Enhanced Personalized Targeting using Augmented Reality

Gaurush Hiranandani*
Adobe Research, India

Kumar Ayush†
IIT Kharagpur, India

Atanu Sinha‡
Adobe Research, India

Sai Varun Reddy Maram
IIT Roorkee, India

Chinnaobireddy Varsha
IIT Guwahati, India

Pranav Maneriker§
Adobe Research, India

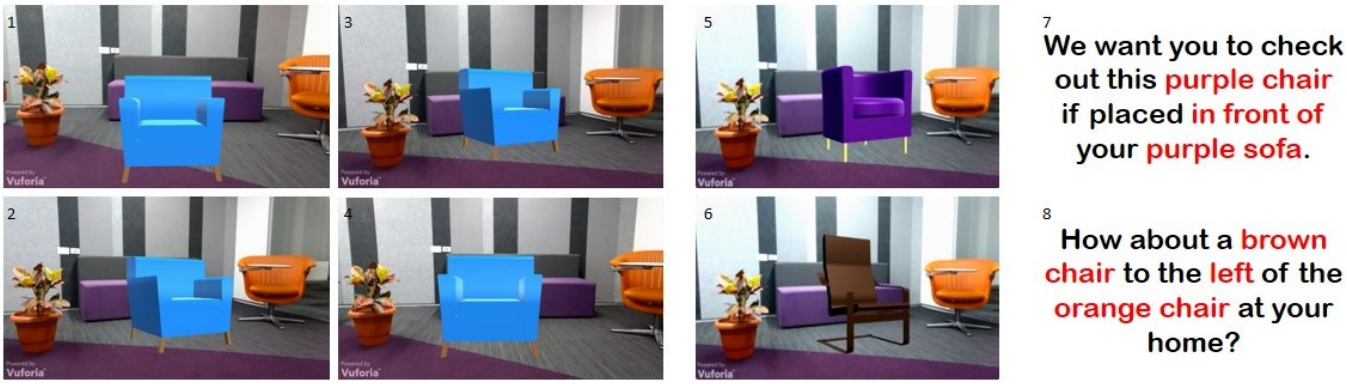

Figure 1: Retargeting solution for augmented reality application users. (1)-(4) shows some screenshot frames from the user's session. (3) has been picked by the viewpoint selection model. (5) and (6) are the recommendations generated based on style and color compatibility. (7) and (8) shows the diverse persuasive content generated based on the viewpoint for the recommendations (5) and (6) respectively. The red colored words in (7) and (8) are generated by the system which are put into predetermined marketing template (shown in black colored text).

Abstract

The trend of harnessing AR-based data has started breeding novel and enriching applications. Though the AR-based apps have been in existence for a long time, its true potential in digital marketing domain has been not exploited yet. In this paper, we bridge this gap through creating a novel consumer targeting system. First, we analyze interactions of consumer on AR-based retail apps to capture rich AR interactions, followed by identifying her purchase viewpoint during the app session. We then target the consumer through a personalized catalog, created by embedding recommended products in the viewpoint visual. The color and style of the embedded product is matched using the visual compatibility with the viewpoint, and personalized text content is created using the visual cues from the AR app data. We then evaluate the system using extensive user studies. The results show that we are able to identify the viewpoint, that our recommendations are better than the tag-based content recommendation system. Moreover, targeting through the recommendations embedded in the viewpoint is significantly better than the usual product catalog based targeting.

Keywords: Augmented reality, viewpoint selection, recommendation, targeting, v-Commerce.

Index Terms: H.5.1 [Information Systems and Presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities; K.4.4 [Computers and Society]: E-Commerce

*e-mail:ghiranan@adobe.com
†e-mail:kumar.ayush@iitkgp.ac.in
‡e-mail:atr@adobe.com
§e-mail:pmanerik@adobe.com

1 Introduction

Embedding reality in consumers' online shopping experience has been heralded as the 'next frontier for retail'. Some have labeled it as the coming of 'v-commerce' [67, 64]. V-commerce provides an engaging way for the consumer to use the virtual product in conjunction with the real world environment to judge a product's compatibility before making a purchase. One example in retail is the use of hand-held devices to virtually 'try on' furniture / shoes before purchase[1]. In the marketing domain, AR advertisement platforms [82, 81] enable brands to promote advertisements to catch customers' attention. In particular, Augmented Reality (AR) layout systems / applications have gained attention in the scientific literature [48, 63, 10, 66] as well as in the industry [84, 80, 83]. They are useful for viewing room and designing room or building layouts without having to buy or move real objects such as furniture [48, 10].

However, these approaches fail to account for consumers behaviors and preferences, which are necessary to make AR relevant to the consumers [25]. If AR can incorporate a consumer's behavior and preference, then it has the potential to improve the shopping experience considerably. Therein underlies the need for scalable data-driven technologies that can make use of rich data AR applications generate about consumers. In this paper, we introduce a system for targeting consumer by making use of the rich visual and interaction data obtained from AR systems. By contrast, existing approaches to consumer targeting used by e-commerce companies use information from users' profiles (demographic filtering) [38], similar neighbors (collaborative filtering) [22], and textual description (content-based model) [35] to make recommendations. These often generate irrelevant suggestions to consumers due to their failure to recognize both consumers intentions, as well as visual simi-

[1] www.youtube.com/watch?=dwt-mgxq_ao, www.youtube.com/watch?v=silya3JSpEU

larity among products [23]. The kinds of feature used in the traditional e-commerce scenario is not sufficient to tell how the product interacts with the consumer's physical environment. This problem can be tackled by using information from AR applications. We posit that the visual image data obtained from such AR apps is very rich and using which the frontiers of targeting can be extended.

Consider an online product-search session in which the consumer uses an AR system to improve her selection process and outcome. Such a rendition has the consumer going through a 'tryout' by placing objects on a background of a real world environment on her device-screen. For example, if the consumer is interested in buying a chair for her living room, the real world environment is the consumer's living room. Broadly speaking she does two things: (i) she places different chairs from the app on the background of her living room, and (ii) she moves the background around to check the compatibility from different viewing perspectives. We define a construct, *viewpoint*, to represent the visual (image) at which the consumer judges the compatibility of the virtual product (3D model) with the surrounding real world environment. The viewpoint holds important information previously unavailable in the data on web-based browsing of products. Based on the consumer's viewpoint data, products having better design and color compatibility with the surrounding real objects can be rendered. Further, to target a consumer, instead of sending images of merely the recommended products, images of recommended products *embedded in viewpoint* can be sent. Moreover, personalization strategies in 'Targeted Content Marketing', ranging from content that include a recipient's name to fully contextual content are key fronts in the battle for customer experience differentiation [41]. Marketer can leverage *viewpoint* data to create content about consumer's physical surroundings achieving greater personalization. This level of personalization has not been achieved before and will make marketing more engaging potentially leading to increased conversions [62].

AR based data enables us to create individual AR contents for each user depending upon context with which improved marketing activities can be achieved [39]. This paper makes three novel contributions in advancing targeting through AR applications data. These are:

- *Viewpoint Selection*: Based on the consumer's interaction data with the AR app, we build a statistical model to select the viewpoint with the highest likelihood of influencing the consumer's purchase.

- *Recommendation System*: We create a recommendation system based on style and color compatibility of the objects present in the viewpoint. We use this system to create product recommendations and embed recommended products in the selected viewpoint.

- *Targeting Content Generation*: We generate diverse personalized targeting content using visual information obtained from the consumer's AR data. This creates persuasive content by relating to the physical surroundings of the consumer.

In Section 2, we review literature regarding AR based retail applications, viewpoint selection, recommendation and targeting in traditional e-commerce scenario and intelligent solutions made over data obtained from AR apps. Section 4 provides description of the datasets used and technical details of the three contributions. We evaluate this system by a user study described in Section 5. The same contains the results as well. Finally, in Section 6, we provide conclusions and the future work.

2 Review of Literature

We start by reviewing the literature which relates to consumers' shopping. Later we mention the industry prior art. AR can provide two benefits to consumer shopping. One, it can ease the consumer's real-time shopping experience through virtual tryout; two, it can help the firm with better information about consumer's browsing / trial for future, improved targeting. Belonging to the former category, [88] automatically customizes an invisible (or partially visible) avatar, based on the consumer's body size and the skin color, for a fitting-room experience for clothes. One study [52] has shown that the use of a mobile screen, such as a tablet, instead of a larger, fixed monitor holds an advantage by allowing to move about the store as well as change the angle of viewing. Another paper [69] presents a demo platform application developed for a real-time shopping experience for shoes to explore the antecedents of consumer's purchase intention.

