141.

Bell RM, Koren Y (2007) Scalable collaborative filtering with jointly derived neighborhood interpolation weights. *Proceedings from the Seventh IEEE International Conference on Data Mining (ICDM)*: 43–52.

Bloch PH (1995) Seeking the ideal form: Product design and consumer response. *Journal of Marketing* 59(3): 16–29.

Bobadilla J, Ortega F, Hernando A, Bernal J (2012) A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-Based Systems* 26: 225–238.

Chavan UB, Kulkarni DB (2013) Facial expression recognition—review. *International Journal of Latest Trends in Engineering and Technology (IJLTET)* 3(1): 237–243.

Cohen I, Sebe N, Garg A, Chen LS, Huang TS (2003) Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding* 91(1): 160–187.

Cordier F, Seo H, Magnenat-Thalmann N (2003) Made-to-measure technologies for an online clothing store. *IEEE Computer Graphics and Applications* 23(1): 38–48.

DeSarbo WS, Ramaswamy V, Cohen SH (1995) Market segmentation with choice-based conjoint analysis. *Marketing Letters* 6(2): 137–147.

Desrosiers C, Karypis G (2011) A comprehensive survey of neighborhood-based recommendation methods. Ricci R, Rokach L, Shapira B, Kantor PB, eds. *Recommender Systems Handbook* (Springer, New York), 107–144.

Ding M, Hauser JR, Dong S, Dzyabura D, Yang Z, Su C, Gaskin S (2011) Unstructured direct elicitation of decision rules. Journal of Marketing Research 48(1): 116-127.

Dong S, Ding M, Huber J (2010) A simple mechanism to incentive-align conjoint experiments. *International Journal of Research in Marketing* 27(1): 25–32.

Dzyabura D, Hauser JR (2011) Active machine learning for consideration heuristics. *Marketing Science* 30(5): 801–819.

Eckman M, Damhorst ML, Kadolph SJ (1990) Toward a model of the in-store purchase decision process: Consumer use of criteria for evaluating women's apparel. *Clothing and Textiles Research Journal* 8(2): 13–22.

Ekman P (1994) Strong evidence for universals in facial expression: A reply to Russell's mistaken critique. *Psychological Bulletin* 115: 268–287.

Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17(2): 124–129.

Fang F, Chen YE, Lu H (2013) Research on building the preference model of seamless badminton sportswear based on conjoint analysis. *Advanced Materials Research* 796: 462–467.

Fasel B, Luettin J (2003) Automatic facial expression analysis: A survey. *Pattern Recognition* 36(1): 259–275.

Franke GR, Park JE (2006) Salesperson adaptive selling behavior and customer orientation: A meta-analysis. *Journal of Marketing Research* 43(4): 693–702.
Frijda NH (1986). *The emotions* (Cambridge University Press, Cambridge, United Kingdom).
Frijda NH, Zeelenberg M (2001) Appraisal: What is the dependent? Scherer KR, Schorr A, Johnstone T, eds. *Appraisal Processes in Emotion Theory, Methods, Research* (Oxford University Press, New York), 141–155.
Georghiades AS, Belhumeur PN, Kriegman D (2001) From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(6): 643–660.
Goldberg D, Nichols D, Oki BM, Terry D (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM* 35(12): 61–70.
Goldberg K, Roeder T, Gupta D, Perkins C (2001) Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval* 4(2): 133–151.
Green PE, Srinivasan V (1990) Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing* 54(4): 3–19.
Grohmann B, Spangenberg ER, Sprott DE (2007) The influence of tactile input on the evaluation of retail product offerings. *Journal of Retailing* 83(2): 237–245.
Gupta P, Goel A, Lin J, Sharma A, Wang D, Zadeh RB (2013). WTF: The who-to-follow system at Twitter. *Proceedings of the 22nd International Conference on the World Wide Web*, 505–514.
Hauser JR, Toubia O, Evgeniou T, Befurt R, Dzyabura D (2010). Disjunctions of conjunctions, cognitive simplicity, and consideration sets. *Journal of Marketing Research* 47(3): 485–496.
Herlocker JL, Konstan JA, Terveen LG, Riedl JT (2004) Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* 22(1): 5–53.
Holbrook MB (1983) Using a structural model of halo effect to assess perceptual distortion due to affective overtones. *Journal of Consumer Research*, 10: 247-252.
Hui SK, Fader PS, Bradlow ET (2009a) Path data in marketing: An integrative framework and prospectus for model-building. *Marketing Science* 28(2): 320–335.
Hui SK, Fader PS, Bradlow ET (2009b) The traveling salesman goes shopping: The systematic deviations of grocery paths from TSP-optimality. *Marketing Science* 28(3): 566–572.
Hui SK, Huang Y, Suher J, Inman JJ (2013) Deconstructing the "first moment of truth:" Understanding unplanned consideration and purchase conversion using in-store video tracking. *Journal of Marketing Research* 50(4): 445–462
Jones MJ, Rehg JM (2002) Statistical color models with application to skin detection. *International Journal of Computer Vision* 46(1): 81–96.
Karypis G (2001) Evaluation of item-based top-n recommendation algorithms. *Proceedings of the Tenth International Conference on Information and Knowledge Management, ACM*, 247–254.
Kim HJ, Park YH, Bradlow E, Ding M (2014). PIE: A holistic preference concept and measurement model. *Journal of Marketing Research* 51(3): 335–351.
Klatzky RL, Loomis JM, Lederman SJ, Wake H, Fujita N (1993). Haptic identification of objects and their depictions. *Perception and Psychophysics* 54(2): 170–178.
Klein J, Gorton I, Ernst N, Donohoe P, Pham K, Matser C (2015). Performance evaluation of NoSQL databases: A case study. *Proceedings of the 1st Workshop on Performance Analysis of Big Data Systems, ACM*, 5–10.
Klems M, Bermbach D, Weinert R (2012). A runtime quality measurement framework for cloud database service systems. *Proceedings of the Eighth International Conference on Quality of Information and Communications Technology (QUATIC), IEEE*, 38–46.
Kovac J, Peer P, Solina F (2003) Human skin color clustering for face detection. *EUROCON 2003, Computer as a Tool, IEEE Region 8* 2: 144–148.
Kozinets RV, Belk WR (2006) Videography. Jupp V, ed. *The Sage Dictionary of Social Research Methods* (SAGE Publications Ltd., London), 318–320.
Krishna A ed. (2009) *Sensory Marketing: Research on the Sensuality of Products* (Psychology Press, New York), 17–63.
Kruppa H, Castrillon-Santana M, Schiele B (2003) Fast and robust face finding via local context. *Proceedings from the Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, 157–164.

