The Accuracy and Value of Machine-Generated Image Tags

Design and User Evaluation of an End-to-End Image Tagging System

Lexing Xie, Apostol Natsev, Matthew Hill, John R. Smith
IBM Watson Research Center, Hawthorne, NY, USA
{xlx, natsev, mh, jsmith}@us.ibm.com

Alex Phillips
IBM Global Business Services, United Kingdom
alex_phillips@uk.ibm.com

ABSTRACT

Automated image tagging is a problem of great interest, due to the proliferation of photo sharing services. Researchers have achieved considerable advances in understanding motivations and usage of tags, recognizing relevant tags from image content, and leveraging community input to recommend more tags. In this work we address several important issues in building an end-to-end image tagging application, including tagging vocabulary design, taxonomy-based tag refinement, classifier score calibration for effective tag ranking, and selection of valuable tags, rather than just accurate ones. We surveyed users to quantify tag utility and error tolerance, and use this data in both calibrating scores from automatic classifiers and in taxonomy based tag expansion. We also compute the relative importance among tags based on user input and statistics from Flickr. We present an end-to-end system evaluated on thousands of user-contributed photos using 60 popular tags. We can issue four tags per image with over 80% accuracy, up from 50% baseline performance, and we confirm through a comparative user study that value-ranked tags are preferable to accuracy-ranked tags.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

Image tagging, social media, user value

1. INTRODUCTION

Social tagging has been popularized by ubiquitous and diverse content sharing services, from bookmarks, articles, to photos and videos. Tag recommendation plays an important role in enhancing user-experience: it reduces the effort of text entry, and assists content organization, searching, and sharing. Unlike social sharing on sites such as bookmarks on del.icio.us or articles on digg.com, tags for user-generated photos and videos do not recur across users. Furthermore, visual tags are not easily extracted from the content directly as in the case of topical keywords in articles, and both of these factors make photo tag recommendation difficult.

Considerable advances have been achieved in several subtopics around image tag recommendation, such as understanding user tagging motivation and behavior, image/video classification and semantic concept detection, and community-based tag recommendation. However, several important questions remain. Which tags shall the system recommend? Despite formal ways of organizing concepts in natural language and images [10, 19], single word plain text descriptors ("visual tags") have taken root as the prevalent form of social tags. We note that visual tags are distinct from "visterms" or "visual words" such as feature vectors often used in quantization. We systematically examine the most popular and most visually salient tags to construct a tag vocabulary, and then hierarchically organize them to assist visual classification. Which tags are the most descriptive for an image? We calibrate and re-rank tags for an image based on the information content of a tag and its perceived utility in a large photo sharing site (ie, Flickr). We conduct a user-study to measure the value of visual tags, and to quantitatively compare the tag ranking schemes as perceived by the user.

This paper has several novel contributions. (1) We have systematically studied the most popular and visual salient tags for constructing a comprehensive visual vocabulary. (2) We have devised ways to estimate tag value from large data collections. (3) We successfully developed methods for calibrating classifiers scores and issuing accurate tags for each image based on tag relationships. (4) We have recruited over 20 pilot users, who helped validate the results both with performance measures on user-contributed test set, and with a comparative study between different re-ranking strategies.

2. RELATED WORK

Problems related to image tagging are addressed in several research communities. The image retrieval and computer vision community has created many algorithms and systems for performing image annotation and object recognition; the web, human-computer interface communities have been engaged in studying user intention and usage of social tagging systems; information from social sharing, media content and ontology are then used to improve and revise photo tags.

Object and scene recognition has been one of the grand challenges in computer vision, and considerable progress has been made in recent years by learning a separate model for each semantic category and evaluating it on each image. The success of such systems have been witnessed by sev-

CIVR '10, July 5-7, Xi'an China

eral ongoing benchmark evaluation campaigns [28, 3]. A variety of approaches represent the state-of-the-art in performance, training and generalization, including discriminative ensembles over many features [30], multiple scales and local regions [32], many data samples [18], models for local parts [11], real-time annotation [15], and association between words and regions [5]. External knowledge and concept relationships can also be used to help, examples include using universal ontologies such as WordNet [10] for additional training data [29], or using task-specific ontologies [21, 26].

Not all tags are created or used equally. A number of categorizations have been proposed based on the motivation and utility of tags: *categorization* vs. *description* [7], *refine* vs. *identify* [13], *personal* vs. *sharing* [20], *organization* vs. *communication* [4]. Bischoff et. al [6] proposed a 8-way segmentation of social tags and evaluated their utility for searching.

Using social knowledge to help re-rank, filter or expand tags is also a very active research problem. Several Flickr image tag recommendation systems are based on tag concurrence and inter-tag aggregation, used in a fully-automatic [27] or interactive [12] setting. Liu et. al [16] have incorporated image similarity in conjunction with tag co-occurrence, evaluated with subjective labeling of tag usefulness; Kennedy et. al [14] use shared visual appearance to improve the retrieval of specific tags such as landmark or proper names.

Our work performs image classification with ensemble binary classifiers [18, 35], calibrates the output with precision estimates, and refines labels with a novel multi-faceted taxonomy. We systematically examine visually salient and popular tags, whereas most prior work does not address the aspect of proper visual tag vocabulary design. We also focus on estimating tag utility in our tag re-ranking design, which was most inspired by [4, 6]. Our tag re-ranking strategy models the inherent usefulness of visual tags, and proposes estimation of such criteria from social and local collections. Our users do not only provide ground-truth for system evaluation, their inputs guide the end-to-end system design and are used for direct comparison of re-ranking algorithms.