AR applications generate massive data, creating opportunities for research. In particular, the deployment of AR in v-commerce facilitates rich consumer experience, which can be leveraged for enhanced personalized targeting, provided the challenge of uncovering insights from the massive data can be overcome [25]. In other areas of big data there has been significant research effort guided toward recommendation using mobile systems' data. For example, [40, 42, 19, 16, 86] use GPS based location from consumer's device as a feature to create location based recommendations. Some include application of AR. One paper [65] proposes a visualization method to recommend the most suitable restaurant in an area based on the preferences of the user and online comments, while [89] presents an aggregated random walk algorithm incorporating personal preferences, location information, and temporal information in a layered graph to improve recommendation in Mobile AR ecosystems. Some of the examples using AR systems' data include [49], which understands visual behaviour of subjects looking at paintings, using eye-tracking technology, in order to define a protocol for optimizing an existing Augmented Reality (AR) application In one paper [50], head tracking data is used to retrieve a person's intended direction of walking in order to increase their immersion. In the category of gesture-based interaction studies, [26] evaluates the feasibility of tracking-based interaction in order to create and edit 3D objects in AR applications. Further, [55] present the results of a guessability study focused on hand gestures in AR.

Investment by industry[2] in AR mobile and desktop apps creates opportunities for research. IKEA has released AR catalog app that allows customers a virtual preview of furniture in their room. Rayban's Virtual Mirror, enable the consumer to see the sunglasses rendered over their faces on their desktop screens. Sephora offers tools for consumers to virtually try on cosmetics. The visual data, if captured from such applications, is very rich and a variety of solutions enhancing consumer experience can be made. Considering the example described in the introduction, the viewpoint is subjective and has to be unearthed from data of a consumer's interaction with the AR system. We now relate to the literature on viewpoint.

In the extant literature in computer vision, as well as in the nascent field of VR, metric for viewpoint is defined differently in different contexts. For Image Based Rendering [73, 74] defines a new measure, viewpoint entropy, to compute good viewing positions automatically. How to automatically select the most representative viewpoint of a 3D model has been shown in [59, 77, 4]. Other approaches include, addressing the problem of finding camera views which capture better views of the scene [61]; Viewpoint Quality Estimation algorithms to know which places in a virtual environment are interesting or informative [18]. Few works recognize the human perception. One such paper [12] evaluates the best view selection algorithm based on experimentally generated ground truth data comprising preferred views of 68 3D models provided from 26 human subjects. The evaluation metric is the most informative view of a 3D model. Yet another paper [48] brings human into the loop by having participants wear head mounted display (HMD) to improve the configuration of furniture in the room. To overcome the potential annoyance associated with HMD [75]

[2] www.ikea.com, http://www.ray-ban.com/, http://www.sephora.com/

proposes an automatic viewpoint-recommending method, based on an assumption that the different trajectory distributions cause a difference in the viewpoint selection according to personal preference. It is noted that none of these methods uses observational data generated from deployed AR systems to do *Viewpoint Selection*. This is significant since we do the viewpoint selection from massive data of consumer's interaction without actually using the large sequence of visual images seen by her.

The viewpoint based visual becomes a big differentiator between the traditional e-commerce recommendation system and the system which we propose. As mentioned before, Recommendation Systems garner a lot of research attention in both academia and industry [21, 70, 37, 54]. In virtual world as well recommendation systems have been proposed under both collaborative based filtering [34]. Other research develops such systems for second life and opensimulator [13]. Visual recommendation proposed by [78] relies only on images and extracts color and texture information in order to find visually similar items. Various other recommender systems use personal data context derived from GPS, gyroscope, compass and accelerometer [38, 24, 2]. Another work proposes a generic system [56] that relies on previously gathered user feedback data (i.e. ratings and clickstream history), context data, and ontology-based content categorization schemes. Their recommendation system identifies the context of the user from clickstream data which is also done in most e-commerce scenario. Our approach can ingest all such data, when available. In addition what is novel is the ability to use viewpoint information coming from consumer browsing on an AR system to enrich the recommendation.

From the consumer targeting content's perspective, self-relevance is a well-established means of increasing message elaboration [76]. People are more persuaded by messages matching aspects of their identity [57]. Highlighting the persuasive / advertising content generation, [60] automate message personalization by inserting adjectives and adverbs mined from social media evoking positive sentiment in specific audience segments, into basic versions of ad messages. In another study, [11] directly model content selection decisions based on a set of psychologically-motivated domain-independent personal traits including personality and basic human values and show how it can be used in automated personalized content selection for persuasive message generation. Marketing researchers are interested in applying theories of personalization and persuasion in promoting consumer products[6, 36, 17].

To the best of our knowledge there is no prior work which uses the consumer interaction data as well as the visual data obtained from AR systems. We leverage these observational data to create novel consumer centric recommendation and targeting system. Our use of observational data emanating from AR systems is a new contribution. That is, we do not create an AR system, where much effort is expended, but make use of the massive data that generates. We perform all three tasks automatically.

3 DATASETS

There are two datasets we have used in this paper. The first step in our approach is viewpoint selection from observational data, which we describe first. Then, we describe a data repository which is used to create the recommendations and textual content. It is used for the evaluation study as well.

3.1 Data for Viewpoint Selection Modeling

This proprietary data is generated by consumers' interaction with the AR mobile application of a large company that designs and sells household products and accessories. The consumers can select and virtually place objects in their rooms using AR. Consumers can change the position and orientation of the object, generating rich, clickstream data capturing their hits / clicks, and which are time stamped in discrete steps of one second. Thirty days of data provided by the app is available to us. Both augmented reality based features and other relevant features are identified from this massive data. Then the data is processed to create aggregated features corresponding to different AR sessions. All the AR interactions of a user with a product within a session constitute one observation. The features in the data are:

1. $\#c$: Number of times an object is chosen (the same object can be chosen more than once by the consumer).
2. $\#o$: Number of objects chosen in a session.
3. $\#r$: Number of times an object's location or orientation is changed.
4. $\#e$: This denotes the total number of AR based events, defined as, $\#e = \#c + \#r$.
5. T_C: Time elapsed while the AR action is 'chosen'. 'Chosen' time is the time duration between an object is chosen and the next event. T_C is the sum of all the 'chosen' time in a session.
6. T_R: Time elapsed while the AR action is 'rendered. 'Rendered' time is the time duration between an object is rendered in a different location or orientation and the next event. T_R is the sum of all the rendered time in a session.
7. T_T: Total time elapsed in AR interaction = T_C + T_R. It excludes the time elapsed due to events other than AR based events.
8. VPI: An indicator variable which shows whether the 'View Product Information' button is clicked or not during the session. It is common for many sites to offer this button to the consumers who seek more detailed information about a product when they become seriously interested in the product. For our model, this variable VPI is very useful. Consumers who click on this button during the session indicate their serious interest about the product. There are other consumers who do not click on this button during the session. Thus, for this latter group we cannot observe their interest since there is no such clicking data during the session. We characterize data from these consumers as censored observations and recognize them as such in the model. It is improper to treat these observations as if they show no interest in a product; rather it is beneficial to think of them as observations who may have interest in a product, but the end of session (for whatever is the reason) censored the observance of their interest in a product.
9. T_P: Time elapsed between start of AR interaction and the first clicking of the VPI button. This is equal to T_T if $VPI = 0$. Note that T_T is the censored random variable when $VPI = 0$.
10. T_i: The time interval between $(i+1)^{th}$ and i^{th} AR action. We consider six such intervals.
11. A_{i-1}: An indicator variable having value 1 if the accelerometer reading is below a pre-determined threshold at the $(i-1)^{th}$ time step. Otherwise, it is 0. This tells us the whether the device is stationary or not before the i^{th} time step. This is useful to have a clear (non-blur) viewpoint image(s).

The visual data i.e. the frames rendered in the session are not observed. Hence, we look for the time point just after rendering of the interesting visual i.e. T_p. We worked with approximately 50,000 session of which 12% of the sessions had $VPI = 1$. This mobile application dataset is represented by `A`.

3.2 Data for Recommendations and Targeting Content

Shapenet [7] is a richly-annotated, large-scale repository of shapes represented by 3D CAD models of objects. ShapeNet contains 3D models from a multitude of semantic categories and organizes them under the WordNet [46] taxonomy. The 3D models have been grouped into the following categories: 'single 3D models', '3D scenes', 'billboards', and 'big ground plane'. We took only 'single 3D models' category. We have used a subset of Shapenet - ShapeNetSem which is a smaller, more densely annotated subset consisting of 12,000 models spread over a broader set of 270 categories. In addition to manually verified category labels and consis-

tent alignments, these models are annotated with real-world dimensions, estimates of their material composition at the category level, and estimates of their total volume and weight. This ensures that any tag / description based recommendation system (baseline) has a fair chance to generate good recommendations.

For the purpose of creating the targeting system and conducting Amazon Mechanical Turk study, we selected a subset of 150 models each from the categories 'armchairs' and 'coffee table'. The models were selected so that the chosen models formed groups based on keyword annotations. This included words corresponding to wordnet synset associated with the model description (which included design name, color name, etc.). The reason for such a selection was done to ensure that there are enough good recommendation candidates if generated from baseline [71] (described later). We denote this dataset by S.