Kuch J, Huang T (1995) Vision-based hand modeling and tracking for virtual teleconferencing and telecollaboration. *Proceedings of the International Conference on Computer Vision, IEEE*, 666–671.
Lee BK, Lee WN (2004) The effect of information overload on consumer choice quality in an online environment. *Psychology and Marketing* 21(3): 159–183.
Lee N, Broderick AJ (2007). The past, present and future of observational research in marketing. *Qualitative Market Research: An International Journal* 10(2): 121–129.
Li C, Kitani KM (2013) Pixel-level hand detection in ego-centric videos. *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3570–3577.
Liu YJ, Zhang DL, Yuen, MMF (2010). A survey on CAD methods in 3D garment design. *Computers in Industry* 61(6): 576–593.
Lyon D (2001) *Surveillance Society: Monitoring Everyday Life* (McGraw-Hill Education, Maidenhead, United Kingdom).
Ma L, Khorasani K (2004) Facial expression recognition using constructive feed forward neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 34(3): 1588–1595.
Manouselis N, Costopoulou C (2007) Experimental analysis of design choices in multiattribute utility collaborative filtering. *International Journal of Pattern Recognition and Artificial Intelligence* 21(2): 311–332.
McCabe DB, Nowlis SM (2003) The effect of examining actual products or product descriptions on consumer preference. *Journal of Consumer Psychology* 13: 431–439.
Miller S (2013) GfK adds facial coding to ad testing system. http://www.research-live.com/news/technology/gfk-adds-facial-coding-to-ad-testing-system/4009252.article
Mitra S, Acharya T (2007) Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 37(3): 311–324.
Narayan V, Rao VR, Saunders C (2011) How peer influence affects attribute preferences: A Bayesian updating mechanism. *Marketing Science* 30(2): 368–384.
Netzer O, Toubia O, Bradlow ET, Dahan E, Evgeniou T, Feinberg FM, ... Rao VR (2008) Beyond conjoint analysis: Advances in Preference Measurement. *Marketing Letters* 19(3/4): 337–354.
Ojala T, Pietikinen M, Harwood D (1996) A comparative study of texture measures with classification based on featured distribution. *Pattern Recognition* 29(1): 51–59.
Oliver N, Pentland A, Berard F (2000) LAFTER: A real-time face and lips tracker with facial expression recognition. *Pattern Recognition* 33: 1369–1382.
Peck J, Childers TL (2003) To have and to hold: The influence of haptic information on product judgments. *Journal of Marketing* 67(2): 35–48.
Peck J, Shu SB (2009) The effect of mere touch on perceived ownership. *Journal of Consumer Research* 36(3): 434–447.
Peck J, Wiggins J (2006) It just feels good: Customers' affective response to touch and its influence on persuasion. *Journal of Marketing* 70(4): 56–69.
Pham MT (1998) Representativeness, relevance, and the use of feelings in decision making. *Journal of Consumer Research* 25(2): 144–159.
Raouzaiou A, Tsapatsoulis N, Tzouvaras V, Stamou G, Kollias S (2002) A hybrid intelligence system for facial expression recognition. *Proceedings of European Symposium on Intelligent Technologies, Hybrid Systems and their Implementation on Smart Adaptive Systems*, 482–490.
Rashid AM, Albert I, Cosley D, Lam SK, McNee SM, Konstan JA, Riedl J (2002). Getting to know you: Learning new user preferences in recommender systems. *Proceedings of the 7th International Conference on Intelligent User Interfaces, ACM*, 127–134.
Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J (1994) GroupLens: An open architecture for collaborative filtering of netnews. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 175–186.
Resnick P, Varian HR (1997) Recommender systems. *Communications of the ACM* 40(3): 56–58.
Roseman IJ, Antoniou AA, Jose PE (1996) Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. *Cognitive Emotion* 10(2): 41–77.
Rowley HA, Baluja S, Kanade T (1998) Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(1): 23–38.

