Large Scale Visual Recommendations From Street Fashion Images

Vignesh Jagadeesh
eBay Research Labs
2065 East Hamilton Avenue
San Jose, California, USA
vjagadeesh@ebay.com

Robinson Piramuthu
eBay Research Labs
2065 East Hamilton Avenue
San Jose, California, USA
rpiramuthu@ebay.com

Anurag Bhardwaj
eBay Research Labs
2065 East Hamilton Avenue
San Jose, California, USA
anbhardwaj@ebay.com

Wei Di
eBay Research Labs
2065 East Hamilton Avenue
San Jose, California, USA
wedi@ebay.com

Neel Sundaresan
eBay Research
2065 East Hamilton Avenue
San Jose, California, USA
nsundaresan@ebay.com

ABSTRACT

We describe a completely automated large scale visual recommendation system for fashion. Our focus is to efficiently harness the availability of large quantities of online fashion images and their rich meta-data. Specifically, we propose two classes of data driven models in the Deterministic Fashion Recommenders (DFR) and Stochastic Fashion Recommenders (SFR) for solving this problem. We analyze relative merits and pitfalls of these algorithms through extensive experimentation on a large-scale data set and baseline them against existing ideas from color science. We also illustrate key fashion insights learned through these experiments and show how they can be employed to design better recommendation systems. The industrial applicability of proposed models is in the context of mobile fashion shopping. Finally, we also outline a large-scale annotated data set of fashion images (**Fashion-136K**) that can be exploited for future research in data driven visual fashion.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Data Driven Models; Fashion, Retrieval; Crowdsourcing

Keywords

visual recommenders; fashion; e-commerce; color modeling; user behavior

1. INTRODUCTION

Everybody Loves Fashion: Fashion has been and continues to be an active area of research in a variety of disciplines. Artists and designers strive to create fashionable apparel that please end users and in turn enjoy wide adoption. Economists and Statisticians have long studied the spending patterns of shoppers to understand and design market structures for fashion. Psychologists have been looking into reasoning about abstract fashion notions on what it means to be visually/aesthetically pleasing, why certain fashion trends become instantly popular while others do not. More recently data scientists have been investing considerable effort into building systems are capable of organizing and making sense of the tremendous amount of fashion data flooding the web. While the different problems appear diverse at the surface, it is fairly evident that the driving force behind each problem is in understanding fashion DNA of the target audience. The target audience could be variable: people of certain age group, people from a certain geographic location, people visiting a website at a given time of year, or at the extreme case a single user. On the other hand, the key difference between bleeding edge research into data driven fashion analytics and more traditional approaches of studying fashion is in leveraging the scale of data. For instance, perceptually motivated color science rules are a result of many small scale psychological experiments, results of which psychologists agree upon. These research studies do not leverage the scale of data which has become increasingly available over the past few years. As will become evident in later sections, purely data driven approaches proposed in this work comprehensively outperform classical rules of color science. This

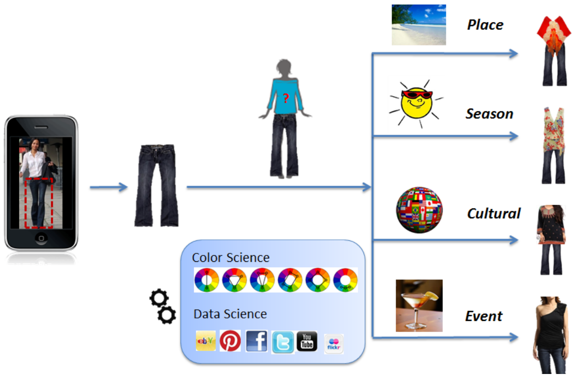


Figure 1: Illustration of context based recommendation. Different tops are recommended for the same pair of blue jeans.

KDD'14, August 24–27, 2014, New York, New York, USA.

http://dx.doi.org/10.1145/2623330.2623332

makes a strong case for re-thinking fashion rules from a bottom up data-driven perspective, from single users to groups of users to modeling fashion tastes of entire populations. **Why Large Scale Fashion Analytics → Fashion DNA ?**

- Large scale datasets increasingly available due to fashion sources from web, mobile and social networks
- Fashion choices are influenced by variety of factors eg: visual, textual, branding, seasonal, demographic, occasional
- Leveraging large scale datasets is the most promising strategy to capture these variable influencers and their correlations
- Thorough understanding of these influencers results in better personalization of fashion choices
- Fashion analytics could shed light on novel interaction mechanisms of the future

We believe there is ample scope to adopt several existing problems in the fashion domain and demonstrate the utility/gains demonstrated by a data driven thinking. We adopt the problem of "upselling in e-commerce" where an online shopper with an item in her shopping cart is recommended other items that complement the cart. This paper proposes a similar application for fashion, where given an image of a fashion item (say "jeans"), the goal is to recommend matching fashion items (say "tops") that complement the given item (Figure 1). In order to learn the recommender we leverage large scale street fashion data, where visual experiences have strong social, cultural and commercial importance. Street fashion images typically contain full view of a street fashion model striking a pose with several fashionable accessories, see Figure 2(c) for an example. During the training phase, each image has a fashion model wearing multiple apparel (street fashion data), while the images used in test phase comprise a single apparel (e-Commerce data). Throughout this work, we stick to image data alone while noting that fusing modalities such as text, video and brands would lead to more comprehensive models.

Complementary Recommenders for e-Commerce: The computational challenges involved in complementary recommenders are two fold. Firstly, the representation of fashion item into visual features is an open problem. This problem is compounded by the task of learning inter-dependencies between such features from different items (i.e. "tops" and "jeans"). Further, a number of practical challenges that need to be addressed include obtaining high quality image data with clean background, scalability issues in terms of memory and speed requirements that can handle large image databases as well as generate real-time predictions. Such constraints are even more crucial in the context of mobile shopping which is limited by the compute, memory and network capacities. We formulate the recommendation problem as follows: Given an image i containing a set of fashion items, also referred to as "parts" (i.e. sunglasses, tops, skirts, shoes, handbags), the holistic description H_i of an image i is given by $\underline{H}_i^T := [\underline{h}_{i1}^T, \underline{h}_{i1}^T, \ldots, \underline{h}_{iP}^T] \in \Re^{PK}, \underline{h}_{ij}^T \mathbf{1}_K = 1, \forall j \in [1, 2, ..P]$, where P is the number of parts, K is the size of code book to represent a part and $\underline{h}_{ij} \in \Re^K$ denotes the representation of part j for image i. Our task is to learn a predictive model $M(\mathbf{H}, \mathbf{q})$ where $\mathbf{H} = [\underline{H}_1, \underline{H}_2, \ldots, \underline{H}_N]^T$ for N number of images in the data set and $\mathbf{q} = [\underline{h}_{q1}^T, \underline{h}_{q2}^T ... \underline{h}_{qm}^T, \ldots, \underline{h}_{qP}^T] \in \Re^{PK}$ represents an input query description with a missing part m. Given an input query vector with a missing column entry, the problem of model learning reduces to predicting the value of empty column in the vector using Model M. Once the model is learned, it can be used to transform the input query which can then be used in applications such as recommendation systems, retrieval and personalization.

Table 1: Summary of fashion data sets in our study.

