Chic or Social: Visual Popularity Analysis in Online Fashion Networks

Kota Yamaguchi
Tohoku University
Sendai, Miyagi, Japan
kyamagu@vision.is.tohoku.ac.jp

Tamara L. Berg
University of North Carolina
Chapel Hill, NC, USA
tlberg@cs.unc.edu

Luis E. Ortiz
Stony Brook University
Stony Brook, NY, USA
leortiz@cs.stonybrook.edu

ABSTRACT

From Flickr to Facebook to Pinterest, pictures are increasingly becoming a core content type in social networks. But, how important is this visual content and how does it influence behavior in the network? In this paper we study the effects of visual, textual, and social factors on popularity in a large real-world network focused on fashion. We make use of state of the art computer vision techniques for clothing representation, as well as network and text information to predict post popularity in both in-network and out-of-network scenarios. Our experiments find significant statistical evidence that social factors dominate the in-network scenario, but that combinations of content and social factors can be helpful for predicting popularity outside of the network. This in depth study of image popularity in social networks suggests that social factors should be carefully considered for research involving social network photos.

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: Perceptual reasoning

General Terms

Experimentation; Human Factors; Measurement

Keywords

Social multimedia; Online fashion networks

1. INTRODUCTION

Nearly every blog or social network utilizes a combination of images, text, and other modalities (e.g. location) to convey information and promote interaction. In many online communities the amount of visual data is quite vast, sometimes representing the main source of content. Despite the active research in social multimedia and social popularity hypothesis [15], it is not well-studied *how much* the content quality and the network is influencing to the resulting visual content popularity. In this paper, we take a big-data approach to quantitatively study social influence in a real-world online social network specialized in fashion.

MM'14, November 03 - 07 2014, Orlando, FL, USA

http://dx.doi.org/10.1145/2647868.2654958.

Figure 1: An example of a Chictopia post.

We purposely choose a fashion-focused online community, Chictopia, for the following reasons: 1) The network is large and real-world, with over 175k users and 600k pictures, 2) Content in this network is mainly visual, consisting of "outfit of the day" pictures uploaded by users, 3) The community is focused on a single topic (fashion), which yields relatively consistent user based popularity in contrast to general photos with diverse categories [7], 4) The relevant data is publicly and readily available online, 5) We can take advantage of state of the art computer vision techniques for recognizing clothing in fashion images [19] to help extract the visual content most relevant to popularity.

In a network focused around fashion and style, one might assume that visual content would be the most influential factor for popularity. However, we find that social factors dominate both visual and textual factors in prediction models, even in a community where outfits are purportedly rated based on their fashion style. Furthermore, studying the effects of content outside of the social network, we find that this social bias does not appear, but rather the social and content information together provide a good predictory for popularity. These insights are useful for multimedia researchers and engineers seeking to exploit human behavior in social network applications.

The following list summarizes our contributions:

- Models to predict popularity of outfit pictures incorporating visual, social, and textual factors.
- A new computer vision feature for representing outfit style based on clothing parsing (semantic segmentation of clothing items)
- A large-scale empirical study of social vs. content influence on popularity in a real-world, uncontrolled fashion network
- A large crowdsourcing effort to simulate the socially isolated condition

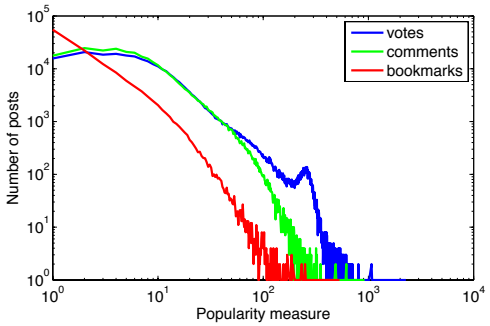


Figure 2: Distribution of voting, commenting, and bookmarked based popularity in Chictopia. The number of votes shows a slight kink perhaps due to front-page highlighting or special promotion by the website.

Type	Name	Modality	Vector	Size
Social	User identity	Network	Sparse	1,000
	Node degrees	Network	Dense	6
	Previous posts	Metadata	Dense	1
Content	Tag TF-IDF	Textual	Sparse	1,000
	Style descriptor	Visual	Dense	441
	Parse descriptor	Visual	Dense	1060
	Color entropy	Visual	Dense	6
	Image composition	Visual	Dense	6
Other	Date bias	Metadata	Sparse	58

Table 1: Summary of the content models.