Russell JA, Fernandez-Dols, JM (1997) *The Psychology of Facial Expression* (Cambridge University Press, Cambridge, United Kingdom).
Sahoo N, Krishnan R, Duncan G, Callan JP (2006) Collaborative filtering with multi-component rating for recommender systems. *Proceedings of the 16th Workshop on Information Technologies and Systems*.
Saxe D, Foulds, R (1996) Toward robust skin identification in video images. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 379–384.
Schafer JB, Frankowski D, Herlocker J, Sen S (2007) Collaborative filtering recommender systems. Brusilovsky P, Kobsa A, Nejdl W, eds. *The Adaptive Web. Lecture Notes in Computer Science 4321* (Springer, Berlin/Heidelberg, Germany), 291–324.
Scheibehenne B, Greifeneder R, Todd PM (2010) Can there ever be too many options? A meta-analytic review of choice overload. *Journal of Consumer Research* 37(3): 409–425.
Schwarz N, Clore GL (1983) Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology* 45(3): 513.
Schwarz N, Clore GL (1988) How do I feel about it? The informative function of affective states. Fiedler K, Forgas J, eds. *Affect, Cognition, and Social Behavior* (Hogrefe, Zurich), 44–62.
Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing* 27(6): 803–816.
Shan S, Gao W, Cao B, Zhao D (2003) Illumination normalization for robust face recognition against varying lighting conditions. *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 157–164.
Shergill GS, Sarrafzadeh A, Diegel O, Shekar A (2008) Computerized sales assistants: The application of computer technology to measure consumer interest—a conceptual framework. *Journal of Electronic Commerce Research* 9(2): 176–191.
Srinivasan V, Park CS (1997) Surprising robustness of the self-explicated approach to customer preference structure measurement. *Journal of Marketing Research* 34(5): 286–291.
Su X, Khoshgoftaar TM (2009) A survey of collaborative filtering techniques. *Advances in Artificial Intelligence, 2009*, 4.
Sung KK, Poggio T (1998) Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1): 39–51.
Swan JE, Bowers MR, Richardson LD (1999) Customer trust in the salesperson: An integrative review and meta-analysis of the empirical literature. *Journal of Business Research* 44(2): 93–107.
Tan X, Triggs B (2010) Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Image Processing, IEEE Transactions on Biometrics Compendium*, 19(6): 1635–1650.
Teixeira T, Picard R, el Kaliouby R (2014) Why, when, and how much to entertain consumers in advertisements? A web-based facial tracking field study. *Marketing Science* 33(6): 809–827.
Teixeira T, Wedel M, Pieters R (2012). Emotion-induced engagement in Internet video advertisements. *Journal of Marketing Research* 49(2): 144–159.
Tian Y, Kanade T, Cohn JF (2011). Facial expression recognition. Jain AK, Li SZ, eds. *Handbook of Face Recognition* (Springer, London, United Kingdom), 487–519.
Turk M, Pentland A (1991) Eigenfaces for recognition. *Journal of Cognitive Neuroscience* 3(1): 71–86.
Valizade-Funder S, Heil O, Jedidi K (2012). Impact of retailer promotions on store traffic—a video-based technology. Working paper.
Viola P, Jones M (2001). Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition* 1: 511–518.
Viola P, Jones M (2004). Robust real-time face detection. *International Journal of Computer Vision* 57(2): 137–154.
Wang H, Ahuja N (2003). Facial expression decomposition. *Proceedings of the Ninth IEEE International Conference, Computer Vision 2003*, 958–965.
Weitz, BA, Sujan H, Sujan M (1986). Knowledge, motivation, and adaptive behavior: A framework for improving selling effectiveness. *Journal of Marketing* 50(4): 174–191.
Wyer RS, Carlston DE (1979) *Social Cognition, Inference and Attribution* (Psychology Press, Hillsdale, NJ), 191–210.
Xiao L, Ding M (2014) Just the faces: Exploring the effects of facial features in print advertising. *Marketing Science* 33(3): 338–352.

Xiao L, Kim H, Ding M (2013) An introduction to audio and visual research and applications in marketing. *Review of Marketing Research* 10: 213–253.
Yazdanparast A, Spears N (2012) Need for touch and information processing strategies: An empirical examination. *Journal of Consumer Behaviour* 11(5): 415–421.
Yoruk E, Konukoglu E, Sankur B, Darbon J (2006) Shape-based hand recognition. *IEEE Transactions on Image Processing* 15(7): 1803–1815.
Zeelenberg M, Pieters R (2004) Beyond valence in customer dissatisfaction: A review and new findings on behavioral responses to regret and disappointment in failed services. *Journal of Business Research* 57(4): 445–455.
Zhang X, Li S, Burke R (2012) Modeling the dynamic influence of group interaction and the store environment on shopper preferences and purchase behavior. Working paper.
Zhang Z, Li Y, Gong C, Wu H (2002) Casual wear product attributes: A Chinese consumers' perspective. *Journal of Fashion Marketing and Management* 6(1): 53–62.
Zhao W, Chellappa R, Phillips PJ, Rosenfeld A (2003) Face recognition: A literature survey. *ACM Computing Surveys* 35(4): 399–458.
Zhu X, Yang J, Waibel A (2000) Segmenting hands of arbitrary color. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 446–453.

