Neuroaesthetics in Fashion: Modeling the Perception of Fashionability

Edgar Simo-Serra[1], Sanja Fidler[2], Francesc Moreno-Noguer[1], Raquel Urtasun[2]

[1]Institut de Robòtica i Informàtica Industrial (CSIC-UPC), [2]University of Toronto

Abstract

In this paper, we analyze the fashion of clothing of a large social website. Our goal is to learn and predict how fashionable a person looks on a photograph and suggest subtle improvements the user could make to improve her/his appeal. We propose a Conditional Random Field model that jointly reasons about several fashionability factors such as the type of outfit and garments the user is wearing, the type of the user, the photograph's setting (e.g., the scenery behind the user), and the fashionability score. Importantly, our model is able to give rich feedback back to the user, conveying which garments or even scenery she/he should change in order to improve fashionability. We demonstrate that our joint approach significantly outperforms a variety of intelligent baselines. We additionally collected a novel heterogeneous dataset with 144,169 user posts containing diverse image, textual and meta information which can be exploited for our task. We also provide a detailed analysis of the data, showing different outfit trends and fashionability scores across the globe and across a span of 6 years.

1. Introduction

"The finest clothing made is a person's skin, but, of course, society demands something more than this."

Mark Twain

Fashion has a tremendous impact on our society. Clothing typically reflects the person's social status and thus puts pressure on how to dress to fit a particular occasion. Its importance becomes even more pronounced due to online social sites like Facebook and Instagram where one's photographs are shared with the world. We also live in a technological era where a significant portion of the population looks for their dream partner on online dating sites. People want to look good; business or casual, elegant or sporty, sexy but not slutty, and of course trendy, particularly so when putting their picture online. This is reflected in the growing online retail sales, reaching 370 billion dollars in the US by 2017, and 191 billion euros in Europe [19].

Computer vision researchers have started to be interested in the subject due to the high impact of the application domain [1, 2, 3, 6, 8, 11, 18, 29]. The main focus has been to infer clothing from photographs. This can enable a variety of applications such as virtual garments in online shopping. Being able to automatically parse clothing is also key in order to conduct large-scale sociological studies related to family income or urban groups [20, 26].

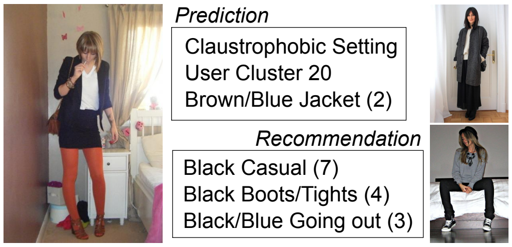
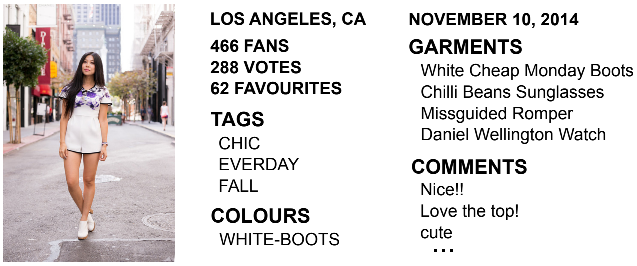


Figure 1: Example of recommendations provided by our model for the post on the left. In this case the user is wearing what we have identified as "Brown/Blue Jacket". This photograph obtains a score of 2 out of 10 in fashionability. Additionally the user is classified as belonging to cluster 20 and took a picture in the "Claustrophobic" setting. If the user were to wear a "Black Casual" outfit as seen on the right, our model predicts she would improve her fashionability to 7 out of 10. This prediction is conditioned on the user, setting and other factors allowing the recommendations to be tailored to each particular user.

In this paper, our goal is to predict how fashionable a person looks on a particular photograph. The fashionability is affected by the garments the subject is wearing, but also by a large number of other factors such as how appealing the scene behind the person is, how the image was taken, how visually appealing the person is, her/his age, etc. The garment itself being fashionable is also not a perfect indicator of someone's fashionability as people typically also judge how well the garments align with someone's "look", body characteristics, or even personality.

Our aim here is to give a rich feedback to the user: not only whether the photograph is appealing or not, but also to make suggestions of what clothing or even the scenery the user could change in order to improve her/his look, as illustrated in Fig. 1. We parametrize the problem with a Conditional Random Field that jointly reasons about several important fashionability factors: the type of outfit and garments, the type of user, the setting/scenery of the pho-



LOS ANGELES, CA
466 FANS
288 VOTES
62 FAVOURITES

TAGS
CHIC
EVERDAY
FALL

COLOURS
WHITE-BOOTS

NOVEMBER 10, 2014

GARMENTS
White Cheap Monday Boots
Chilli Beans Sunglasses
Missguided Romper
Daniel Wellington Watch

COMMENTS
Nice!!
Love the top!
cute
...

Figure 2: Anatomy of a post from the Fashion144k dataset. It consists always of at least a single image with additional metadata that can take the form of tags, list of nouns and adjectives, discrete values or arbitrary text.

tograph, and fashionability of the user's photograph. Our model exploits several domain-inspired features, such as beauty, age and mood inferred from the image, the scene type of the photograph, and if available, meta-data in the form of where the user is from, how many online followers she/he has, the sentiment of comments by other users, etc.

Since no dataset with such data exists, we created our own from online resources. We collected 144,169 posts from the largest fashion website chictopia.com to create our *Fashion144k* dataset[1]. In a post, a user publishes a photograph of her/himself wearing a new outfit, typically with a visually appealing scenery behind the user. Each post also contains text in the form of descriptions and garment tags, as well as other users' comments. It also contains votes or "likes" which we use as a proxy for fashionability. We refer the reader to Fig. 2 for an illustration of a post.