- Studies of within-network and outside of network scenarios, including empirical findings of asymmetry in content popularity prediction for these scenarios

Related Work: Our work is related to a growing interest in popularity prediction of online content. Work in this direction has mainly looked into early social reaction to content and the prediction of popularity growth in videos [5, 16, 14, 3, 4], news [10, 18], and discussion forums [9]. Some very recent work has looked into visual influence on popularity or behavior [2, 6, 1, 7], or in the reverse direction, to categorize visual contents using social information [8]. Social influence on popularity seen in these studies is also consistent with social browsing behavior on Flickr [11, 17]. We distinguish ourselves from the previous work in that: we focus on a community devoted to fashion, and in particular to rating outfits, our main goal is to evaluate *h*ow much social influence affects popularity for in and out of network scenarios, we take advantage of recent state-of-the-art computer vision approaches to recognize clothing (rather than general visual features or content), and we take a big-data approach using real-world data from a social networking site focused on a single topic (fashion).

2. DATASET

We collect data from `chictopia.com`, a social fashion network where users post pictures of their daily outfits along with a title, description, and several labels. Figure 1 shows an example of picture and metadata we can observe. We initially collect 617,708 posts from Chictopia. To compare visual features consistently across images, we run a state of the art pose detector to automatically select pictures with a standing person detected. This leaves us with 328,604 pictures, dating from March 2008 to Dec 2012, with 34,327 unique users.

Figure 3: Style descriptor (left) and parse descriptor (right). Style descriptor extracts visual information from patches while parse descriptor extracts information from the predicted clothing parse (semantic assignment of pixels to garment labels).

The observable popularity measures in this data are the number of *votes*, *comments*, and *bookmarks* associated with each post. As is the case with any web content, Chictopia popularity reveals a long-tailed distribution. Figure 2 plots the popularity histogram from all 617K posts. To deal with the difficulty in such distribution, we consider log-votes as the popularity of posts in our experiments.

3. CONTENT REPRESENTATION

We represent a post as a vector of the quantized information sources available for the post. Two major components are modeled for each post: social factors and content factors. Social factors capture information related to the user and their social status within the network, whereas content factors are to capture the quality of the content. In addition to social and content factors, we extract timestamp information to account for the popularity change due to seasons and site growth over time. We use a sparse indicator to represent the month of the post (a 58 dimensional vector in our experiments). Table 1 summarizes all of the terms of our model. All factors are appropriately normalized.

3.1 Social factors

User identity: We represent the identity of users using a sparse indicator vector. To constrain feature dimensionality we restrict this indicator vector to the top-1000 most frequent users. Posts from the same user will all have the same feature vector.

Node degree: We use counts and log-counts of friends, followers, and followees of the user as a six-element feature vector. Posts from the same user will all have the same feature. Note that we empirically tested other network features, but none produced a better prediction than node degree.

User Expertise: We use the number of previous posts from the same user as a scalar feature, related to user expertise.

3.2 Content factors

Tag TF-IDF: In Chictopia, a user can label each individual post with various structured tags indicating general style, occasion, colors, brands, clothing types, in addition to free-form words. We first extract unigrams and bigrams from all the tags and treat them as a document, and compute TF-IDF weights. As in the case of user identity, to constrain the dimensionality of this feature, we only consider the 1000 most frequent n-grams found in the training samples.

Style descriptor: The style descriptor is a state-of-the-art visual clothing representation proposed in [19] for fashion image retrieval, and considered a comprehensive representation of the overall fashion style. The descriptor is based

Factors	In-network				Out-of-network			
	R^2	Spearman	Accuracy		R^2	Spearman	Accuracy	
			25%	75%			25%	75%
Social	0.491 ±0.005	0.682 ±0.004	**0.847** ±0.002	**0.779** ±0.003	0.423 ±0.011	0.634 ±0.007	0.845 ±0.004	0.787 ±0.005
Content	0.248 ±0.005	0.485 ±0.005	0.778 ±0.003	0.737 ±0.003	0.428 ±0.012	0.647 ±0.008	**0.888** ±0.004	**0.862** ±0.004
Social+Content	**0.493** ±0.005	**0.685** ±0.004	0.845 ±0.002	0.775 ±0.002	**0.473** ±0.014	**0.686** ±0.008	0.884 ±0.004	0.858 ±0.004

Table 2: Regression and top-K% classification on the observed popularity with accompanying 95% bootstrapped confidence intervals on error. For cleaner presentation, the tiny asymmetric difference in confidence intervals are rounded.

on image patches localized on the person's body by pose estimation. At each patch, various features including color, texture, shape, and skin-hair probability are extracted and pooled to produce a high-level description of fashion.