Figures and Tables

Figure 1. General Design of the VAR System

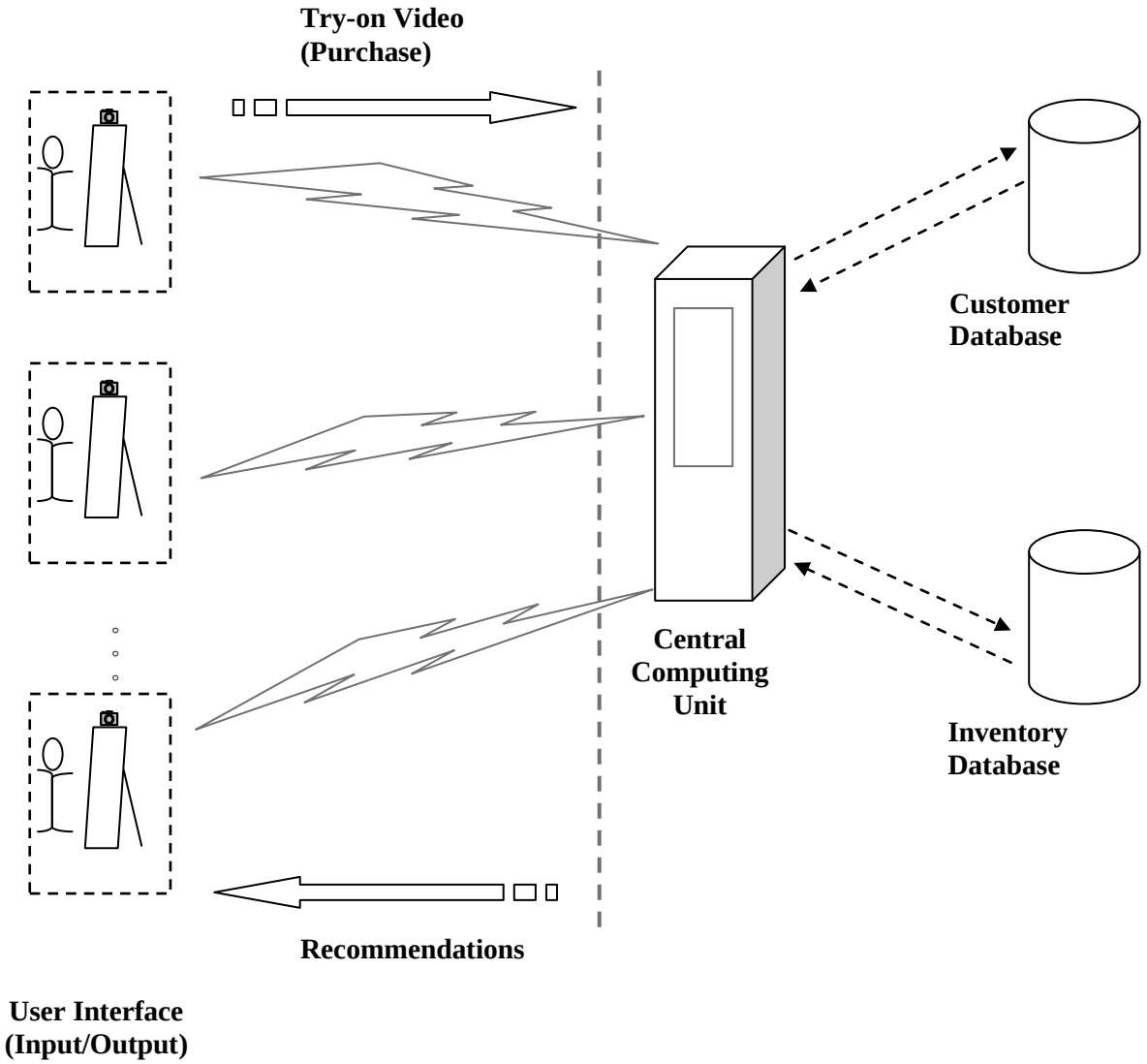
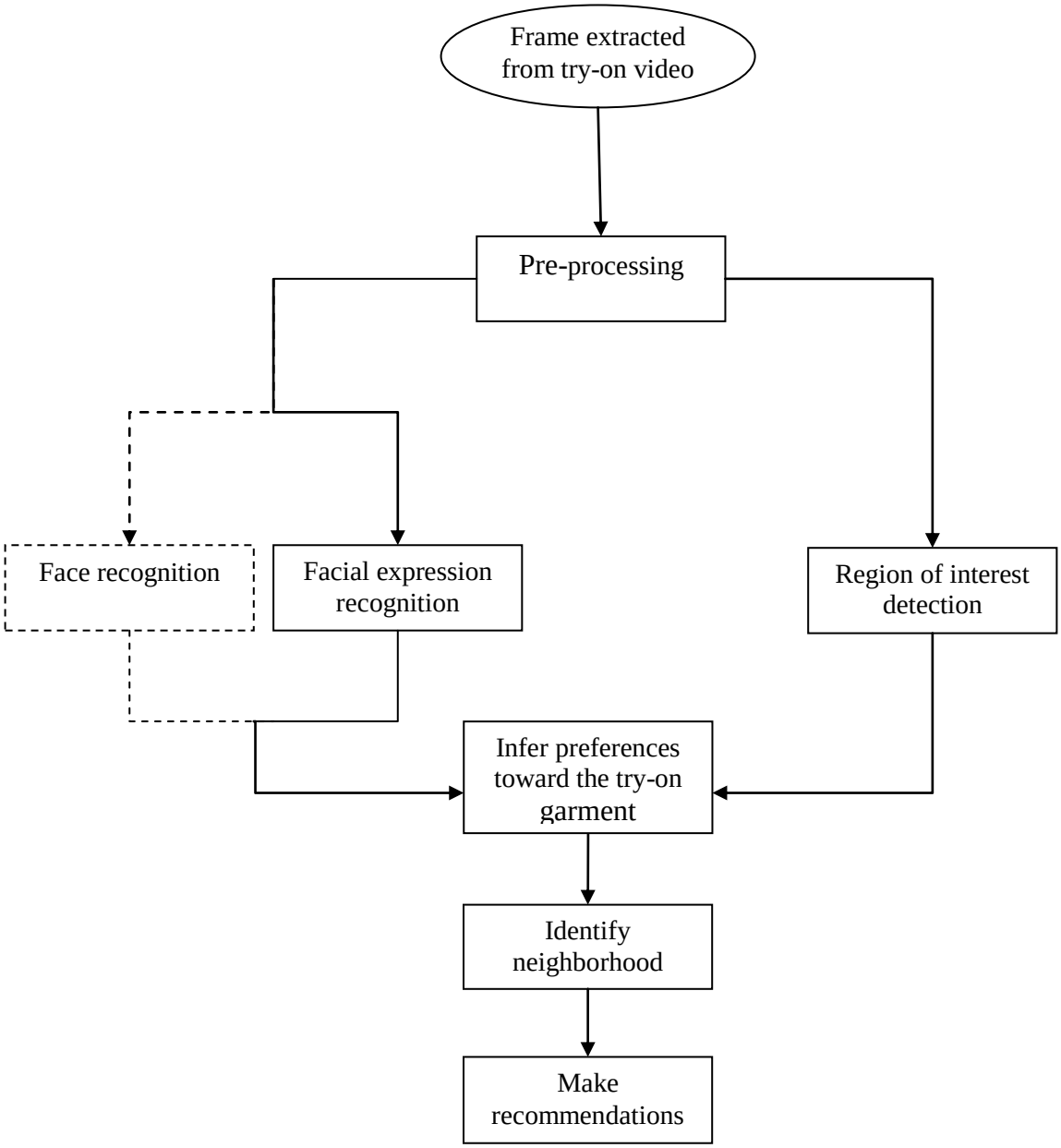


Figure 2. General Process of Video Analysis in the VAR System[28]



[28] We did not use face recognition to track customers in our empirical studies. However, we have checked our codes and the VAR system recognizes faces with high accuracy (see discussion in Section 6.1.1).

