ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/320751075

[POSTER] Enhanced Personalized Targeting Using Augmented Reality

Conference Paper · October 2017

DOI: 10.1109/ISMAR-Adjunct.2017.34

CITATIONS
4

READS
180

6 authors, including:

Gaurush Hiranandani
Adobe Research Labs
20 PUBLICATIONS **48** CITATIONS
SEE PROFILE

Kumar Ayush
Indian Institute of Technology Kharagpur
28 PUBLICATIONS **1,147** CITATIONS
SEE PROFILE

Atanu Sinha
University of Colorado Boulder
32 PUBLICATIONS **210** CITATIONS
SEE PROFILE

All content following this page was uploaded by Kumar Ayush on 04 September 2018.

The user has requested enhancement of the downloaded file.

[POSTER] 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

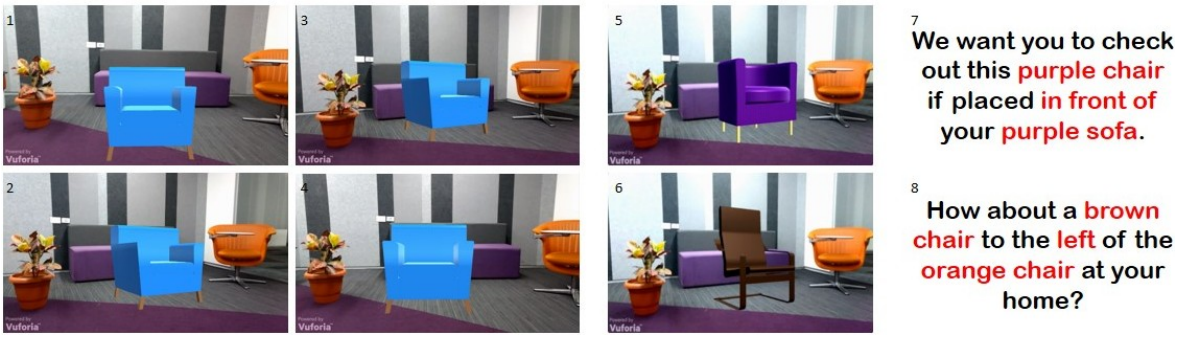

Figure 1: Targeting from AR-based data. From user session frames (1)-(4), the viewpoint selection model picks (3). Style and color compatibility recommends (5) and (6), while (7) and (8) show diverse targeting text for the recommendations (5) and (6) respectively. Words in red in (7) and (8) are system generated and put into predetermined marketing template (words in black).

Abstract

Augmented Reality (AR) based applications have existed for some time; however, their true potential in digital marketing remains unexploited. To bridge this gap we create a novel consumer targeting system. First, we analyze consumer interactions on AR-based retail apps to identify her preferred purchase viewpoint during the session. We then target the consumer through a personalized catalog, created by embedding recommended products in her viewpoint visual. The color and style of the embedded product are matched with the viewpoint to create recommendations, and personalized text content is created using visual cues from the AR data. Evaluation with user studies show that our system is able to identify the viewpoint, our recommendations are better than tag-based recommendations, and targeting using the viewpoint is better than that of usual product catalogs.

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

1 Introduction

Embedding reality in consumers' online shopping experience has been heralded as the 'next frontier for retail' and the coming of 'v-commerce'. V-commerce enables a consumer to overlay a virtual product on the real-world environment to judge its compatibility prior to purchase. Examples include the use of hand-held devices to virtually 'try on' furniture / shoes before purchase.[1] AR applications have drawn significatant attention in academics [6] and industry[1]. However, these works ignore consumers' preferences necessary to enhance user experience in AR [9]. The proposed approach introduces a robust statistical framework to model interaction as well as visual data generated by AR-based retail apps for targeting. Prior targeting approaches only use information from users' profiles [16], and textual description (content-based model) [14].