Data set	Comments
Fashion-136K	135893 street fashion images with annotations by fashionistas, brand, demographics.
Fashion-Toy	Subset of Fashion-136K. $n = 600$, $n_{train} = 500, n_{test} = 100$.
Fashion-63K	Subset of Fashion-136K. $n = 63K$, $n_{train} = 53K, n_{test} = 10K$
Fashion-350K	350K images of just tops & blouses to test retrieval performance using fashionistas.
Fashion-Q1K	1K images of skirts used to retrieve images from Fashion-350K. Skirts have one of the following different patterns: animal-print (100), floral (200), geometric (100), plaids & checks (150), paisley (50), polka dots (100), solid (200), stripes (100).

Product Deployment: The techniques described in this work are currently being deployed for internal testing through the eBay Mobile Fashion App, building on the eBay Image Swatch [1], see (https://itunes.apple.com/in/app/ebay-fashion/id378358380?mt=8). We estimate that the next version of the Fashion App with visual recommendation technology will be available for download later this year through the App Store for users in the United States. The system is implemented in such a way that the user first takes a picture of either top/bottom clothing using a mobile phone, and the algorithms described in this work respond with real time recommendations of complementary bottom/top clothing after searching through millions of images indexed by our system.

2. DATA DRIVEN VISUAL RECOMMENDERS

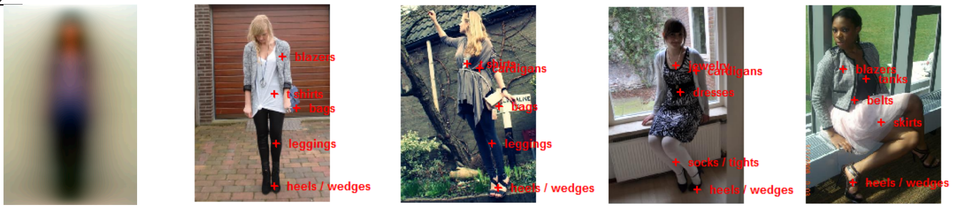
In order to place our ideas in context, we have divided the scope of data driven fashion analytics into four main stages where one stage naturally leads to another.

1. The Datasets: We use 3 datasets in our experiments, namely: *Fashion-136K*, *Fashion-350K* and *Fashion-Q1K*. *Fashion-136K* is a data set created by crawling the web for photographs of fashion models. It consists of 135893 street fashion images with annotations by fashionistas, brand, demographics. Hence, all images comprise top and bottom clothing co-occurring in the same image. This dataset is used for the training phase. See Table 2 for the list of tags present in *Fashion-136K*, and their mapping to corresponding human body parts. *Fashion-Q1K* dataset consists of 1000 skirt images which are used as input queries to recommender. Given these queries, the recommender searches and returns matches from *Fashion-350K*. The result of retrieval is a ranked list of the Fashion-350K images sorted by relevance to a query from Fashion-Q1K. *Fashion-350K* images are from a clothing inventory containing only top clothing (without model or mannequin). In a mobile setting, it is assumed that the user takes a photograph of a single apparel of interest (or guides the recommender by drawing a bounding box around the apparel), say bottom clothing image patches used as queries for top clothing. With advancements in computer vision, it might be possible in the near future to automatically detect query apparel patches from images of people wearing fashion accessories.

Low-Level Representation: We utilize two low-level representations of image, namely (i) HSV Histograms: Given P parts of interest characterized by P distinct bounding boxes, a normalized 40

Item	Quantity
Initial size of corpus	196974
Final size after cleanup $(H \geq 400, H/W > 1)$	135893
Number of unique tags	71
Number of unique users	8357
Number of unique brands	20110
Number of known cities	115
Number of known regions	136

(a) Images, tags, geo-location.



(b) Average image.

(c) Exemplars showing various pose and complex background.

Figure 2: Basic summary of street fashion corpus Fashion-136K. Images were posted and tagged by fashion designers and fashionistas. There are about 3-4 annotations per image, with each user posting anywhere from 8 to 524 images. Most contain full view of model.

Table 2: From accessory tags to semantic attributes of human body. The tags highlighted in bold are those which are most popular, and are listed with the number of occurrences in the dataset.

Semantic Group	Tags in Corpus
1 - Head (7069)	beanies / berets, hats, headbands, headdress
2 - Face (13738)	earrings, glasses / sunglasses
3 - Neck (32278)	**jewelry-25255**, necklaces, scarves / echarpes
4 - Chest (1514)	bodysuits, bustiers / bras, corsets, swimwears
5 - Hand (60584)	**bags-43864**, bracelets, clutches, gloves, rings, watches
6 - MidTorso (113506)	**dresses-33253**, jumpsuits, little black dresses, little white dresses, long dresses, minidresses, **shirt / blouses - 33923**, sweaters, **t-shirts-24907**, tanks, tartan, tunics, turtlenecks
7 - SideTorso (49961)	blazers, brooches, capes, cardigans, coats, hoodies sweatshirts, **jackets-18016**, trenches, vests
8 - Waist (11669)	belts
9 - UpperThigh (34906)	miniskirts, **skirts-20732**, tutus, shorts
10 - LowerThigh (48028)	boyfriend pants, harem / baggy pants, **jeans-21663**, jeggings, leggings, over the knee, **pants-17928**, rolled-up, skinny pants
11 - Knee (5881)	color tights, socks / tights
12 - Foot (84745)	ankle boots / booties, ballerinas, biker boots, boots, cowboy boots, flats, gladiators, **heels / wedges-38447**, lace ups, mary janes, oxfords / derbies, peep toes, sandals, sneakers

dimensional color histogram is computed by quantizing Hue Saturation and Value into $24, 8, 8$ uniformly spaced bins respectively, (ii) Color BoW: A K-dimensional feature vector is obtained by randomly sampling fixed size (15×15) patches from the bounding box and quantizing them to a learned code book of K code book entries. We performed extensive experiments with texture features such Gabor Wavelets [10], Textons [12], Dense-SIFT[3] and found them to be unstable for texture classification in clothing. A preliminary analysis suggests that deformable nature of clothing often leads to unreliable texture estimation where even wrinkles and folds on solid clothes can be misinterpreted as texture. Further, we also looked at style based features based on attributes of clothing like cuts, sleeve etc., but were unable find a robust style descriptor in the vision literature. Hence, in this paper we focus our attention to using only color based image representations.

Parsing Fashion Accessories using Computer Vision: Given a database of images of fashion models (see Figure 2(c)), we are required to parse the image of the fashion model into constituent apparel being worn. In other words, the aim is to identify or localize every apparel from a label set purse, sunglass, shirt, pant etc ... Attempting to get these labels through a crowdsourcing would quickly turn out to be infeasible since the number of images are fairly large along with large label sets. Automating this by computer vision is the only feasible method to leverage the utility of this large scale data. Over the past two years there have been a couple of efforts to approach this very problem of automating image parsing. While the progress shown by these successive papers are indeed very impressive they are still not scalable to very large datasets. In this work, we utilize the x,y coordinates of top and bottom clothing tagged by fashionistas as a *weak label* to extract swatches from the image. Ensuing discussions will focus on modeling pairwise relationships between top and bottom clothing using color descriptors.

2. Preliminary Data Analysis: Once the data has been acquired, figuring out the right questions to ask from the data would lead to mechanisms for understanding the data well. For instance, how many red shirts occur with blue pants in street fashion images? These questions can be trivially answered by database queries on the collected datasets. Since we predominantly focus on color modeling, the question we ask is "Is there signal in the empirical co-occurrence of color patterns between top and bottom clothing of fashion models in images we collect?" The empirical co-occurence of color patterns street fashion dataset (*Fashion-136K*) in Figure 3. The result is intuitively satisfying since there is a clear signal of some co-occurrence of color patterns being preferred a lot more than others. Estimating these patterns would help one understand fashion tastes better, and sits at the foundation of discussions to follow.

