Building User Profiles from Shared Photos

ABSTRACT

In this paper, we analyze the association between a social media user's photo content and their interests. Visual content of photos is analyzed using state-of-the-art deep learning based automatic concept recognition. An aggregate visual concept signature is thereby computed for each user. User tags manually applied to their photos are also used to construct a tf-idf based signature per user. We also obtain social groups that users join to represent their social interests. In an effort to compare the visual-based versus tag-based user profiles with social interests, we compare corresponding similarity matrices with a reference similarity matrix based on users' group memberships. A random baseline is also included that groups users by random sampling while preserving the actual group sizes. A difference metric is proposed and it is shown that the combination of visual and text features better approximates the group-based similarity matrix than either modality individually. We also validate the visual analysis against the reference inter-user similarity using the Spearman rank correlation coefficient. Finally we cluster users by their visual signatures and rank clusters using a cluster uniqueness criteria.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Indexing methods

Keywords

Visual content analysis, tf-idf, inter-user similarity

1. INTRODUCTION

Social media enables users to easily share information about all aspects of human life. While Flickr was one of the earliest photo-based social networks, in the last several years, many other photo-centric social services have emerged (e.g. Instagram, Snapchat, Tumblr, and Path). Twitter started primarily as a text-based microblog service, but it also increasingly includes images and video tweets. In short, photos are becoming ubiquitous within social media. As a consequence, services that make product recommendations or advertisements based on users' social content have to inevitably shift their attention to photos.

Brisbane '15 Australia

Figure 1: Sample photos and highest ranked concept classes (with their corresponding scores) obtained using *Caffe*.

For years, the multimedia and computer vision communities have grappled with the challenge of correctly interpreting the content of photos. The recent advances of deep learning based methods in concept recognition [6] can enable new experiences around socially shared photos that require reliable automatic content analysis.

From the point of view of advertisement, product recommendation, and other services, it is valuable to understand the association between a user's interests and their social media content. When the content in question is photos, the challenge has been exacerbated in the advent of the digital camera, as people capture even mundane things in their photos. Sifting useful visual information amid the visual noise is imperative to effectively mine their interests. We have observed that summarizing visual content across a number of a user's pictures can have a filtering effect and often tends to highlight some of their primary visual interests.

Our goal in this paper is to systematically ascertain the extent to which the "thousand words" contained in photos (expressed as user tags or as visual concepts extracted from photos collectively) are representative of photographers' interests at large. In other words, we study the consistency between the results of automatic analysis of a user's photos and their interests as explicitly expressed in terms of memberships in user groups. Instead of focusing on budding photo-centric social networks mentioned above, we turn to Flickr, one of the earliest and most trusted photo centric

social networks. Specifically, we focus on a subset of the Yahoo Flickr Creative Commons 100 Million dataset which was released in 2014 [12].

A critical feature of Flickr for our purposes is the ability of users to join groups in which relevant photos may be shared and pooled among group members[1]. We view the explicit actions of users to join groups as a direct reflection of their interests. We build a profile of each user in our experiments based on their group memberships and use this information to quantify pairwise inter-user similarity. In this way, we measure inter-user interest similarity in terms of users' actions rather than any indirect inference or assumptions. We employ this inter-user similarity as the "ground truth" in our experiments and use it to assess and compare analogous similarity measures constructed based on automatic analysis of users' photos, and text analysis of the tags manually applied to users' photos, as well as similarity measures that combine the two modalities. We thus establish a relative sense of the utility of modern visual analysis for modeling users' interests and also compare it against a typical user tag-based baseline.

2. RELATED WORK

There is a growing literature that studies both the behavior and practices of Flickr users. Negoescu and Gatica-Perez [9] performed a comprehensive examination of user groups on Flickr and built tools to help organize the groups into related clusters while also examining photo-sharing practices. De Choudhury, *et al.* [1] presented a group recommendation system that analyzed visual content, user tags, and interaction features derived from comments. Their experiments showed both that groups were more thematically coherent than individual users, and recommending photos to groups performed better than trying to match photos to users. Most users' interests can't be well represented by any single group's theme, but instead span multiple topics.