A typical AR-based v-commerce app would enable customer to 'tryout' the desired product like chair on a background of her living room. She can either (i) place different chairs on the background, or (ii) move the background around to check the compatibility from different viewing angles. We define *viewpoint*, to represent the visual at which the consumer judges the compatibility of the virtual product with the surrounding real world environment. The viewpoint holds information previously unavailable from the web-based browsing data, and provides the basis to suggest products having better compatibility with the surrounding real objects. For example, for enhanced targeting, images of recommended products *embedded in viewpoint* can be sent. Moreover, marketers can use *viewpoint* to create content about consumer's physical surroundings achieving greater personalization. This paper makes three novel contributions in advancing targeting through AR applications data. These are:

- **Viewpoint Selection:** A statistical model to select the viewpoint with the highest likelihood of influencing the consumer's purchase.
- **Recommendation System:** A system based on 3D design style and color compatibility to create product recommendations and embed recommended products in the selected viewpoint.
- **Targeting Content Generation:** A diverse personalized targeting content generator that uses visual information to create persuasive content regarding the physical surroundings of the consumer.

2 Review of Literature

The deployment of AR in v-commerce enhances consumer experience, as well as provides rich interaction data. The source of

*Corresponding author e-mail: ghiranan@adobe.com

[1] www.tinyurl.com/ycvydl89, www.tinyurl.com/yca5krvm

978-0-7695-6327-5/17 $31.00 © 2017 IEEE
DOI 10.1109/ISMAR-Adjunct.2017.34

IEEE computer society

AR-based data could be eye-tracking [20], head tracking [21], hand gestures [23] or GPS locations [17]. There has also been significant investment by industry[2] in AR apps. While the IKEA AR catalog app allows customers to have a virtual preview of furniture, Rayban's Virtual Mirror enables the consumer to try virtual sunglasses. The rich visual data collected by these apps would help in enhancing consumer experience.

In particular, customer viewpoint during an AR app session offers several insights into her preferences. The metric for viewpoint has varied definitions in the literature across different contexts. Vazquez *et al.* [28] define viewpoint entropy to compute good viewing positions automatically, while [2] shows how to automatically select the most representative viewpoint of a 3D model. An evaluation of the view selection algorithms has been conducted in [8]. However, none of these methods use data from AR-enabled systems for viewpoint selection. The proposed approach models the interactions in AR-based apps, without looking into the bulky visuals generated by the apps.

The *customer viewpoint* provides a unique advantage to the proposed system over the traditional recommendation systems [1]. The contextual recommendation in [24] exploits users ratings and ontology-based content categorization schemes. Wroblewska *et al.* [29] rely on images and extract color and texture information to find visually similar items. Our approach can ingest all such data, when available. In addition, the novelty lies in the ability to use viewpoint information to enrich the recommendation.

In the marketing literature, self-relevance is a well-established means of increasing message elaboration and personalization [15]. Roy *et al.* [26] automate message personalization by inserting adjectives evoking positive sentiment into the messages. A content selection method tailored to the customer has been proposed by [7] to automate personalized content selection. To the best of our knowledge there is no prior work which uses the consumer's viewpoint data from AR systems to automatically enrich marketing messages.

3 DATASETS

3.1 Data for Viewpoint Selection Modeling

This proprietary data, denoted by `A`, is generated by consumers' interaction with the AR mobile app of a multinational company that designs and sells household products. Using AR, consumers can virtually place objects in their rooms and change their configuration, generating clickstream data with time stamps in discrete steps of one second. Thirty days of data are available to us. We process the data to create aggregated features. All the AR interactions of a user within a session constitute one observation. The features are:

1. $\#c$: Number of times an object is chosen.
2. $\#o$: Number of distinct objects chosen in a session.
3. $\#r$: Number of times an object's configuration is changed.
4. $\#e$: Number of AR based events, defined as, ($\#e = \#c + \#r$).
5. T_C: Sum of all 'chosen' times ('chosen' time is the time duration between choosing an object and the next event).
6. T_R: Sum of all 'rendered' times ('rendered' time is the duration between changing an object's configuration and the next event).
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 denoting whether the *View Product Information* button is clicked or not during the session. E-commerce sites offer this button, or one similar, to consumers seeking more information when they become seriously interested in the product. Consumers who click on *VPI* during the session indicate serious interest about a product. Other consumers do not click on this button. We cannot observe their interest since there is no clicking-data during the session. Data from the latter consumers are recognized as censored observations. It is incorrect to treat these observations as if they show no interest in a product; it is beneficial to treat them as observations who may have interest in a product, but the end of the session (for whatever reason) censored the observance of their interest in a product.
9. T_P: Time elapsed between start of AR interaction and the first click on the *VPI* button. This is equal to T_T if $VPI = 0$. Note that T_P is the censored random variable when $VPI = 0$.
10. T_i: 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 predetermined threshold at the $(i-1)^{th}$ time step. It tells us whether or not the device is stationary before the i^{th} time step, and the device has a clear (non-blur) viewpoint image(s).