4 METHODOLOGY

We describe the approach in the same sequence as the three primary contributions: (a) Viewpoint Selection, (b) Catalog Creation, and (c) Targeting Content Creation. Then we explain how these solutions are merged to create an email based targeting system.

4.1 Viewpoint Selection

In Section 1 we defined *viewpoint* as representing the visual (image) at which the consumer judges the compatibility of the virtual product (3D model) with the surrounding real world environment. In other words, *a viewpoint is an augmented visual image*, or, in short, an augmented visual. For clarity, an image is a single frame. A *preferred augmented visual(s)* is defined as a frame (image) [frames / images] that interests the consumer most in the augmented product in the presence of background and surrounding objects.

It is crucial that the data on augmented visuals are properly captured and relevant visuals identified, in order to deliver personalized consumer-preferred augmented visual to her device as a recommendation. However, there are at least two challenges that make this problem difficult: (i) the high volume of images that result from a consumer's session interacting with the AR app, and (ii) identification of augmented visual(s) from among these sequentially viewed images, that the consumer prefers. Using dataset A, we built a statistical model to uncover the preferred augmented visual for the consumer. The novelty of our model is that we select the preferred augmented visual by analyzing the time stamps at which images (frames) are rendered on the app during a session. This is developed based on the idea that there is a natural time sequence of frames / images viewed by a consumer, in an AR session on her device, and we can exploit this time sequence to identify her preferred frame(s) / image(s). In a departure from the previous literature, we do not use the visual data generated during the session.

1. **Trigger of Interest.** Since we are using time stamps, we need to define an event in time that represents the consumer's preferred augmented visual which is defined above. This time-based event is labeled the 'trigger of interest' for the consumer during her AR app session. Following the *Awareness-Interest-Desire- Action (AIDA)* model in e-commerce [8], we posit that the trigger of interest is the stable single image (frame) at the time epoch just before the consumer clicks the button 'View Product Information' (VPI). VPI has been explained in Section 3.1. It is noted that VPI is not a restrictive feature relevant only to our dataset. All apps for product search have a button which allows the consumer to click on it to obtain detailed information about a product during a search session. The label used may vary across apps. For a different dataset, the trigger of interest will be the time epoch just before the consumer clicks on such a button. Thus, trigger of interest is a general concept in the context of AR apps.

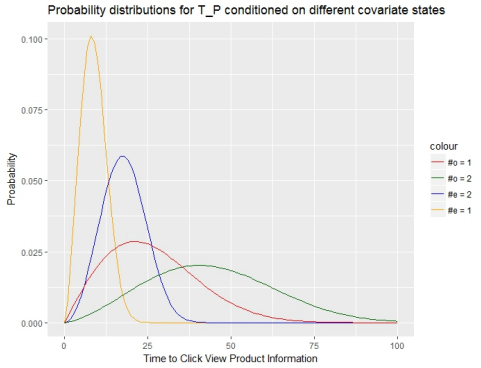

Figure 2: Fitted distributions for different parameter settings.

Having defined the event 'trigger of interest' we now explain how we use this event in the model. There are two outcomes of any consumer session: session in which the consumer selects the VPI button at least once, and session in which the consumer does not select VPI. For sessions in which VPI is selected, the image (frame) at the time epoch just before selection of VPI is the consumer's preferred augmented visual. If more than one VPI is chosen in a session, the last clicked VPI is used. The time epoch just prior to VPI is observable in the data and hence the frame is observable in the data. However, that is not the case when the consumer does not click on VPI in the session. For these sessions, we probabilistically assign the time epoch at which the consumer is most likely to click on VPI. That is, the time epoch and hence the image (frame) representing the preferred augmented visual are inferred from the data through the model. We can do this by recognizing that in the latter sessions the time epoch is censored due to the end of session, which could have happened for any number of reasons. Our model can thus infer preferred augmented visuals even in the absence of observed trigger of interest and is one of the novelties we present. The model estimates $f(T_P|\tilde{Z})$ where T_P is the time to view product information from the start of the AR session, $\tilde{Z}$ is the vector of covariates and $f(.)$ is the probability density function.

Accelerated Failure Time Model (AFT). We observe that the empirical distributions of T_P conditioned on many covariate states ($\#o = 1, \#e = 2, etc.$) are positively skewed. Empirical distribution of T_P corresponding to some covariate states is shown in Figure 2. This supports findings in the response time literature [72] from which we borrow our model. We find that the response time to click on VPI, is positively skewed for almost all the groups of data points, and provides confidence in our choice of model. For concreteness, we calculate the sample skewness [31] for many groups of data points created based on different covariate states and find them to be less than -1.

The censored nature of T_P leads to accelerated failure time model [9] as a good approach to use. The Generalized Gamma distribution, defined by two shape and one scale parameters is estimated using maximum likelihood estimation. Generalized Gamma distribution is flexible to respond to the characteristics of each group of data points for which different estimates can result. We find that the Generalized Gamma distribution is a good choice for the distribution of T_P given $\tilde{Z}$. With parameters a, b and c, the pdf of Generalized Gamma distribution is given by:

$$f(T_P) = \frac{a}{\tau(c)b}\left(\frac{T_P}{b}\right)^{ac-1} e^{-\left(\frac{T_P}{b}\right)^a} \quad (1)$$

The problem reduces to estimating a, b and c which are functions

of the covariates. Hence, we have,

$$f(T_P|\tilde{Z}) = \frac{a(\tilde{Z})}{\tau(c(\tilde{Z}))b(\tilde{Z})}(\frac{T_P}{b(\tilde{Z})})^{a(\tilde{Z})c(\tilde{Z})-1}e^{-(\frac{T_P}{b(\tilde{Z})})^{a(Z)}} \quad (2)$$

2. **Functional Form of Parameters.** We assumed that any parameter in the distribution function of T_P potentially depends on the covariates (features) which are described in Section 3.1. For example, in the case of the generalized gamma distribution, defined by three parameters, let shape parameters be denoted by a and c and scale parameter be denoted by b. Then,

$$a = K(Z_1, Z_2, ..., Z_p) \quad (3)$$
$$b = G(Z_1, Z_2, ..., Z_p) \quad (4)$$
$$c = H(Z_1, Z_2, ..., Z_p) \quad (5)$$

where p is the number of covariates and K, G and H are some functions of the covariates. For generality, we allow the parameters to be functions of the covariates, and let the data inform the specific nature of the relationship. Each parameter can have a different functional relationship with the covariates. We fit the candidate functional forms and select the functional form which gives the maximum R^2 value. This is important since usual libraries will allow the same functional form of the parameters. Later we find that indeed the functional forms derived from the data are not the same for all three parameters.

3. **Fit Distribution.** We fit the generalized gamma distribution to the training data. In the likelihood function of the generalized gamma, we replaced the distribution parameters with the functional form described in the previous step. We tried various functional forms like logarithmic, exponential, linear, etc. for the three parameters. For example, noting that a, b and c have exponential, linear and constant functional forms respectively for both *#o* and *#e*, based on Equation 2, we defined the following:

$$a(\#objects, \#events) = e^{\alpha_a + \beta_{1a}\#objects + \beta_{2a}\#events} \quad (6)$$
$$b(\#objects, \#events) = \alpha_b + \beta_{1b}\#objects + \beta_{2b}\#events \quad (7)$$
$$c(\#objects, \#events) = \alpha_c \quad (8)$$

The parameters $(\alpha_a, \beta_{1a}, \beta_{2a}, ...)$ are estimated by `flexsurv` package in R [28] using Nelder-Mead method [53].

4. **Frame Selection.** Once the scale and shape parameters were estimated, the empirical distribution of $f(T_P|\tilde{Z})$ was obtained and the mode of the empirical distribution was judged to be the value of T_P. The logic follows from the definition of mode as being the most likely outcome. According to the assumption made, the visual perspective of likely interest to the consumer is the frame at the time point just before this estimated value of T_P.

5. **Visual perspective for New Session.** We observe that both in training and testing dataset (explained later), the frame selected as the viewpoint is one of the frames when the accelerometer reading was negligible for some duration. This is indeed natural and necessary to obtain clearer (not blurred) visual perspectives. Now, we can use the above method for a new consumer session, by computing the propensity at each time point. We store the frame (image) at the time point just before T_P - the point when the maximum propensity to click on 'View Product Information' is achieved (and the accelerometer value reading is below the predetermined threshold). If both the conditions are satisfied at some later stage in the consumer session, then we update the viewpoint (image) accordingly. So, at the end of the session there will be only one image sent to the marketer from the app denoting the viewpoint. This viewpoint frame / image can be stored in the marketer's server and can be used for recommendations and targeting as discussed later.

Model	Par	a, b, c	1-s	3-s	5-s
SA	T1, T2	lin, lin, lin	4.68	17.36	30.62
SA-FF	#o, #e	exp, lin, lin	12.34	23.45	35.76
W-Mode	#o, #e	exp, lin, -	18.33	27.03	39.64
W-Mean	#o, #e	exp, lin, -	16.18	25.64	38.57
LR	#o, #e	-	16.42	26.1	38.81
RF-SA	T1, T2	-	17.52	26.40	39.35

Table 1: Results for 1, 3 and 5 second windows from the discussed models. Par denotes the parameters. The entries for 1, 3 and 5 second windows are in percentages.