As another contribution, we provide a detailed analysis of the data, in terms of fashionability scores across the world and the types of outfits people in different parts of the world wear. We also analyze outfit trends through the last six years of posts spanned by our dataset. Such analysis is important for the users, as they can adapt to the trends in "real-time" as well as to the fashion industry which can adapt their new designs based on the popularity of garments types in different social and age groups.

2. Related Work

Fashion has a high impact on our everyday lives. This also shows in the growing interest in clothing-related applications in the vision community. Early work focused on manually building composite clothing models to match to images [4]. In [11, 23, 31, 32, 33], the main focus was on clothing parsing in terms of a diverse set of garment types. Most of these works follow frameworks for generic segmentation [27, 34] with additional pose-informed potentials. They showed that clothing segmentation is a very challenging problem with the state-of-the-art capping at 12% intersection over union [23].

Property	Total	Per User	Per Post
posts	144169	10.09 (30.48)	-
users	14287	-	-
locations	3443	-	-
males	5%	-	-
fans	-	116.80 (1309.29)	1226.60 (3769.97)
comments	-	14.15 (15.43)	20.09 (27.51)
votes	-	106.08 (108.34)	150.76 (129.78)
favourites	-	18.49 (22.04)	27.01 (27.81)
photos	277537	1.73 (1.00)	1.93 (1.24)
tags	13192	3.43 (0.75)	3.66 (1.12)
colours	3337	2.06 (1.82)	2.28 (2.06)
garments	-	3.14 (1.57)	3.22 (1.72)

Table 1: Statistics of the dataset.

More related to our line of work are recent applications such as learning semantic clothing attributes [3], identifying people based on their outfits, predicting occupation [26] and urban tribes [20], outfit similarity [28], outfit recommendations [17], and predicting outfit styles [16]. Most of these approaches address very specific problems with fully annotated data. In contrast, the model we propose is more general, allowing to reason about several properties of one's photo: the aesthetics of clothing, the scenery, the type of clothing the person is wearing, and the overall fashionability of the photograph. We do not require any annotated data, as all necessary information is extracted by automatically mining a social website.

Our work is also related to the recent approaches that aim at modeling the human perception of beauty. In [5, 7, 10, 15] the authors addressed the question of what makes an image memorable, interesting or popular. This line of work mines large image datasets in order to correlate visual cues to popularity scores (defined as e.g., the number of times a Flickr image is viewed), or "interestingness" scores acquired from physiological studies. In our work, we tackle the problem of predicting fashionability. We also go a step further from previous work by also identifying the high-level semantic properties that cause a particular aesthetics score, which can then be communicated back to the user to improve her/his look. The closest to our work is [14] which is able to infer whether a face is memorable or not, and modify it such that it becomes. The approach is however very different from ours, both in the domain and in formulation.

3. Fashion144k Dataset

We collected a novel dataset that consists of 144,169 user posts from a clothing-oriented social website chictopia.com. In a post, a user publishes one to six photographs of her/himself wearing a new outfit. Generally each photograph shows a different angle of the user or zooms in on different garments. Users sometimes also add a description of the outfit, and/or tags of the types and colors of the garments

[1] http://www.iri.upc.edu/people/esimo/research/fashionability/

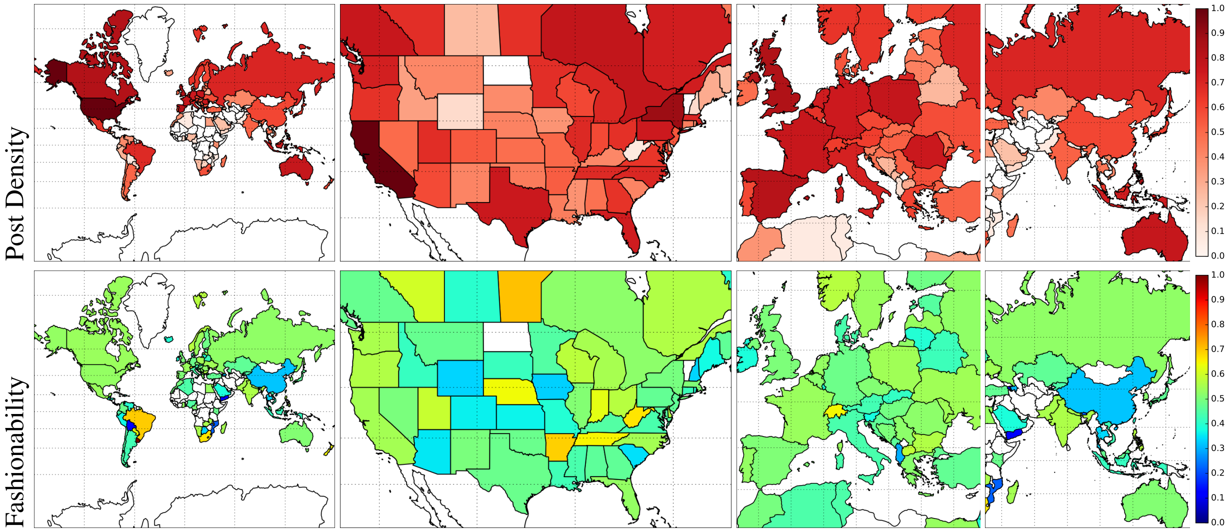


Figure 3: Visualization of the density of posts and fashionability by country.