The visual 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 sessions of which 12% sessions had $VPI = 1$.

3.2 Data for Recommendations and Targeting Content

Shapenet [3] is a repository of 3D CAD models of objects. We have used the 'single 3D models' subset ShapeNetSem, which is a more densely annotated subset consisting of 12,000 models spread over 270 categories. The dense annotations about real-world dimensions ensures that any tag / description based recommendation system (baseline) has a fair chance to generate good recommendations. For our purpose, we selected a subset of 150 models each from the categories 'armchairs' and 'coffee tables'. The models were selected to form groups based on keyword annotations (design name, color name, etc.) to ensure good recommendation candidates from baseline [27] (described later). We denote this dataset by `S`.

4 METHODOLOGY

First, we describe the three primary contributions: (a) Viewpoint Selection, (b) Catalog Creation, and (c) Targeting Content Creation. We then explain the targeting system which uses these solutions.

4.1 Viewpoint Selection

In Section 1 we defined *viewpoint* as the visual (image) at which the consumer judges the compatibility of the virtual product (3D model) with the real world surroundings. There are two challenges that make viewpoint selection difficult: (i) the high volume of images that result from a consumer's session, and (ii) identification of augmented visual(s) from among these sequentially viewed images that the consumer prefers. Using dataset `A`, we build a statistical model to uncover the preferred viewpoint for the consumer. The novelty of our model is that we select the preferred augmented visual by analyzing the interaction of the consumers and the time stamps at which images (frames) are rendered on the app during a session. 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. We call this time-based event as the 'trigger of interest'. Following the Awareness-Interest-Desire-Action (AIDA) model in e-commerce [4], 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*), explained in Section 3.1. *VPI* is not a restrictive feature only to our dataset. All apps for product search have a button to obtain detailed information about a product. The trigger of interest will be the time epoch just before the consumer clicks on such a button. We use 'trigger of interest' in the following manner. There are two outcomes of any consumer session: session in which the consumer selects the *VPI* 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* for the

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

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

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

first time is the consumer's preferred augmented visual. The time epoch just prior to *VPI* is observable and hence the frame is observable in the data. However, time epoch is unobservable when the consumer does not click on *VPI* in the session. For these sessions, the model provides the estimate of the time epoch and hence, the image for which the consumer is most likely to click on *VPI*. We do this by recognizing that the time epoch is censored due to the end of session, as explained earlier. Our model thus infers the preferred augmented visual even in the absence of observed trigger of interest. The model estimates $f(T_P(\tilde{Z})$ where T_P is the time to view product information from the start of the session, $\tilde{Z}$ is the vector of covariates, and $f(.)$ is the probability density function.

2. **Accelerated Failure Time Model (AFT).** The model is borrowed from the response time literature [5], where positively skewed response times form the basis. We visually inspected that the time to click on *VPI* is positively skewed for almost all the groups (e.g., T_P given $\#o = 1$, T_P given $\#e = 2$, etc.) of data points. The empirical skewness for such groups is also less than -1. The censored nature of T_P leads to AFT models [5], for which the Generalized Gamma distribution, defined by two shape (say, a and b) and one scale (say, c) parameters is used. The flexibility of this distribution to respond to the characteristics of each group of data points justifies its use. The problem reduces to estimating a, b and c which are functions of the covariates. Hence, we have,

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

Each parameter can have a different functional relationship like logarithmic, exponential, linear, etc. with the covariates. We select the functional form which gives the maximum R^2 value. This is novel since most libraries only allow the same functional form to be used for all parameters. Empirically we find that the functional forms are different across parameters.

