Predicting Participants in Public Events using Stock Photos

Neil O'Hare, Luca Maria Aiello, Alejandro Jaimes
Yahoo! Research, Avinguda Diagonal 177, Barcelona, Spain
nohare@yahoo-inc.com, alucca@yahoo-inc.com, ajaimes@yahoo-inc.com

ABSTRACT

Pictures taken by journalists for distribution and for inclusion in stock photo collections are often enriched with metadata. One key aspect of such photos is that they focus largely on events and feature celebrities and other public figures. They may provide interesting insights into how such public figures are related to each other in terms of the events they attend, and in their social proximity in terms of how often they are photographed together. In this paper, we study a corpus of approximately 9 million stock photographs taken over a 10 year period and, using their metadata, we extract a social network from co-appearance of public figures in events depicted in the photographs. We exploit this latent social information and combine it with the rich image metadata to explore the possibility of predicting attendees at future events, showing promising performance for this task.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Miscellaneous; H.3.3 [**Information Search and Retrieval**]: Miscellaneous

Keywords

stock photos, public events, celebrities, social networks

1. INTRODUCTION

Many news articles are accompanied by photographs that depict events, and very often such photographs include public figures (e.g., politicians, celebrities, athletes, etc.). The photos, mostly taken by journalists, are usually manually enriched with textual metadata, including a short caption describing the image, and additional information such as the location of the photo and possibly an event related to the photo. They are distributed via wire services (e.g, AFP), and included in stock photo collections for future use[1].

In aggregate, large collections of news stock images can provide interesting insights into how different celebrities and public figures are "related" to each other in terms of the events they attend and also in terms of whether, and how often, they are photographed together. Such information may be useful for marketing (e.g., an athlete who often appears in non-sports events might appeal to a different demographic group than one who is only photographed with other athletes) and for search (e.g., finding related celebrities[2]). Such relationships could also be used in data-driven journalism tasks (e.g., a recent report by the L.A. Times analyzed the demographics of Oscars' voters[3]). Additionally, such collections could be leveraged to predict who will attend future events and which individuals may be photographed together.

In this paper, we study a corpus of almost 9 million stock photographs, taken over a 10 year period. Using metadata from the photos, we extract a social network based on two forms of co-appearance of public figures: co-appearance in events, and in individual photos. We exploit this latent social information and combine it with the rich textual metadata attached to the images to explore the possibility of predicting attendees at future events. We show that textual and social network information can both be exploited for this task, and that they can be fruitfully combined. The main contributions of our work are:

- The extraction and analysis of a large public figure social networks from a large image corpus.
- Using this network, in combination the photo metadata, to predict attendees at public events.

To the best of our knowledge, this is the first work that has attempted to predict attendees at public events using images and their metadata.

In the next section we discuss related work, followed by a description of the dataset and the social network in Section 3. We then introduce methods for predicting event attendees in Section 4, and evaluate these approaches in Section 5. Section 6 concludes the paper.

2. RELATED WORK

The problem of event recognition from multimedia has been studied extensively in the last few years [13], assuming even greater importance with the increasing amount of

[1] There are hundreds of such services, but the largest ones are Getty Images, Corbis, and Sipa Press

[2] Yahoo! Image search provides this functionality.

[3] http://graphics.latimes.com/towergraphic-la-et-academy-tower

MM'12, October 29–November 2, 2012, Nara, Japan.