Figure 3. An Example of a Garment Region Map

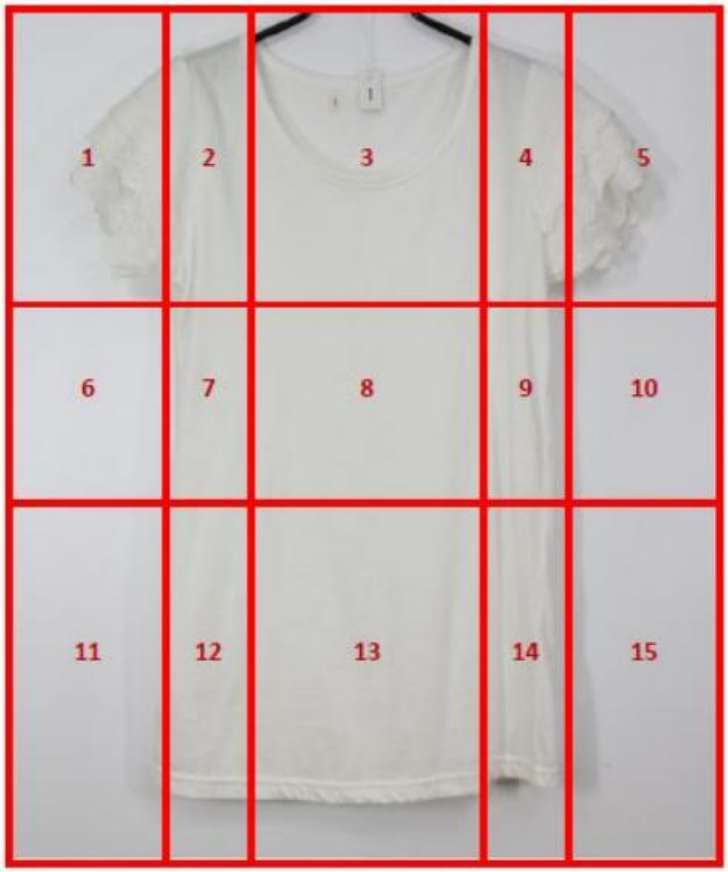


Figure 4a. Evaluation Time per Try-on Item

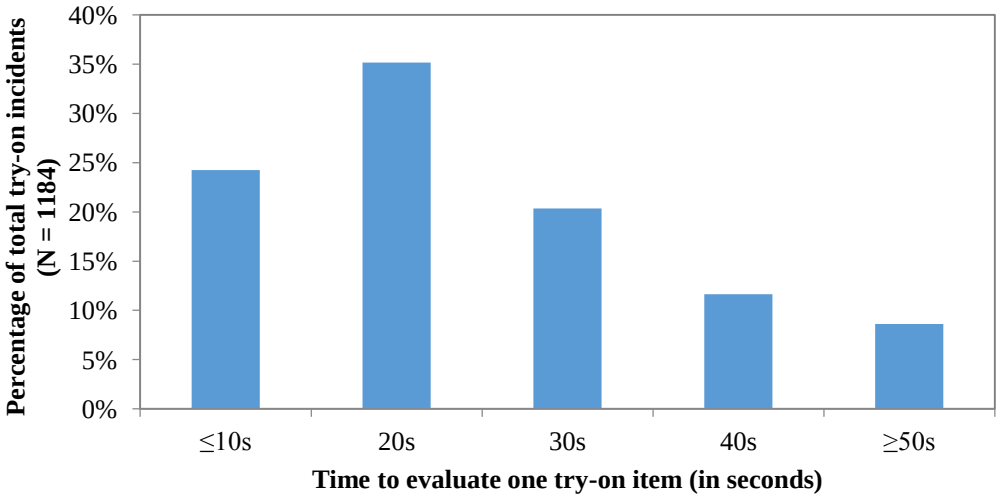


Figure 4b. Total Evaluation Time per Participant

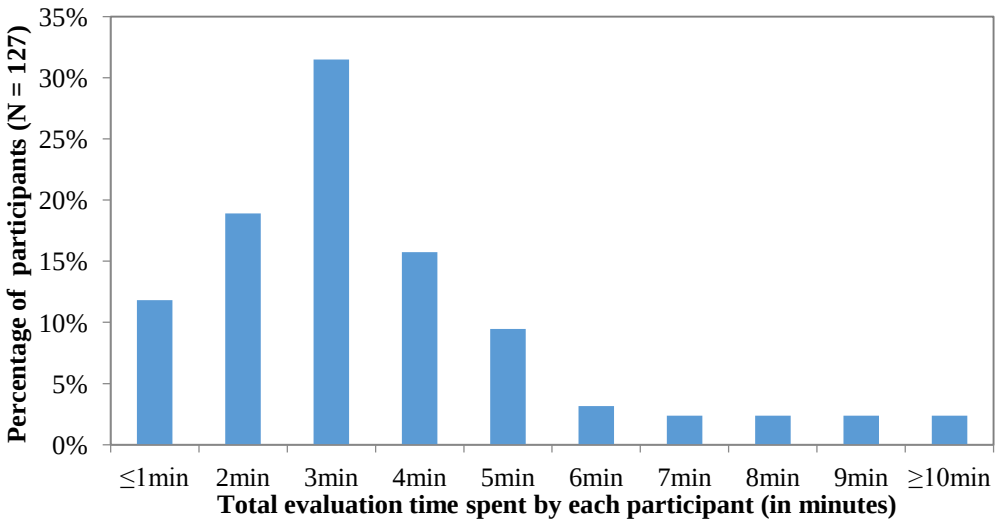