3. Predictive Modeling: The goal of predictive modeling would be to forecast, or predict what might be interesting to the user based on past history. Our assumption in this paper is that street fashion images contain co-occurrence patterns which reflect the current fashion tastes of shoppers. This constitutes historical (training) data from which a visual recommender is to be learnt for predicting co-occurring patterns that are potentially interesting to shoppers. Further, we design two classes of models DFR (Deterministic Fashion

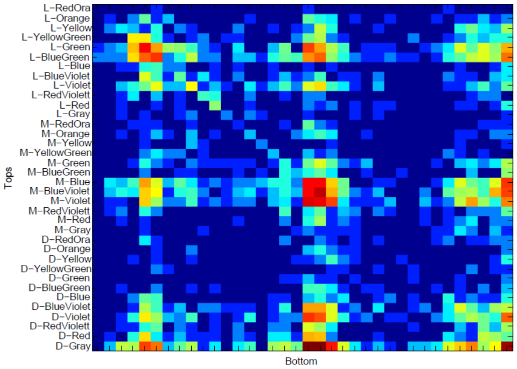


Figure 3: Empirical co-occurrence matrix of tops along rows, and bottom clothing along columns. Some intuitive patterns that can be read off the above matrix are green and blue tops go well with green and blue bottoms respectively. Further, the last column indicates that dark gray (black) bottoms go well with any colored top.

Recommender) and SFR (Stochastic Fashion Recommender). As the names suggest, results from DFR are reproducible, while those from SFR are not. Our intuition behind designing SFR was that fashion is inherently exploratory, and injecting randomness into recommenders would serve to capture this exploratory nature. We propose a suite of algorithms under each category, and comprehensively benchmark their relative pros and cons. Section 4 discusses these issues in detail.

4. User Interaction with Predictive Models: Finally, it is of considerable interest to understand how the predictive models interact with users. In other words, the aim is to perform extensive validation of these predictive models through end user studies. More on this in Section 5.

3. BACKGROUND AND RELATED WORKS

There have been only handful attempts to solve the problem of visual fashion recommendation. We are aware of only two existing works in this area. Iwata et al. [6] study this problem in isolation where they propose a topic model based approach to solve this problem. However, given the small dataset size of their experimentation it is difficult to ascertain the relative merits of their system. Liu et al. [8] propose a latent SVM based formulation to handle both “wear properly" and “wear aesthetically" criteria in their model. However most of their experimentation is tailored towards solid colored clothing and their qualitative analysis fails to demonstrate the efficacy of system performance on a variety of clothing patterns. Other related papers on fashion domain [4, 9, 13, 14, 15] work on the problem of fashion parsing and similarity retrieval where the goal is to retrieve similar fashion image for a given query. These methods either employ the mixture of parts based pose estimator or use the poselets based part detectors after applying the deformable parts based detector for person detection. Yamaguchi et al.[14] utilize a superpixel labeling approach on a CRF for inferring part labels from a labeled dataset of fashion apparel. The Street to Shop system [9] attempts to solve the cross domain discrepancy between catalogue images and images captured in the wild sent in as a query. Gallagher et al. [4] utilize attribute based classifiers regularized by a CRF that models interactions between attributes. We believe such methods are limited by their ability to accurately parse humans from real-world images since the problem of person detection and subsequent pose estimation is still largely unsolved [16].

There exists a vast amount of literature on learning aesthetically pleasing color contexts which can provide a great insight for a deeper understanding of this problem. In these papers [5, 7] the goal is to present the user with matches that are perceptually pleasing along with the query item. As a result of this research, the Matsuda templates have emerged as a popular choice for representing color basis functions. This allows for a straightforward retrieval mechanism by rotating the color wheel according to some preconceived rule. For instance, complementary color retrieval simply shifts the hue wheel by 180 degrees and subsequently retrieves nearest neighbors. In this paper, we propose to use these techniques as baselines and compare their performance against multiple data driven approaches.

4. PREDICTIVE MODELS FOR FASHION RECOMMENDATION

In order to set the context, imagine a shopper who has bought an apparel (say pants) and is not sure about what shirt would go well with the pant they have bought. It would be very useful is a trained model could leverage the scale of data and recommend relevant items from the online inventory.

We think about recommenders in two main ways: 1. Determinisitic Fashion Recommenders 2. Stochastic Fashion Recommenders

Since very limited research exists for investigating the full spectrum for solving the problem of visual recommendation, we propose the following set of algorithms to address this issue. This also allows to exploit the relative merits of each algorithm for designing a better solution. We split a data set of n images into n_{train} training and n_{test} testing images where each image has a set of parts, such as head, foot, torso, etc. Let us assume part indices to be represented by $p = \{1, 2, 3, \ldots, P\}$. For example, $p = 1, 2, P$ could correspond to head, torso and foot, respectively. Some of these parts are observable to the algorithm, while the rest are hidden. The aim of the algorithm is to infer features related to the hidden parts. On a similar note, let us assume visible part indices to be $p_v = \{1, 2, ..., |p_v|\}$ and hidden part indices to be $p_h = \{1, 2, ...|p_h|\}$. Note that $|p_v| + |p_h| = P$.

We assume each part j of image i to have an associated part descriptor $\underline{h}_{ij} \in \Re^K$. The feature $\underline{h}_{ij}$ can be color, texture, or any other visual descriptor such as bag of words (BoW). An i^{th} image can now be described by the concatenation of all part descriptors:

$$\underline{H}_i^T := [\underline{h}_{i1}^T, \underline{h}_{i2}^T, \ldots, \underline{h}_{iP}^T] \tag{1}$$

In case of a K-dimensional HSV histogram, K is the number of bins, while in case of bag of words, K is the size of code book.

4.1 Deterministic Fashion Recommenders (DFR)

Deterministic Fashion Recommender (DFR) aims to harness the power of data in recommending co-ordinating fashion items. For each co-ordinating clothing piece in our training data (e.g. *skirts*,*tops*), we extract color features in form of a K-dimensional HSV histogram [1]. This generates 40-dimensional feature vector for each clothing piece. Features from <*skirt*,*top*> tuple are then indexed in a database. During testing time, when a query *skirt* image is presented, a 40-dimensional query feature vector is computed and searched across the *skirt* portion of all indexed tuples. Finally, *top* portions of all nearest tuples are returned as the recommended *tops*. Since, the proposed model is based on a nearest-neighbor principle, the quality of recommendation depends on the size and quality of the data. With large and diverse data, DFR can provide better recommendations and can also be interpreted as being more objective with its fashion choices.

4.1.1 Tuned Perceptual Retrieval (PR)

This baseline uses the popular perceptually motivated Matsuda templates [5] and the transformation is either a complementary or triad pattern for the hue histogram. The saturation and value channels are simply reflected on the intuition that people prefer wearing contrasting styles.

4.1.2 Complementary Nearest Neighbor Consensus (CNNC)

Let us define a fixed metric on the space, say $dist$. The metric could be $||.||_1, ||.||_2$, KL-divergence, or Earth Mover's distance. We stick to $||.||_1$ for simplicity. The problem we are faced with is predicting the hidden P^{th} part. Denoting a single test query sample as q,

$$N_q = \underset{i:1 \leq i \leq n_{train}}{\text{argmin}_{\mathcal{K}}} \sum_{j \in p_v} dist(\underline{h}_{qj}, \underline{h}_{ij}) \tag{2}$$

The above equation simply picks those images that are similar to the input query image q in the visible parts. If a simple $argmin$ was defined, it would return the closest neighbor. However, we want the $\mathcal{K}$ closest neighbors and hence the notation $argmin_{\mathcal{K}}$.