		Compatriot	Mean Sentiment Score	
Country	Posts	Comments	Compatriots	Total
United States	28.0%	14.86%	3.78	3.76
Unknown	21.8%	-	-	-
United Kingdom	5.1%	2.67%	3.80	3.75
Philippines	5.1%	14.54%	3.61	3.72
Canada	4.5%	2.95%	3.68	3.76
Spain	3.9%	1.52%	3.06	3.75
Poland	2.5%	1.07%	3.63	3.80
Australia	2.4%	1.76%	3.62	3.75
France	2.3%	0.46%	3.23	3.75
Romania	2.0%	6.83%	3.73	3.77

Table 2: Number of posts by country and the percentage of comments that come from users of the same country. We also show the mean score on a scale of 1 to -5 of the sentiment analysis [25] for both compatriots and all commentors.

City Name	Posts	Fashionability
Manila	4269	6.627
Los Angeles	8275	6.265
Melbourne	1092	6.176
Montreal	1129	6.144
Paris	2118	6.070
Amsterdam	1111	6.059
Barcelona	1292	5.845
Toronto	1471	5.765
Bucharest	1385	5.667
New York	4984	5.514
London	3655	5.444
San Francisco	2880	5.392
Madrid	1747	5.371
Vancouver	1468	5.266
Jakarta	1156	4.398

Table 3: Fashionability of cities with at least 1000 posts.

they are wearing. Not all users make this information available, and even if they do, the tags are usually not complete, i.e. not all garments are tagged. Users typically also reveal their geographic location, which, according to our analysis, is an important factor on how fashionability is being perceived by the visitors of the post. Other users can then view these posts, leave comments and suggestions, give a "like" vote, tag the post as a "favorite", or become a "follower" of the user. There are no "dislike" votes or "number of views" making the data challenging to work with from the learning perspective. An example of a post can be seen in Fig. 2.

We parsed all information for each post to create Fashion144k. The oldest entry in our dataset dates to March 2nd in 2008, the first post to the chictopia website. The last crawled post is May 22nd 2014. We refer the reader to Table 1 for detailed statistics of the dataset. We can see a large diversity in meta-data. Perhaps expected, the website is dominated by female users (only 5% are male). We also inspect dataset biases such as users voting for posts from the users of the same country of origin. Since there is no information of who gave a "like" to a post, we analyze the origin of the users posting comments on their compatriot's posts in Table 2. From this we can see that users from the Philippines seem to be forming a tight-knit community, but this does not seem to bias the sentiment scores.

Measuring Fashionability of a Post. Whether a person on a photograph is truly fashionable is probably best decided by fashion experts. It is also to some extent a matter of personal taste, and probably even depends on the nationality and the gender of the viewer. Here we opt for leveraging the taste of the public as a proxy for fashionability. In particular, we base our measure of interest on each post's number of votes, analogous to "likes" on other websites. The main issue with votes is the strong correlation with the time when the post was published. Since the number of users fluctuate, so does the number of votes. Furthermore, in the first months or a year since the website was created, the number of users (voters) was significantly lower than in

Feature	Dim.	Description
Fans	1	Number of user's fans.
ΔT	1	Time between post creation and download.
Comments	5	Sentiment analysis [25] of comments.
Location	266	Distance from location clusters [24].
Personal	21	Face recognition attributes.
Style	20	Style of the photography [13].
Scene	397	Output of scene classifier trained on [30].
Tags	209	Bag-of-words with post tags.
Colours	604	Bag-of-words with colour tags.
Singles	121	Bag-of-words with split colour tags.
Garments	1352	Bag-of-words with garment tags.

Table 4: Overview of the different features used.

the recent years.

As the number of votes follows a power-law distribution, we use the logarithm for a more robust measure. We additionally try to eliminate the temporal dependency by calculating histograms of the votes for each month, and fit a Gaussian distribution to it. We then bin the distribution such that the expected number of posts for each bin is the same. By doing this we are able to eliminate almost all time dependency and obtain a quasi-equal distribution of classes, which we use as our fashionability measure, ranging from 1 (not fashionable) to 10 (very fashionable). Fig. 3 shows the number of posts and fashionability scores mapped to the globe via the user's geographic information. Table 3 reveals some of the most trendy cities in the world, according to chictopia users and our measure.

4. Discovering Fashion from Weak Data

Our objective is not only to be able to predict fashionability of a given post, but we want to create a model that can understand fashion at a higher level. For this purpose we make use of a Conditional Random Field (CRF) to learn the different outfits, types of people and settings. Settings can be interpreted as where the post is located, both at a scenic and geographic level. Our potentials make use of deep networks over a wide variety of features exploiting Fashion144k images and meta-data to produce accurate predictions of how fashionable a post is.

More formally, let $u \in \{1, \cdots, N_U\}$ be a random variable capturing the type of user, $o \in \{1, \cdots, N_O\}$ the type of outfit, and $s \in \{1, \cdots, N_S\}$ the setting. Further, we denote $f \in \{1, \cdots, 10\}$ as the fashionability of a post $\mathbf{x}$. We represent the energy of the CRF as a sum of energies encoding unaries for each variable as well as non-parametric pairwise potentials which reflect the correlations between the different random variables. We thus define

$$\begin{aligned} E(u,o,s,f) &= E_{user}(u) + E_{out}(o) + E_{set}(s) + E_{fash}(f) \\ &+ E_{np}^{uf}(u,f) + E_{np}^{of}(o,f) + E_{np}^{sf}(s,f) \\ &+ E_{np}^{uo}(u,o) + E_{np}^{so}(s,o) + E_{np}^{us}(u,s) \qquad (1) \end{aligned}$$

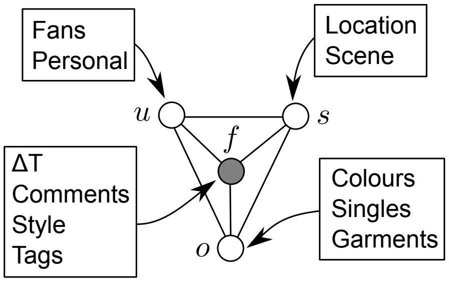


Figure 4: An overview of the CRF model and the features used by each of the nodes.