3. TAGGING SYSTEM OVERVIEW

We have built an end-to-end image tagging system to acquire and process photos from social media sources, as illustrated on Figure 1. It builds upon the IBM Multimedia Analysis and Retrieval System (IMARS) [2], and adds components for score calibration, tag refinement and re-ranking, which are the focus of this paper. We also created a Flickr *AutoTagr* app, built using the Flickr API [1], which downloads images from Flickr and uploads machine-generated tags back to Flickr. We first classify each of the downloaded images with a set of pre-built visual classifiers using IMARS, which uses approximately 100 visual features to create ensemble models of SVM classifiers [18, 35] for each of the target visual categories, defined in Sec 4. We calibrate the output of these classifiers to estimated accuracy (Sec 5), refine the top tags with a multi-faceted visual taxonomy (Sec 6), and perform importance-based tag re-ranking with a number of different tag value estimates (Sec 7). We use the outcome for quantitative evaluation and comparative user study (Sec 8).

4. TAG VOCABULARY CONSTRUCTION

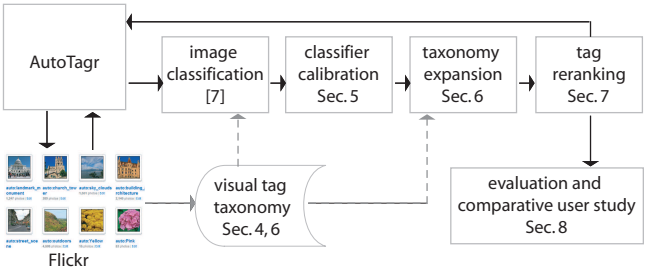

Figure 1: Image tagging system overview

A good set of target tags is the foundation of an automated tagging system. There are three desired properties for visual tags—they must be: (a) *popular*, to cover as many images and users as possible; (b) *visually observable*, to focus on the image content, rather than subjectivity or context. (c) *machine learnable*, to ensure high accuracy for automated tagging systems. The outcome of our vocabulary design is 5000 concepts, and a subset 60-node tag taxonomy.

We use Flickr tag frequency to select the most *popular* tags. Since there is no publicly available tag vocabulary, we start with over 4 million unique tags from the CoPhiR collection [22] of 100M Flickr photos. We case-normalize the tags, filter out non-alphanumeric characters, and exclude rare tags with counts less than 100. We collect frequency counts of the remaining 196,391 tags using the Flickr API [1]. We further prune the vocabulary off rare tags (tagged less than 10,000 times on Flickr, or $1/1000^{th}$ of the most popular tag)—this leaves a set of 24,444 tags of top popularity, which are the tags perceived as most important or useful by users. Note that many of these popular tags refer to contextual information that may not be present or observable in the image itself (e.g., 2008, july, family, fun, usa),

We use an ESP Game [33] data set to further select *visually observable* tags. The ESP game is designed to record words that two independent users agree on for the same image, which naturally favors the most relevant and non-subjective visually observable tags. The most frequently occurring ESP game tags are good representatives of unambiguous tags purely from visual appearance, and do not require contextual information on when, where, how, or why a photo was taken. This dataset[1] contains 100K images and 27,629 unique tags, out of which 6,092 appear at least 10 times. We keep tags that are popular in both the Flickr and ESP lists (by taking the max. rank), and this results in a final prioritized list of 5,060 tags, which are both frequently appearing, popularly used, and visually observable.

In order to study the effect of such tag filtering and selection strategy, we manually analyze the top 500 tags, and categorize them into visual, contextual (e.g., events, holidays), dates/times, named entities (e.g., locations, brands, people), and other non-visual tags (e.g., too abstract, ambiguous, emotional, or subjective), similar to existing tag classification schemes [6]. Table 1 shows a comparison of the distribution of tag categories before and after tag re-ranking, and we can see that the fraction of visual tags increases substantially, from 42% in the original Flickr top 500 to 70%, or ˜350 visual tags, in the re-ranked top-500 set. Meanwhile, the fraction of named entities drops from 36% to only

[1] `http://www.cs.cmu.edu/~biglou/resources/`

Figure 2: Comparison tag cloud for original top-100 Flickr tags (left, in red) vs. re-ranked top-100 tags (right, in blue) using the proposed tag selection approach.

7%. Figure 2 visually illustrates the difference in the top-100 tags before and after re-ranking using a comparison tag cloud. We can see that many of the original top-100 Flickr tags (in red) are subjective or non-visual in nature, and are replaced largely by visual tags in the re-ranked set (in blue).

Table 1: Comparison of original top-500 Flickr tags vs. top-500 re-ranked tags based on the proposed method.

Tag Category	Top-500 Flickr tags	Top-500 Re-ranked tags
Visual	41.6%	69.6%
Contextual/events	4.0%	2.2%
Dates/times/seasons	4.6%	1.8%
Locations	30.2%	5.0%
Brands/people	5.4%	2.4%
All named entities	35.6%	7.4%
Other/non-visual	14.2%	19.0%

We further select a *machine learnable* subset in this vocabulary. We start from the visual tags in the top 500 list, we group synonyms, as well as related but visually indistinguishable categories such as *cat* and *dog* [9]. We trained 105 classifiers, and selected a subset of 60 classifiers that perform well and have sufficient coverage on an independent validation set. Details on the final taxonomy structure and tag value can be found in Section 6 and 8.1.