Once nearest neighbors are picked, the goal is to infer the hidden parts. For this purpose, we accumulate representations for hidden part as $\{\underline{h}_{ij} | i \in N_q, j \in p_h\}$.

We infer the missing part as, $\underline{h}_{qj} = C_j(\underline{h}_{ij}); i \in N_q, j \in p_H$. Here $C_j : \Re^{|N_q \times K|} \rightarrow \Re^K$ is the consensus function for part j. We use the simplest average consensus function:

$$\{\underline{h}_{ij} | i \in N_q, j \in p_h\} \xmapsto{C_j(\underline{h}_{ij})} \underline{h}_{qj} = \frac{1}{N_q} \sum_{i \in N_q} \underline{h}_{ij} \tag{3}$$

Corollary: *KNN Consensus-Diversity* - In a generic shopping experience, a shopper would not like to be presented with the same type of clothing multiple times. As a result, a "diverse" retrieval is required. We propose the following generic optimization for generating diverse transformed queries, each of which can be used to query the inventory for similar images. The following optimization is proposed for the purpose:

$$\mathbf{I}_q^{diverse} = Div(\{\underline{h}_{ij} | i \in N_q, j \in p_H\}) \tag{4}$$

Div is an operator that returns a subset $\mathbf{I}_q^{diverse}$ of images from the set N_q that are as different as possible from one another on the hidden part features. In our case Div is a non-linear operator that clusters the data points and samples points, one from each cluster.

4.2 Stochastic Fashion Recommender (SFR)

Fashion choices can sometimes be highly subjective. For instance, a typical user may like polka dots so he or she may always prefer a match with polka dots over other matches. Such behavior cannot be modeled with DFR-like approaches. To address this issue, we propose Stochastic Fashion Recommender (SFR) that aims to explore the space of user biases in recommending matches. Computationally, the space of all such possible biases is huge and impractical to model. Hence, we constrain these biases to a smaller set of fashion choices. In our study, this constraint is achieved by the common notion that *solid* and *patterned* clothing co-ordinate well together. In other words, having busy patterns in both top and bottom clothing is less popular. To model this, we parameterize our desired output space (e.g. *solids*) with a probability distribution. This ensures that given a *patterned* query, all output choices from the model will be *solids*. Given this particular output space, user biases can be modeled by a distribution sampling process. In our case, we perform a uniform sampling of the distribution (i.e. randomly pick a solid color). Hence, not only does this model ensure "stochasticity" for the same input, it also relaxes computational complexity and allows us to closely mimic the human subjectivity in fashion.

4.2.1 Texture Agnostic Retrieval (TAR)

We propose the Texture Agnostic Retrieval technique, based on the notion of consumers preferring *solid* colors to go well with *patterned* clothing. In other words, having busy patterns in the top and bottom clothing seems counter-intuitive. The recommender can be stated as follows: $\underline{h}_{ij} \sim P(\underline{h}, \alpha)$ where α parameterizes the distribution which can be sampled efficiently. Since the bias has to be towards solid colors, we sample with a constraint that

$$\underline{h}_{ij} \sim P(\underline{h}, \alpha), \text{ s.t. } \underline{h}_{ij}^T \mathbf{1}_K = 1 \tag{5}$$

For our application of interest, we select $P(\underline{h}, \alpha) = U([0, 1])$, while we note any other distribution that can be efficiently sampled can be employed.

4.2.2 Mixture Models

Gaussian Mixture Models (GMM): An alternative way of viewing the above problem is using mixture models. Assuming each dimension of the space in which the Gaussian mixture model (GMM) is defined to correspond to part features, the GMM intuitively captures the part features that co-occur most frequently. At test time, the goal is to efficiently sample these high density regions using features from visible parts to recommend clothing on the hidden parts.

The learning module aims to learn the parameters of a mixture of multivariate Gaussian distribution parameterized by a random variable $\underline{H}$, and denoted by:

$$p(\underline{H}|\lambda) = \sum_{i=1}^{M} w_i g(\underline{H}\mu_i, \mathbf{\Sigma}_i) \tag{6}$$

In the above equation, $g(\underline{H}|\mu_i, \Sigma_i) = \mathcal{N}(\mu_i, \mathbf{\Sigma}_i), \lambda = (\underline{\mu}, \mathbf{\Sigma}, \underline{w})$ denotes the parameters of the Gaussian mixture distribution having M mixture components, $\underline{\mu} = [\underline{\mu}_1 ... \underline{\mu}_M], \mathbf{\Sigma} = [\mathbf{\Sigma}_1 ... \mathbf{\Sigma}_M]$ and $\underline{w} = [w_1 ... w_M]$ respectively denote the means, co-variance matrices and mixture weights of the M Gaussian mixture components.

Assume query vector $\underline{H}_q^T = [\underline{h}_{q1}^T, ..., \underline{h}_{qj}^T, ..., \underline{h}_{qP}^T]$, with a missing part m, the goal is to infer the most probable value $\underline{h}_{qm}$ of missing part m.

We cast it as the following problem:

$$\hat{\underline{h}}_{qm} = \underset{\underline{h}_{qm} : \underline{H}_q^T = [\underline{h}_{q1}^T, ..., \underline{h}_{qm}^T, ..., h_{qP}]^T}{\text{argmax}} p(\underline{H}|\lambda) \tag{7}$$

This is a constrained conditional maximization and will fix values of parts that have already been observed and will only search over unknown variables for a maximum likelihood assignment score. Since the mixture model defined above requires high dimensional density estimation, we first vector quantize the part feature space. Assuming that P parts share a common code book of size, a mixture model in P dimensions results. In other words, we utilize the code word to which a part feature is most closely associated with as an input to learn the mixture model.

Markov Chain - LDA (MCL): An approach for retrieval using inherent cluster structure of the data is to utilize topic models. In essence, what we require is a topic model where words have a specific structure. In other words, even though words are drawn

Table 3: Comparison of different models: + (Easy), o (Medium), x (Hard)

Model:	CNNC	GMM	TAR	MCL	PR
Learn Ease	+	+	+	x	+
Test Ease	x	x	+	+	+
Scalability	o	+	+	o	+
Generalization	+	o	o	+	x

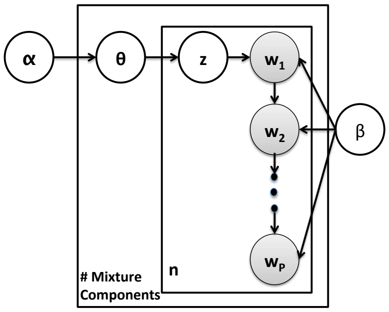


Figure 4: Markov Chain LDA model for learning fashion co-occurrences, with w_i as the word for part i.

from a code book, certain word combinations (say, color combinations) have a much higher probability of occurrence than others. We propose to model the structure in word combinations by a Markov Chain (Figure 4). It is useful to observe that a Markovian assumption may not reflect the true dependencies in fashion co-occurrence. However, the conditional independence properties yielded by a Markov Chain significantly reduces the computational needs for structured word generation.