3. **T_P Estimation.** In fitting the model, parameters with the functional forms described in the previous step are replaced in the likelihood function of the generalized gamma (Equation 1). For example, let a and b have exponential and linear functional forms respectively for both $\#o$ and $\#e$. Then, we replaced the parameters in the likelihood functions by:
$a(\#o,\#e) = e^{\alpha_a+\beta_{1a}\#o+\beta_{2a}\#e}$, $\quad b(\#o,\#e) = \alpha_b+\beta_{1b}\#o+\beta_{2b}\#e$
The parameters (α_a, β_{1a}, ...) are estimated by flexsurv package in R using Nelder-Mead method [11]. Once the parameters are estimated, the empirical distribution of $f(T_P(\tilde{Z}))$ is obtained, and the mode of the distribution is judged to be the value of T_P. The logic follows from the definition of mode as being the most likely point.

4. **Viewpoint Selection for New Session.** In the dataset, we observe that 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 clear (not blurred) visual perspectives. We use the above method for a new consumer session as follows. 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 is below the predetermined threshold. If both the conditions are satisfied at some later stage in the session, the viewpoint is updated. So, there is only one viewpoint at the end of the session.

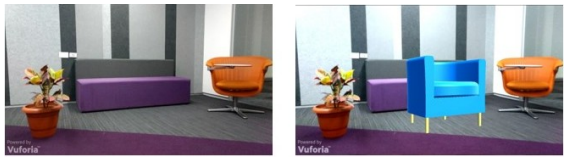

Figure 2: Viewpoint camera (left) and screenshot (right) frames.

4.1.1 Evaluation of Viewpoint Selection

`A` is divided into two parts: 80% for training and 20% for testing. We present results for 1, 3, and 5 seconds time windows that can potentially contain the viewpoint. We denote our model by `SA-FF` (Survival Analysis with different Functional Forms) and compare it against various baselines. For each model, the results are reported after selecting the best combination of features. The baselines are:

(a) **Standard Survival Analysis (`SA`):** This model fits generalized gamma assuming that the parameters of the distribution depends on the covariates but retain similar functional forms for all of them.

(b) **Weibull with Observed Functional Form:** We follow the same process as described above except that, instead of generalized gamma, we fit Weibull distribution with two parameters - shape and scale. The location parameter is taken to be 0. We tried two variants of this model: `W-Mode` and `W-Mean` which return the mode and the mean of the distribution as the estimated time point of T_P respectively. `W-Mode` is our method only as Weibull is a special case of generalized gamma class of distributions.

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

(d) **Random Forest for Survival Analysis (`RF-SA`):** For a non-linear model, we fit random forest for survival analysis [10] with target being T_P and the regressors being covariates ($\tilde{Z}$).

As per Table 1, `W-Mode` provides the best fit. We achieve 18.33%, 27.03% and 39.64% accuracy in 1, 3 and 5 seconds window respectively which is a significant improvement over `LR`. These accuracy results are good numbers as we only report small windows containing viewpoint from a session which can contain large number of such windows. Fitting Weibull is better than generalized gamma due to: (a) library dependency, (b) relationship observed between features and parameters. Further, the results justify the use of mode as the estimate. `LR` provides results similar to those from returning the mean of Weibull distribution (`W-Mean`). The improvement for our method decreases as the window length increases.

4.2 Catalog Creation

After obtaining the viewpoint, the second step is the catalog creation. For illustration purposes, let the final outcome of our viewpoint selection model be the two images shown in Figure 2. On the left is the background viewpoint (the camera image). On the right is the AR viewpoint which embeds the virtual product (chair) on that background (screenshot image). The workflow of the recommendation system is as follows:

1. **Location and Pose Identification.** To create a catalog with different embedded objects, the location and pose of the virtual object is required in the viewpoint. We designed our system so that it captures the location and pose of the virtual object in the camera coordinates throughout the consumer's session and then uses them for the time point when the viewpoint is selected.

Figure 3: Some of the candidate images having embedded product.

2. **Shape Style Similarity.** A consumer may prefer objects that are similar in physical design to the product she has tried. Unlike the usual e-commerce setting, where the similarity in design is determined by meta-tags, we leverage a structure-transcending method for evaluating the stylistic similarity of 3D shapes [18]. It is a more sophisticated way to determine the shape style similarity which can be computed for all the pairs of 3D models belonging to the same class (e.g. chairs) present in the marketer's repository. This method returns a distance measure between two objects. Let the style distance between objects i and j be $\alpha_{i,j}$. This is transformed to similarity (say $s_{i,j}$) by $s_{i,j} = \frac{1}{1+\alpha_{i,j}}$ which lies in $[0,1]$.