5. CLASSIFIER SCORE CALIBRATION

We use an ensemble of bagged SVM classifiers to generate initial classification scores [18, 35]. Calibrating SVM output scores is a problem well known among SVM users. It includes estimating output probability from the score, or choosing a cut-off threshold. Let $s \in \mathcal{R}$ denote classification score, $y = \pm 1$ denote the binary class label. Vapnik [31] parameterizes y and s in the feature space with a cosine series expansion in a direction orthogonal to the separating hyperplane. Directly parameterizing the output score in the feature space requires solving a linear system for every evaluation of the SVM, which is difficult to carry out in high-dimensional nonlinear feature space. Platt [23] proposed a logistic model for the class posterior given a score, i.e. $p(y = 1|s) = \frac{1}{1+exp(as+b)}$. Platt's model is learnable

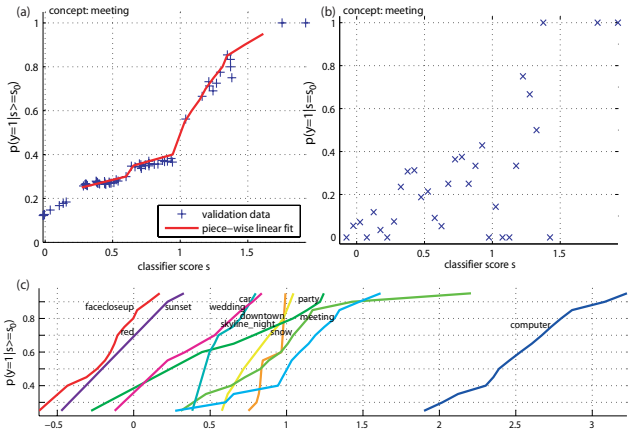

Figure 3: Classifier score calibration. (a) Validation data and non-parametric estimate of accumulated precision for concept *meeting*. (b) Local precision (as modeled by [23]) for concept *meeting*. (c) An overview of non-parametric estimates for 11 concepts.

from data, simple, and widely used. However, noise is often too great for the learning to yield meaningful results, and a single logistics model has difficulty fitting an SVM ensemble when each unit classifier has different probability mappings.

We propose the following smoothed non-parametric estimate to calibrate SVM scores. We obtain empirical estimates of $P(y = 1|s \geq s_0)$ from a separate validation set, and we perform piece-wise linear fit to the observations. For each positive data point with score s_i in the validation set, $P(y = 1|s \geq s_i)$ is taken as the fraction of positive points with scores no less than s_i:

$$P(y = 1|s \geq s_i) = \frac{\#(y = 1, s \geq s_i)}{\#(s \geq s_i)}.$$

We choose a series of control points p_c at every 0.05 interval in $[0.25, 0.95]$. We then estimate value of s_c that satisfies $P(y = 1|s \geq s_c) = p_c$. We use a five-point local triangle window to smooth the estimates, ensuring that the (p_c, s_c) pairs are monotonically increasing. Using such a non-parametric model on the cumulative probability of y is more robust to outliers since the precision is computed over a range of s instead of a local neighborhood, and it can adapt to uneven local gradients over s vs p, commonly seen with a classifier ensemble. Furthermore, $P(y = 1|s \geq s_0)$ gives a direct estimate of accuracy if we were to threshold scores at s_0.

Examples of such probability calibrations can be found in Figure 3. Figures 3(a) and (b) show the validation data points on the same visual concept *meeting*, plotting accumulated or local precision estimates vs. raw classifier scores. We can see that this is unsuitable for fitting local precision estimates [23], while we can still get a smoothed accumulated precision. Figure 3(c) gives a snapshot of calibrated scores of 11 concepts. Note that the classifier scores are in very different numeric ranges, making calibration an essential step for tag ranking. Detailed evaluation setup and validation data information can be found in Section 8.

6. TAXONOMY-BASED TAG REFINEMENT

We use a faceted tag taxonomy, which encodes external knowledge about the relationships and structure between the target visual tags, and which can be used to eliminate

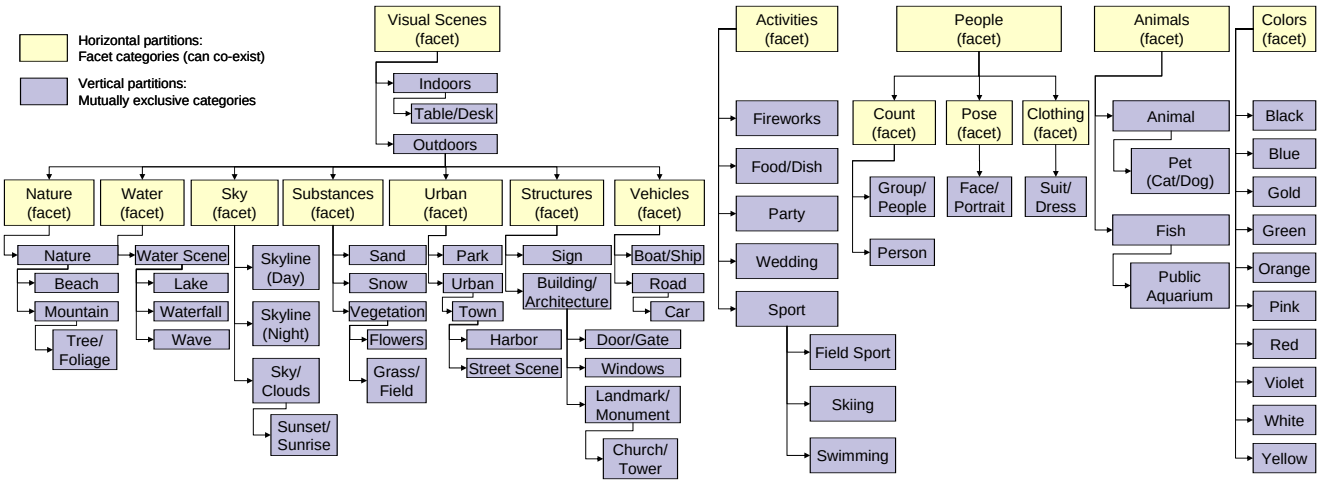

Figure 4: Hierarchical multi-faceted taxonomy of 60 target visual tags. Yellow nodes represent *faceted tags* that capture independent visual aspects and can co-occur; blue nodes represent *mutually exclusive tags*.

conflicting tags and to automatically infer additional correct ones. A key question is how to encode the tag relationship information so that it can be used for automatic reasoning.