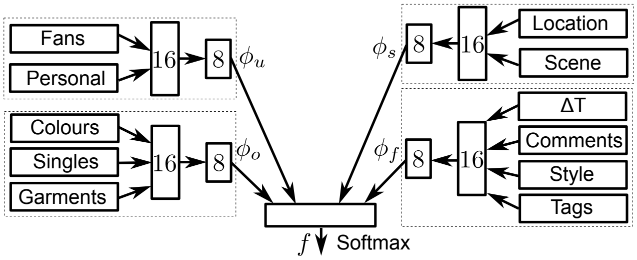


Figure 5: Illustration of the type of deep network architecture to learn features. We can see that it consists of four network joined together by a softmax layer. The output of the different networks ϕ_f, ϕ_o, ϕ_u, and ϕ_s are then used as features for the CRF.

We refer the reader to Fig. 4 for an illustration of the graphical model. We now define the potentials in more detail.

User: We compute user specific features encoding the logarithm of the number of fans that the particular user has as well as the output of a pre-trained neural network-based face detector enhanced to predict additional face-related attributes. In particular, we use rekognition[2] which computes attributes such as ethnicity, emotions, age, beauty, etc. We run this detector on all the images of each post and only keep the features for the image with the highest score. We then compute our unary potentials as the output of a small neural network with two hidden layers that takes as input the user's high dimensional features and produces an 8D feature map $\phi_u(x)$. We refer the reader to Fig. 5 for an illustration. Our user unary potentials are then defined as

$$E_{user}(u = i, \mathbf{x}) = \mathbf{w}_{u,i}^T \phi_u(\mathbf{x})$$

with $\mathbf{x}$ all the information included in the post. Note that we share the features and learn a different weight for each user latent state.

Outfit: We use a bag-of-words approach on the "garments" and "colours" meta-data provided in each post. Our dictionary is composed of all words that appear at least

[2] https://rekognition.com

50 times in the training set. This results in 1352 and 604 words respectively and thus our representation is very sparse. Additionally we split the colour from the garment in the "colours" feature, e.g., red-dress becomes red and dress, and also perform bag-of-words on this new feature. We then compute our unary potentials as the output of a small neural network with two hidden layers that takes as input the outfit high dimensional features and produces an 8D feature map $\phi_o(\mathbf{x})$. We refer the reader to Fig. 5 for an illustration. Our outfit unary potentials are then defined as

$$E_{out}(o = i, \mathbf{x}) = \mathbf{w}_{o,i}^T \phi_o(\mathbf{x})$$

with $\mathbf{x}$ all the information included in the post. Note that as with the users we share the features and learn a different weight for each outfit latent state.

Setting: We try to capture the setting of each post by using both a pre-trained scene classifier and the user-provided location. For the scene classifier we have trained a multilayer perceptron with a single 1024 unit hidden layer and softmax layer on the SUN Dataset [30]. We randomly use 70% of the 130,519 images as the training set, 10% as the validation set and 20% as the test set. We use the Caffe pre-trained network [12] to obtain features for each image which we then use to learn to identify each of the 397 classes in the dataset, corresponding to scenes such as "art_studio", "vineyard" or "ski_slope". The output of the 397D softmax layer is used as a feature along with the location. As the location is written in plain text, we first look up the latitude and longitude. We project all these values on the unit sphere and add some small Gaussian noise to account for the fact that many users will write more generic locations such as "Los Angeles" instead of the real address. We then perform unsupervised clustering using geodesic distances [24] and use the geodesic distance from each cluster center as a feature. We finally compute our unary potentials as the output of a small neural network with two hidden layers that takes as input the settings high dimensional features and produce an 8D feature map $\phi_s(\mathbf{x})$. Our outfit unary potentials are then defined as

$$E_{set}(s = i, \mathbf{x}) = \mathbf{w}_{s,i}^T \phi_s(\mathbf{x})$$

with $\mathbf{x}$ all the information included in the post. Note that as with the users and outfits we share the features and learn a different weight for each settings latent state.

Fashion: We use the time between the creation of the post and when the post was crawled as a feature, as well as bag-of-words on the "tags". To incorporate the reviews, we parse the comments with the sentiment-analysis model of [25]. This model attempts to predict how positive a review is on a 1-5 scale (1 is extremely negative, 5 is extremely positive). We used a pre-trained model that was trained on the rotten tomatoes dataset. We run the model on all the comments and sum the scores for each post. We also extract features using the style classifier proposed in [13] that is pre-trained on the Flickr80k dataset to detect 20 different image styles such as "Noir", "Sunny", "Macro" or "Minimal". This captures the fact that a good photography style is correlated with the fashionability score. We then compute our unary potentials as the output of a small neural network with two hidden layers that takes as input the settings high dimensional features and produce an 8D feature map $\phi_f(\mathbf{x})$. Our outfit unary potentials are then defined as

$$E_{fash}(f = i, \mathbf{x}) = \mathbf{w}_{f,i}^T \phi_f(\mathbf{x})$$

Once more, we shared the features and learn separate weights for each fashionability score.