The generative model for the original LDA model [2] for words w, topics θ and hyper parameters α and β is given by:

$$p(w|\alpha,\beta) = \int p(\theta|\alpha)\prod_{i=1}^{n}\sum_{z_i} p(z_i|\theta)p(w|z_i,\beta)d\theta \qquad (8)$$

We propose the following generalization of basic LDA to Markov chains on words, thus modifying the model to:

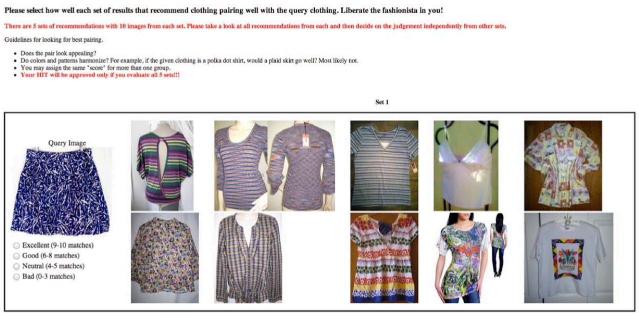
$$p(\underline{w}|\alpha,\beta) = \int p(\theta|\alpha)\prod_{i=1}^{n}\sum_{z_i} p(z_i|\theta)p(\underline{w}|z_i,\beta)d\theta \qquad (9)$$

where $\underline{w} = [w_1, w_2, \ldots, w_P]$. Inference over the joint distribution is computationally expensive. Hence we make the following simplifying assumption:

$$p(\underline{w}|\alpha,\beta) = \int p(\theta|\alpha)\prod_{i=1}^{n}\sum_{z_i} p(z_i|\theta)\cdot \underbrace{p(w_1)\prod_{j=2}^{P} p(w_j|w_{j-1},\beta,z_i)\,d\theta}_{p(\underline{w}|z_i,\beta)} \qquad (10)$$

This modified model is now learned offline using training data. When a new query comes in, say a shirt, the topic most likely to have this word is now picked and the Markov chain is sampled to generate a new transformed sample that can be utilized for retrieving complementary nearest neighbors.

The methods described till now are summarized based on ease of training, ease of testing, scalability, and generalization in Table 3. The requirements for a specific application would dictate the method of choice for recommendation. We also attempted adopting the collaborative filtering formulation of [11], but do not describe it in detail due to practical difficulties we encountered in scaling the algorithm to large data sets. We note that there is rich literature on scaling collaborative filtering approaches to large scale datasets, which we do not investigate in this work.



Please select how well each set of results that recommend clothing pairing well with the query clothing. Liberate the fashionista in you!

There are 5 sets of recommendations with 10 images from each set. Please take a look at all recommendations from each and then decide on the judgement independently from other sets.

Guidelines for looking for best pairing.

- Does the pair look appealing?
- Do colors and patterns harmonize? For example, if the given clothing is a polka dot shirt, would a plaid skirt go well? Most likely not.
- You may assign the same "score" for more than one group.
- **Your HIT will be approved only if you evaluate all 5 sets!!!**

Set 1

Query Image

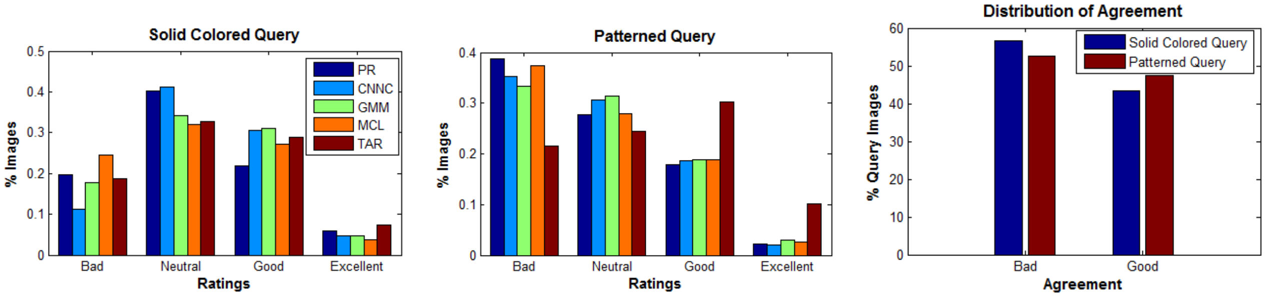
○ Excellent (9-10 matches)
○ Good (6-8 matches)
○ Neutral (4-5 matches)
○ Bad (0-3 matches)

Figure 5: Illustration of crowdsourcing evaluation interface

5. USER BEHAVIOR/INTERACTION WITH PREDICTIVE MODELS

Crowd-sourcing Fashion: Since the primary objective is to understand user preferences in fashion, we require user ratings on recommendations provided by DFR and SFR. We propose to use crowd-sourcing to obtain these ratings. This scheme also aligns well with our goal of understanding "Street Fashion", where a layman user's (i.e. fashionista on crowd sourcing platform) rating can be used as a ground-truth dataset. This validation procedure is in line with the established validation protocols for image retrieval [17]. Figure 5 outlines the proposed crowd-sourcing experiment. A user is presented with an input query image on left side (polka skirt) and the top-10 recommended matches (tops) on the right side. User rates these matches as one of -1 (bad match), 0 (neutral match), 1 (good match), 2 (excellent match). To ensure consistency in ratings, each query was evaluated by 5 different users leading to a total of 5000 users ratings. A total of 140 unique users participated in the experiment of which majority of users had a 100% approval ratings from their previous experiments. However, since fashion tastes vary greatly across people, it is usually hard to get a consensus on which algorithm performs the best from all 5 fashionistas. As a result, we only retain ratings from fashionistas who *agree* on algorithm performance. Since there are inadvertent errors by fashionistas, the total number of ratings that result after a preliminary filtering of ratings was 937, split across 187 solid queries and 750 patterned queries. We note here that since fashion is very subjective, a rigorous evaluation using principles from psychology is a very interesting line of future work.

Agreement Among Users: Understanding users is a logical first step in designing a user-driven fashion recommendation system. This can be done by gathering useful and interesting signals from user preferences that can be captured to replicate user choices in fashion. We study this effect by introducing the notion of disagreement between user preferences over visual recommendations. Our hypothesis is that presence of strong signals in user preferences will lead to lower disagreement among their preferences. Denoting each rating of 2 algorithms (DFR, SFR) for a query q using a 2-dimensional vector $\underline{\zeta}_{qi} \in \{-1, 0, 1, 2\}^2$, where each entry is a rating from -1 to +2, the score for disagreement across users is defined as, $\gamma_{qi} = \sum_{j=1}^{2} ||\underline{\zeta}_{qi} - \underline{\zeta}_{qj}||_1$. Further, an agreement threshold is defined to be the median of disagreement scores across all



(a)

(b)

(c)

Figure 6: Performance comparison of algorithms. CNNC performs the best for solid colored queries (20% of query set Fashion-Q1K) and TAR for patterned queries. (a) Ratings by fashionistas for solid colored queries. (b) Ratings for patterned queries. (c) Overall performance for the complete query set (includes both solid colors and patterns). The hybrid approach uses CNNC for solids and TAR for patterned queries, based on the decision of a solid vs pattern classifier.