While there are various large concept ontologies based on linguistic relationships [25, 10, 17], visual vocabularies are much scarcer, The Large-Scale Concept Ontology for Multimedia, or LSCOM [19], is arguably the largest visual ontology, consisting of over 2,000 visual concepts related to broadcast news video that are linked in to a subset of Cyc [25]. However, LSCOM was primarily designed to meet the needs of broadcasters and analysts of professionally produced content rather than social media users. Furthermore, traditional hierarchical tree taxonomies are not very suitable for automatic reasoning and refinement of visual categories since the latter are inherently fuzzy. For example, a photo from inside a room, looking through a window, depicts both indoor and outdoor aspects, which is a virtual impossibility if we consider only the semantic relationships of these two concepts. Similarly, a photo can depict many concepts at the same time, and is therefore unlike words, which is what linguistic taxonomies are designed to organize.

We tackle the above problems by introducing a *faceted taxonomy* of visual concepts, where faceted nodes represent independent visual aspects that can co-occur within an image, and regular category nodes represent mutually exclusive concepts that rarely co-occur in an image. The proposed faceted taxonomy of the 60 target visual concepts is illustrated in Figure 4. Note that taxonomy design is subjective, and there are always examples that violate the encoded relationships, frequently mutual exclusivity. The above taxonomy is simply a tool to minimize tagging errors and improve overall tagging quality, even if some of the relationships can clearly be violated (e.g., we intentionally force a choice between *door* and *windows* in order to emphasize the *dominant* aspect of images).

The proposed faceted taxonomy allows us to improve the set of recommended tags for an image in a number of ways:

- **Precision:** We can eliminate conflicting tags that appear as mutually exclusive siblings in the taxonomy by selecting at most one such sibling (e.g., choose *indoor* vs. *outdoor*, *sand* vs. *snow* vs. *vegetation*, etc.).
- **Recall:** We can augment the set of tags by propagating tag confidence scores bottom-up in the taxonomy (e.g., *mountain* implies *nature*).
- **Clarity:** We can disambiguate meaning of otherwise similar tags by explicitly encoding parent-child relationships between them (e.g., *urban* → *town*).
- **Usability:** We can prioritize and re-rank tags based on their depth in the taxonomy, since users would typically prefer to see more specific tags (e.g., *church/tower* ≫ *landmark* ≫ *building* ≫ *urban*).

In Section 8, we report the results of a 24-user evaluation on various aspects related to the visual tag vocabulary and taxonomy, including tag usability and perceived user value (Section 8.1), accuracy of the taxonomy-refined tags (Section 8.2), and user preference assessment of several tag re-ranking approaches (Section 8.3).

7. VALUE-BASED TAG RE-RANKING

The best tag for an image isn't necessarily the one with the highest precision estimate. Tag recommendation under uncertainty tries to find a trade-off between the accuracy of tags being issued, the perceived usefulness of each tag, and the risk of issuing a wrong tag. In the following, we present several approaches for obtaining an overall *tag importance value*, and we multiply the precision-normalized tag scores by the estimated tag values to re-rank suggested tags based on a combination of tag relevance and tag importance. Here we use c to denote a "visual tag", i.e. one that is distinguishable by a visual classier here, and u to denote its social tag counterpart such as those found on Flickr.

7.1 Re-ranking by perceived tag value

We surveyed a group of users, asking each of them to rate the subjective value of each tag on a scale of 1-5 (Section 8.1). We take the mean value across all users, denoted as $f_{user}(u)$. While this is a direct way to measure perceived tag values, it would be difficult to scale up the number of tags and the number of representative users.

7.2 Re-ranking by tag information content

One well-known method for measuring the usefulness of observing a probabilistic event (e.g. observing visual concept c) is the information content (IC) [8], measured as the negative logarithm on the probability of the event. The IC re-ranking factor for each visual concept is simply:

$$f_{ic}(c) = IC(c) = \log \frac{1}{P(c)} = -\log(P(c)). \quad (1)$$

Information content is a measure of specificity, and will naturally boost more specific tags, whose presence in a given image is considered more informative than the presence of common or generic tags. We estimate $P(c)$ from a partially-labeled validation set by thresholding probability-calibrated classification scores (Section 5) and estimating the corresponding concept frequency. The information content of our tag vocabulary ranges from 2 bits (e.g. *outdoors, nature*) to 9 bits (e.g *boat/ship, fireworks*). This estimate can be noisy due to the incompleteness of ground-truth, and assigning a higher value to rarer concepts does not take into account the inherent preferences users may have for some topics.