Correlations: We use a non-parametric function for each pairwise and let the CRF learn the correlations. Thus

$$E_{np}^{uf}(u = i, f = j) = w_{i,j}^{uf}$$

Similarly for the other pairwise potentials.

4.1. Learning and Inference

We learn our model using a two step approach: we first jointly train the deep networks that are used for feature extraction to predict fashionability as shown in Fig 5, and estimate the initial latent states using clustering. Our network uses rectified linear units and is learnt by minimizing cross-entropy. We then learn the CRF model (2430 weights) using the primal-dual method of [9]. We use the implementation of [22]. As task loss we use the L1 norm for fashionability, and encourage the latent states to match the initial clustering. We perform inference using message passing [21].

5. Experimental Evaluation

We perform a detailed quantitative evaluation on the 10-class fashionability prediction task. We also provide a qualitative evaluation on other high level tasks such as visualizing changes in trends and outfit recommendations.

5.1. Correlations

We first analyze the correlation between fashionability and various features in our model. We consider the effect of the country on fashionability: in particular, we look the effect of economy, income class, Gross Domestic Product (GDP) and population. Results are in Table 5-left. A strong relationship is clear: poorer countries score lower in fashionability than the richer, sadly a not very surprising result.

We also show face-related correlations in Table 5-right. Interestingly, but not surprising, younger and more beautiful users are considered more fashionable. Additionally,

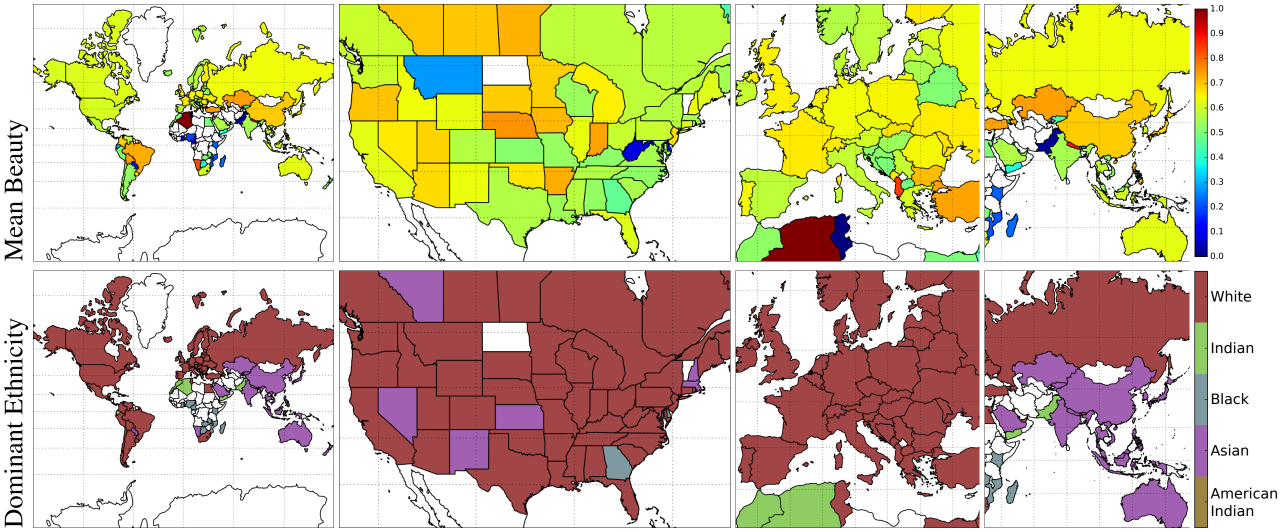


Figure 6: Visualization of mean beauty and dominant ethnicity by country.

Attribute	Corr.
Economy class	-0.137
Income class	-0.111
log(GDP)	0.258
log(Population)	0.231

Attribute	Corr.
age	-0.025
beauty	0.066
eye_closed	0.022
gender	-0.037
smile	-0.023
asian	0.024
calm	0.023
happy	-0.024
sad	0.023

Table 5: Effect of various attributes on the fashionability. Economy and Income class refer to a 1-7 scale in which 1 corresponds to most developed or rich country while 7 refers to least developed or poor country. For the face recognition features we only show those with absolute values above 0.02. In all cases we show the Pearson Coefficients.

we show the mean estimated beauty and dominant inferred ethnicity on the world map in Fig. 6. Brazil dominates the Americas in beauty, France dominates Spain, and Turkey dominates in Europe. In Asia, Kazakhstan scores highest, followed by China. There are also some high peaks which may be due to a very low number of posts in a country. The ethnicity classifier also seems to work pretty well, as generally the estimation matches the ethnicity of the country.

5.2. Predicting Fashionability

We use 60% of the dataset as a train set, 10% as a validation, and 30% as test, and evaluate our model for the fashionability prediction task. Results of various model instantiations are reported in Table 6. While the deep net obtains slightly better results than our CRF, the model we propose is very useful as it simultaneously identifies the type of user, setting and outfit of each post. Additionally, as we show later, the CRF model allows performing much more flexible tasks such as outfit recommendation or visualization of trends. Since the classification metrics such as accuracy, precision, recall, and intersection over union (IOU) do not capture the relationship between the different fashionability levels, we also report the L1 norm between the ground truth and the predicted label. In this case both the CRF and the deep net obtain virtually the same performance.