4.1.1 Evaluation of Viewpoint Selection

`A` was divided into two parts: 80% for training and 20% for testing. For the purpose of evaluation we try different time windows that can potentially contain the viewpoint. We present the accuracy results for 1 second, 3 second and 5 second time windows. We compare our method against other commonly used methods as baselines. For each model, we report the best accuracy obtained after selecting the best combination of features. The models tried on this data are described below.

(a) **Standard Survival Analysis Models:** This baseline model fits generalized gamma assuming that the parameters of the distribution depends on the covariates but retain similar functional forms for all the features / convariates. We denote this method `SA`.

(b) **Generalized Gamma with Different Functional Forms:** This is the method described in the previous section. We denote this method by SA-FF.

(c) **Observed functional Form and Fitted Weibull:** Empirical method is used to figure out the functional form of the parameters in the case of Weibull, which has two parameters - shape and scale. The location parameter is taken to be 0. We allowed different functional forms for the parameters and then fitted a Weibull distribution using those parameters. There are two variants for this method. In one, we return the mode of the distribution as the estimated time point of T_P. We denote this method by `W-Mode`. This is our method except we take the Weibull distribution (Weibull belongs to generalized gamma class of distributions) in consideration instead of generalized gamma. The other variant is the one where we return mean of the distribution as an estimated value of T_P. We denote this method by $W-Mean$.

(d) **Linear Regression:** We fit linear regression on the data with target being T_P and the regressors being covariates ($\tilde{Z}$). We denote this model by *LR*.

(e) **Random Forest for Survival Analysis:** We fit random forest for survival analysis [27] with target being T_P and the regressors being covariates ($\tilde{Z}$). We denote this model by $RF-SA$.

4.1.2 Conclusions from Viewpoint Selection

Some comments about the results on the test data mentioned in Table 1. The Weibull distribution with mode provides the best fit overall. We achieve 18.33%, 27.03% and 39.64% accuracy in 1, 3 and 5 second window respectively for the `W-Mode` method. This let us achieve 11.63%, 3.56% and 2.14% improvement from the linear regression model in 1, 3 and 5 window respectively. Note that, these accuracy results are good numbers as we only report 1, 3 or 5 second windows containing augmented visual perspective from a session which can contain large number of such windows. Increasing the size of the window leads to increase in percentage values denoting the presence of viewpoint in that window. Fitting Weibull is better than generalized gamma due to: a) library dependency, and b) relationship observed between features and parameters. The use of the mode as the estimate is justified because we are interested in the time point at which the consumer's interest is the highest. We

Figure 3: Viewpoint images for illustration. Left one is the camera frame. Right one is the screenshot frame.

tried using the mean (model `W-Mean`) and Table 1 show that linear regression provides similar answers to the method when mean is considered in the Weibull distribution. The improvement for our method is less for the 5 second window suggesting that the improvement in accuracy decreases as the window length increases.

4.2 Catalog Creation

Recall that the primary contribution is to build a recommendation system which exploits data from AR apps, with viewpoint selection being the first step. The second step is catalog creation, which leverages the rich visual data from AR based retail apps, and combine with methods used in Computer Vision (CV) to create catalogs that embed recommended products in the preferred viewpoint selected in step one. To the best of our knowledge, our work is among the first to bring together AR visual data into CV algorithms for recommendation. This section describes the work flow of the recommendation system. For illustration purposes, let the final outcome of our viewppoint selection model be the two images shown in Figure 3. Here we depict both the viewpoint images on the left is the background viewpoint (the camera image), as well as the AR viewpoint which embeds the virtual product (chair) on that background (screenshot image).

1. **Location and Pose Identification.** To create the catalog with different embedded objects, the location and pose of the virtual object is required in the viewpoint. We use exemplar, part-based 2D-3D alignment [3] in the screen-shot frame. The approach is a mixture of part-based discriminative models [15] and exemplar-based matching [44]. Analogous to part-based models, [3] represents objects using a set of connected appearance parts. However, like exemplar-based methods, [3] avoids explicitly training an object model, instead relying on a large dataset of object instances that serve as their own model, both in 2D as well as in 3D. The latter is be useful for a marketer because the model can be readily run on a limited repository. It gives us the probabilistic estimates of the location and pose of the virtual object in the camera coordinates.

Alternatively, as a deterministic solution, one can design the augmented reality system so that it captures the location and pose of the object throughout the consumer's session and then uses the location and pose for the time point when the viewpoint is selected. We implemented both the solutions on our dataset. For the probabilistic approach, the error was within the bounds mentioned in [3]. Thus, either method can be chosen. The probabilistic method requires less information from the consumer by trading off accuracy in determining the location and pose of the virtual product.

2. **Shape Style Similarity.** We posit that the consumer may prefer objects that are similar in overall physical design to the product she has tried [87]. We operationalize this through shape style similarity. In the e-commerce setting, similarity in design is determined by some form of meta-tags (as discussed in Section 2). However, the meta tags do not capture the shape style nuances in the design or other aesthetic features of the products. We leverage a more sophisticated way to determine the shape style similarity by using a structure-transcending method for evaluating the stylistic similarity of 3D shapes [43]. The proposed measure is well aligned with the human perception of style, is motivated by art history literature, and is learned from and validated against crowd-sourced data. Although it works well even while measuring style similarity in different classes of objects, we use it to evaluate shape style similarity within a particular class of objects with similar overall structure (e.g. chairs). This method returns a distance measure between two objects. Let the style distance between object i and object j be $\alpha_{i,j}$ which is transformed to similarity (say $s_{i,j}$) by $s_{i,j} = \frac{1}{1+\alpha_{i,j}}$. This returns the similarity measure ranging in $[0,1]$.

Figure 4: Some of the candidate images having embedded product. Few of them will be selected as the final recommendation images.

3. **Embed Products in Viewpoint.** Using the location and pose from Step 1 above, we embed the products in the camera image of the viewpoint. It is achieved by using BlenderVR [33]. This creates a candidate set of images with embedded recommendations. All candidate images are normalized such that they have the same reference in in the camera coordinates. Some of the candidate images are shown in Figure 4.

4. **Color Compatibility of Images.** A criterion for choosing products by offline shoppers is how good the color compatibility of the product is with other objects in the room [87]. For each image created in the previous step, we extract a theme of five dominant colors (represented in hex codes) by using [51]. Each theme of colors is transformed to a vector of 326 features including sorted colors, differences, PCA features, hue probability, hue entropy, etc. (similar to [51]). Then, vectors representing the themes is passed to a lasso regression model [51]. It is a color compatibility model which provides a rating to the theme on a scale of 1 - 5, where the weights are learned from a large scale crowd-sourced data. Let r_i be the rating for each image i formed in the previous step. The ratings are normalized to lie in $[0,1]$ by the transformation, $c_i = \frac{r_i - 1}{5-1}$, where c_i is the color compatibility of image i.

5. **Overall Score.** We need to find an overall score of a candidate recommendation product embedded in the viewpoint, depending upon the quantitative scores obtained in the above steps. To achieve this, we conducted a survey of 120 participants. We produced a collection of 6 lists of images with 6 unique starting products, each capturing a different viewpoint. For each product, we embedded 9 candidate products at the same location and pose as that of the starting product. The scores s and c were then calculated for each of the $6 \times 9 = 54$ candidate recommendations with respect to their starting products. The participants were asked to rank the names in a list from 1 to 9. On average, the Kendall-Tau correlation (allowing ties) between the average ranks and individual ranks are 0.66, 0.68, 0.62, 0.72, 0.68 and 0.70 for the six lists. These numbers suggested that participants tend to indicate similar rankings given an image of the starting product embedded in a viewpoint. We took

the average rank and then ranked the averages to get the final ranking. In this way, we obtained the ground truth rankings for the six lists.

Further, we find that the Pearson Correlation between 2 scores corresponding to images generated for the above experiment is 0.23. Additionally, the ranks observed from individual scores has a Kendall-Tau correlation (allowing ties) of 0.21. These values do not suggest a strong relation among scores. Therefore, the appeal $A(:)$ of an image is defined to be weighted linear combination of the above mentioned scores. That is,

$$A(i) = w_1 s_i + w_2 c_i \tag{9}$$

Where, $\vec{w} = (w_1, w_2)$ is the weight vector. After getting the ground truth ranking for each list by the above rank aggregation method, we have a total of $6 * \binom{9}{2} = 216$ pairwise comparisons. We perform 4 : 1 : 1 split for training, validation and testing. Then we apply rank-svm [30] algorithm which use the obtained pairwise comparisons to learn the weights for different features. Validation data is used to achieve an optimal cost parameter required in the rank-SVM. The weights show the importance of the corresponding feature in deciding the ranks of the names. The learned weights are:

$$\vec{w} = (0.19, 1.66) \tag{10}$$

Interestingly participants indicated *color compatibility* as more important for comparing candidate recommendations than *style similarity*. For example, for the bottom left image of Figure 4, $s_i = 0.56$ and $c_i = 0.7$. Hence, $A(i) = 1.2684$.