7.3 Re-ranking by tag popularity on Flickr

Another heuristic for tag ranking is that tags used more frequently by a large number of users are more valuable. For example, *wedding* and *party* are both among the most popular tags on Flickr. *Wedding* is more specialized than *party*, occurring much less frequently, yet photos tagged with *wedding* outnumber those with *party* at 12.2 million to 8.7 million. Therefore tag counts on Flickr can serve as a "Flickr measure" of tag importance. Although this approach may favor some of the more frequent tags, such as *nature* with 7.8 million tagged images, such tags are still outnumbered by the less frequent but more important concepts such as *wedding*. To compute the Flickr value, we map each visual tag to one or more Flickr tags, accounting for word morphology, synonyms and most related tags, such as *animal* and *animals*, *flower* and *blossom*. We collect the tag counts using the Flickr API, and take the sum of the counts from different tag variants. We use the log of tag counts as the ranking factor in order to smooth out noise and large scale differences in a scale-free tag collection.

$$f_{flickr}(u) = \log(\#(u)). \quad (2)$$

7.4 Re-ranking by tag posterior probability

Ideally a re-ranking scheme should be able to automatically incorporate both the "unexpectedness" of a visual tag and its "usefulness", as vetted by users. The less expected a tag is and the higher its utility, the larger its importance for ranking purposes. We examine the probability $P(u|I)$ of a tag u being assigned to an image I by an arbitrary user. We unroll this conditional probability on the actual presence of a corresponding visual tag c in the image:

$$P(u|I) = P(c|I)\cdot P(u|c,I) + P(\bar{c}|I)\cdot P(u|\bar{c},I) \approx P(c|I)\cdot P(u|c)$$

We can further assume that u is independent of I once c is given, which simplifies the first term to $P(c|I)\cdot P(u|c)$. This is a reasonable assumption to make, essentially stating that once the relevance of a tag to a given image is known, the image itself is no longer required to determine if the user will apply the corresponding tag, and the latter becomes a function only of the user tagging preferences. Furthermore, the second component in the above formulation is a product of two terms involving the absence of the visual tag, $\bar{c}$. Note that $P(u|\bar{c},I)$ encodes the chance that a user will apply the tag to the image even though the tag is not relevant to the image, essentially producing a semantic "false alarm". While this is unfortunately a frequent phenomenon in user photo tagging (e.g., bulk-tagging all vacation photos as *beach*), it is certainly an undesirable behavior for an automatic visual tagging system. We therefore set this factor to zero and only model the first component of $P(u|I)$. For the remaining part, we apply Bayes rule to further factorize $P(u|c)$, yielding three terms, $P(u)$, $P(c)$, and $P(c|u)$:

$$P(u|I) \approx P(c|I)\cdot\frac{P(u)}{P(c)}\cdot P(c|u) \approx \alpha P_c \frac{P(u)}{P(c)} \quad (3)$$

$P(c)$ is the prior probability of observing visual tag c as visually present in a random image, which can be estimated from large corpus statistics. $P(u)$ is the prior probability of an arbitrary user applying tag u to an arbitrary image, which can be estimated as the fraction of Flickr photos bearing tag u. For $P(c)$ we use the estimated validation set prior as in Section 7.2; for $P(u)$ we use the the tag count from Section 7.3 divided by 4 billion, the last published Flickr photo count.$P(c|u)$ is the fraction of photos tagged with a given tag that are actually relevant to that tag. For a large photo pool, this should be a tag-agnostic constant (denoted as α), as noted earlier [24]. We denote the photo-specific confidence $P(c|I)$ with shorthand P_c, and we estimate it from the calibrated classifier scores, as detailed in Section 5.

We name the ratio $P(u)/P(c)$ the likelihood-ratio factor, and use it to re-weight the calibrated confidence values,P_c, for tag re-ranking purposes. Note that this re-weighting factor combines the heuristics from f_{ic} and f_{flickr} above as it gives more weight to popular tags, yet de-emphasizes tags that are not specific and discriminant, performing a trade-off with the estimated classification accuracy.

$$f_{lr}(u,c) = \frac{P(u)}{P(c)} \quad (4)$$

Each of the four factors f_{ic}, f_{flickr}, f_{lr} and f_{user} serve as a re-scoring factor for the classifier confidence among the different tags within the same image. We simply multiply these factors with the classifier confidence.

$$P_s^*(u) = P_s \cdot f(u)$$

Note that a few pre-processing steps and assumptions have made the construction of our weighting factor easy. Mapping both the visual and text tags to the same vocabulary makes each of the f_{ic}, f_{flickr}, f_{lr} and f_{user} a single score list on all 60 concepts. We have not considered multiple word senses (e.g. *apple* being both fruit and a class of electronic products) as they did not seem prevalent in our controlled vocabulary of 60 visual tags, while models from prior research [34] can be used to disambiguate tags.

8. EVALUATION

We performed several user studies in order to evaluate each aspect of the proposed tagging system: the usefulness of the chosen tag vocabulary, the accuracy of the generated tags, and the quality of the tag re-ranking approaches. The user studies comprised of 24 volunteers and a total of 5245 photos donated by them for this evaluation. Travel pictures, celebrations, natural scenery, tourism, and city street scenes were the most common elements, representing typical consumer imagery. We received multiple forms of feedback from the users, which included responses to a detailed survey questionnaire, over 35,000 manually-provided or ver-

ified tags, and over 11,000 personal preference judgments comparing pairs of tagging results.

8.1 Tag vocabulary evaluation

In order to evaluate the usefulness of the tag vocabulary, we conducted a user survey. The participants were asked to assess the perceived value of each target visual tag on a 5-point scale, ranging from 1 (useless) to 5 (very useful). A summary of the user responses, aggregated across all 25 responders, is presented in Figure 5(a). On average, the target tags received a value score of 3.6, with a standard deviation of 1. Over 40% of the target tags received average value scores ≥ 4 (i.e., rating of *Quite Useful* or above), and over 80% of the tags received scores ≥ 3 (i.e., *Somewhat Useful* or above). The remaining ˜20%, or 11 tags, were predominantly the color tags, which were perceived as not very useful by our users, even though they are in the top-100 most popular tags both on Flickr and the ESP game.