3. **Color Compatibility of Products Embedded in the Viewpoint.** We use Unity3D[3] to embed the products (3D models) in the camera image of the viewpoint using the location and pose obtained from Step 1 above. This creates a candidate set of images with embedded recommendations. Some of them are shown in Figure 3. Offline shoppers often use color compatibility of the product with the objects in the room. Thus, for each image, we extract a theme of five dominant colors by using [22]. This is passed to a crowd-sourcing based lasso regression model [22] to get a rating to the theme on a scale of 1 - 5. Let r_i be the rating for image i. 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.

4. **Recommendations.** To define an overall score of a candidate recommendation product embedded in the viewpoint, we conducted a survey over 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 calculated for each of the $6 \times 9 = 54$ candidate recommendations. 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 were 0.66, 0.68, 0.62, 0.72, 0.68 and 0.70 for the six lists. This 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 ground truth rankings for the six lists. Further, we found that the Pearson Correlation between the two scores corresponding to images in the experiment was 0.23. Additionally, the ranks observed from individual scores had a Kendall-Tau correlation (allowing ties) of 0.21. These values do not suggest a strong relation among scores. Hence, we define appeal $A(:)$ of an image i as weighted linear combination of the two scores. That is, $A(i) = w_1 s_i + w_2 c_i$. Here, $\vec{w} = (w_1, w_2)$ is the weight vector. After getting the ground truth ranking for each list, we have $6 * \binom{9}{2} = 216$ pairwise comparisons. We perform $4:1:1$ split for training, validation and testing. Then we apply rank-SVM [13] algorithm which use the obtained pairwise comparisons to learn the weights for different features. Validation data is used to achieve an optimal cost parameter as required in rank-SVM. The weights show the importance of the corresponding feature in deciding the ranks of the images. The learned weights are: $\vec{w} = (0.19, 1.66)$. For example, for the bottom left image of Figure 3, $s_i = 0.56$ and $c_i = 0.7$. Hence, $A(i) = 1.2684$. The recommendations embedded in images are ranked in decreasing order of their appeal. We select a predetermined number of top ranked images for the final catalog.

[3] www.unity3d.com

Figure 4: Labels with confidence, bounding boxes for real objects in camera frame (left) and virtual chair in screenshot frame (right).

4.3 Text Content Creation

In the recommendation system so far, the content created focused on shape, style, color and location of the objects in the viewpoint. To round off the recommendation, in this section we show how to incorporate textual content in it. We emphasize on diversity and persuasiveness of the text for the recommendations. For illustration, let the final outcome of the recommendation system, for which the text content is to be created, be the images shown in Figure 3.

1. **Objects, Color, and Location Identification.** We use Faster R-CNN (Region-based Convolutional Neural Network) [25] which takes as input an image and returns object proposals (bounding boxes) and object label with confidence score. To identify objects in viewpoint (camera image), we use the same parameters as used by [25]. Further, Step 1 of Section 4.2 gives the location of the virtual object in the camera coordinates as well (see Figure 4). Next, we identify the color of each object present in the background (viewpoint camera image) using the method in [22]. One, we take the above bounding boxes and resize them in the desired shape as required by [22]. Two, in a deviation from the work in [22], instead of looking for a 5 color theme we confine to a 1 color theme based on dominance. This is easy to do by just adding a constraint of all theme palettes to be equal in the objective function of [22]. Further, we name colors at two granularities: hue and shade names. For example, Crimson and Salmon are different shades of the Red hue. The 1-color theme is obtained in hex code by solving the above optimization process. We look for the color name of the hex code which is nearest (*L1* distance) to the hex code of the identified object, according to the hash function[4].

Relative position of the virtual object *wrt* each identified object in the viewpoint is determined. We use coordinates of the bounding boxes to determine if they intersect. If they do not intersect, we have either a vertical or horizontal separating axis for them (axis algined boxes). If the axis is vertical (horizontal), we give 'left'/'right' ('front'/'behind') label depending on their relative location. When the boxes intersect, we use a heuristic based on area of intersection to determine which box is in front of (or behind) the other. Moreover, various synonyms can be used 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 a relative position than *beside*.