	Photo	Event
Nodes	38,846	43,323
Edges	404,349	8,798,056
Density	$5.4 \cdot 10^{-4}$	$9.4 \cdot 10^{-3}$
Avg. degree	20.8	406.2
Avg. edge weight	4.5	2.0

Table 1: Statistics on latent social networks, based on co-occurrence in photos, or in events.

multimedia content on online social platforms. Approaches based on metadata such as tags or geo-location [10, 1], as well as content-based techniques [6], have been explored. Identification of people from videos and images has been explored in depth [14] and it has been shown that combining content-based techniques with image metadata [3, 8] or social information [15] can improve performance. More recently, rather than individual person detection from multimedia, some efforts have been spent in the extraction of social connections from images and video. Recognition of social clusters [12] and prediction of social relationship types between individuals in photos [11] are examples of possible tasks in this context. More generally, much latent information is hidden in richly annotated multimedia collections, and this can be leveraged for prediction and classification tasks [5]. Other related work on person finding (e.g., [7]) creates textual representations of people based on the text they create, somewhat similarly to our work, which describes people with the metadata of images they appear in.

The work of Devezas et al. [4] is the closest to ours. They focus on a much smaller collection, and limit their study to a preliminary exploration of the structural clusters in the social graph. Their observations are not used for any prediction task and events are not taken into account.

3. DATASET

We collected the publicly available metadata of approximatively 9 million images from a well-known stock photo agency, covering a time frame from 2000 to 2011. Metadata includes the *timestamp* of when the picture was taken, a *title* and *caption*, a set of *keywords* defining the image type, content and context (e.g., sports, election debate), and possibly the *event* the image depicts. In addition, there are two different sources of information about *people* appearing in the photos: a set of names automatically extracted from the caption, matched with a database of known public figures, and a set of manually annotated person names. We use the automatically extracted names as the main source of people information, but when a photo contains manual person annotations, we consider the intersection between the two sets, and discard names that do not appear in both. At the end of this filtering process we have more than $45K$ unique person names. For the purpose of analysis only, we complemented the description of people with their Wikipedia categories, crawled from the pages dedicated to them. Among all the people in our corpus, the 78% have a Wikipedia page.

3.1 A Social Network extracted from Photos

The rich image metadata enables inference of social connections between people. Co-occurrence of people in a photo or an event can be interpreted as ties in a social graph, where nodes are people and edges represent co-occurrence, with the edge weight proportional to the number of co-occurrences.

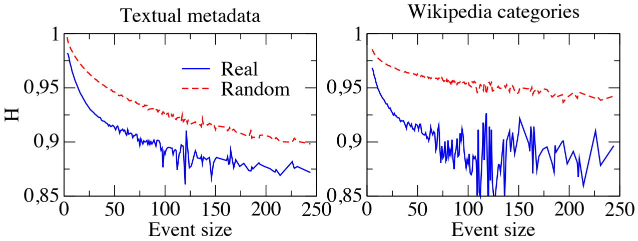

Figure 1: Average entropy H of keywords in photo metadata and Wikipedia categories for people at the same event, at fixed event size (# of attendees).

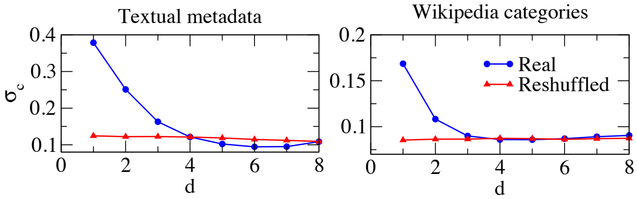

Figure 2: Cosine similarity σ_c between keyword/category vectors of pairs of people at distance d in the social graph.

Intuitively, photo co-occurrence implies a stronger social tie than event co-occurrence. We report basic statistics of these social graphs in Table 1. Since the event-based graph expresses weaker interpersonal relations, it is much denser than the photo-based network. Given its more manageable density and stronger social significance, we focus the following analysis on the photo co-occurrence network.

3.2 Data Overview

Latent social information emerging from the images and their additional textual metadata can be exploited to make inferences on the data. In particular, we focus on the task of predicting people in future events, given knowledge of past events. First, we observe that events tend to aggregate people with similar profiles. This can be quantified by describing each person with a vector of terms and measuring the entropy of the overall term distribution inside the event, compared to the entropy of an "artificial" event of the same size composed by random people. We can build such person profiles from the aggregation of all the textual metadata associated to the images they appear in, or from the categories in their Wikipedia pages. Figure 1 shows that, in both cases, the entropy in random events is appreciably higher than in real events, meaning that events tend to aggregate homogeneous people to some extent.

