Detecting Friday Night Party Photos: Semantics for Tag Recommendation*

Philip J. McParlane[1], Yelena Mejova[2], and Ingmar Weber[3]

[1] The University of Glasgow
[2] Yahoo! Research
[3] Qatar Computing Research Institute
p.mcparlane.1@research.gla.ac.uk, ymejova@yahoo-inc.com,
ingmarweber@acm.org

Abstract. Multimedia annotation is central to its organization and retrieval – a task which tag recommendation systems attempt to simplify. We propose a photo tag recommendation system which automatically extracts semantics from visual and meta-data features to complement existing tags. Compared to standard content/tag-based models, these automatic tags provide a richer description of the image and especially improve performance in the case of the "cold start problem".

1 Introduction

Multimedia retrieval heavily relies on finding quality textual annotations for content. For this reason, sites such as YouTube and Flickr encourage users to tag their content. We study the problem of tag recommendation where users provide a (possibly empty) list of input tags and are provided with a ranked list of suggested output tags.

Existing models offer new suggestions by finding highly co-occurring tags with those present in the ground truth [1], thereby ignoring *when* an image is taken and its *scene*. For example, for an image taken indoors on a Friday night, which is tagged with `people`, we should consider `party` as a new recommendation with higher probability than `office`. Our system introduces two kinds of automatically-generated tags based on (i) meta-data such as the time a picture was taken, and (ii) visual content of the image such as the number of detected faces. Using an existing tag co-occurrence based model [2], we incorporate these tags into the tag ranking process, obtaining significant improvement.

The strengths of our approach lie in (i) help with the "cold start problem" before the user enters textual tags, (ii) its simplicity as it can be implemented as an extension of tag co-occurrence based techniques, and (iii) its efficiency compared to using high-dimensional nearest neighbour search. Thus, unlike works focusing solely on image content [3] or solely on the tags [1], our approach provides a middle ground while allowing for standard tag retrieval algorithms.

* This research was supported by the the European Community's FP7 Programme under grant agreements nr 288024 (LiMoSINe).

P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 756–759, 2013.

2 Methodology

Let n denote the number of tags in our vocabulary and m denote the number of images in our collection. $m^{(x)}$ is the number of images tagged with x. We introduce a number of matrices which model the tag-image and tag-tag relationships:

- G is an mxn matrix with each row containing 1's for the presence of a tag in the given image's ground truth.
- C is an nxn co-occurrence matrix where C_{ij} counts how many images the tags i and j co-occur in.

The overall goal of our system is to predict G_{i*} as accurately as possible for each image i, given a number of initial tags q from the user. We want to compute a *ranking* of tags such that for a tag j which is high in the ranking, $G_{ij} = 1$.

In addition to the user-defined (*textual*) tags, we introduce automatic *contextual* and *content* tags which we extract from meta-data and image contents.

Textual tags are defined by the user and provide a valuable insight into the nature of the image. However, these may be difficult to obtain.

Contextual tags extracted, for example, from the time the image was taken, provide further information about the circumstances in which the image was created. Here, we extract the *time of day*, *day of the week*, and *season information.* Additionally we consider whether an image is shot in landscape or portrait.

Content tags are extracted from the visual content of the image using machine learning techniques. We consider distinctions between *city* or *landscape*, *day* or *night*, *indoor* or *outdoor* (using SVMs trained on the Edge Direction Coherence [4], HSV histogram [5] and Colour Moments features respectively [6]), and the *number of faces* (using the technique described in [7]). Accuracies of the given classifiers are shown in Table 1.

To recommend new tags for an image, given a number of *input* tags from the ground truth, we adopt a state-of-the-art (SOTA) approach as described in Algorithm 2 of [2]. It derives a new matrix $\hat{C}$ from C in two steps. First, all diagonal values of C are set to zero. Second, each column of this new matrix is scaled, so that the maximum in each column is 1. The vector of scores s_q is then computed as $s_q = (\hat{C}q) \cdot \text{idf}$. Here, the "·" stands for the component-wise product of two vectors and $\text{idf}(x) = \log(m/m^{(x)})$ is a vector of "inverse document frequencies". Note that $\hat{C}$ can be seen as a normalized version of a standard document-term matrix, so that this scheme is just a simple tf-idf retrieval model.

Table 1. Image Classifiers and Accuracies

Classifier	**Kernel**	**Cost**	**Gamma**	**Accuracy**
Day/Night	Linear	2^{-1}	N/A	88.3%
Indoor/Outdoor	RBF	2^{5}	2^{-5}	71.1%
City/Landscape	RBF	2^{7}	2^{-3}	77.3%

In our experimental approach we introduce a weighting function for the input tags, due to *textual* tags containing more accurate information regarding an image's contents, in comparison to the *automatic* tags. Therefore, we weight each entry of $\hat{C}$ with respect to the keywords popularity by multiplying by $idf(x)^2$, where x is the keyword for the given column. By doing so, we avoid the problem of suggesting popular tags due to the high co-occurrence scores of automatic tags (which exist in every image) with popular textual tags (e.g. `nature`, `art`).