queries, $A_T = median(\gamma_{qi}); q \in [1, 1000], i \in [1, 2]$. User ratings are retained only if $\gamma_{qi} < A_T$. This is done to filter outlier ratings from observations to account for spam in crowd-sourcing experiments. Finally, a query rating confidence is computed as $C_q = \frac{\text{Number of Users Retained for a Query}}{\text{Total Number of Users for a Query}}$. The final rating R_m for a model m is computed as:

$$R_m = \sum_{q=1}^{1000} C_q \frac{\sum_{i=1}^{5} \zeta_{qi}(m)\delta(\gamma_{qi} < A_T)}{\sum_{i=1}^{5} \delta(\gamma_{qi} < A_T)} \tag{11}$$

where $\delta(x)$ is the Dirac-delta function, $\zeta_{qi}(m)$ refers to the rating provided by fashionista i on query q for model m. As shown in Figure 7(a), majority of the users have lower disagreement score (below 0.5) suggesting that there is enough consensus among their fashion choices that can be investigated in more detail. We now take a closer look at all these ratings where users have sufficient agreement as determined empirically through a threshold over disagreement score. These ratings are further split across solid and patterned queries to further isolate the particular section of queries which have greater rating agreements. Results shown in figure 7(b) suggest that users agree more on patterned queries, evidenced by the higher magnitude of red bars on the good agreement bin. In the case of solids, the users agreed that the retrieval was either *Neutral* or *Good* (evidenced by the higher weights to the yellow and blue bars). However, on patterned queries, users overwhelmingly agreed that the retrievals were not favorable (evidenced by the higher weights on the red bar). This result has two major conclusions: (i) *There is a strong consensus among seemingly arbitrary fashion choices made by users* (ii) *It is possible to identify a specific category of visual recommendations (i.e. solid colored bottom and patterned top combination) where such consensus is stronger.*

5.1 Retrieval Experimentation

In the past sections, we presented methods to learn the missing part of a query image. For instance, given a descriptor for skirt, what is the recommended descriptor for blouse? Thus the query descriptor is essentially mapped to a new space (say, from skirt to blouse). This mapped query now serves as an input query to a content based image retrieval (CBIR) system, for retrieving images similar to the mapped query.

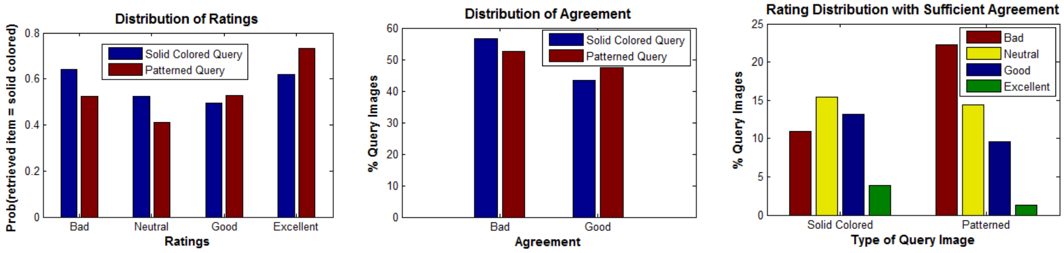
Training: The Fashion-136K is a data set created by crawling the web for photographs of fashion models. Hence, all images comprise top and bottom clothing co-occurring in the same image. Further, their spatial location (x, y) coordinates are also manually annotated by fashionistas. We use this data set for training all data-driven models described previously. While studying the data, we observed interesting co-occurrence relationships between top and bottom clothing such as dark gray/black bottoms going well with any colored top. It is such interesting patterns that the data driven models attempt to learn. In scenarios that require recommendations to be driven by context (Figure 1), it is possible to learn a separate co-occurrence matrix for every context.

Retrieval: Images of skirts (bottom clothing) from the Fashion-Q1K dataset are utilized as queries to retrieve top clothing from the Fashion-350K data set. Fashion-350K images are from a clothing inventory containing only top clothing (without model or mannequin). As a result, we utilize Fashion-136K where top and bottom clothing *co-occur* to learn models that are used while querying Fashion-350K. It is useful to note that this procedure trains the retrieval system on images on a data set (Fashion-136K) that is completely different from the test data (Fashion-350K). The result of retrieval is a ranked list of the Fashion-350K images sorted by relevance to a query from Fashion-Q1K. Since there are multiple SFR/DFR algorithms proposed in this work, the results of retrieval are multiple ranked lists where each list corresponds to a different algorithm's output. The running time and scalability of the system is dominated by the time taken for retrieval [1]. Efficient schemes for indexing and retrieving image features enable real time responses in a mobile setting.

As shown in results from Figure 6(a-c), data driven models outperform perceptual retrieval. We study the performance of algorithms on solid colored and patterned clothing separately to gain better insight into the workings of each algorithm. Further, it is useful to note that we had initially given the fashionistas an option to rate on a scale from Bad-Excellent (-1 to +2). Visual results of the various retrieval algorithms are presented in Figure 9.

Retrieval for Solid Colored Query: On solid clothing queries, it was observed that CNNC had the best normalized score of 0.418 on a scale from 0 to 1, followed by TAR 0.402, GMM 0.398, PR 0.383 and MCL 0.370. As seen in Figure 9, CNNC tends to retrieve patterned clothing. This is in fact favored by fashionistas in our study, for solid colored queries. The strength of adding diversity to a method like CNNC, as in Equation 4, is also shown qualitatively in Figure 8.

Retrieval for Patterned Query: On pattern clothing queries, TAR performed the best with a normalized score of 0.40 on a scale from 0 to 1, followed by GMM 0.31, CNNC 0.30, MCL 0.30 and PR 0.29. The impressive performance of TAR can be attributed to the general preference for solid colored clothing to go with another patterned clothing. Recall that by construction TAR is agnostic to texture and recommends only solid colors. Further, the simplicity



(a) (b) (c)

Figure 7: (a) A solid vs pattern classifier was used to estimate probability of solid colored clothing in the top retrievals. Fashionistas prefer patterned recommendations for solid colored queries and vice versa. (b) Illustration of rating agreement on solid colored and patterned queries. Fashionistas tend to agree more on retrievals of patterned queries, than they agree on solid queries. (c) This expands the second half of (b). Depicts common agreement amongst fashionstas based on retrieved results. Results for solid colored queries are generally more favorable than for patterned queries.

of TAR leads to a fairly robust query transformation. Finally, the query for solid colors by TAR leads to a lot more relevant retrievals from Fashion-350K, since it is much easier to match and retrieve solid clothing from an inventory in comparison to patterned clothing. See Figure 10 for examples of multiple query-retrieval pairs.

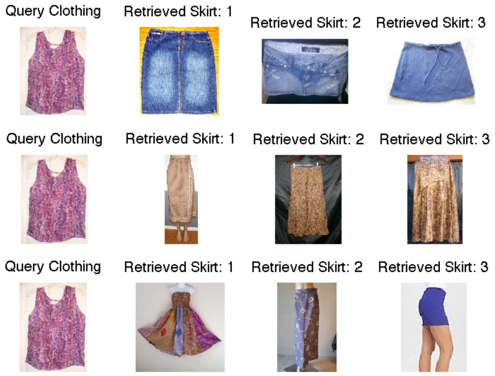


Figure 8: Illustration of diverse retrievals using CNNC, corresponding to a sample query. Each row corresponds to different set of results, while being relevant to the query.

The Hybrid Classifier: Based on our findings, we observe that CNNC performs the best for solid queries, while TAR performs best on pattern queries. Since a query could either be a solid or patterned, we propose using a single hybrid recommender that resorts to using CNNC for solid queries and TAR for patterned queries. The switching can be determined at the beginning of the retrieval procedure using a simple pattern classifier that labels a query as either solid or patterned. Experimental results on the overall query set (solid+patterns) indicate that the hybrid classifier which switches between CNNC and TAR yields the highest overall performance of 0.403, see Figure 6(c).