2. **Tuple and their Rewards.** For each identified object, we generate tuples of the form <object type, object color, relative position>.

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

As per discussion above, multiple tuples are created using multiple color and relative position words for each object. Depending on the number of identified objects, suitably many tuples can be generated. Marketers may want to give preference to some tuples over others. For example, it is better to talk about an object is which is identified with more confidence instead of one which is not. In order to decide which tuples to use in the predefined template of sentences, we define *reward* for tuples which is based on the following properties:

(a) *Object Proposal Confidence (OPC)*: We seek objects which are identified with high confidence. An example is mentioned above. $OPC \in [0,1]$ is obtained from Step 1 of this section.
(b) *Association Value (AV)*: We seek objects having high association value with the recommended product. For example, a sofa has a higher association value with a table than with a painting. So, if the identified object is associated with the endorsed product class, according to [30], then we assign $AV = 1$, otherwise $AV = 0.5$.
(c) *Location Synonym Weight (LS)*: Different weights are given to exact or approximate location words of the recommended product *wrt* identified object. For example, *left* is an exact location word, whereas *beside* can be used for both *right* and *left*. We have taken $LS = 0.7$ for *right*, *left*, *front*, and *behind* and $LS = 0.3$ for others.
(d) *Color Detail Weight (CD)*: Different weights are given to shade (finer color) and hue (coarser color) names. We have taken $CD = 0.7$ and $CD = 0.3$ for the shade and hue names respectively.
(e) *Color Compatibility (CC) of Objects*: Since we have the dominant color of the identified object (Step 2 above) and the recommended object (present in the repository), we create a five color palette by using the remaining three colors as white. A white background gives the appearance of color on paper, which makes it easier to compare and judge the combination. The order in which the colors are arranged in the palette matters [22]. Therefore, we compute the theme score (Step 4 of Section 4.2) of all possible permutations (i.e ${}^5P_3 = 20$). We define the maximum out of the twenty scores as the *color compatibility* score between the two objects.

Finally, we define, $Tuple\ Reward = OPC * AV * LS * CD * CC$.

3. **Final Sentences.** The text content should contain different sentences for different recommended products. Thus, we follow a graph based approach to select diverse tuples as done in the summarization algorithm [19]. Each tuple (i) is a node (v_i) with reward (r_i). An edge (e_{ij}) between two nodes has weight ($w_{ij} \in [0,1]$) indicating similarity between the nodes. We define the similarity as:

$$w_{ij} = \mathbb{1}_{obj}(0.6\mathbb{1}_{col} + 0.1\mathbb{1}_{loc} + 0.2\mathbb{1}_{col}\mathbb{1}_{loc} + 0.1) \qquad (2)$$

where, $\mathbb{1}_k$ denotes the indicator for k being same in nodes i and j. Thus, a fully connected graph $G(V,E,W)$ is created with budget B. B denotes the number of tuples which the marketer wants. 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 [19]). For example, the tuples selected for the mentioned recommendations shown in Figure 3 are:

- <sofa, purple, front>
- <potted-plant, brown, next>
- <chair, orange, left>
- <chair, pumpkin-orange, right>

We embed the selected tuple elements in predefined sentence templates (commonly used in targeting) to generate the final content corresponding to each recommendation. For example, for each of the four products recommended, the corresponding sentences are:

- 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** looks great too.

The words in bold above are generated from the algorithm.

Table 2: Average ratings for the viewpoint 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

4.4 Final Targeting System

We create an AR system using Vuforia[5] SDK and Unity 3D[3]. The application tracks the features discussed in Section 3.1. The *Viewpoint Selection* model (Section 4.1) is plugged into the app. Some 3D models are provided in the app from Shapenet. The repository of 150 armchairs and 150 coffee tables is maintained on the server. After an app session, the *Viewpoint Selection* model sends the viewpoint images along with the location and pose of the virtual object in the camera coordinates to the server. Using this data, recommendations are created and embedded in the viewpoint. Lastly, targeting content is created for the recommended products.