Figure 5. Percentage of Valid Frames per Video Clip

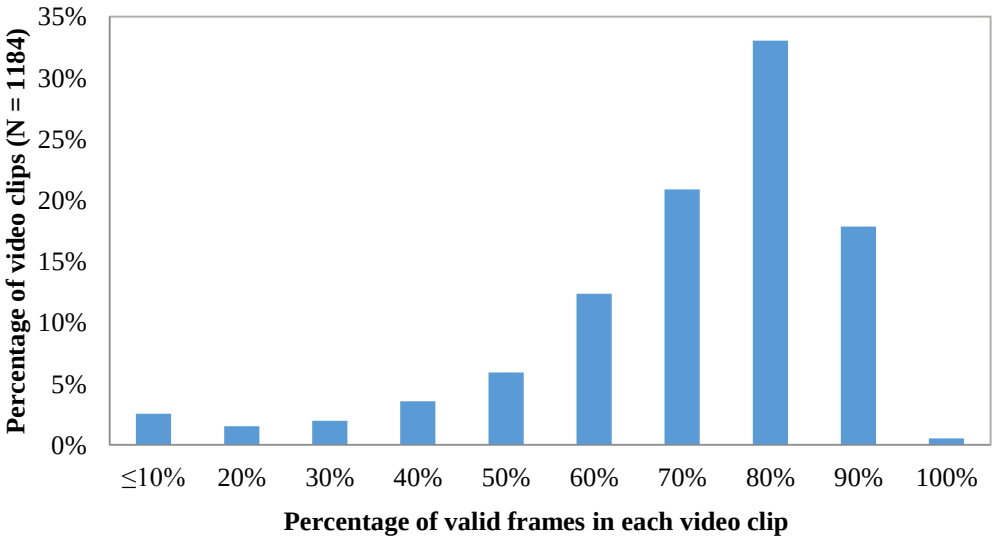


Figure 6. An Example of Hand Position Detection

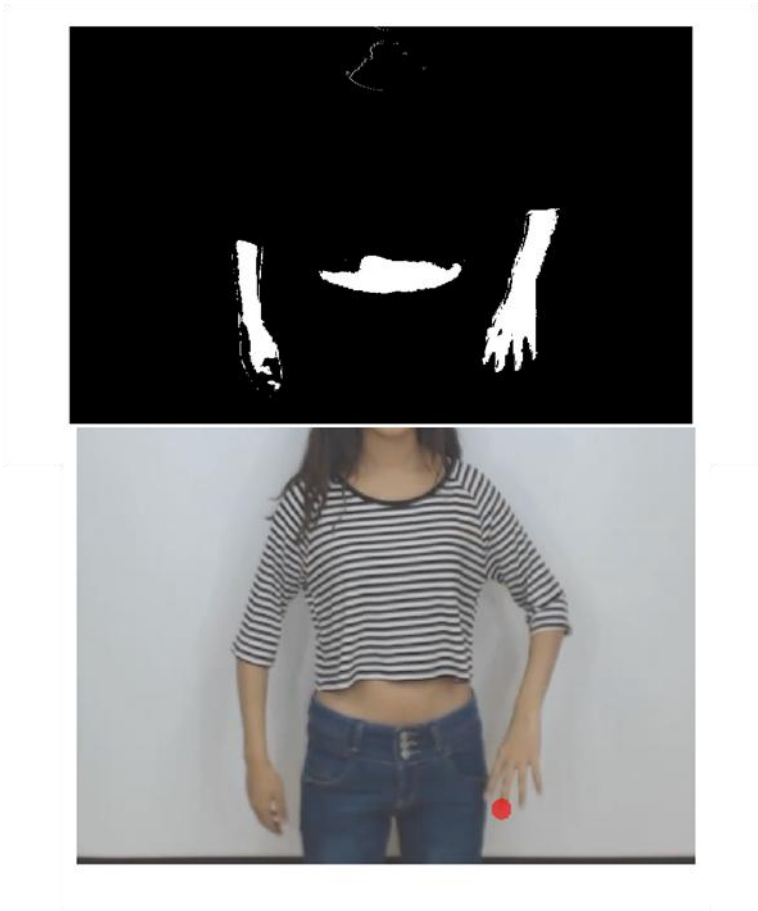


Figure 7. Distribution of Number of Would-Have-Tried Garments across All Recommendation Sets