Parse descriptor: In addition to the style descriptor, we develop a new fashion-focused image descriptor based on clothing parsing [19], which we call *parse descriptor*. Clothing parsing is the task of assigning a semantic garment (or skin or background) label to each pixel in the image. The parse descriptor is designed to represent the appearance of individual garment items found in a picture, and we experimentally verified that the parse descriptor in combination with the style descriptor gives a strong prediction. We consider the parse descriptor to be the most important representation in content analysis, because the parse descriptor specifically captures the appearance of a person's garment items. Figure 3 illustrates the style and parse descriptors.

We compute the parse descriptor in the following steps: 1) Compute clothing parse using [19], and obtain 10 masks corresponding to specific garment groups, such as *outer top*, *dress*, or *footware*. Note that we map the original 56 garment categories [19] to 10 garment sets to improve robustness. 2) Extract RGB color, Lab color, Texture response, HOG descriptor, distance from image border, and probability of skin and hair at every pixel. 3) Compute mean-std pooling of the extracted features in each region. 4) Concatenate all pooled features over 10 regions (1060 dimensions).

Color entropy: We compute the entropy of RGB and Lab color from the image. This feature helps distinguish drawings from natural photos.

Image composition: Given a bounding box encompassing the person (estimated by the pose detector), we measure the overall composition of how the person is depicted relative to the image frame as, 1) normalized width, height, and area; 2) normalized x and y displacement from the center of the image; and 3) normalized distance from the image center.

4. IN-NETWORK POPULARITY

Experimental protocol: We first apply a linear regression analysis on the log-votes of the posts using social, content, and a combination of social and content factors. For this analysis, we adopt R^2 and Spearman coefficients as measures of fitness of prediction to the truth. These measures are evaluated on a statistical bootstrapping protocol with our 328K posts; We randomly resample posts, subsample this dataset to 10,000 posts (for computational tractability), and evaluate the above measures with a 90%-10% train-test split. This process is repeated 100 times to derive statistical significance.

We also apply a classification based analysis, in which we predict a binary indicator of being Top K% in popularity. In this experiment, we vary our threshold for 25% and 75% quantiles of the votes in the training samples to see the difference between the most popular and least popular posts. The performance is measured in terms of accuracy.

Results: In-network column of Table 2 shows the results of regression analysis. The regression models fit significantly better when social factors are present, suggesting a user's social connections largely dominates the popularity of their posts over the content itself. However, we should note that social factors may also be highly correlated with content quality – users with many followers may tend to wear highly fashionable outfits. Nevertheless, our results indicate that the influence of content quality is considerably smaller than network influence.

Classification reveals an asymmetry between top 25% and top 75% prediction, indicating the prediction of the most popular posts is easier than predicting the least popular posts. We suspect that this is partly due to the consistently better quality of top-rated pictures and partly due to social bias more strongly affecting popular posts. Also, the slightly smaller difference between the social-only and the content-only model in top 75% prediction suggests that popularity is less affected by social influence in least popular posts.

5. OUT-OF-NETWORK POPULARITY

To examine the effects of social and content factors more deeply, we utilize *crowdsourcing* to emulate a content network without social relationships. We run the popularity voting process in Amazon Mechanical Turk, where no social network exist.

Our task shows the crowd worker 50 random large-resolution images in sequence, and asks them to vote on the picture if they find it *chic*. For quality control, we measure how long it takes each worker to complete this first step and reject a worker's votes if they completed too quickly or didn't display enough variation in their voting procedure. We perform our experiments by randomly selecting 3,000 posts (from our dataset of 328k) and instantiating the above tasks 60 times. We assign 25 workers to each of the 60 resulting tasks. Therefore, any post on each task can obtain up to 25 *chic* votes.

Using the voting data from the crowd, we apply the same analysis. Our main interest here is, however, the influence of the social factors observed in Chictopia on crowd popularity. We use the same bootstrap method from the 3,000 posts to compare the social-only, content-only, and combined models in this experiment. Here, social factors are taken from Chictopia dataset which has no relationship to crowd popularity.

Results: Results are shown in The out-of-network column of Table 2 shows the result. Given our previous results, we initially expected the social factors to lead to much weaker

predictors. However, the results suggest that social factors (from Chictopia) still lead to comparable predicators in regression. We conjecture that the solid regression result of the social factors is due to the user and content quality correlation.

Also, the combined model (social + content) is significantly better than any single-factored model in regression. One possible explanation to this result is that the social factors are actually providing complementary information to the content factors in predicting the *unbiased* popularity from the crowd, as opposed to the biased popularity in the network where social influence is by far the stronger predicator of popularity. We observed the asymmetry of prediction also in the out-of-network condition, but to weaker extent perhaps due to no social bias in voting.