6. **Final Recommendations and Image Enhancement.**: The recommendations are ranked in decreasing order, according to the overall appeal. We select a predetermined number of top ranked embedded images for inclusion in the final catalog. The catalogs can be used to target potential customers via various marketing channels like emails, push notifications, etc. Further, to correct for poor camera in the customer's device or viewpoint that includes irrelevant background, the images in the catalogs are enhanced by contrasting, sharpening and auto-cropping using available tools[3].

4.3 Text Content Creation

In creating the recommendation system, so far the content created focused on shape style of objects, color of objects, and location of the virtual object with respect to the preferred viewpoint. To round off the recommendation, in this section we show how to incorporate textual content in it. We emphasize diversity and persuasiveness of the text corresponding to the recommended images. For illustration, we use Figure 4 as the recommended set of images for which text content is to be created. Again for illustration purposes, let the final outcome of our viewppoint selection model be the two images shown in Figure 3.

1. **Objects Identification.** We use Region-based Convolutional Neural Network (R-CNN) [20] which takes as input an image and returns object proposals (bounding boxes) and object label with confidence score. The output of the RPN is used by Fast R-CNN [58] for detection. This gives the labels of the real world objects in the viewpoint with their bounding boxes. We also get the confidence score for the labels by this method which is used later in our system. Further, Step 1 of Section 4.2 gives the location of the virtual object in the camera coordinates. This step enables us to mark bounding boxes for the virtual object as well (see Figure 5).

2. **Object Color Identification.** Next, we identify the color of each object present in the background (camera image of the viewpoint). To achieve this we adapt the method presented in [51]. One,

[3]https://pypi.python.org/pypi/Pillow/2.1.0

Figure 5: Bounding boxes for all the objects with labels and confidence score from the camera frame (left). Bounding box for the virtual chair (3D model) from the screenshot frame (right).

we take the above bounding boxes and resize them as images in the desired shape as required by [51]. Two, in a deviation from the work in [51], instead of looking for a 5 color theme we confine to a 1 color theme based on dominance. Our objective function is shown in Equation 11. We extract the dominant color (hex code) of each bounding box denoting the color of the object present in it. Dominant colors are obtained with an objective function that attempts to represent or suggest an image while also being highly rated.

$$\max_{\substack{\boldsymbol{t} \\ s.t. \quad t_1=t_2=t_3=t_4=t_5}} \left(\alpha r(\boldsymbol{t}) - \frac{1}{N} \sum_i \max(||c_i - t_1||_2, \sigma) - \frac{\beta}{M} \sum_{j \in \mathcal{N}(t_1)} max(||c_j - t_1||_2, \sigma) \right) \tag{11}$$

In the above expression, $r(\boldsymbol{t})$ is the rating of theme t, pc_i is a pixel color, $\boldsymbol{t}$ a theme color, N is the number of pixels, σ is the threshold for distance defined by the norm, α and τ are the learning rate parameters. The first term measures the quality of the extracted theme. The second term penalizes dissimilarity between each image pixel c_i and the most similar color $\boldsymbol{t}$ in the theme. The third term penalizes dissimilarity between theme colors $\boldsymbol{t}$ and the M most similar image pixels $\mathcal{N}(\boldsymbol{t})$, to prevent theme colors from drifting from the image. We use $M = \frac{N}{20}$, $\tau = 0.025$, $\alpha = 3$ and $\sigma = 5$. The DIRECT algorithm [32] is used for solving the optimization problem as done in [51].

We name colors at two granularities depending upon the hue name and the shade name. For example, Crimson, Indian Red and Salmon are different shades having the hue of Red. We determine the shade of the colors. The 1-color theme is mentioned in hex code by solving the above optimization process. We determine the color name from the hex code using a hash function [4] which maps a hex code to the color name. Since the hex code of the extracted color can take any value from the set of $256 \times 256 \times 256$ values and since the hash function does not map all the hex codes to their color name, we look for the color name of the hex code which is nearest to the hex code of the identified object. We use the following $L1$ distance metric to approximate the RGB code for the color,

$$\texttt{distance} = |R_i - r| + |G_i - g| + |B_i - b| \tag{12}$$

Here $< R_i, G_i, B_i >$ is the RGB triplet of a hex code in the hash function and $< r, g, b >$ is the RBG triplet corresponding to the hex code of the identified object.

3. **Relative Position Determination.** Now we determine the relative surface position (left, right, front, back) of the virtual object with respect to each identified object in the viewpoint. Using the bounding boxes' coordinates obtained in the first point of this method, locations are determined. We use the hyperplane separation theorem [5] to determine if the bounding boxes intersect. If they do not intersect, we will have either a vertical or horizontal separating axis (since the boxes are axis-aligned). If the axis

[4]https://www.w3schools.com/colors/colors_groups.asp

Table 2: Surface Location based Preposition Synonyms

Relative Position	Alternate words
Front	front, after, anterior
Back	back, before, behind, rear, posterior
Left	left, next, beside, by
Right	right, next, beside, by

is vertical, the box with lower x-coordinates is labeled as 'left' of the other. Similarly, the label 'top'/'bottom' is assigned when the axis is horizontal. However, when the rectangles intersect, we use a heuristic based on the relative areas and the area of intersection of the two boxes to determine which box is in front of (or behind) the other. Moreover, various synonyms can be used to denote relative position in order to ensure syntactic diversity in the generated content. For example, *left* can be written as *next*, *beside*, *by* etc. Note that, *left* is a more detailed description of relative position than *next*, *beside* or *by*.

4. **Tuple Creation.** For each identified object, we generate tuples of the form <object type, object color, relative position of the virtual object with respect to the identified object>. Per discussion above, two tuples are generated for color. Similarly, multiple tuples are created using the relative position words mentioned in Table 2. Depending on the number of identified objects, suitably many tuples can be generated.

5. **Tuple Reward.** Marketer may want to give preferences to some tuples over another. For example, a marketer may like to talk about an object is which is identified with more confidence instead about one for which the object detection algorithm is not sure. In order to decide which tuples are important to use in the predefined template of sentences, we define *reward* for tuples. The reward is based on the following intrinsic properties:

(a) *Object Proposal Confidence (OPC)*: Our goal is to detect objects with high confidence in the background. For example, if the probability of a chair lying in the background is low then it is better not to mention it in the recommendation text. The measure of $OPC \in [0,1]$ is obtained from Step 1 of this section.

(b) *Association Value (AV)*: We seek objects which have high association value with the recommended product. Association value is taken from association rule mining [79, 45]. For example, a sofa has a higher association value with a coffee table than with a wall painting. Here we use a simple measure. If the identified object is associated with the endorsed product class, according to [79], then we assign $AV = 1$. Otherwise we assign $AV = 0.5$. More complex ways of defining association score can be used.

(c) *Location Synonym Weight (LS)*: Exact location of the recommended product with respect to identified object or an approximate location synonym can be used. For example, *left* can be used as exact location, whereas prepositions like *beside*, *next*, *by* can be used for both right and left. We give more weight to exact location words. For words in the left column of the Table 2 $LS = 0.7$ and for the words in the right column $LS = 0.3$.

(d) *Color Detail Weight (CD)*: Here too a finer label or a coarser label describing the color of the recommended product may be used. For example, *Red* is coarser, whereas *Salmon* can be considered finer being a specific shade of red. We give higher weight to shade names of the hex codes. For shade name $CD = 0.7$ and for the hue names $CD = 0.3$.

(e) *Color Compatibility (CC) of Objects*: The last piece of the puzzle lies in describing the color compatibility between the recommended product and the dominant color in the bounding box. This is achieved by measuring the color compatibility between the recommended object and the color of the dominant object.

Since we have the dominant color of the identified object and the recommended object, we create a five color palette by using other three colors as white. The reason for choosing white is that most color palettes are designed to be printed on white paper. A white background gives the appearance of color on paper, which makes it easier to compare and judge the combination [68]. The order in which the colors are arranged in the palette matter [51]. Therefore, we compute the theme score of all possible permutations (i.e ${}^5P_3 = 20$) by using a theme scoring model [51] which takes as input a five color theme/palette and returns a score between 1 and 5. We consider the maximum out of the twenty scores as the *color compatability* score between the two objects

Finally, we defined the tuple reward as:

$$Tuple\ Reward = OPC * AV * LS * CD * CC \tag{13}$$

For example, consider a tuple $< sofa, purple, front >$ corresponding to the purple sofa in Fig 5. Let the candidate recommendation be the blue chair shown in the same figure. The reward for this tuple is, $Tuple\ Reward = 0.896 \times 1 \times 0.3 \times 0.7 \times 0.3085 = 0.058$ Here, *front* being the exact location synonym yields $LS = 0.7$; purple being the hue name results in $CD = 0.3$. Since, we have a set of recommended products, each product has its own set of tuples with rewards calculated using the above reward function. This reward value is treated as an intrinsic property for every tuple.