Li, *et al.* [7] argued that user tags can be mined to determine user interests and can often be more useful to characterize web documents than keywords. They examined the `del.icio.us` bookmarking database to first verify that user tags were generally consistent with the content to which they were applied. Additionally, they used the tags to group URLs by topics according to tag usage patterns. The Argo advertising system utilizes an ontology to jointly represent both the content in users' photos as well as a set of advertisements [13]. In this way, ads can be matched to users based on their interests as inferred from shared photos. The open directory project (ODP) ontology[2] is used as a space of topics for representing the users interests. Feng and Xian [2] also used the ODP ontology in a hierarchical manner and integrate user tags with automatic annotation to build a distribution over a set of topics to represent user interests. Most of these related works also use tf/idf measures [8] to weight the tags in their processing. Xie, *et al.* [14] developed a latent variable model that encompasses both users and photos. The approach is purely image based and represents photo content hierarchically in terms of pixels, visual words, regions, and themes. Users' interests are modeled by the distribution of their photos over the theme layer of the model. This work also included a clever evaluation which relied on the assumption that photos that a user marked as "favorites" could be used to represent their interests. This idea is similar in spirit to our evaluation in which we also infer user interests from their explicit behavior on Flickr.

Our focus here is user profiling based on user's shared photos and tags contributed to Flickr. We ground inter-user similarity derived from established methods for modeling users' visual content and tags in terms of the groups that users themselves form on Flickr. We assume that the explicit act of joining a group is rooted in a user's interest in that group's theme. Thus we build a "ground truth" inter-user similarity measure from the pairwise Jaccard similarity in (1). We then compare the inter-user similarity computed from per-user profiles of photo content, per-user profiles of tag usage, and their combination against this group-based ground truth. We present results that show that on Flickr, the visual analysis based method using deep learning and tag based approach both outperform random baselines. As well, we show a joint method offers further improvements.

3. USER PROFILING

We profile users' interests based on the photos they share. Analysis of the photos and their tags generates profiles for each user in our pool. We further compute pairwise inter-user similarity with the downstream aim of clustering users according to their interests. To understand the basic efficacy of content analysis towards this end, we compare pairwise inter-user similarities computed using content analysis with the reference inter-user similarity based on group memberships. We first review the users profiles we construct from users' photos.

3.1 Visual content analysis

The visual interest signature for every user consists of a summary of visual concepts discovered in their photos. Concept detection has been extensively studied in computer vision for several decades now. We used the state-of-the-art deep learning based image annotation framework Caffe [4] for concept discovery. Figure 1 shows example concept classes and corresponding scores for four images in our dataset. The concepts used in our experiments conform to the 1000 ImageNet categories [10] and produce a 1000 dimensional score vector for each image. The visual feature vector for each user captures the average distribution of concept scores across their images. An average of scores (for all concepts) was performed across all images with the aim to highlight dominant or recurring visual interests and suppress less prevalent visual concepts.

3.2 Tag-based analysis

Flickr images include tags given by owners and third party users. In order to compare users in terms of tag-based characteristics of their photos, we adopt the standard information retrieval term frequency inverse document frequency (tf-idf) methodology [8]. We aggregate tags from each user's photos in a synthetic document. Thus our tag-based interest signature is a standard tf-idf based score vector across the vocabulary of tags.

We compare two variations for constructing the tag vocabulary. The first (Large Vocabulary in Table 1 includes tags that have been used by at least two users, whereas the second (Small Vocabulary in Table 1) includes tags that have been used by at least five users. In both experiments,

[1]https://www.flickr.com/groups/
[2]http://www.dmoz.org

Table 1: Evaluating inter-user similarity mined from different modalities versus the group-based ground truth reference using (4). For the randomly created groups, we list mean (standard deviation) of values across five random runs.