From a network perspective, social proximity is related to person profile similarity. Figure 2 shows that people tend to be more similar to people who are closer in terms of number of hops in the network than to people residing far away. The similarity decay with the distance is evident for both types of user profiles. The difference of the similarity decay applied on a reshuffled version of the network, where links are rewired at random, shows that profile alignment of close people is not determined by chance or by pure assortativity. Edge strength is also correlated with profile similarity, meaning that people often co-appearing in a photo share many profile features (Figure 3). This also holds for the Wikipedia

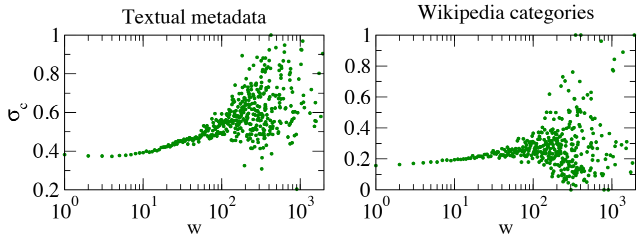

Figure 3: Average cosine similarity of profiles of people connected with a edge of weight w. Positive slopes show positive correlation.

categories, showing that it is not simply an artifact of people "sharing" the annotations of photos they co-appear in.

4. PEOPLE PREDICTION IN EVENTS

We propose two ways to use this photo data for predicting the people in events. The first strategy, *text-based*, predicts attendees by retrieving people from past, similar events, given the description of the target event. The second, *network-based* approach, predicts people that are connected with strong ties or that reside near one of the event participants, assuming that at least one person is known.

4.1 Text-based Prediction

We use the titles, captions and keywords associated with photos to capture the shared semantics between people and the events they attend. For a given person, we represent them by the text associated with the photos depicting them (photos with duplicate titles, captions and keywords are only counted once), and build a language model to represent each person. Given the textual representation of an event, we rank candidate persons by the likelihood that their language model "created" the event description T [9]:

$$P(Per|T) = \frac{P(T|\theta_{Per})P(Per)}{P(T)} \tag{1}$$

where $P(Per)$ is the prior probability of person Per, and $P(T|\theta_{Per})$ is the probability of the text T, given the person language model, calculated as the product of the probabilities of the individual terms in T. The probabilities of the individual terms are estimated as the maximum likelihood estimate, smoothed with a background model using Jelinek Mercer smoothing. We can estimate $P(Per)$ using the relative frequency of the candidate person in the collection. We also discuss in Section 4.3 how we can use the social network information to refine the estimate of this prior. $P(T)$, the prior probability of the event description text, is constant for all people being ranked, and can be ignored.

4.2 Network-based Prediction

For the network-based approaches we consider both photo and event co-occurrence networks (*Photo Net* and *Event Net*). We assume that a single event attendee is known, namely the one with the highest degree in the social network, and consider that person's social neighbors to be good candidates for attending the same event. For this simple baseline, we rank candidates based on their edge weight between the seed person. We also propose a method that also takes into account nodes that do not belong to the seed's neighborhood, since not all people in the same event are neighbors in the photo co-occurrence network. We apply a personalized version of PageRank (PR), where every random walk starts from the seed node (see Boldi et al. [2]), to the *Photo Net* network. Intuitively, the higher the PageRank of a node, the more reachable it is from the seed node. The PR is applied to a directed version of the network, where arcs from node v to node u are weighted by $w(v, u)/\sum_{x\in\Gamma(v)}(w(v, x))$, where $\Gamma(v)$ is the set of v's neighbors and w is the weight of an edge in the original undirected network. We refer to this approach as *Photo Net PR*

4.3 Hybrid Approach

We explore two alternative approaches to combining textual and network based prediction. The first is to expand the textual event description with the name of the seed attendee, and use the standard text approach. Since names are not removed from the textual representations of people, this returns people who have the seed person in their description, and therefore co-occur with the seed person.