Model	Acc.	Pre.	Rec.	IOU	L1
CRF	29.27	30.42	28.69	17.36	1.46
Deep Net	30.42	31.11	30.26	18.41	1.45
No Metadata	19.63	17.06	17.47	8.31	2.31
Log. Reg.	23.92	22.54	22.99	12.55	1.91
Baseline	16.28	-	10.00	1.63	2.32
Random	9.69	9.69	9.69	4.99	3.17

Table 6: We show results for classification for random, a baseline that predicts only the dominant class, a standard logistic regression on our features, a deep network without data-specific metadata (comments, fans, and time offset), the deep network used to obtain features for the CRF and the final CRF model. We show accuracy, precision, recall, intersection over union (IOU), and L1 norm as different metrics for performance.

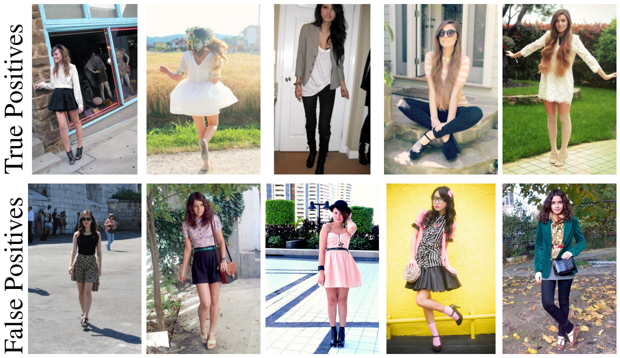


Figure 7: Examples of true and false positives for the fashionability classification task obtained with our CRF model.

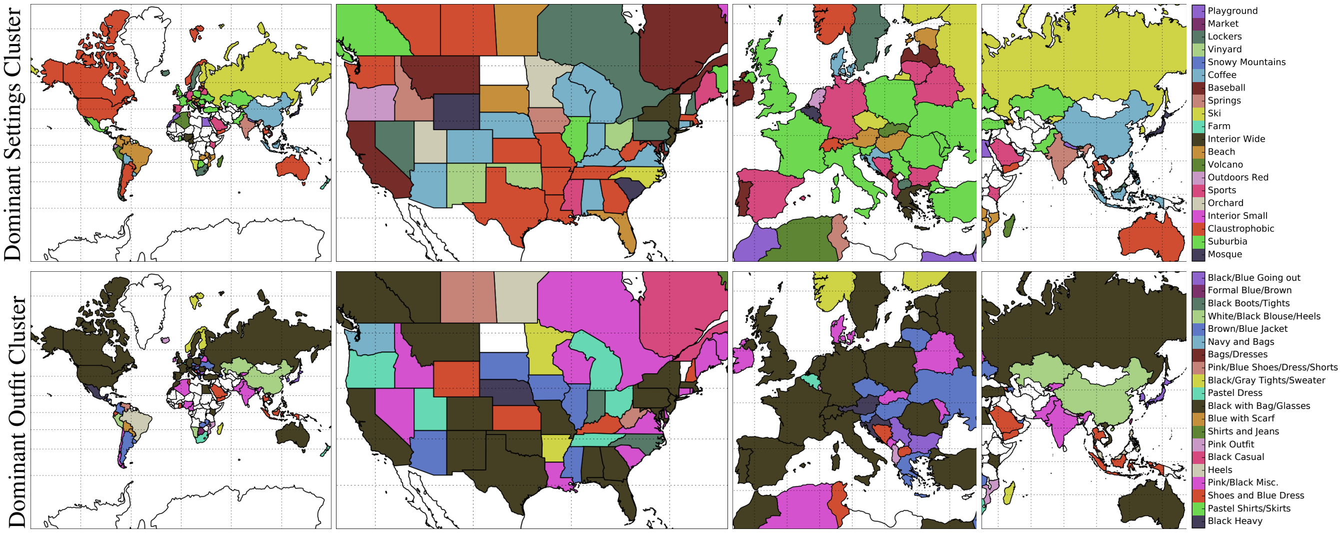


Figure 8: Visualization of the dominant latent clusters for the settings and outfit nodes in our CRF by country.

Feature	Single feature	Leave one out
Baseline	16.3	23.9
Comments	19.7	21.6
Tags	17.4	23.7
ΔT	17.2	23.4
Style	16.3	23.4
Location	16.9	23.3
Scene	16.1	23.3
Fans	18.9	23.2
Personal	16.3	23.1
Colours	15.9	23.0
Singles	17.2	22.8
Garments	16.2	22.7

Table 7: Evaluation of features for the fashionability prediction task using logistic regression. We show two cases: performance of individual features, and performance with all but one feature, which we call leave one out.

Furthermore, we show qualitative examples of true positives, false positives, true negatives and false positives in Fig. 7. Note that while we are only visualizing images, there is a lot of meta-data associated to each image.

In order to analyze the individual contribution of each of the features, we show their individual prediction power as well as how much performance is lost when a feature is removed. The individual performances of the various features are shown in the second column of Table 7. We can see that in general the performance is very low. Several features even perform under the baseline model which consists of predicting the dominant class (Personal, Scene, and Colours).The strongest features are Comments and Fans, which, however, are still not a very strong indicator of fashionability as one would expect. In the leave one out case shown in the third column, removing any feature causes a drop in performance. This means that some features are not strong individually, but carry complementary information to other features and thus still contribute to the whole. In this case we see that the most important feature is once again Comments, likely caused by the fact that most users that comment positively on a post also give it a vote.