Ratings for Solid Colored Retrievals: The previous quantitative experiments seem to offer an insight that fashionistas prefer solid retrievals for patterned queries (evidenced by the impressive performance of TAR), which we validate in this experiment by exploring how fashionistas respond to solid colored retrievals. In other words, we run a binary classifier on the retrieval set to obtain the average probability of the retrieval set to be solid. When considering solid queries in Figure 7(a), we infer that fashionistas do not favor solid colored retrievals. This is evidenced by the higher magnitude of the blue bars on *Bad* and *Neutral* ratings. On the other hand, when considering patterned queries in Figure 7(a), we infer that fashionistas tend to favor solid colored retrievals. Observe the higher magnitude of the red bars on *Good* and *Excellent* ratings. This provides an intuitive justification behind the impressive performance of TAR on the pattern queries, since solids are retrieved by construction.

Distribution of Rating Agreement: Next, we study the rating agreement on solid and patterned queries. In other words, we measure how *agreeable* the recommendation algorithm's results are across solid and patterned queries, see Figure 7(b). It can be readily observed that fashionistas tend to agree more on patterned queries, evidenced by the higher magnitude of red bars on the good agreement bin. On the other hand, fashionistas tend to disagree more on the solid queries in comparison to patterns. It would be interesting to study whether rating agreement was on retrievals that were rated as *Good*, or on retrievals they thought were *Bad*. Figure 7(c) shows the split of among the ratings that fashionistas agreed on. In the case of solids, the fashionistas agreed that the retrieval was either *Neutral* or *Good* (evidenced by the higher weights to the yellow and blue bars). However, on patterned queries, the fashionistas overwhelmingly agreed that the retrievals were not favorable (evidenced by the higher weights on the red bar). This finding indicates that the performance of the algorithms on pattern queries can be further enhanced. A major reason for this result is the fact that the proposed system is purely color based, and integration of novel texture features devoid of drawbacks mentioned in previous sections could possibly yield a significant performance boost. This is a promising avenue for future research.

Insights into Street Fashion: Our experimental analysis leads to the following insights on street fashion, (i) Fashion Cues: Analyzing fashionista ratings for different algorithms suggest an intuitive fashion cue that a pair of patterned and solid colored clothing is more perceptually pleasing than other combinations. It also suggests that even among matches for patterned clothing, color is more important than the type of texture (i.e. plaids, polka dots), which also underscores the importance of simple yet visually strong descriptors such as color, (ii) Among all the patterns, we observed that fashionistas agree the most on recommending matches for paisley and stripes. However, fashionistas have more well-defined preferences for stripes as compared to paisley. Our initial analysis suggests this behavior can be attributed to the structure of pattern: stripes have strong structural information as compared to paisley. Hence, it is easier to recommend matches for stripes than paisley.

5.2 Understanding Users

Insights From Visual Recommendations: We observe from the previous section that CNNC gives superior performance among the DFR class of methods, and TAR performs best among the SFR

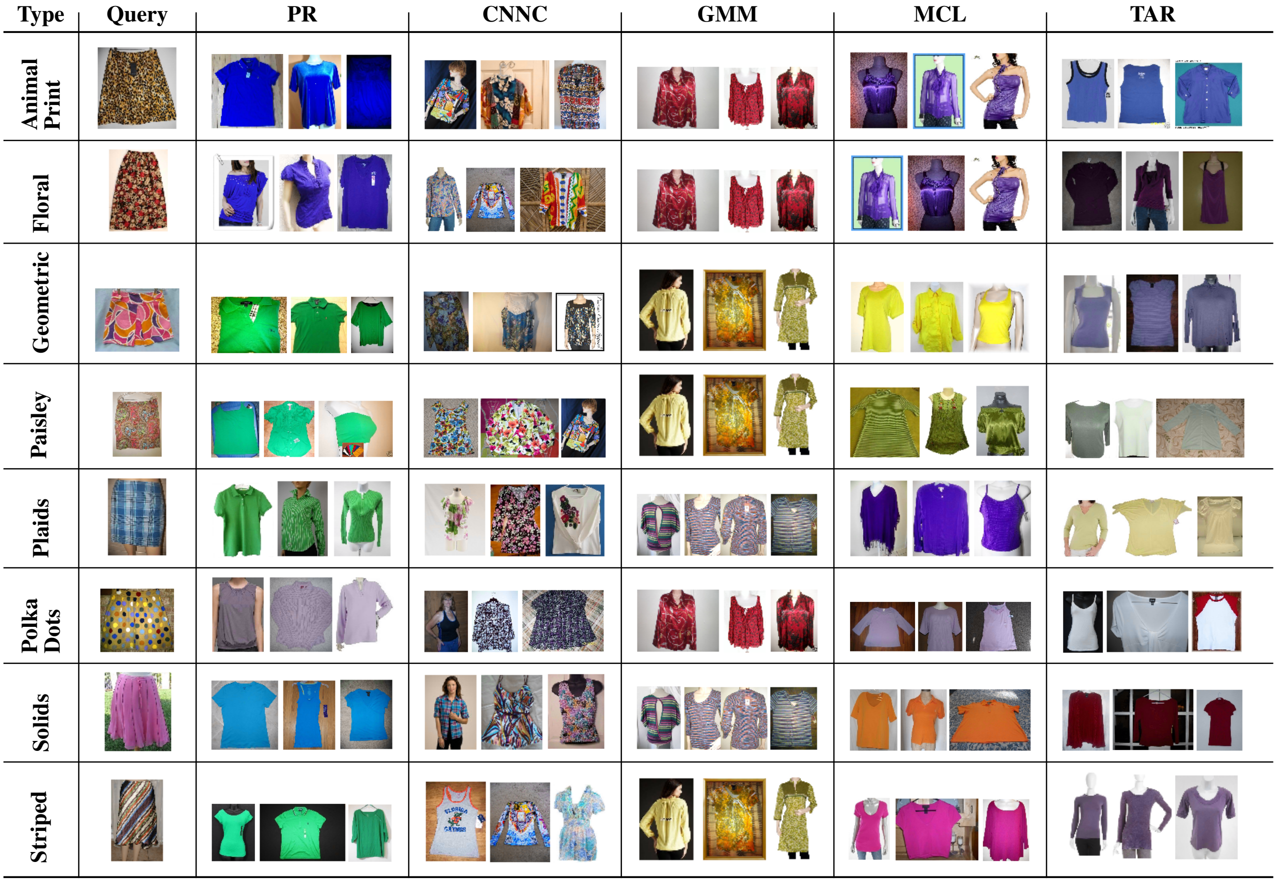


Figure 9: Top 3 retrieved items from “tops", recommended by each algorithm for query “skirts". Rows correspond to a query from a given pattern. Recommendations by TAR are more preferred than those by PR and MCL. Recommendations by CNNC and GMM have more patterns, even for patterned queries. Note that code words and modes in GMM yield identical retrievals for multiple queries.