The second approach is to use the network predictions to estimate the prior probability of a person, $P(Per)$, in Equation 1. The network approaches (with the exception of the random walk approach) will give a 0 estimate to any people who do not share a direct connection with the seed person, which means such non-neighbors will have zero prior probability. To improve the network estimate of the prior, we smooth it with the frequency estimate discussed in Section 4.1, using Jelinek Mercer smoothing.

5. EVALUATION

We evaluate our approaches to person prediction using the public photo corpus described in Section 3. First, we filter out small events with less than 5 people or less than 20 photos, giving approximately 51K events. We split our corpus into a training set (approx. 50K events and 6.5M photos) used for building our language models and social networks, and a test set (1K) on which we evaluate the ability of these models to predict the participants in an event, given its' metadata. The corpus was split based on time, with all test set events occurring after the training events. We represent each event by the unique terms in the event (with known people removed). For the textual approaches, we rank candidate people by the likelihood that their language model created this event. For the social network approaches (and for the "seed name" text approaches) we use the person with the highest degree as a seed. For all approaches, this "seed name" is removed from the ground truth and the result predictions. For the language models, we use a standard default collection smoothing parameter ($\lambda = 0.85$), and for smoothing the social network prior with the frequency prior, we give equal weight to each. We evaluate using Mean Average Precision (MAP) and Precision at K (P@K).

The results of the evaluation are presented in Table 2. The baseline text-based results show that it is possible to achieve reasonably high precision early in the ranking, with a baseline P@1 of 0.444, which can be improved to 0.465 if we incorporate a frequency-based prior. Adding the name of one known person to the query does not improve performance for most evaluation measures, showing that this naive method of exploiting the social network does not improve prediction performance. All of the explicit social network approaches outperform the naive approach of using the seed person as a query for the semantic approach ("seed name").

Of the social network methods, network based on event co-occurrence gives the best results, even for precision at very high ranks. Although this seems somewhat surprising due to the fact that photo co-occurrence should be a stronger indicator of relatedness, it suggests that the event-based network is more robust, possibly because it is less affected by data sparseness. Although worse for P@K, the *Photo Net PR* method has higher MAP than the *Photo Net* approach, most likely because it can predict also people residing at distance ≥ 2 from the seed node. Using the output of the event-based network as a prior in the semantic model harms performance considerably, which is unsurprising since many persons will have a prior of 0. Smoothing this prior with the frequency-based model, however, leads to an improvement over the semantic approach (from 0.219 to 0.225 for MAP, a relative improvement of almost 3%). This result is quite encouraging, showing that relatively simple social network approaches, when reliable seed information is available, can offer an improvement over a strong baseline, whereas standard information retrieval style approaches to exploiting the same information fail to give a similar improvement. We plan, in future work, to investigate whether pseudo relevance feedback could provide reliable seeds for such approaches.

Method	MAP	P1	P3	P5	P10	P20
Text- Based Prediction						
No Prior						
Textual	0.216	0.444	0.372	0.325	0.263	0.020
Frequency Prior						
Textual	0.219	0.465	0.381	0.331	0.266	0.020
Seed Name	0.063	0.143	0.107	0.097	0.087	0.079
Textual+Name	0.219	0.462	0.382	0.332	0.266	0.199
Social Network- Based Prediction						
Photo Net	0.077	0.225	0.169	0.147	0.122	0.099
Event Net	0.096	0.232	0.178	0.158	0.136	0.109
Photo Net PR	0.084	0.185	0.159	0.143	0.115	0.095
Text and Social Network- Based Prediction						
Network Prior	0.181	0.371	0.309	0.270	0.214	0.164
Smoothed Prior	**0.225**	**0.467**	**0.385**	**0.338**	**0.270**	**0.204**

Table 2: Results for Event Attendance Prediction. Best results are in bold.