5.3. Identifying Latent States

In order to help interpreting the results we manually attempt to give semantic meaning to the different latent states discovered by our model. For full details on how we chose the state names please refer to the supplemental material. While some states are harder to assign a meaning due to the large amount of data variation, other states like, e.g., the settings states corresponding to “Ski” and “Coffee” have a clear semantic meaning. A visualization of the location of some of the latent states can be seen in Fig. 8.

By visualizing the pairwise weights between the fashionability node and the different nodes we can also identify the “trendiness” of different states (Fig. 9). For example, the settings state 1 corresponding to “Mosque” is clearly not fashionable while the state 2 and 3 corresponding to “Suburbia” and “Claustrophobic”, respectively, have positive gradients indicating they are fashionable settings.

5.4. Outfit Recommendation

An exciting property of our model is that it can be used for outfit recommendation. In this case, we take a post as an input and estimate the outfit that maximizes the fashionability while keeping the other variables fixed. In other words, we are predicting what the user should be wearing in order to maximize her/his look instead of their current outfit. We show some examples in Fig. 10. This is just one example of the flexibility of our model. Other tasks such what is the least fitting outfit, what is the best place to go to with the current outfit, or what types of users this outfit fits the most, can also be done with the same model.

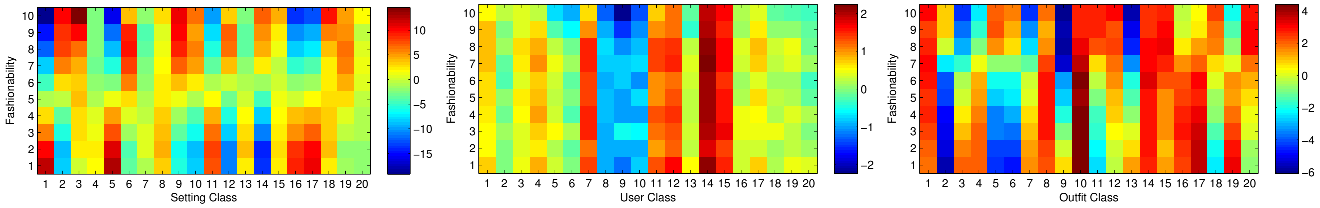
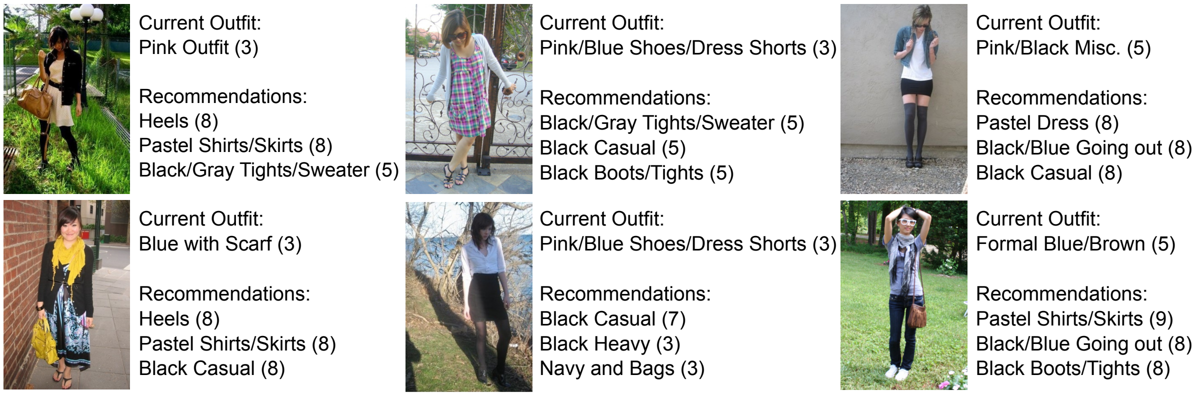


Figure 9: Visualizing pairwise potentials between nodes in the CRF. By looking at the pairwise between fashionability node and different states of other variables we are able to distinguish between fashionable and non-fashionable outfits and settings.



Current Outfit:
Pink Outfit (3)

Recommendations:
Heels (8)
Pastel Shirts/Skirts (8)
Black/Gray Tights/Sweater (5)

Current Outfit:
Pink/Blue Shoes/Dress Shorts (3)

Recommendations:
Black/Gray Tights/Sweater (5)
Black Casual (5)
Black Boots/Tights (5)

Current Outfit:
Pink/Black Misc. (5)

Recommendations:
Pastel Dress (8)
Black/Blue Going out (8)
Black Casual (8)

Current Outfit:
Blue with Scarf (3)

Recommendations:
Heels (8)
Pastel Shirts/Skirts (8)
Black Casual (8)

Current Outfit:
Pink/Blue Shoes/Dress Shorts (3)

Recommendations:
Black Casual (7)
Black Heavy (3)
Navy and Bags (3)

Current Outfit:
Formal Blue/Brown (5)

Recommendations:
Pastel Shirts/Skirts (9)
Black/Blue Going out (8)
Black Boots/Tights (8)

Figure 10: Example of recommendations provided by our model. In parenthesis we show the predicted fashionability.

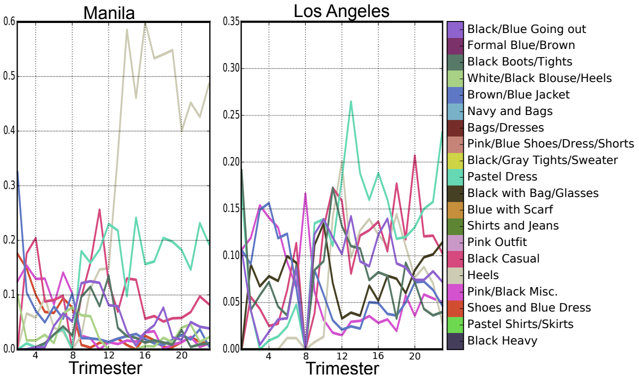


Figure 11: Visualization of the evolution of the different trends in Manila and Los Angeles. The less significant clusters have been manually removed to decrease clutter.