5 MTurk Evaluation Studies and Results

Different user studies evaluate viewpoint selection, recommendation system and improvement in targeting, by using a between subjects design where responses are based on a discrete Likert scale.

5.1 Viewpoint Selection Study and Results

The goal is to evaluate if the viewpoint outputted by the system agrees with human judgment. The study emulated a session where a consumer is trying to select a chair for a room. From parts of the session video having low accelerometer reading, we select 3 distinct images A, B and C. We use three videos: Video 1: Images A (10 seconds), B (5s), C (5s); Video 2: Images A (5s), B (10s), C (5s) and Video 3: Images A (5s), B (5s), C (10s). For each video, three recommended images are created. For A, B and C, the focal object (chair) is replaced with a recommended object keeping location and pose identical, resulting in three recommended images corresponding to the three videos. A different group of 30 participants saw each of nine conditions so as to reduce biases. Following the use of multiple measures, the three questions asked were: The chair **Q1.** is unattractive (-3) - attractive (3), **Q2.** does not (-3) - does fit (3) in the room and **Q3.** is a poor (-3) - good (+3) choice.

In Table 2, *V1*, *V2*, and *V3* denote the videos, and *R1*, *R2* and *R3* the recommended images. Per our hypothesis, when *V1* (or *V2*, or *V3*) is shown, the highest preferred choice is *R1* (or *R2*, or *R3*). Table 2 shows average ratings per question. First, we use test of means (*t-test*) to verify whether the average ratings for *R1*, *R2*, and *R3* corresponding to *V1*, *V2*, *V3* are greater than 0, the mid-point of the scale. We find that aggregated across *Q1-Q3*, *R1* is rated not significantly greater than 0 ($t = 1.15$, $n = 30$, $p = 0.13$) for *V1*, while *R2* and *R3* are rated significantly greater than 0 for *V2* ($t = 6.94$, $n = 30$, $p < 0.001$) and *V3* ($t = 5.43$, $n = 30$, $p < 0.001$), respectively. Thus, as a second step we confine testing to *V2* and *V3*. We compute the proportion of ratings that are positive as opposed to negative. Aggregated across *Q1-Q3*, our 1-sided *Chi-Square* test of proportions for *V2* and *V3* combined yields $\chi_2 = 4.17$, $n = 180$, $p = 0.02$. That signifies for *V2* and *V3*, our hypothesis is validated. We find support for our identified viewpoint versus that of human judgment, except in the *V1-R1* case. We expect that deploying the system for participants will provide stronger statistical support.

5.2 Background Relevance Study and Results

The goal is to check whether the use of background helps in influencing consumers preference. We use the same product (chair) and the recommended product (chair) on a white background as used

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

Table 3: 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

in the viewpoint study. Participants compared the two images: focal chair and recommended chair with no (white) background. The questions asked were: **Q1.** The chair is unattractive (-3) - attractive (3). **Q2.** The chair is a poor choice (-3) - a good choice (+3). For the 'no background' case, both *Q1* and *Q2* got mean ratings of 0.83, whereas 'with background' produces mean ratings of 1.33 and 1.11 for *Q1* and *Q2* respectively. The 'with background' condition comprises the three backgrounds used in Section 5.1. Aggregating *Q1* and *Q2*, *Mann-Whitney U* test to verify the 1-sided hypothesis that background is preferred over no background gives $U = 1151.5$, $n = 120$, $p = 0.098$. The study in Section 5.1 shows that background *V1* has weaker results than *V2* and *V3*. Thus, considering only *V2* and *V3*, we get $U = 679.5$, $n = 90$, $p = 0.017$. Overall, the 'background' is preferred over that of 'no background', providing an empirical basis for our investigation.

5.3 Recommendation Evaluation Study and Results

The goal is to have humans compare recommendations from our model with a baseline recommendation [27] based on description similarity. The baseline takes the browsed products as input. It then finds recommendations similar to the input based on attributes such as model, weight, etc. Using 2 chairs and 2 tables we have four conditions, with 30 different participants allocated to each condition. All the participants see the same input product as the browsing image. The recommended products are placed on the same viewpoint as the browsing image, keeping location and pose constant. This controls for unintended variations with respect to which the recommended image is judged. For each condition, 6 recommended images result, 3 from the proposed method and 3 from the baseline. The participants then rank the 6 recommended images without ties, through distributing 100 points among the recommended images.