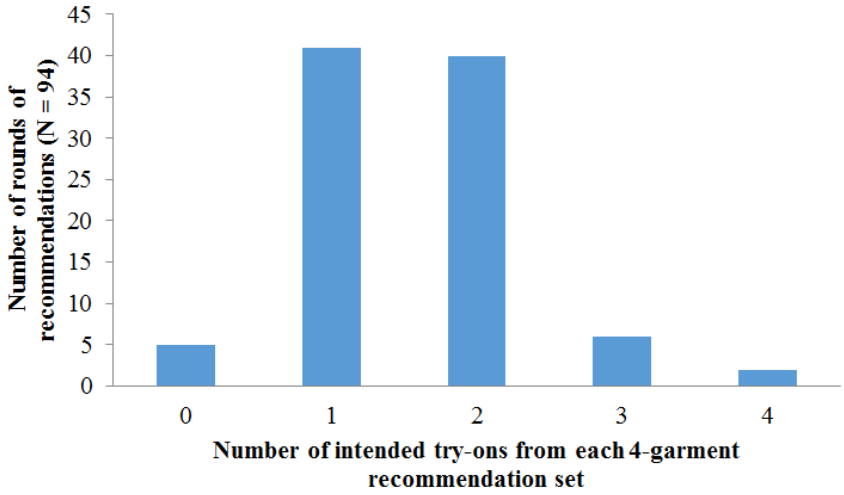


Figure 8. A Comparison of Video Processing Time

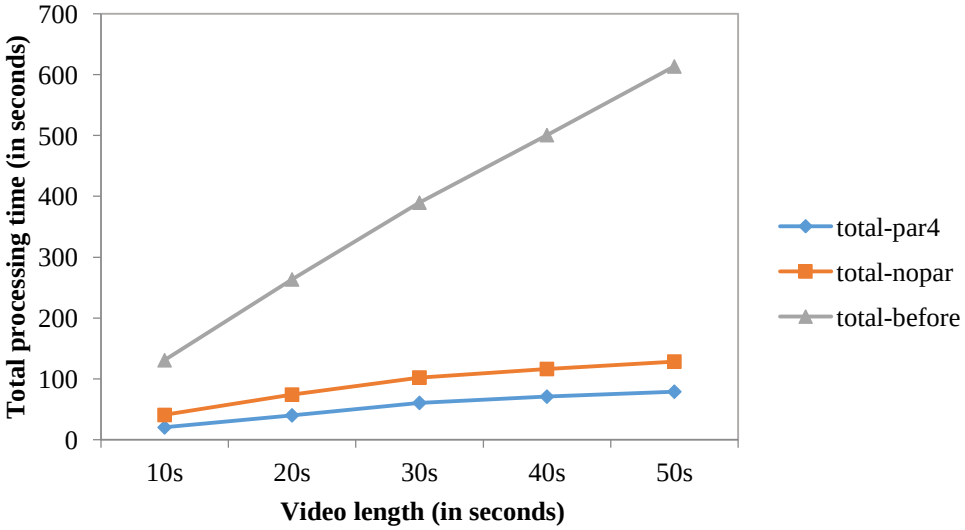


Figure 9. An Example of a Hand Heat Map

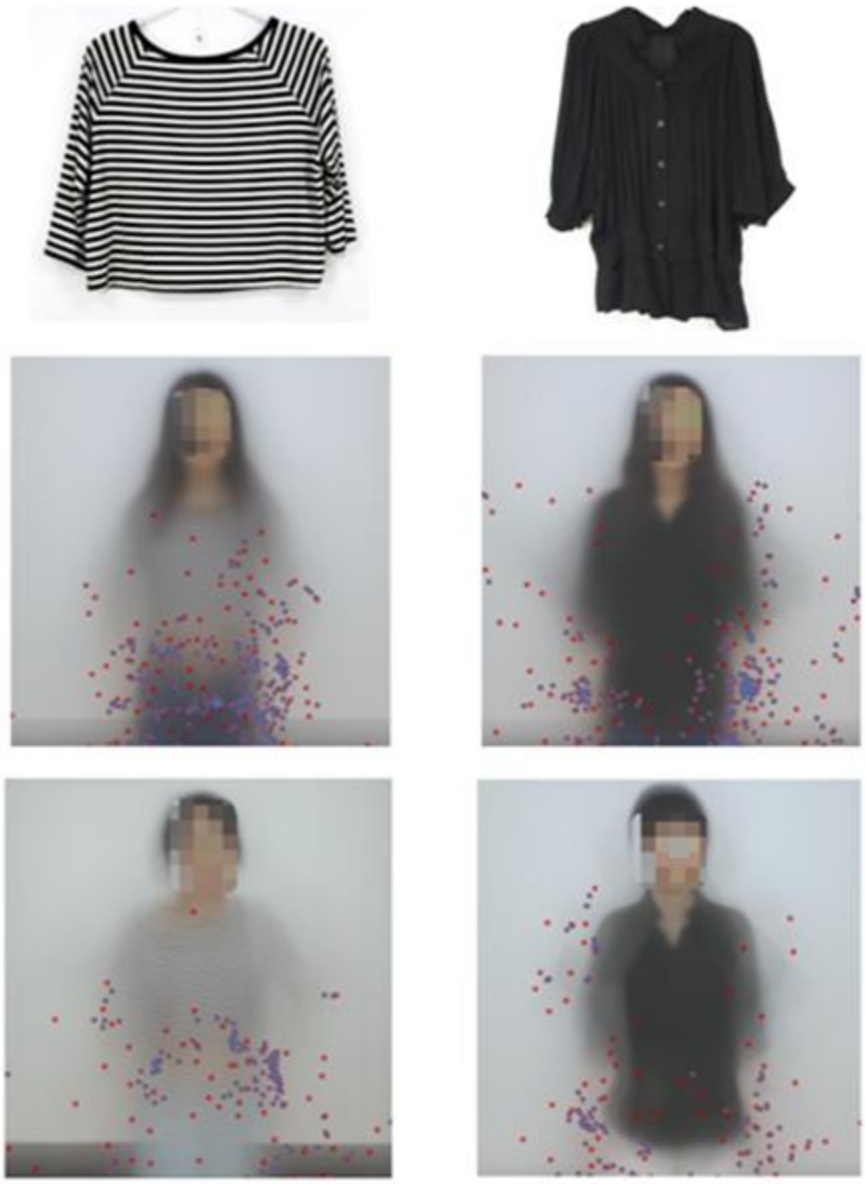


Table 1. Comparison of the Three Models

Model	SEC	SET	VAR
Data source	Customer	Customer	Camera
Preference inference	Self-reported, one-time measurement	Item-based, self-stated feedback	Item-based, video-inferred preference
Recommendation mechanism	Utility ranking	Collaborative filtering	Collaborative filtering
Cost to retailers	Computer/tablet	Computer/tablet	Computer/tablet with camera
Cost to customers	Time and effort to answer the questions	Time and effort to answer the questions; scan bar code	Leaving the fitting room to evaluate garments; scan bar code