Comparison Types	Matrix Norm (Large Vocab.)	Matrix Norm (Small Vocab.)
Visual vs. True Groups	745.6	742.0
Tags vs. True Groups	720.7	718.0
Visual+Tags vs. True Groups	685.1	682.5
Visual∗Tags vs. True Groups	666.1	663.5
Random vs. True Groups	782.5 (1.44)	779.2 (1.48)

we eliminate users that do not have any relevant vocabulary tags in their representation (i.e. users who use tags not in the vocabulary). This results in the user populations of sizes 1962 and 1954 for the Large and Small Vocabulary experiments respectively.

We used the `tfidfvectorizer` class within the Python scikit-learn library's[3] feature extraction module to compute the tf-idf signature for each user. We performed stop word elimination and Porter stemming on the raw user tag sets. After preprocessing the sizes for the Large and Small Vocabulary experiments was $34,446$ and $11,714$ respectively.

4. EXPERIMENTS

We obtained SFO Bay Area photos from within the Yahoo Flickr Creative Commons 100 Million dataset (YFCC100M). A rectangular grid roughly representing the SFO Bay Area was defined and a total of $1,448,554$ photos lying within this geo-grid were obtained. We performed further filtering to eliminate users with fewer than 100 photos in this region. In other words, we primarily target people who live in the SFO Bay Area using photos explicitly geotagged within the area. Finally our dataset consisted of a total of 2170 users and a total of $1,249,098$ photos. For each of these users, we obtained the Flickr groups to which they belong using the Flickr API. The 2170 users belonged to a total of $54,610$ groups. The maximum number of groups that a user belonged to was $2,980$ while the median number of group memberships per-user was 23. Removing users who do not belong to any group and further excluding users whose photos do not have any tags in our vocabularies produces user populations of 1962 and 1954 in the two experimental conditions as mentioned above.

4.1 Inter-user Similarity

In the first experiment, we characterize on a macro-scale how patterns detected in shared photos relate to the interests of a large number of users. Our intention is to study the extent to which content-based similarity scores or ranks reliably model pairwise interest similarity (here represented by shared groups). Emphasis on pairwise comparison stems from the fact that common processing such as collaborative filtering or unsupervised clustering of users usually builds on pairwise similarity. The overall reliability of a similarity comparator can be gauged from a comparison across all available pairs of users. To study the relative strengths of visual content analysis vis-a-vis manual tags given by users in the aforementioned task, we compare pairwise similarities using different modalities to one another over the user population.

[3]http://scikit-learn.org/stable/index.html

In the absence of absolute ground truth profiles representing user interests, we utilize group memberships to represent user interests. For two users, denoted u, v we compute the Jaccard similarity to assess the relative overlap of the users' group memberships:

$$S_{GT}(u,v) = \frac{|G_u \cap G_v|}{|G_u \cup G_v|} \quad , \tag{1}$$

where G_u denotes the set of groups for user u. This equation specifies a pairwise similarity matrix with rows and columns indexed by the set of users. We use this matrix as a ground truth reference for comparison with other content-based similarity matrices below.

For content-based similarity, the standard cosine measure is computed across pairs of users for both their visual and tag-based signatures (as shown respectively in (2) and (3)):

$$S_{vis}(u,v) = \frac{f_{vis}(u) \cdot f_{vis}(v)}{\| f_{vis}(u) \| \| f_{vis}(v) \|} \quad , \tag{2}$$

$$S_{tag}(u,v) = \frac{f_{tag}(u) \cdot f_{tag}(v)}{\| f_{tag}(u) \| \| f_{tag}(v) \|} \quad . \tag{3}$$