3 Experiments

For our experiments we test our approach on a subset of a larger collection containing, 1,857,46 images (with 21,139 tags), which are crawled from Flickr[1]. Initially 2000 *nouns* (collected from categories such as `animal` and `artifact`) are extracted from WordNet [8], which are used to query the Flickr API. We then collect the top 2000 images returned for each noun. Flickr-specific tags such as those denoting awards achieved on Flickr, camera meta data, and tags which were used by fewer than 150 users are removed. Using this approach ensures a balanced, unbiased collection. Finally, we select a random subset of this collection containing 7,000 train images and 139 tags, with 500 images used for testing.

For every image in the test set, we select a number of *input* tags from the ground truth (ranging from 0 to 4), and attempt to predict the *other* tags in the ground truth. Our baseline, uses only the user tags in the ground-truth whereas our experimental approach also uses the range of "automatic tags", which are extracted from an image's visual contents and meta data, as input.

4 Results

The findings of our experiments are summarised in Table 2. By using either tags extracted from image contents or contextual meta-data (or a combination) we are able to achieve statistically significant improvements to prediction accuracy over our SOTA baseline. As the number of user tags increases, however, exploiting the content and contextual tags reduces recommendation accuracy (see Figure 1), due to the noise present in these automatic tags. Therefore, exploiting image contents and context is of most use in a cold start scenario (i.e. where an image has no or 1 initial tag(s)). Also note that because with increasing number of input tags, the number ground truth decreases, making evaluation more difficult (and the scores lower).

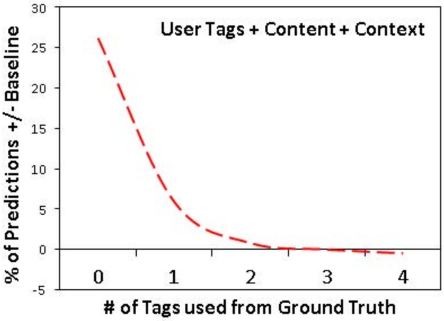


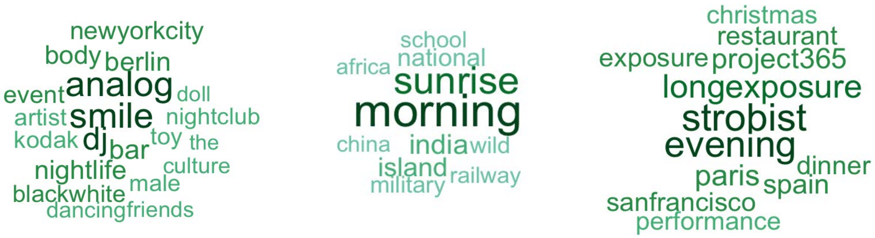
Fig. 1. Cold Start Performance

Using matrix C, we can further show the semantic cohesiveness of our automatic tags. For example, Figure 2 shows the top distinct tags co-occurring with tags *night*, *morning*, and *evening*. These tags exemplify the activities and places associated with each of the time of the day. Thus, although general, these automatic tags divide the dataset into semantically meaningful segments.

[1] This data is available at `http://www.dcs.gla.ac.uk/~philip/`

Table 2. Tag recommendation performance (P@5). Paired t-test statistical significance comparing our experimental approach against the baseline are denoted as * being $p < 0.05$, ** being $p < 0.01$ and *** being $p < 0.001$. † predicting 5 most popular tags.

	Number of input tags				
Input	**0**	**1**	**2**	**3**	**4**
User's tags	0.093†	0.164	0.147	0.105	0.063
+ content	0.121***	0.169*	0.147	0.103	0.059*
+ context	0.125**	0.171*	0.148	0.105	0.060
+ content & context	0.149***	0.174**	0.149	0.105	0.060



(a) night (b) morning (c) evening

Fig. 2. Tag clouds for contextual tags *night*, *morning*, and *evening*

Thus, we show that "semantic" tag information derived from an image's content or meta information can improve tag recommendation, especially when no or few textual input tags are given. As our approach is agnostic of the underlying co-occurrence based recommendation algorithm, we plan to experiment with other algorithms in the future. We will also explore image classes that benefit more, as well as evaluate performance on larger collections with richer co-occurrence information.

References

1. Sigurbjörnsson, B., Van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: WWW, pp. 327–336 (2008)
2. Garg, N., Weber, I.: Personalized, interactive tag recommendation for flickr. In: RecSys, pp. 67–74 (2008)
3. Sun, A., Bhowmick, S., Nam Nguyen, K., Bai, G.: Tag-based social image retrieval: An empirical evaluation. JASIST (2011)
4. Vailaya, A., Jain, A.K., Zhang, H.: On image classification: city images vs. landscapes. Pattern Recognition 31(12), 1921–1935 (1998)
5. Wan, Y., Hu, B.G.: Hierarchical image classification using support vector machines. In: ACCV (2002)
6. Vailaya, A., Member, A., Figueiredo, M.A.T., Jain, A.K., Zhang, H.J., Member, S.: Image classification for content-based indexing. TIP 10, 117–130 (2001)
7. Hare, J., Samangooei, S., Dupplaw, D.: Openimaj and imageterrier: Java libraries and tools for scalable multimedia analysis and indexing of images. In: MM (2011)
8. Miller, G.A.: Wordnet: A lexical database for english. CACM 38, 39–41 (1995)