5.5. Estimation Fashion Trends

By incorporating temporal information we can try to visualize the changes in trends for a given location. In particular we look at the trendiest cites in the dataset, that is Manila and Los Angeles, as per Table 3. We visualize these results in Fig. 11. For Manila, one can see that while until the 8th trimester, outfits like "Pastel Skirts/Shirts" and "Black with Bag/Glasses" are popular, after the 12th trimester there is a boom of "Heels" and "Pastel Dress". Los Angeles follows a roughly similar trend. For LA however, before the 8th trimester, "Brown/Blue Jacket" and "Pink/Black Misc" are popular, while afterwards "Black Casual" is also fairly popular. We'd like to note that in the 8th trimester there appears to have been an issue with the chictopia website, causing very few posts to be published, and as a consequence, results in unstable outfit predictions.

6. Conclusions

We presented a novel task of predicting fashionability of users photographs. We collected a large-scale dataset by crawling a social website. We proposed a CRF model that reasons about settings, users and their fashionability. Our model predicts the visual aesthetics related to fashion, and can also be used to analyze fashion trends in the world or individual cities, and potentially different age groups and outfit styles. It can also be used for outfit recommendation.

This is an important first step to be able to build more complex and powerful models that will be able to understand fashion, trends, and users a whole in order to improve the experience of users in the modern day society. We have made both the dataset and code public[3] in hopes that this will inspire other researchers to tackle this challenging task.

Acknowledegements: This work has been partially funded by the Spanish MINECO under project PAU+ DPI2011-27510 and ERA-Net Chistera project ViSen PCIN-2013-047; and by the EU project ARCAS FP7-ICT-2011-287617. The authors also thank Orbeus, Inc. for kindly providing us access to their face analysis API (rekognition.com).

[3] http://www.iri.upc.edu/people/esimo/research/fashionability/

References

[1] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. V. Gool. Apparel classifcation with style. In *ACCV*, 2012. 1

[2] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, 2011. 1

[3] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *ECCV*, 2012. 1, 2

[4] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. Zhu. Composite templates for cloth modeling and sketching. In *CVPR*, 2006. 2

[5] S. Dhar, V. Ordonez, and T. Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, 2011. 2

[6] A. C. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *CVPR*, 2008. 1

[7] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Gool. The interestingness of images. In *ICCV*, 2013. 2

[8] B. Hasan and D. Hogg. Segmentation using deformable spatial priors with application to clothing. In *BMVC*, 2010. 1

[9] T. Hazan and R. Urtasun. A primal-dual message-passing algorithm for approximated large scale structured prediction. In *NIPS*, 2010. 5

[10] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *TPAMI, in press*, 2014. 2

[11] N. Jammalamadaka, A. Minocha, D. Singh, and C. Jawahar. Parsing clothes in unrestricted images. In *BMVC*, 2013. 1, 2

[12] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013. 5

[13] S. Karayev, A. Hertzmann, H. Winnemoeller, A. Agarwala, and T. Darrell. Recognizing image style. In *BMVC*, 2014. 4, 5

[14] A. Khosla, W. A. Bainbridge, A. Torralba, and A. Oliva. Modifying the memorability of face photographs. In *ICCV*, 2013. 2

[15] A. Khosla, A. D. Sarma, and R. Hamid. What makes an image popular? In *International Conference on World Wide Web*, 2014. 2

[16] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, 2014. 2

[17] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, X. Changsheng, and S. Yan. Hi, magic closet, tell me what to wear! In *ACMMM*, 2012. 2

[18] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-toshop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, 2012. 1

[19] F. Magazine. US Online Retail Sales To Reach $370B By 2017; €191B in Europe. http://www.forbes.com, 2013. [Online; accessed 14-March-2013]. 1

[20] A. C. Murillo, I. S. Kwak, L. Bourdev, D. Kriegman, and S. Belongie. Urban tribes: Analyzing group photos from a social perspective. In *CVPR Workshops*, 2012. 1, 2

[21] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Distributed message passing for large scale graphical models. In *CVPR*, 2011. 5

[22] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Efficient structured prediction with latent variables for general graphical models. In *ICML*, 2012. 5

[23] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. A High Performance CRF Model for Clothes Parsing. In *ACCV*, 2014. 2

[24] E. Simo-Serra, C. Torras, and F. Moreno-Noguer. Geodesic Finite Mixture Models. In *BMVC*, 2014. 4, 5

[25] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013. 3, 4, 5

[26] Z. Song, M. Wang, X. s. Hua, and S. Yan. Predicting occupation via human clothing and contexts. In *ICCV*, 2011. 1, 2

[27] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013. 2

[28] S. Vittayakorn, K. Yamaguchi, A. C. Berg, and T. L. Berg. Runway to realway: Visual analysis of fashion. In *WACV*, 2015. 2

[29] N. Wang and H. Ai. Who blocks who: Simultaneous clothing segmentation for grouping images. In *ICCV*, 2011. 1

[30] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 4, 5

[31] K. Yamaguchi, M. H. Kiapour, and T. L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, 2013. 2

[32] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. 2

[33] W. Yang, P. Luo, and L. Lin. Clothing co-parsing by joint image segmentation and labeling. In *CVPR*, 2014. 2

[34] Y. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 2