We assess different modalities (tags, visual, and their combination) by comparison with the ground truth across all pairs of users. Intuitively, this is computing the norm of difference matrices between a content-based similarity matrix and the ground truth similarity matrix. We use the Frobenius matrix norm:

$$\mathcal{D}(Sim, GT) = \left[\sum_{u,v} (S_{Sim}(u,v) - S_{GT}(u,v))^2 \right]^{\frac{1}{2}} \quad . \tag{4}$$

Applying this metric directly to the various similarity measures is difficult. Most of the elements of S_{gt} from (1) are zero, while the cosine similarity measure ranges in $[0,1]$ for non-negative features. To normalize the various measures, we map each element to its percentile within the entire matrix. This has the effect of emphasizing the ordering of the elements within the matrices and provides a common scale for elementwise comparison.

To provide a baseline for the experiment, we constructed five different random groupings of users This was done by randomly sampling users with replacement and grouping them according to the observed group sizes. Table 1 shows the comparison metric of (4) for visual, tag, and combined similarity as well as the random baselines. Values across all random baselines are shown in the form of means and standard deviation values for the matrix norm metric over all random runs.

We consider two similarity measures combining tags and visual analysis. The first is additive:

$$S_{\text{vis+tag}}(u,v) = (\lambda \cdot S_{vis}(u,v)) + ((1-\lambda) \cdot S_{tag}(u,v)) \ , \tag{5}$$

where $\lambda = 0.5$ for our experiments. The second is a multiplicative combination:

$$S_{\text{vis}*\text{tag}}(u,v) = (S_{vis}(u,v))^{\gamma} \cdot (S_{tag}(u,v))^{(1-\gamma)} , \quad (6)$$

where we set $\gamma = 0.6$ based on experimentation.

The results show that visual content analysis (Visual) and tag-based similarity (Tags) outperform the random baselines (Random) in both experiments. We also note that the combined Visual+Tags, Visual∗Tags variants have the best performance in both experiments. This is in agreement with research on fusion of multimedia streams (classifier scores or ranked lists) in machine learning. While potentially a more advanced text representation (such as using topic models) could improve tag-based performance here, the emphasis of our experiments is benchmarking modern visual analysis. At the same time, more advanced methods can also be applied to visual based concept features with corresponding improvements.

4.2 Spearman Rank Correlation

We also examined the Spearman rank correlation coefficient [11] as an alternate approach to validating the visual analysis against the reference ground truth inter user similarity of (1). The Spearman's rho is a nonparametric measure of how monotonic the relationship between two variables is. For our purposes in modeling users' interests using content analysis, we use it to assess how consistent the orderings induced by the content-based similarity measures under consideration are relative to the ordering induced by our ground truth reference. Its computation involves converting raw observations in to their ranks and processing the differences between rank positions by observation. More details can be found in [5].

Table 2: Spearman's rho computed over the entire data set using the small vocabulary condition.

Modality	Spearman's ρ	p-value
Visual	0.0863	0.0
Random	0.0048	2.826e-11
Tag	0.175	0.0
Visual + Tag	0.1062	0.0
Visual ∗ Tag	0.1846	0.0

Here, we performed two experiments using the Small Vocabulary data. First, we again examine the dataset globally across users, and treat each pairwise similarity matrix as a single set of inter-user measurements. The results in Table 2 show that visual analysis (Visual) shows a modest correlation with the ordering of the ground truth reflecting the user's group membership. The tag-based (Tag) results show a stronger correlation than the visual inter-user similarity. In contrast, the baseline (Random) results are essentially uncorrelated with the ground truth. As before, the fusion condition (Visual∗Tag) does better than Tag and Visual, but in this case, the additive combination (Visual+Tag) performs somewhat worse than the tag analysis alone.