In addition to tag usefulness, the survey participants were also asked to assess how many incorrect tags they would tolerate per image as a fraction of the total number of auto-generated tags. We asked this question in two different ways: 1) *How many incorrect tags (out of 10) would you tolerate and still consider the results useful?* and 2) *If you would tolerate 1 incorrect tags along with X correct tags, what would X be?* We then converted the responses into corresponding precision ranges, resulting in minimum user-acceptable precision of $74 \pm 6\%$ on question 1 and $82 \pm 5\%$ on question 2, indicating a user-acceptable precision range of $74\% \sim 82\%$. Based on this finding, we set the target precision of our system to 80% as a good trade-off between precision and recall.

8.2 Tag accuracy evaluation

We evaluate tag accuracy on 5245 user-donated photos. The photos were uploaded to a Flickr account, the users enter tags they deem useful, as well as verify the machine-generated tags as correct or not, by deleting the latter. The user-entered tags were then manually mapped to the common tag vocabulary, where possible. We also required users to delete a special *NotYetReviewed* tag for each of their photos in order to ensure that each image had been manually inspected. The resulting mix of ˜30K manual and auto-generated but manually verified tags formed the basis of our ground-truth for accuracy evaluation. The distribution of user-entered vs. auto-generated-but-manually-verified tags was 46% to 54%, indicating that we more than doubled the completeness of the ground-truth by assisting users in the tagging process. We further propagated relevance judgments from specific tags to more generic ones using the taxonomy described in Section 6 (i.e., *lake* implies *water scene* and *outdoors*), adding ˜6K extra tags to the ground-truth.

Based on the above ground-truth, we quantitatively evaluate the proposed classifier score calibration (Section 5) and taxonomy expansion (Section 6) methods. We collate all (score, tag, image) tuples, sort them by score in a descending order, and compute "micro-precision" on this ranked list, i.e., the fraction of tags that are correct at any given depth n (divided by the total number of images). Figure 6 shows the comparison among two baseline runs and our methods. The purple square run is generated by sorting the raw classifier scores, equivalent to the "unsupervised" probability estimation method in [23], this run performs poorly mainly due to the significantly different ranges among different classifiers (c.f. Figure 3(c)). The blue circle run is another baseline with manually selected calibration threshold for each concept (corresponding roughly to .5 local precision in a neighborhood of 50 images), and then range-normalizing the scores above the threshold to between 0 and 1. The green diamond run uses scores calibrated with the non-parametric precision estimate from Section 5, and the red star run is generated using the green diamond run after taxonomy-based refinement (Section 6). From the performance comparison, we can see that the probability score calibration is a critical step, and the proposed approach outperforms even the manually chosen thresholds and calibration points. The taxonomy-based run further improves the results by producing less errors (due to eliminating conflicting tags) and more complete tags (due to taxonomy-based expansion), resulting in the best precision for any given number of issued tags. Overall, the system achieves 83% precision with an average of four tags issued per image, which meets or exceeds the threshold of minimum user-acceptable accuracy for image tagging applications, as described in Section 8.1.

8.3 Tag re-ranking evaluation

We conduct a user study on the tag re-ranking methods (Section 7), to validate the selection of more *relevant* or *valuable* tags, and not just the *correct* ones. The baseline run here is the score-calibrated and taxonomy-expanded tags as is, with no re-ranking. Specifically, we created a web application which presents the user with an image and the top 4 tags assigned to the image, from each of two target runs being compared. Any tags that were in common between the two sets were grouped together, and the tags which differed were presented in a column on the left or right. We randomized the right / left placement of the differing tags with each image in order to eliminate any unintentional user bias. The user would then choose either the left or right side as being superior, based on their own subjective judgment of the usefulness of the tags. There was a third choice as well, to indicate indifference between the two sets. We had 5 users participate in this study, giving their preferences for all possible pairs of 5 runs, a total of ten pairwise comparisons. We did this for a subset of 300 images in which there was significant variation in the tags among the 5 runs. The 5 users provided 11,700 preference choices, and each pairwise comparison was evaluated by 4 users on average.

We consider aggregate statistics across all 10 pairwise comparisons in order to derive overall rankings of the 4 tag re-ranking methods and the baseline. For each pairwise comparison, we first measure fraction of votes cast for each of

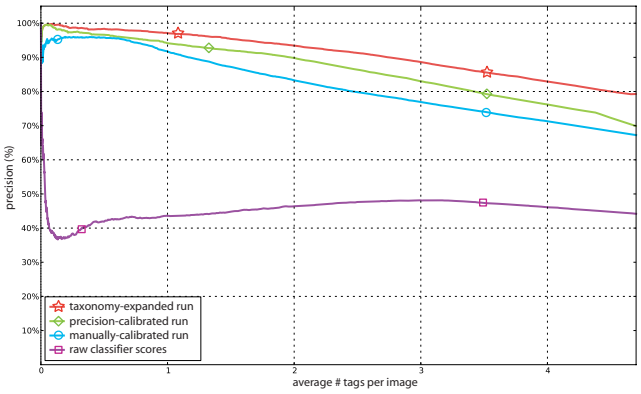

Figure 6: Comparison of tagging approaches.