6. **Selection of Tuples.** The text content is expected to contain different sentences for different recommended products. The approach we use is inspired from the diversity component of summarization algorithm [47]. To sentence diversity we follow the graph based approach to selecting diverse tuples. Each tuple is represented by a node v_i with node value as tuple reward r_i. There exists an edge e_{ij} between every two nodes. Edge e_{ij} has a weight $w_{ij} \in [0,1]$ which indicates the similarity between v_i and v_j. For our purpose we define similarity in the following way:

- If all the tuple elements for v_i and v_j are same then $w_{ij} = 1$
- If the object type is different then $w_{ij} = 0$
- If object type is same then three cases arise.
  - Same object color but different relative position, $w_{ij} = 0.7$
  - Same relative position but different object color, $w_{ij} = 0.2$
  - Both object color and relative position are different, $w_{ij} = 0.1$

Thus, an undirected fully connected graph is created $G(V,E,W)$, with some edges having zero weights, reward set for nodes R and budget B. Here budget B represents the number of tuples to be selected. The marketer can decide how many diverse sentences to generate. Corresponding to a recommended product, the tuple with the highest reward different from the already selected tuples is selected using an iterative approach aimed at exhausting the budget (see reference [47]). For example, the following tuples are selected for the mentioned recommended products shown in Figure 4.

- Purple Chair - $< sofa, purple, front >$
- Brown Chair - $< chair, orange, left >$
- Red Chair - $< potted-plant, brown, next >$
- Brown Chair - $< chair, pumpkin-orange, right >$

7. **Final Sentences.** After selection of tuples, we embed the tuple elements in predefined sentence templates (commonly used in targeting / recommendation emails) to generate the final content corresponding to each recommended product. The text content is added to the object embedded image and can be sent in an email or push notification. For example, for each of four products recommended, the corresponding sentences could be (each email contains one recommendation and one sentence; or an email can contain a catalog of several recommendations):

- We want you to check out this **purple chair** if placed **in front of** your **purple sofa**.
- How about a **brown chair** to the **left** of the **orange chair** at your home?
- In fact, this **red chair** will look amazing if placed **next** to your **brown potted-plant**.
- A **brown chair** to the **right** of your **pumpkin-orange chair** is a perfect combination too.

In these examples, the bold words are generated from the algorithm and the rest come from predefined marketing templates.

4.4 Final Targeting System

We create an AR system using Vuforia [14] SDK and Unity 3D [85]. The application is able to track the features discussed in Section 3.1. The Viewpoint Selection model described in Section 4.1 is plugged into the application. Some of the sample 3D models are provided in the app from Shapenet. We maintain the repository of 300 models (150 armchairs and 150 coffee tables) on the server. After an app session, the *Viewpoint Selection* model sends the viewpoint images (screen shot and camera) along with the location and pose of the virtual object in the camera coordinates to the server. Then, using those two images recommendations are created and embedded in the viewpoint. These generate the final catalogs having embedded recommendations. Further, the two viewpoint images and the products chosen from recommendation system are used to generate recommendation text.

5 MTurk Evaluation Studies and Results

Three different studies are conducted to evaluate the viewpoint selection model, the recommendation system, and whether or not images embedded in background are better for targeting compared to only product images. The object identification, location and color identification, and diversity measurement are well within the accuracy mentioned in the corresponding references. Hence, we do not evaluate text creation part. We explain the studies' task and the methodologies below. All the data, code and experimental results can be found in the supplementary material. The MTurk template [1], is used to create the user studies.

5.1 Viewpoint Selection Study

The goal is to evaluate whether the viewpoint our model captures agrees with the human judgment about the compatibility of the product with the surrounding objects. Implementing a full-scale study which exposes participants to a lot of viewpoints in different orderings to mimic real life is a huge resource requirement. Short of that we create a limited evaluation study as described below.

We examine the scenario of a session where a consumer is trying to select a chair for a meeting room. From that session we manually select three distinct images / frames for which the accelerometer reading (the most discriminating feature in the model as discussed in 4.1) is almost 0 for some duration. We label these images as, A, B, and C. Each image is shown for a longer duration and a shorter duration. With three images, two levels of duration, and orderings in which the images are depicted, we have many combinations if we run a full scale-study. For resource constraint, we keep the ordering constant.

The study implemented had three videos: Video 1 Image A (10 seconds), Image B (5 seconds), Image C (5 seconds); Video 2 Image A (5 seconds), Image B (10 seconds), Image C (5 seconds). Video 3 Image A (5 seconds), Image B (5 seconds), Image C (10 seconds). For each video, which is meant to mimic consumer AR based browsing, we create three recommended images as follows: In image A, B and C, the focal object (chair) is replaced with a recommended object keeping orientation, pose and location identical. The recommended object is kept same in all the images and is chosen randomly from the repository. This forms the three recommended images corresponding to the three videos. With three videos and three recommended images for each video, we have nine conditions for the study. Keeping with best practices in human based studies, this study was run as between-subjects that is, a different group of participants saw each condition so as not to compromise the purpose of the study and reduce biases. We had nine groups of MTurk participants each seeing one video and one recommendation image. We used a stringent criterion to screen MTurk participants, as well as embedded test questions (dummy questions unrelated to the subject matter) to ensure attentive participation. Thirty participants contributed to each condition. For details of the scenario used and questions asked please see supplemental materials.

Table 3: Average ratings for the viewpoint selection study questions.

	R1			R2			R3		
	Q1	**Q2**	**Q3**	**Q1**	**Q2**	**Q3**	**Q1**	**Q2**	**Q3**
V1	0.9	-0.2	0.4	1.2	0.7	0.9	1.4	0.9	1.3
V2	1.2	0.9	1.2	1.4	1.4	1.5	1.1	0.7	0.9
V3	1.3	0.8	1.3	1.7	0.9	1.2	1.9	0.9	1.6

5.2 Results Viewpoint Selection Study

The participants answered three questions on a scale of -3 to +3, with a middle point of 0. Using multiple measures for evaluation is a sound approach. The questions were: **Q1.** The chair is unattractive (-3) . attractive (3). **Q2.** The chair does not fit in the room (-3) . fits in the room (3). **Q3.** The chair is a poor choice (-3) a good choice (+3). An additional input field was provided for them to express qualitative reasons for their responses.

In Table 3 $V1$, $V2$ and $V3$ denotes the three videos; $R1$, $R2$ and $R3$ denotes the three recommended images. Per our hypothesis, we expect, when $V1$ (or $V2$, or $V3$) is shown, the highest preferred choice is $R1$ (or $R2$, or $R3$). Each cell denotes the average ratings over 30 responses for the particular question. In Table 3 the average ratings on all of $Q1$, $Q2$ and $Q3$ show that for both $V2$ and $V3$, our hypothesis is validated since when $V2$ is shown, the best performer is $R2$ and when $V3$ is shown, the best performer is $R3$. Only when $V1$ is shown, the best performer is not $R1$ for any question. So, 2 out of 3 times (67%) we see that recommendation shown in the viewpoint selected by our model (using only most discriminating feature i.e. accelerometer) gets higher ratings. Further, we also see that $R1$ is rated consistently lower than $R2$ regardless of the video shown and lower than $R3$ except when $V2$ is shown. There may be an inherent disliking towards image $R1$ by the participants. Hence, the results are not as expected in the $V1-R1$ case.

This provides reasonable support for conformance of our identified viewpoint with human judgment. We note that the study is difficult to execute without deploying the system for a large consumer base, although the results are promising.

5.3 Background Relevance Study

This study is a basic check of whether use of background is beneficial for influencing consumer's preference. We use the same product (chair) and the recommended product (chair) on a white background as used in the viewpoint study. Thirty MTurk participants compared the two images focal chair and recommended chair with white background. The supplement shows the details of the scenarios.

5.4 Results Background Relevance Study

The participants answered two questions on a scale of -3 to +3, with a middle point of 0. Using multiple measures for evaluation is a

Table 4: Background vs No Background

Questions	No-Background	Background
Q1	0.83	1.33
Q2	0.83	1.11

sound approach. The questions were: **Q1.** The chair is unattractive (-3) . attractive (3). **Q2.** The chair is a poor choice (-3) a poor choice (+3). An additional input field was provided for them to express qualitative reasons for their responses. Further, recall that these questions were also asked in the viewpoint selection study.

See Table 4 for results. For each question the room-background condition is preferred substantially over the no background condition, providing an empirical basis of our premise for doing the investigation. This tells us that recommending images embedded in suitable background is more engaging for a consumer instead of sending just the images of the product.