In a second experiment, we consider the pairwise similarities as before, but in this case compute Spearman's rho for each of the users in our data set individually. Thus we fix a user u and consider the set of observations $\{S(u,v) : v \in \text{users}\}$ and again compare the orderings of the content-based and ground truth similarities. We summarize the resulting set of 1954 experiments in Table 3 using the means and standard deviations computed over the set of all users. The trend exhibited in the previous global analysis recurs here. The standard deviations show that the agreement with the ground truth varies considerably per user, as expected. Figure 2 shows histograms depicting the distribution of the Spearman's rho for different modalities over the set of users.

Table 3: Spearman's rho averaged over the data set per-user using the small vocabulary condition.

Modality	Spearman's ρ mean	Spearman's ρ std. deviation
Visual	0.075937	0.061978
Random	0.002002	0.022317
Tag	0.122359	0.072729
Visual + Tag	0.090269	0.066904
Visual ∗ Tag	0.13480	0.075892

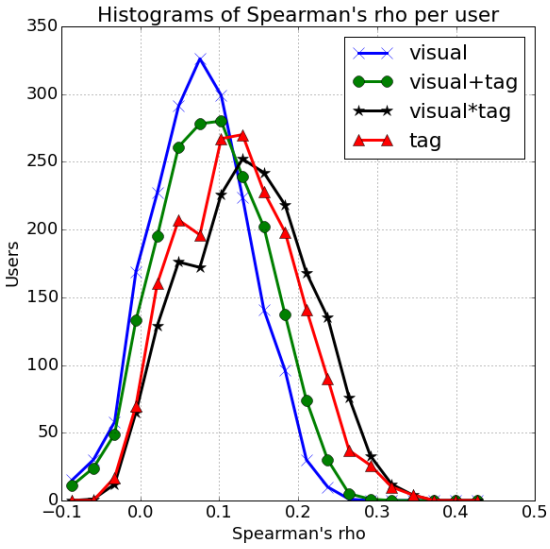

Figure 2: Plots showing histograms of the Spearman's rho per-user computed using different modalities in comparison against the reference group-based similarity measure.

4.3 Visual Clustering of Users

Our experiments have shown that visual information combined with tags outperforms any individual modality. However, meaningful tags are becoming increasingly rare in recent photo-centric social networks, such as Pinterest, Instagram etc. where the trend is shifting from content oriented tags (such as in Flickr) to *#hashtags* that indicate a social movement or phenomena. As well, tags are not always available, even on Flickr. In the experiments above, restricting the tag vocabulary in turn reduced the user population that could be compared using tags from 2033 to 1954 in the

Table 4: Describing the content of five exemplar user clusters obtained using visual analysis. The rightmost column shows high ranking concepts from Caffe and their corresponding scores.

Total Inter-cluster centroid distance	# of users	Highest Ranking Concepts
212.3	5	['junco, snowbird', 0.04], ['kite', 0.039], ['red-backed sandpiper, dunlin, Erolia alpina', 0.037], ['dowitcher', 0.031], ['red-breasted merganser, Mergus serrator', 0.029]
211.6	2	['daisy', 0.28], ['bee', 0.038], ['cardoon', 0.032], ['sea anemone, anemone', 0.032], ['pinwheel', 0.03]
209.0	2	['knee pad', 0.335], ['rugby ball', 0.037], ['unicycle, monocycle', 0.036], ['football helmet', 0.036], ['puck, hockey puck', 0.03]
206.1	2	['convertible', 0.169], ['beach wagon, station wagon, wagon, estate car, beach waggon, station waggon, waggon', 0.126], ['pickup, pickup truck', 0.111], ['grille, radiator grille', 0.055], ['car wheel', 0.038]
205.9	3	['mountain bike, all-terrain bike, off-roader', 0.283], ['bicycle-built-for-two, tandem bicycle, tandem', 0.161], ['crash helmet', 0.05], ['unicycle, monocycle', 0.046], ['football helmet', 0.027]

Small Vocabulary condition. Hence in our final experiment, we qualitatively explore using visual analysis alone to discover clusters of users with similar interests. Such clusters are expected to emphasize common visual themes within a social context.