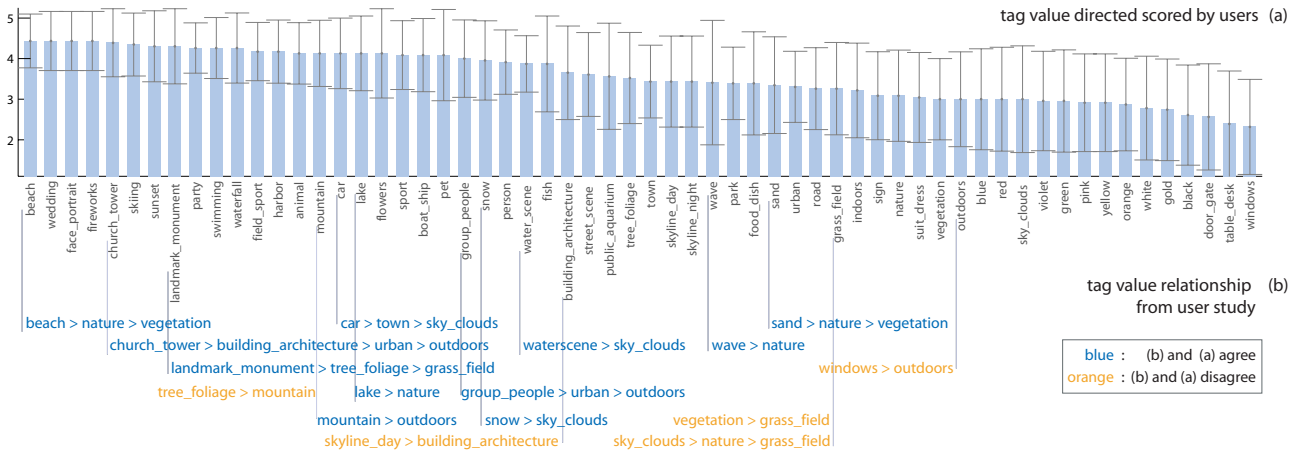

Figure 5: Tag values. (a,top) From user survey (Section 8.1), tag are sorted in descending value, the error bars denote variance. (b,bottom) Discovered relationships in tag re-ranking experiments (Section 8.3).

the two runs being compared, and then average these fractions across all comparisons each run participates in. This gives us an overall measure of how frequently each run "wins" when compared against any of the other runs, similar to a group tournament ranking system. We deal with ties in two different ways: a) in one alternative, we split the ties equally among the two runs (i.e., give half votes to each run for each tie); and b) we consider only the subset of votes where voting users agree on the winner unanimously, and we discard tie votes and votes where users contradict each other. We note that consensus is reached in approximately 57% of the total votes.

The results of the aggregate user preferences are summarized in Table 2. Results for option a), which counts ties as half votes, are presented in the third column of Table 2, while option b), which considers only consensus non-tie votes, is reflected in the middle column. We note that the overall ranking is the same according to both aggregate measures, although the consensus column can discriminate a bit more effectively. From the results, we can conclude (with statistical significance at $p = 0.05$) that 1) any re-ranking approach is better than no re-ranking (the baseline); 2) the user value-based, information content-based and Flickr popularity-based methods are not statistically different, that is, we can recover the performance of explicitly provided user value-based re-ranking with approaches based on automatically computed value estimates; and 3) the posterior probability-based approach outperforms all other approaches. These aggregate performance observations are generally confirmed when looking at individual pairwise comparisons, which are omitted here due to space limitations.

We also use the ten pairwise run comparisons to extract individual tag relationships. We extract the subset of top-four tags that differ, and we add one "winning" vote to the set being preferred when there is a consensus among the annotators. Specifically, we only accumulate votes when both sets are in the ground-truth (1,535 pairs total, out of which 784 have annotator consensus), eliminating cases where one set of tags is being preferred simply because the other set is incorrect. For any image I, let the sets of tags that differ be $\mathcal{T}^*{}_1$, and $\mathcal{T}^*{}_2$, each with size k tags. We add $1/k$ votes to tag pair ordering (t_1, t_2), and subtract $1/k$ to tag pair ordering (t_2, t_1), $\forall\ t_1 \in \mathcal{T}^*{}_1,\ t_2 \in \mathcal{T}^*{}_2$. The votes are normalized by the total number of times a tag-pair is voted on. From this result we extract both pairwise tag preferences and the accumulated votes on any tag to obtain a sort order. Out of the 60 tags, 25 tags and 83 pairs (with votes ≥ 0.1) can be meaningfully compared.

Table 2: Comparisons of baseline and 4 tag re-ranking methods from Section 7. Results aggregated across 11,700 votes from 5 users for 10 pairwise comparisons. Each number represents fraction of votes received by the corresponding re-ranking approach when compared against the other methods.

Re-Ranking Method	%Winning Votes	
	(consensus)	(all votes)
Baseline (no re-ranking)	39.0%	44.7%
User perceived value	49.9%	49.8%
Information Content	51.7%	49.9%
Flickr popularity	52.7%	50.5%
Posterior probability	56.7%	55.1%

Figure 5(b) shows a few examples of tag preferences—the full result set has 62 pairs that agree with Figure 5(a) and 21 pairs that do not. We can see that the users tend to prefer more specific tags in both scoring and image comparison (e.g. *beach*>*nature*, *church/tower*>*building*), and that the preference over specific tag pairs may be ambiguous or may also vary depending on image context (e.g. *windows* vs. *outdoors*, *tree/foliage* vs. *mountain*). Figure 7 shows the 25 tags sorted in descending order of the total votes they received (against each other). While the gross ballpark of tags agrees with the user scores, e.g. *beach, snow* among the most preferred, and *outdoors, urban* among the least, individual tag order may vary, suggesting there may be a few equivalence tag classes.

9. CONCLUSIONS

In this paper, we have presented a detailed case study in the design and evaluation of an end-to-end image tagging system for consumer photos. We proposed a methodology for extracting meaningful visual tag vocabularies for image auto-tagging systems, and defined a sample 5000-tag vocabulary and a subset 60-tag taxonomy. We proposed a novel faceted taxonomy structure for capturing both co-occurrence and mutual exclusivity relationships across tags, and proposed methods for automatic taxonomy-based tag

refinement, which can increase both recall and precision. We presented a classifier score calibration method based on non-parametric precision estimates, which boosts overall system accuracy from less than 50% to over 80%. Finally, we proposed and evaluated four tag re-ranking approaches based on various estimates of perceived tag values. All aspects of the system are evaluated with several user studies, including over 20 users, 5,000 photos, 35,000 tag judgments, and 11,000 tagging preference judgments. The experiments validate the utility and accuracy of the chosen tags, and confirm that value-ranked tags are preferable to accuracy-based tags.