5.5 Recommendation Evaluation Study

The goal is to have humans compare recommendations from our model which uses style similarity and color compatibility with the baseline recommendation method which is based on content / description similarity. For the baseline, we use attribute specific similarity functions proposed by [71]. The idea is to create recommendations for a consumer who has browsed a product, which forms the input. The algorithm finds products similar to the product provided as input, from a set of similar products, where each product is described by a number of attributes such as name of the model, tags, weight, etc. For technical details of recommendations created by the baseline see supplemental material. To allow for generality, as inputs, two products each for two classes of products, are used. We use two different chairs and two different tables for the study, yielding four products and giving us four conditions for the study. Going with the between-subject study justified earlier, 30 participants are allocated to each condition.

In each condition, all participants see the same input product representing the browsing image. The background on which the chair is seen is identical in the browsing image as well as in all the recommendation images. These are done for maximum experimental control to disallow variations in browsed images with respect to which the recommended image is judged. For the specific focal product, our recommendation system and the baseline system each returns top three products. We thus get six products. Thus, for one chair we get six recommended chairs, for one table we obtain six recommended tables. Each recommended product is placed on the same background image as the browsing image keeping location, pose and viewpoint all constant. Thus, in each of the four conditions, we get six recommended images, half from our method and half from the baseline method. Our goal is to have participants in each condition evaluate their respective six images on preference ordering. For details of the study scenario see supplemental materials.

5.6 Results Recommendation Evaluation Study

Each participant ranked the six recommendation images from one to six, with no ties-in-ranking being allowed. As an additional metric, each distributed $100 amongst the six recommendation images, so that total added up to 100. This allows us to capture the degree to which one recommendation is preferred over another. Finally, an additional input field was provided for them to express qualitative reasons for their evaluation.

In Table 5, $O1$, $O2$ and $O3$ represent the recommendation images generated by our method (in ranked order), $B1$, $B2$ and $B3$ denote the recommendation images generated by the baseline (in ranked order). Cell (i, j) of Table 5 shows the proportion of the total number of times O_j is ranked above B_i. The metric is calculated over all

Table 5: Relative ratings for recommended images.

Images	O_1	O_2	O_3
B_1	0.59	0.49	0.35
B_2	0.67	0.67	0.42
B_3	0.68	0.65	0.52

Table 6: nDCG for the two methods across the surveys.

Method	Surv1	Surv2	Surv3	Surv4	Mean
Ours	1.0	0.67	1.0	1.0	0.92
Baseline	0.95	1.0	0.5	1.0	0.86

120 responses, that is, the total number of times responses obtained for the four focal products.

Looking at the lower diagonal matrix, we observe that the entries are all greater than 0.5. This provides support that the top recommendations generated by our method ($O1$ and $O2$) are preferred over the top recommendations generated by the baseline method ($B1$ and $B2$). The third recommendation of ours $O3$ is similar to the third from the baseline $B3$ (0.52).

We also calculate the normalized Discounted Cumulative Gain (nDCG) [29] for the images shown in the four conditions. The *relevance* (required for nDCG calculation) for each image is taken to be the average '$' assigned by the participants for that image. See supplement for the technical details of this metric. Table 6 shows the nDCG values for the four conditions. The last column shows the mean of the rows. A value so close to 1 for both the methods show that ranking given by both methods concurs with human rankings. That is, the mean nDCG value of 0.92 shows that the ordering done by humans for the recommended images was same as the ordering done by our method for those images. This also holds true for the baseline method but to a lower extant. This tells us that the ranking done by our recommendation system is more in concordance with the human rankings relative to that done by the baseline method.

6 Conclusions and Future Work

Augmented Reality (AR) has been used to significantly enhance marketing experiences in recent times. However, there is a dearth of study about the interaction and visual data from the AR applications in marketing domain. The visual data is quite rich, effectively defining the consumer context and would be useful to convert potential buyers into customers. In this work, we have created a novel consumer targeting system through addressing several challenges in harnessing the AR data. First, based on the consumer's interaction with the AR applications, we build a statistical model to identify augmented visual influencing the consumer purchase (viewpoint). Secondly, we create recommendations based on style and color compatibility of the objects present in the viewpoint. Further, we create personalized catalogs by embedding recommended products in the viewpoint. Third, we generate diverse personalized-persuasive targeting content talking about the physical surroundings of the consumer using the viewpoint cue. We have extensively evaluated this system through user studies. The results show that we are able to identify the viewpoint, our recommendations are better than the baseline method based tag similarity, and the recommendations embedded in the background is significantly better method for targeting than the usual product image-based targeting.

In future, we plan to deploy this system and evaluate the viewpoint selection model along with the effectiveness of the final targeting system. Further, we plan to create a recommendation system which uses both the textual description (tags) and the visual information. Incorporating more properties about the real world surroundings while creating persuasive content is another direction where we would like to venture into. Privacy of visual data is another concern which we would like to address in future.

REFERENCES

[1] M. Alper. Turksuite template generator. *http://mturk.mit.edu/template.php*, 2014.

[2] X. Amatriain. Mining large streams of user data for personalized recommendations. *ACM SIGKDD Explorations Newsletter*, 14(2):37–48, 2013.

[3] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3762–3769, 2014.

[4] X. Bonaventura Brugués et al. Perceptual information-theoretic measures for viewpoint selection and object recognition. 2015.

[5] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[6] M. C. Campbell and A. Kirmani. Consumers' use of persuasion knowledge: The effects of accessibility and cognitive capacity on perceptions of an influence agent. *Journal of consumer research*, 27(1):69–83, 2000.

[7] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[8] A. Charlesworth. *Key concepts in e-commerce*. Palgrave Macmillan, 2007.

[9] D. R. Cox and D. Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.

[10] D. Dai. Stylized rendering for virtual furniture layout. In *Multimedia Technology (ICMT), 2011 International Conference on*, pages 780–782. IEEE, 2011.

[11] T. Ding and S. Pan. Personalized emphasis framing for persuasive message generation. *arXiv preprint arXiv:1607.08898*, 2016.

[12] H. Dutagaci, C. P. Cheung, and A. Godil. A benchmark for best view selection of 3d objects. In *Proceedings of the ACM workshop on 3D object retrieval*, pages 45–50. ACM, 2010.

[13] J. Eno, G. Stafford, S. Gauch, and C. W. Thompson. Hybrid user preference models for second life and opensimulator virtual worlds. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 87–98. Springer, 2011.

[14] Q. C. Experiences. Inc., qualcomm vuforia developer portal (2015).

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[16] G. Ference, M. Ye, and W.-C. Lee. Location recommendation for out-of-town users in location-based social networks. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 721–726. ACM, 2013.

[17] C. M. Ford. Speak no evil: targeting a population's neutrality to defeat an insurgency. 2005.

[18] S. Freitag, B. Weyers, A. Bönsch, and T. W. Kuhlen. Comparison and evaluation of viewpoint quality estimation algorithms for immersive virtual environments. In *ICAT-EGVE*, pages 53–60, 2015.

[19] H. Gao, J. Tang, X. Hu, and H. Liu. Exploring temporal effects for location recommendation on location-based social networks. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 93–100. ACM, 2013.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[21] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[22] S. H. Ha. Helping online customers decide through web personalization. *IEEE Intelligent systems*, 17(6):34–43, 2002.

[23] J.-H. Hsiao and L.-J. Li. On visual similarity based interactive product recommendation for online shopping. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 3038–3041. IEEE, 2014.

[24] S. H. Hsu, M.-H. Wen, H.-C. Lin, C.-C. Lee, and C.-H. Lee. Aimed-a personalized tv recommendation system. In *European Conference on Interactive Television*, pages 166–174. Springer, 2007.

[25] Z. Huang, P. Hui, and C. Peylo. When augmented reality meets big data. *arXiv preprint arXiv:1407.7223*, 2014.

[26] W. Hürst and C. Van Wezel. Gesture-based interaction via finger tracking for mobile augmented reality. *Multimedia Tools and Applications*, 62(1):233–258, 2013.

[27] H. Ishwaran and U. Kogalur. Randomforestsrc: random forests for survival, regression and classification (rf-src). *R package version*, 1(0), 2014.

[28] C. H. Jackson. flexsurv: a platform for parametric survival modelling in r. *Journal of Statistical Software*, 70(8):1–33, 2016.

[29] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[30] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[31] D. Joanes and C. Gill. Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):183–189, 1998.

[32] D. R. Jones, C. D. Perttunen, and B. E. Stuckman. Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Applications*, 79(1):157–181, 1993.

[33] B. F. Katz, D. Q. Felinto, D. Touraine, D. Poirier-Quinot, and P. Bourdot. Blendervr: Open-source framework for interactive and immersive vr. In *Virtual Reality (VR), 2015 IEEE*, pages 203–204. IEEE, 2015.

[34] K. Kawase, B. H. Le, and R. Thawonmas. Collaborative filtering for recommendation of area in virtual worlds. In *Proceedings of the 13th Annual Workshop on Network and Systems Support for Games*, page 12. IEEE Press, 2014.

[35] P. Kazienko and M. Kiewra. Integration of relational databases and web site content for product and page recommendation. In *Database Engineering and Applications Symposium, 2004. IDEAS'04. Proceedings. International*, pages 111–116. IEEE, 2004.