Table 2. Garment Design Attribute Space

Attribute	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Fabric	Cotton	Not cotton	-	-	-	-
Color	White	Red	Yellow	Green	Blue	Black
Silhouette	A-type	H-type	X-type	S-type	-	-
Stripes	No	Yes	-	-	-	-
Fashion element	None	Lace	Ruffle	-	-	-
Neckline	Scoop	Crew	V-neck	Boat neck	Flat collar	Shirt collar
Sleeve design	No	Puff	Bat wing	-	-	-
Sleeve length	No	Short	Medium	Long	-	-
Hemline length	Short	Medium	Long	-	-	-

Table 3. Confusion Matrix for Facial Expression Recognition

Ground truth	**Recognized**			**Correct recognition rate (%)**
	Positive	Neutral	Negative	
Positive (25)	22	1	2	88
Neutral (26)	0	21	5	81
Negative (28)	4	5	19	68

Table 4a. Empirical Comparison of Model Performance for Predicting Considerations

	Recall/hit rate (%)[29]	Precision (%)[30]	K-L[31]	K-L percentage (%)[32]
SEC Model				
Recommendation set size = 1	6.9	24.4	0.6045	7.7
Recommendation set size = 2	12.7	22.4	0.6041	7.4
Recommendation set size = 3	18.2	21.5	0.6051	7.3
Recommendation set size = 4	24.0	21.3	0.6031	7.7
SET Model				
Recommendation set size = 1	7.8	27.6	0.6025	7.7
p-value (SET vs. SEC)	(.29)	(.29)	(.46)	(.50)
Recommendation set size = 2	14.2	25.2	0.6019	7.8
p-value (SET vs. SEC)	(.25)	(.25)	(.46)	(.40)
Recommendation set size = 3	20.9	24.7	0.5984	8.3
p-value (SET vs. SEC)	(.12)	(.12)	(.37)	(.25)
Recommendation set size = 4	26.0	23.0	0.6025	7.6
p-value (SET vs. SEC)	(.25)	(.25)	(.49)	(.46)
VAR Model				
Recommendation set size = 1	9.3*	33.1*	0.5977*	8.3*
p-value (VAR vs. SEC/SET)	(.07, .17)	(.07, .17)	(.37, .41)	(.36, .36)
Recommendation set size = 2	16.7*	29.5*	0.5849*	10.3*
p-value (VAR vs. SEC/SET)	(.04, .14)	(.04, .14)	(.17, .20)	(.03, .045)
Recommendation set size = 3	22.4*	26.5*	0.5887*	9.7*
p-value (VAR vs. SEC/SET)	(.04, .28)	(.04, .28)	(.21, .32)	(.049, .17)
Recommendation set size = 4	27.8*	24.6*	0.5919*	9.1*
p-value (VAR vs. SEC/SET)	(.098, .28)	(.098, .28)	(.29, .30)	(.15, .13)

* Best in models with the same recommendation set size.

[29] Recall (or hit rate) is the number of items predicted correctly divided by the total number of considered items for all 127 participants (i.e., 450).

[30] Precision is the number of items predicted correctly divided by the number of recommendations made.

[31] K-L refers to the Kullback-Leibler divergence with natural scaling, a commonly-used information theoretic measure that balances false positives and false negatives. Here we calculate divergence from perfect prediction, so a smaller K-L is better. The detailed formulae for calculating K-L can be found in the web appendix of Ding et al. (2011) and Dzyabura and Hauser (2011).

[32] K-L percentage is a metric calculated relative to the K-L divergence of a null model (random recommendation). The K-L percentage is 0% for the null model and 100% for perfect prediction. Hence, a larger K-L percentage is better. The detailed formulae for calculating K-L percentage can be found in the web appendix of Ding et al. (2011).

Table 4b. Empirical Comparison of Model Performance for Predicting Choices

	Recall/hit rate (%)[33]	Precision (%)	K-L	K-L percentage (%)
SEC Model				
Recommendation set size = 1	10.1	16.5	0.3402	9.0
Recommendation set size = 2	17.4	14.2	0.3357	10.3
Recommendation set size = 3	22.2	12.1	0.3385	9.6
Recommendation set size = 4	27.1	11.0	0.3361	10.6
SET Model				
Recommendation set size = 1	9.2	15.0	0.3414	9.4
p-value (SET vs. SEC)	(.37)	(.37)	(.48)	(.45)
Recommendation set size = 2	16.9	13.8	0.3350	10.9
p-value (SET vs. SEC)	(.45)	(.45)	(.49)	(.41)
Recommendation set size = 3	26.1	14.2	0.3257	13.1
p-value (SET vs. SEC)	(.16)	(.16)	(.30)	(.07)
Recommendation set size = 4	30.4	12.4	0.3294	12.3
p-value (SET vs. SEC)	(.25)	(.25)	(.39)	(.20)
VAR Model				
Recommendation set size = 1	12.1*	19.7*	0.3355*	10.0*
p-value (VAR vs. SEC/SET)	(.26, .16)	(.26, .16)	(.42, .40)	(.36, .41)
Recommendation set size = 2	21.7*	17.7*	0.3228*	12.9*
p-value (VAR vs. SEC/SET)	(.14, .12)	(.14, .12)	(.29, .30)	(.16, .23)
Recommendation set size = 3	27.5*	15.0*	0.3222*	13.2*
p-value (VAR vs. SEC/SET)	(.095, .38)	(.095, .38)	(.25, .44)	(.07, .48)
Recommendation set size = 4	32.4*	13.2*	0.3238*	13.3*
p-value (VAR vs. SEC/SET)	(.15, .35)	(.15, .35)	(.30, .41)	(.09, .31)

* Best in models with the same recommendation set size

[33] Recall (or hit rate) is the number of items predicted correctly, divided by the total number of chosen items for 127 participants (i.e., 207).