Acknowledgment

This work builds upon the foundations of the IBM Multimedia Analysis and Retrieval System (IMARS) [2]. While we focus on orthogonal aspects for this paper, the work would not have been possible without the efforts of the extended IMARS team, to which we are deeply grateful. We would like to acknowledge Rong Yan in particular for his key contributions to the current IMARS classification system [18, 35], and for creating the visual classifiers used in this paper. We also thank Quoc-Bao Nyugen for system design input, and Ambreen Javed for annotation. Last but not least, we are very grateful to our enthusiastic group of users for sharing photos and providing feedback.

10. REFERENCES

[1] Flickr API. `http://www.flickr.com/services/api/`.
[2] IBM Multimedia Analysis and Retrieval System. `http://www.alphaworks.ibm.com/tech/imars`.
[3] The PASCAL visual object classes homepage. `http://pascallin.ecs.soton.ac.uk/challenges/VOC/`.
[4] M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. In *Proc. CHI*, pages 971–980, 2007.
[5] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, 2003.
[6] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu. Can all tags be used for search? In *Proc. ACM CIKM*, pages 193–202, 2008.
[7] T. Coates. Two cultures of fauxonomies collide... `http://www.plasticbag.org/archives/2005/06/`, 2005.
[8] T. M. Cover and J. Thomas. *Elements of information theory*. Wiley, 1991.
[9] J. Elson, J. R. Douceur, J. Howell, , and J. Saul. Asirra: A captcha that exploits interest-aligned manual image categorization. In *ACM CCS '07*, 2007.
[10] C. Fellbaum et al. *WordNet: An electronic lexical database*. MIT press Cambridge, MA, 1998.
[11] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. *Journal of Artificial Intelligence Research*, 29, 2007.
[12] N. Garg and I. Weber. Personalized tag suggestion for Flickr. In *Proc. WWW*, pages 1063–1064, 2008.
[13] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2), 2006.
[14] L. Kennedy, M. Slaney, and K. Weinberger. Reliable tags using image similarity: mining specificity and expertise from large-scale multimedia databases. In *WSMC '09: Proc. of workshop on Web-scale multimedia corpus*, pages 17–24, 2009.
[15] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *IEEE Trans. PAMI*, 30(6):985–1002, 2008.
[16] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *Proc. WWW*, pages 351–360, 2009.
[17] H. Liu and P. Singh. ConceptNet: a practical commonsense reasoning toolkit. *BT Tech. Journal*, 22(4):211–226, 2004.
[18] M. Campbell et. al. IBM research TRECVID-2007 video retrieval system. *TREC Video Retrieval Evaluation Online Proceeding*, 2007.
[19] M. Naphade et. al. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86–91, 2006.
[20] C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proc. the 17th conference on Hypertext and hypermedia*, page 40, 2006.
[21] M. Marszałek and C. Schmid. Semantic hierarchies for visual object recognition. In *Proc. CVPR*, jun 2007.
[22] P. Bolettieri et. al. CoPhIR: a test collection for content-based image retrieval. *CoRR*, 2009.
[23] J. Platt. Probabilities for SV machines. *Advances in Neural Information Processing Systems*, pages 61–74, 1999.
[24] T. Rattenbury, N. Good, and M. Naaman. Towards extracting Flickr tag semantics. In *Proc. WWW '07*, pages 1287–1288, 2007.
[25] S. Reed and D. Lenat. Mapping ontologies into cyc. In *Proc. AAAI Conference 2002 Workshop on Ontologies for the Semantic Web*, pages 02–11, 2002.
[26] R. Shi, C.-H. Lee, and T.-S. Chua. Enhancing image annotation by integrating concept ontology and text-based bayesian learning model. In *Proc. ACM Multimedia*, pages 341–344, New York, NY, USA, 2007.
[27] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. WWW '08*, pages 327–336, 2008.
[28] A. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, page 330, 2006.
[29] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.
[30] K. Van De Sande, T. Gevers, and C. Snoek. Evaluation of color descriptors for object and scene recognition. In *Proc. IEEE CVPR*, page 1, 2008.
[31] V. Vapnik. *The nature of statistical learning theory*. Springer Verlag, 2000.
[32] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *Proc. of the Intl. Conf. on Computer Vision (ICCV)*, 2009.
[33] L. Von Ahn and L. Dabbish. Labeling images with a computer game. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pages 319–326, 2004.
[34] K. Q. Weinberger, M. Slaney, and R. Van Zwol. Resolving tag ambiguity. In *Proc. ACM Multimedia*, pages 111–120, New York, NY, USA, 2008.
[35] R. Yan, J. Tesic, and J. R. Smith. Model-shared subspace boosting for multi-label classification. In *Proc. ACM KDD*, pages 834–843, 2007.

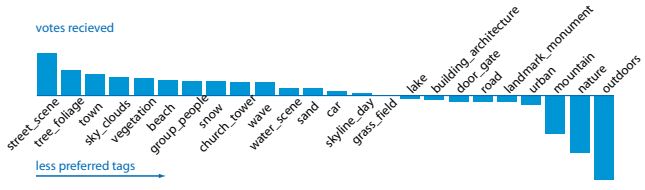

Figure 7: Tag order derived from the user re-ranking evaluation. x-axis: tags sorted in descending preference. y-axis: average score of a tag over all contending tags.