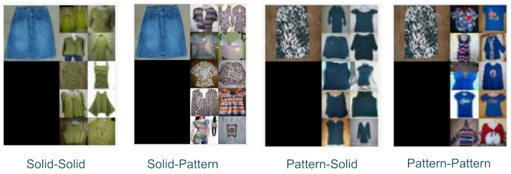


Figure 10: Illustration of the different combinations of retreivals, namely Solid-Solid, Solid-Pattern, Pattern-Solid and Pattern-Pattern

class of methods. Henceforth, comparisons of DFR/SFR methods utilize CNNC/TAR as exemplar techniques. The previous results provide an interesting cue that deeper insights can be revealed by segmenting fashion choices into classes such as solids or patterned queries. Our goal is to further analyze this behavior to ascertain its possible causes. One possible explanation can be found in similar results studied in computational neuroscience and vision science. These results directly correlate the complexity of an image to the visual processing time as well as image aesthetics. Motivated by this intuition, we look at the amount of time it takes to rate queries corresponding to different pattern. We hypothesize that since each pattern has a varying level of visual complexity, it should also reflect in the fashion preferences of the users. Figure 11(b) illustrates the cumulative histogram of time taken to annotate queries from

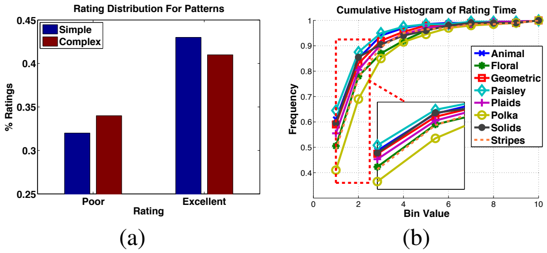
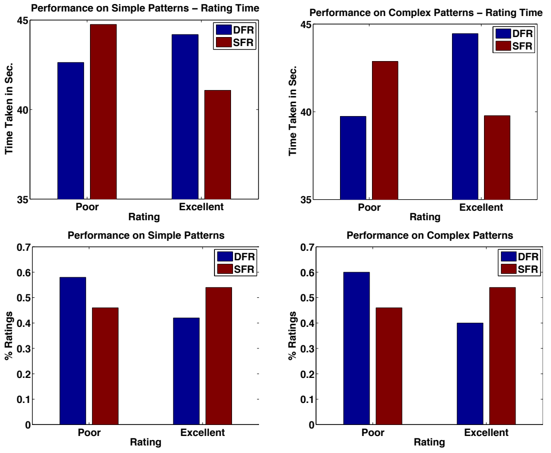


Figure 11: (a) Rating Distribution. Simple patterns are rated higher than complex patterns. (b) Cumulative Histogram of Rating Time for Each Pattern. Simple Patterns take more time due to more relevant matches as compared to complex patterns.

each pattern class. As shown, two patterns - Polka and Paisley occupy the ends of the spectrum. Annotations from Polka take more time, whereas it's relatively faster to annotate Paisley queries. This behavior can also be understood in terms of quality of matches for each pattern. Our qualitative results suggest that in general Polka queries tend to return more relevant matches as compared to Paisley which has mostly obvious incorrect matches. A typical user takes more time to carefully rate higher number of relevant matches in Polka and hence higher annotation time as compared to Paisley. Further, it can also be concluded that quality of matches are a func-



(a) (b)

Figure 12: (a) Rating Time and Model Performance on Simple Patterns. Matches from SFR are visually more appealing and hence take much less time for rating as compared to DFR. (b) Rating Time and Model Performance on Complex Patterns. SFR matches for complex patterns are more appealing as compared to simple patterns.

tion of pattern complexity - Polka is a simple pattern that Paisley, hence it returns better matches. This leads us to generate a weak categorization of all 8 patterns into 2 categories - Simple Patterns (Polka, Solids, Stripes, Plaids) and Complex Patterns (Animal, Floral, Geometric, Paisley) inspired by simple and complex cells in visual cortex. Using this scheme, we further analyze user ratings on these two categories. Figure 12(a) demonstrates that a larger proportion of users provide excellent ratings for matches corresponding to simple patterns. This result has the following conclusion: (i) *Fashion preferences tend have a strong correlation to associated complexities in visual perception.*

Enhancing Recommendations With User Understanding: All the above results provide a great insight into the role of users in fashion recommendation. However, they still do not answer if computer models can be built to replicate this understanding. To demonstrate this aspect, we evaluate two proposed models, DFR and SFR over the simple and complex pattern classes. Specifically, we look at two primary aspects, *time to rate recommendations* and *rating of recommendation* to illustrate this result. As shown in the top row of Figure 12(a&b), for all excellent ratings, users take much less time to rate recommendations from SFR as opposed to DFR. This is a strong indicator of the fact that visual recommendations from SFR have lower visual complexity and hence faster processing time for users. Moreover, the gain in time between SFR and DFR is even more pronounced for complex pattern classes. This again strengthens the argument that in the case of complex classes (when the visual complexity of the images are higher), more elastic modeling techniques such as SFR provide better aesthetically pleasing visual recommendations as opposed to pure data-driven techniques like DFR. Hence, this result illustrates that users prefer visual recommendations from SFR over DFR models. Next, we look at the distribution of user ratings for both these models over different fashion classes. The bottom row in Figure 12(a),(b) illustrates the performance of both the models. As shown, SFR outperforms DFR as it gets a much higher ratio of excellent user ratings and lower ratio of poor ratings for all the pattern classes. This result has following conclusions: (i)*Incorporating user preference by incorporating stochasticity in recommendation model makes the results visually more pleasing and relevant.*(ii) *Stochastic models are able to generalize well to different fashion segments as recommendations from both simple and complex patterns improve from it.*

6. CONCLUSION

In summary, this work approached the problem of visual recommendation by learning from street fashion data and transferring learnt knowledge to e-Commerce data. Two classes of recommenders, namely Deterministic (DFR) and Stochastic (SFR) recommenders were proposed. Initially, a thorough benchmarking of retrieval performance was presented. The results observed motivated the creation of a hybrid classifier which combines the best of SFR and DFR. Finally, the interaction of users with the recommender in terms of rating scores, user agreement, and query complexity were presented. Future work includes utilization of style based visual descriptors to enhance the existing color based descriptors. On the data curation side, an interesting line of investigation would be in collecting images at different times of the year to capture seasonal trends. Further, extension of recommendations from top-bottom clothing to multiple fashion apparel is an interesting avenue for future investigation.

7. REFERENCES

[1] A. Bhardwaj, A. Das Sarma, W. Di, R. Hamid, R. Piramuthu, and N. Sundaresan. Palette power: Enabling visual search through colors. In *KDD*, 2013.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[3] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification via plsa. In *ECCV*, pages 517–530, 2006.

[4] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, pages 609–623, 2012.

[5] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y. Xu. Color harmonization. In *SIGGRAPH*, 2006.

[6] T. Iwata, S. Wanatabe, and H. Sawada. Fashion coordinates recommender system using photographs from fashion magazines. In *IJCAI*, pages 2262–2267, 2011.

[7] C. Li and T. Chen. Aesthetic visual quality assessment of paintings. *Selected Topics in Signal Processing, IEEE Journal of*, 3(2):236–252, 2009.

[8] S. Liu, J. Fen, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan. Hi, magic closet, tell me what to wear! In *ACMMM*, 2012.

[9] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, 2012.

[10] W.-Y. Ma and B. S. Manjunath. Texture features and learning similarity. In *CVPR*, pages 425–430, 1996.

[11] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *NIPS*, 2008.

[12] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *IJCV*, 62(1), 2005.

[13] M. Weber and M. Bauml. Part-based clothing segmentation for person retrieval. In *AVSS*, pages 361–366. IEEE, 2011.

[14] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012.

[15] M. Yang and K. Yu. Real-time clothing recognition in surveillance videos. In *ICIP*. IEEE, 2011.

[16] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392, 2011.

[17] B. Yao, A. Khosla, and L. Fei-Fei. Classifying actions and measuring action similarity. *ICML*, 2011.