In Table 3, $\{O1, O2, O3\}$ denote the recommended images from our method (in ranked order), $\{B1, B2, B3\}$ are images from the baseline (in ranked order). Cell (i, j) shows the proportion when *Oj* is ranked above *Bi*, calculated over 120 responses obtained across the 4 conditions. In Table 3 the top recommendations from our method (*O1* and *O2*) are preferred over the top recommendations by the baseline (*B1* and *B2*). We run a 1-sided *Chi-Square* one sample proportions test to check whether our recommendation is better than the baseline in more than 50% of the cases. Comparisons *O1-B1*, *O1-B2*, *O1-B3*, *O2-B2*, and *O2-B3* have $\chi_2 \in (3, 15)$, $n = 120$, $p < 0.05$. *O3* has similar preference as *B3* ($\chi_2 = 0.14$, $n = 120$, $p = 0.31$). Further, we computed nDCG [12] for the 4 conditions corresponding to our recommendations. The *relevance* (required for nDCG) for each image is taken to be the average points assigned by the participants for that image. We got the mean nDCG of 0.92 across the 4 conditions which shows that the human ordering concurs with ordering given by our method.

6 Conclusions and Future Work

We create a novel consumer targeting system through modeling the AR-based data. Firstly, using the time-stamps of consumers interaction, the proposed statistical model identifies the augmented visual influencing consumers purchase (viewpoint). Secondly, a personalized catalogue of recommendations based on the style and color compatibility of other objects in the viewpoint is offered. Thirdly, we generate diverse personalized targeting content accounting for the physical surroundings as inferred from the viewpoint. Evaluation through user studies shows good accuracy in identifying the viewpoint, recommendations being better than the baselines, and improved targeting using viewpoint. In future, we plan to deploy this system for comprehensive evaluation, as well as study other context parameters to further enrich the experience.

References

[1] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-based systems*, 46:109–132, 2013.

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

[3] A. X. Chang, T. Funkhouser, L. Guibas, et al. Shapenet: An information-rich 3d model repository. *arXiv:1512.03012*, 2015.

[4] A. Charlesworth. *Key concepts in e-commerce*. P. Macmillan, 2007.

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

[6] D. Dai. Stylized rendering for virtual furniture layout. In *ICMT*, 2011.

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

[8] 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.

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

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

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

[12] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *Transactions on Information Systems*, 20(4), 2002.

[13] T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD*, pages 133–142. ACM, 2002.

[14] P. Kazienko and M. Kiewra. Integration of relational databases and web site content for product and page recommendation. In *Database Engineering and Applications Symposium*. IEEE, 2004.

[15] 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.

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

[17] B. Liu et al. Point-of-interest recommendation in location based social networks with topic and location awareness. In *ICDM*, 2013.

[18] Z. Lun, E. Kalogerakis, and A. Sheffer. Elements of style: learning perceptual shape style similarity. *Transactions on Graphics*, 34, 2015.

[19] N. Modani et al. Summarizing multimedia content. In *WISE*, pages 340–348. Springer, 2016.

[20] 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 *AVR*. Springer, 2016.

[21] T. Nescher and A. Kunz. Using head tracking data for robust short term path prediction of human locomotion. In *TCS*. Springer, 2013.

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

[23] T. Piumsomboon, A. Clark, M. Billinghurst, and A. Cockburn. User-defined gestures for augmented reality. In *CHI*, pages 955–960, 2013.

[24] C. Rack, S. Arbanowski, and S. Steglich. A generic multipurpose recommender system for contextual recommendations. In *ISADS*, 2007.

[25] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[26] 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. *CICLing*, 2015.

[27] A. d. S. Urique Hoffmann and M. Carvalho. Finding similar products in e-commerce sites based on attributes. In *AMW*, 2015.

[28] 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, 2003.

[29] A. Wroblewska and L. Raczkowski. Visual recommendation use case for an online marketplace platform: allegro. pl. In *ACM SIGIR*, 2016.

[30] T. Wu, Y. Chen, and J. Han. Association mining in large databases: A re-examination of its measures. In *ECMLPKDD*. Springer, 2007.

View publication stats