Apart from factorizing content and social influence, we can also use our learned models to predict the popularity of photos. Figure 4 shows an example of the most and least popular pictures predicted by one of our models. There is clearly a distinction in visual quality between the most and the least popular pictures. Perhaps it is also possible to build a system that can predict *unbiased* content popularity. Such prediction could be useful for many e-commerce applications, such as automatic outfit quality feedback [12], socially-aware fashion recommendation [13]. Stable popularity prediction can benefit in online ad optimization and traffic balancing. It is our future work to use this insight to build a socially-aware multimedia system.

In summary, the out-of-network popularity analysis yield insights that suggest 1) social factors contain not only network information but also some aspect of content evaluation, 2) content factors capture different aspects of popularity than social, and 3) their combination yields better predictions for content popularity.

6. CONCLUSIONS

We presented a vision-based approach to quantitatively evaluate the influence of social and content factors on fashion pictures. Our content representation takes advantage of various sources of information from computer vision, natural language processing, and the network itself. Through experiments, we showed statistical evidence of dominant social influence in networked media and the strength of social factors in unbiased content evaluation. One lesson from our experiments is that any attempt to learn subjective measures such as aesthetics from social content should explicitly consider social influence.

7. ACKNOWLEDGEMENTS

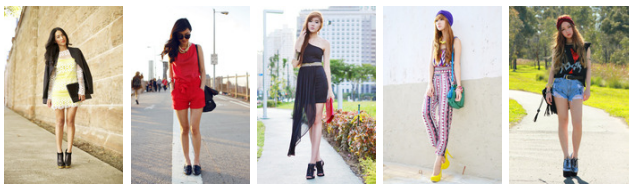
Research supported by Google Faculty Award, "Seeing Social: Exploiting Computer Vision in Online Communities," NSF Career Award #1054133 and #1054541.

8. REFERENCES

[1] S. Bakhshi, D. Shamma, and E. Gilbert. Faces engage us: Photos with faces attract more likes and comments on instagram. *CHI*, 2014.

[2] J. Biel and D. Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *Multimedia*, 15(1):41–55, 2013.

[3] Y. Borghol, S. Ardono, and N. Carlsson. The untold story of the clones: Content-agnostic factors that impact youtube video popularity. *KDD*, 2012.



(a) Predicted most popular

(b) Predicted least popular

Figure 4: Prediction examples from our combined model.

[4] A. Brodersen, S. Scellato, and M. Wattenhofer. Youtube around the world: geographic popularity of videos. In *WWW*, pages 241–250. ACM, 2012.

[5] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *IMC*. ACM, 2007.

[6] S. Chang, V. Kumar, E. Gilbert, and L. G. Terveen. Specialization, homophily, and gender in a social curation site: Findings from pinterest. In *CSCW*, pages 674–686. ACM, 2014.

[7] A. Khosla, A. D. Sarma, and R. Hamid. What makes an image popular? *WWW*, 2014.

[8] A. Kimura, K. Ishiguro, M. Yamada, A. Marcos Alvarez, K. Kataoka, and K. Murasaki. Image context discovery from socially curated contents. In *ACM Multimedia*, pages 565–568, 2013.

[9] J. G. Lee, S. Moon, and K. Salamatian. An approach to model and predict the popularity of online contents with explanatory factors. In *WI-IAT*, volume 1, pages 623–630. IEEE, 2010.

[10] K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. In *WWW*, pages 621–630. ACM, 2010.

[11] K. Lerman and L. A. Jones. Social browsing on flickr. *ICWSM*, 2007.

[12] L. Liu, H. Xu, J. Xing, S. Liu, X. Zhou, and S. Yan. Wow! you are so beautiful today! In *ACM Multimedia*, pages 3–12, 2013.

[13] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan. Hi, magic closet, tell me what to wear! In *ACM Multimedia*, pages 619–628. ACM, 2012.

[14] L.-P. Morency, R. Mihalcea, and P. Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *ICMI*, pages 169–176. ACM, 2011.

[15] M. Slaney. Web-scale multimedia analysis: does content matter? *MultiMedia, IEEE*, 18(2):12–15, 2011.

[16] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[17] M. Trevisiol, L. Chiarandini, L. M. Aiello, and A. Jaimes. Image ranking based on user browsing behavior. In *SIGIR*, pages 445–454. ACM, 2012.

[18] H. M. Ung. Social influence, popularity and interestingness of online contents. In *ICWSM*, 2011.

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