[36] A. Kirmani and M. C. Campbell. Goal seeker and persuasion sentry: How consumer targets respond to interpersonal marketing persuasion. *Journal of Consumer Research*, 31(3):573–582, 2004.

[37] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434. ACM, 2008.

[38] B. Krulwich. Lifestyle finder: Intelligent user profiling using large-scale demographic data. *AI magazine*, 18(2):37, 1997.

[39] M. Lengheimer, G. Binder, and T. Rosler. Content management systems for mobile, context-dependent augmented reality applications. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, pages 1521–1526. IEEE, 2014.

[40] K. W.-T. Leung, D. L. Lee, and W.-C. Lee. Clr: a collaborative location recommendation framework based on co-clustering. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 305–314. ACM, 2011.

[41] H. P. Levy. Five key trends in gartners 2015 digital marketing hype cycle. *www.gartner.com/smarterwithgartner/five-key-trends-in-gartners-2015-digital-marketing-hype-cycle/*, 1995.

[42] B. Liu and H. Xiong. Point-of-interest recommendation in location based social networks with topic and location awareness. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 396–404. SIAM, 2013.

[43] Z. Lun, E. Kalogerakis, and A. Sheffer. Elements of style: learning perceptual shape style similarity. *ACM Transactions on Graphics (TOG)*, 34(4):84, 2015.

[44] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 89–96. IEEE, 2011.

[45] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development*

in Information Retrieval, pages 43–52. ACM, 2015.

[46] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[47] N. Modani, P. Maneriker, G. Hiranandani, A. R. Sinha, V. Subramanian, S. Gupta, et al. Summarizing multimedia content. In *International Conference on Web Information Systems Engineering*, pages 340–348. Springer, 2016.

[48] M. Mori, J. Orlosky, K. Kiyokawa, and H. Takemura. A transitional ar furniture arrangement system with automatic view recommendation. In *Mixed and Augmented Reality (ISMAR-Adjunct), 2016 IEEE International Symposium on*, pages 158–159. IEEE, 2016.

[49] S. Naspetti, R. Pierdicca, S. Mandolesi, M. Paolanti, E. Frontoni, and R. Zanoli. Automatic analysis of eye-tracking data for augmented reality applications: A prospective outlook. In *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*, pages 217–230. Springer, 2016.

[50] T. Nescher and A. Kunz. Using head tracking data for robust short term path prediction of human locomotion. In *Transactions on Computational Science XVIII*, pages 172–191. Springer, 2013.

[51] P. O'Donovan, A. Agarwala, and A. Hertzmann. Color compatibility from large datasets. *ACM Transactions on Graphics (TOG)*, 30(4):63, 2011.

[52] M. Ohta, S. Nagano, H. Niwa, and K. Yamashita. [poster] mixed-reality store on the other side of a tablet. In *Mixed and Augmented Reality (ISMAR), 2015 IEEE International Symposium on*, pages 192–193. IEEE, 2015.

[53] D. M. Olsson and L. S. Nelson. The nelder-mead simplex procedure for function minimization. *Technometrics*, 17(1):45–51, 1975.

[54] M. J. Pazzani and D. Billsus. Content-based recommendation systems. In *The adaptive web*, pages 325–341. Springer, 2007.

[55] T. Piumsomboon, A. Clark, M. Billinghurst, and A. Cockburn. User-defined gestures for augmented reality. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 955–960. ACM, 2013.

[56] C. Rack, S. Arbanowski, and S. Steglich. A generic multipurpose recommender system for contextual recommendations. In *Autonomous Decentralized Systems, 2007. ISADS'07. Eighth International Symposium on*, pages 445–450. IEEE, 2007.

[57] A. Reed. Activating the self-importance of consumer selves: Exploring identity salience effects on judgments. *Journal of consumer research*, 31(2):286–295, 2004.

[58] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[59] D. Roberts and A. D. Marshall. Viewpoint selection for complete surface coverage of three dimensional objects. In *BMVC*, pages 1–11, 1998.

[60] R. S. Roy, A. Padmakumar, G. P. Jeganathan, and P. Kumaraguru. Automated linguistic personalization of targeted marketing messages mining user-generated text on social media. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 203–224. Springer, 2015.

[61] D. Rudoy and L. Zelnik-Manor. Viewpoint selection for human actions. *International Journal of Computer Vision*, 97(3):243–254, 2012.

[62] N. S. Sahni, S. C. Wheeler, and P. K. Chintagunta. Personalization in email marketing: The role of non-informative advertising content. ., 2016.

[63] H. Sasanuma, Y. Manabe, and N. Yata. Diminishing real objects and adding virtual objects using a rgb-d camera. In *Mixed and Augmented Reality (ISMAR-Adjunct), 2016 IEEE International Symposium on*, pages 117–120. IEEE, 2016.

[64] J. Shen and L. B. Eder. Exploring intentions to use virtual worlds for business. *Journal of Electronic Commerce Research*, 10(2):94, 2009.

[65] Z. Shi, H. Wang, W. Wei, X. Zheng, M. Zhao, J. Zhao, and Y. Wang. Novel individual location recommendation with mobile based on augmented reality. *International Journal of Distributed Sensor Networks*, 12(7):1550147716657266, 2016.

[66] S. Siltanen, H. Saraspää, and J. Karvonen. [demo] a complete interior design solution with diminished reality. In *Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on*, pages 371–372. IEEE, 2014.

[67] A. Slobin. Virtual reality is the next frontier for retail. *www.adage.com/article/digitalnext/virtual-reality-frontier-retail/300061/*, 2015.

[68] M. Stone. Choosing colors for data visualization. *Business Intelligence Network*, 2, 2006.

[69] J. Stoyanova, R. Goncalves, A. Coelhc, and P. Brito. Real-time augmented reality shopping platform for studying consumer cognitive experiences. In *Experiment@ International Conference (exp. at'13), 2013 2nd*, pages 194–195. IEEE, 2013.

[70] S. Trewin. Knowledge-based recommender systems. *Encyclopedia of library and information science*, 69(Supplement 32):180, 2000.

[71] A. d. S. Urique Hoffmann and M. Carvalho. Finding similar products in e-commerce sites based on attributes. In *Alberto Mendelzon International Workshop on Foundations of Data Management*, page 46, 2015.

[72] T. Van Zandt. How to fit a response time distribution. *Psychonomic bulletin & review*, 7(3):424–465, 2000.

[73] P.-P. Vázquez, M. Feixas, M. Sbert, and W. Heidrich. Automatic view selection using viewpoint entropy and its application to image-based modelling. In *Computer Graphics Forum*, volume 22, pages 689–700. Wiley Online Library, 2003.

[74] P.-P. Vázquez, M. Feixast, M. Sbert, and W. Heidrich. Image-based modeling using viewpoint entropy. In *Advances in Modelling, Animation and Rendering*, pages 267–279. Springer, 2002.

[75] X. Wang, K. Hara, Y. Enokibori, T. Hirayama, and K. Mase. Personal multi-view viewpoint recommendation based on trajectory distribution of the viewing target. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 471–475. ACM, 2016.

[76] S. C. Wheeler, R. E. Petty, and G. Y. Bizer. Self-schema matching and attitude change: Situational and dispositional determinants of message elaboration. *Journal of Consumer Research*, 31(4):787–797, 2005.

[77] L. M. Wong, C. Dumont, and M. A. Abidi. Next best view system in a 3d object modeling task. In *Computational Intelligence in Robotics and Automation, 1999. CIRA'99. Proceedings. 1999 IEEE International Symposium on*, pages 306–311. IEEE, 1999.

[78] A. Wroblewska and L. Raczkowski. Visual recommendation use case for an online marketplace platform: allegro. pl. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 591–594. ACM, 2016.

[79] T. Wu, Y. Chen, and J. Han. Association mining in large databases: A re-examination of its measures. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 621–628. Springer, 2007.

[80] www.augmentedfurniture.com. Augmented furniture.

[81] www.aurasma.com. Aurasma.

[82] www.blippar.com/en/. blippar.

[83] www.homestyler.com. Autodesk homestyler.

[84] www.ikea.com. Ikea.

[85] www.unity3d.com. Unity 3d.

[86] M. Ye, P. Yin, and W.-C. Lee. Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 458–461. ACM, 2010.

[87] S.-Y. Yoon and J. Y. Cho. Understanding furniture decision making process and design preference using web-based vr technology. In *Annual Conference of IDEC, St. Louis, Missouri, March*, pages 25–28, 2009.

[88] M. Yuan, I. R. Khan, F. Farbiz, S. Yao, A. Niswar, and M.-H. Foo. A mixed reality virtual clothes try-on system. *IEEE Transactions on Multimedia*, 15(8):1958–1968, 2013.

[89] Z. Zhang, S. Shang, S. R. Kulkarni, and P. Hui. Improving augmented reality using recommender systems. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 173–176. ACM, 2013.