6. CONCLUSIONS

In this paper we build social networks based on photo and event co-occurrence in a corpus of 9M stock photos. We propose methods to predict attendees at public events based on the shared semantics of people and events, and based on social connections between people. Our results show that the two approaches can be profitably combined, and that social network measures outperform naive information retrieval-based approaches to combining these two modalities.

These types of models can be used to predict attendees for any events for which we have some metadata available, which should prove very useful in event-based multimedia indexing. These semantic descriptions of events and people could also be used, of course, to compare events with events, people with people, and to help discover previously unknown links between events and people. Also, although this paper has focused on predicting attendees at events, the results may be useful even when the people are not actually event attendees, in that they may often be strongly related to the event. In future work, we plan to investigate such more general applications of these models. We also plan to explore using content-based analysis to improve the prediction.

7. ACKNOWLEDGMENTS

This research is partially supported by European Community's Seventh Framework Programme FP7/2007-2013 under the ARCOMEM and Social Sensor projects, by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037 "Social Media", and by Grant TIN2009-14560-C03-01 of the Ministry of Science and Innovation of Spain.

8. REFERENCES

[1] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *WSDM*, pp. 291–300. ACM, 2010.

[2] P. Boldi, R. Posenato, M. Santini, and S. Vigna. Algorithms and models for the web-graph. chapter Traps and Pitfalls of Topic-Biased PageRank, pp. 107–116. Springer-Verlag, Berlin, Heidelberg, 2008.

[3] M. Davis, M. Smith, J. Canny, N. Good, S. King, and R. Janakiraman. Towards context-aware face recognition. In *ACM Multimedia*, pp. 483–486. ACM, 2005.

[4] J. Devezas, F. Coelho, S. Nunes, and C. Ribeiro. Studying a personality coreference network in a news stories photo collection. In *ECIR*, pp. 485–488, 2012.

[5] X. Jin, A. Gallagher, L. Cao, J. Luo, and J. Han. The wisdom of social multimedia: using flickr for prediction and forecast. In *ACM Multimedia*, pp. 1235–1244. ACM, 2010.

[6] J. Luo, J. Yu, D. Joshi, and W. Hao. Event recognition: viewing the world with a third eye. In *ACM Multimedia*, pp. 1071–1080. ACM, 2008.

[7] A. Mantrach and J.-M. Renders. A general framework for people retrieval in social media with multiple roles. In *ECIR*, pp. 495–498, 2012.

[8] N. O'Hare and A. F. Smeaton. Context-aware person identification in personal photo collections. *IEEE Transactions on Multimedia*, 11(2):220–228, Feb. 2009.

[9] J. M. Ponte and W. B. Croft. A Language Modeling Approach to Information Retrieval. In *SIGIR*, pp. 275–281, Melbourne, Australia, 1998.

[10] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from Flickr tags. In *SIGIR*, pp. 103–110. ACM, 2007.

[11] P. Singla, H. Kautz, J. Luo, and A. Gallagher. Discovery of social relationships in consumer photo collections using markov logic. In *CVPR*, pp. 1–7. IEEE, 2008.

[12] P. Wu and D. Tretter. Close & closer: social cluster and closeness from photo collections. In *ACM Multimedia*, pp. 709–712. ACM, 2009.

[13] L. Xie, H. Sundaram, and M. Campbell. Event mining in multimedia streams. *Proc. of the IEEE*, 96, 2008.

[14] M. Zhao, Y. W. Teo, S. Liu, T.-S. Chua, and R. Jain. Automatic person annotation of family photo album. In *CIVR*, pp. 163–172, Berlin, Heidelberg, 2006. Springer-Verlag.

[15] T. Zickler, Z. Stone, and T. Darrell. Autotagging facebook: Social network context improves photo annotation. In *CVPR Workshops*, pp. 1–8, 2008.