We perform clustering using the affinity propagation algorithm [3]. We used the default setting of the affinity propagation implementation in `scikit-learn` to obtain a total of 222 clusters for the original set of 2033 users. Clusters were ranked in terms of their aggregate distance (total inter-cluster distance) from all other cluster centroids to promote visually distinct clusters. In Table 2, we describe the content of the five highest ranked clusters using the top concepts from their cluster centroid signatures. It is quite clear from the descriptors that the clusters are focused on (1) birds, (2) flowers, (3) sports, (4) cars, and (5) biking.

This cluster ordering emphasizes smaller clusters that are tightly focused around visual themes by design. In future work, we expect to more systematically explore how to trade-off cluster size and the concentration of the cluster's shared content around one or more distinct themes. At the same time, we would like to point out that a small cluster of size 2 or 3 representing a distinct interest (such as shown in the table) in a population of 2033 users could potentially scale to a few thousand people in a population of millions of users thus representing a potentially focused interest group.

5. CONCLUSIONS

In this paper, we have attempted to benchmark modern visual- and tag-based content processing with respect to modeling user interests. We used a subset of the Yahoo Flickr Creative Commons 100 Million dataset for our experiments. Group membership information obtained independently from Flickr was used as proxy for user interests. We studied the relative performance of visual- and tag-based content similarity methods in modeling pairwise user interest information and compared their performance to random baselines. Our experiments showed that overall both visual analysis and tag-based user similarity methods outperform random baselines at the task of modeling pairwise user interest similarities. Additionally a combination of visual and tags based similarity had the overall best performance. We also performed a visual clustering experiment and discussed some of the high ranked clusters (in terms of a uniqueness criteria).

By comparing content-based user profiling methods against measures grounded in explicit user action, we can both build confidence in the broader applicability of automatic processing and better understand its limitations. Here we have focused on pairwise inter-user similarity as it is the foundation for user clustering and other applications. The performance of state of the art deep learning image classification methods in our experiments is both competitive with an established tag-based baseline and also provides complementary information when combined with tags. In future work, we intend to build further on our initial explorations of clustering user profiles with the broader goal of improving both group and content recommendation.

6. REFERENCES

[1] M. De Choudhury, H. Sundaram, Y.-R. Lin, A. John, and D. Seligmann. Connecting content to community in social media via image content, user tags and user communication. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 1238–1241, June 2009.

[2] H. Feng and X. Qian. Mining user-contributed photos for personalized product recommendation. *Neurocomputing*, 129:409–420, 2014.

[3] B. Frey and D. Dueck. Clustering by passing messages between data points. *Science* **315**(5814): 972-976, 2007.

[4] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, MM '14, pages 675–678, New York, NY, USA, 2014. ACM.

[5] M. Kendall. A new measure of rank correlation. *Biometrika*, 30(81-93), 1938.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[7] X. Li, L. Guo, and Y. E. Zhao. Tag-based social interest discovery. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 675–684, New York, NY, USA, 2008. ACM.

[8] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[9] R. A. Negoescu and D. Gatica-Perez. Analyzing flickr groups. In *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, CIVR '08, pages 417–426, New York, NY, USA, 2008. ACM.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, pages 1–42, April 2015.

[11] C. Spearman. The proof and measurement of correlation between two things. *American Journal of Psychology*, 15(72-101), 1904.

[12] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

[13] X.-J. Wang, M. Yu, L. Zhang, R. Cai, and W.-Y. Ma. Argo: Intelligent advertising by mining a user's interest from his photo collections. In *Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising*, ADKDD '09, pages 18–26, New York, NY, USA, 2009. ACM.

[14] P. Xie, Y. Pei, Y. Xie, and E. P. Xing. Mining user interests from personal photos. In *Proc. of the 29th AAAI Conference on Artificial Intelligence*, 2015.