Personalizing Automated Image Annotation using Cross-Entropy

Xirong Li Efstratios Gavves Cees G.M. Snoek
Marcel Worring Arnold W.M. Smeulders

Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam
Science Park 904, 1098XH Amsterdam, The Netherlands
{x.li,e.gavves,c.g.m.snoek,m.worring,a.w.m.smeulders}@uva.nl

ABSTRACT

Annotating the increasing amounts of user-contributed images in a personalized manner is in great demand. However, this demand is largely ignored by the mainstream of automated image annotation research. In this paper we aim for personalizing automated image annotation by jointly exploiting personalized tag statistics and content-based image annotation. We propose a cross-entropy based learning algorithm which personalizes a generic annotation model by learning from a user's multimedia tagging history. Using cross-entropy-minimization based Monte Carlo sampling, the proposed algorithm optimizes the personalization process in terms of a performance measurement which can be flexibly chosen. Automatic image annotation experiments with 5,315 realistic users in the social web show that the proposed method compares favorably to a generic image annotation method and a method using personalized tag statistics only. For 4,442 users the performance improves, where for 1,088 users the absolute performance gain is at least 0.05 in terms of average precision. The results show the value of the proposed method.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*; I.2.6 [**Artificial Intelligence**]: Learning—*Concept learning*

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Automated image annotation, Personalization, Personal multimedia tagging history, Cross-entropy optimization

Area chair: Wu-chi Feng

1. INTRODUCTION

Annotating large personal collections of pictures on smart phones, personal computers, and the web is of great social importance. With the size of such collections growing so rapidly, full manual annotation is unfeasible. Thus, automatic image annotation is crucial, but this is challenging due to the well-known semantic gap: "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for *a user in a given situation*" [24]. Much work has been conducted to (partially) bridge the gap by learning a mapping between visual features and objective semantics [6,9,20,22,26,27]. However, in the above efforts the *user* factor in the semantic gap is completely ignored. Clearly, users have personal preferences for image subjects. For instance, some users collect pictures of flowers, while others may favor images of cars. This real-world phenomenon suggests that an off-the-shelf image annotation system is unlikely to be universally applicable to the large variations in personal albums. The absence of personal information in devising image annotation models results in unsatisfactory annotations.

Some research has been conducted towards personalizing automated image annotation [4,5,14,21]. Sawant et al. were among the first to leverage user tagging preferences, with their novel observation that "a user's previously used tags can be the best determinants of her future uploads" [21]. Interestingly, their study revealed that simply annotating a user's new images with tags most frequently used by the same user in the past yields a much higher accuracy than several automated methods. This result leads the authors to conclude that prediction by personal tag statistics is a reasonable upper bound on personalized image annotation performance.

In this paper we study the problem of *personalizing automated* image annotation in a social web context. We tackle the problem by proposing a generic framework which jointly exploits generic content-based image annotation and personal multimedia tagging history. We present a learning algorithm which optimizes the personalization process in terms of a performance measurement which can be arbitrarily chosen. Due to this technical advantage, we go beyond the performance upper bound of personalized image annotation established in [21].

MM'11, November 28–December 1, 2011, Scotsdale, Arizona, USA.

2. RELATED WORK

Instead of proposing a new generic image annotation model, this paper studies the personalization of automated image annotation. We first review recent progress in generic image annotation, and then we discuss related work on image annotation personalization.

2.1 Generic Image Annotation

A considerable amount of papers have been published for generic image annotation [6, 9, 13, 15, 22, 23, 25, 26]. We divide existing work into content-based methods and content-context-based methods.

The content-based methods predict tags purely based on image content analysis [6, 9, 13, 22, 25, 26]. Li and Wang [9] train a multivariate Gaussian mixture model for each tag, while Support Vector Machines are used in [22]. Liu et al. [13] annotate images by maximizing the joint probability of images and tags. The authors in [26] perform image annotation by learning a mapping into a common feature space where both images and tags are represented. They rank tags in terms of their distance to a test image in the common space. In contrast to per-tag modeling, k-nearest-neighbors based methods make predictions by propagating tags to the unlabeled image from its visual neighbors [16]. Weighted nearest neighbors are considered in [6]. Sparse reconstructions are employed in [25] to reduce the chance of incorrectly including neighbors which are semantically irrelevant to the unlabeled image. To enhance content-based image annotation, contextual information on the creation of an unlabeled image has been investigated [15, 23]. In [15], GPS data, indicating where the image was captured, is employed, whereas in [23], camera metadata such as shutter speed and focal length describing how the image was captured is studied. In both content-based and content-context-based methods, all users are treated equally, without taking personal preferences into account.

Work such as [3, 12, 29] studies learning image annotation models from social-tagged images. Datta et al. [3] treat user tags as positive feedback to incrementally update an existing model. Liu et al. [12] and Zhu et al. [29] measure both image-wise visual similarity and tag-wise semantic similarity to refine existing annotations. These methods learn from the social community but do not consider personalized image annotation.

2.2 Personalized Image Annotation

Recently some papers have appeared towards automated approaches to personalized image annotation [4, 5, 14, 21]. According to whether user interaction is needed, we divide existing work into two types of methods, automatic methods and semi-automatic methods.

The automatic methods achieve personalization by inferring from personal digital calendars [5] or multimedia tagging history [21], or training a generic model on personal collections [4]. In [5], Gallagher et al. explore the possibility of using personal calendar event annotations to label images. The rationale for the idea is based on the coincidence between the calendar event and image capture time. Calendar annotations are not always available. More importantly, tagging a calendar event is different from tagging an image. Therefore, personalizing image annotation based on the calendar tagging history seems questionable. Assuming that context-constrained images such as those captured at the same location have a similar visual style, Duan et al. [4] propose a probabilistic model where styles are viewed as latent variables. While learning from visual features of high dimensionality requires many labeled examples, the number of personal images for a specific user is relatively small and many of them are unlabeled. Hence, learning models using personal collections alone seems problematic. To overcome the problem, Liu et al. [14] propose a semi-automatic method, by first learning a generic model for each tag using images from a professional photo forum. They then solicit user feedback to adapt the learned model to personal collections. The difficulty in obtaining user feedback for thousands of tags puts the scalability of the semi-automatic method into question. Moreover, in [4,5,14], personal tagging history, a strong clue for building personalized annotation models, is untouched.

In [21], Sawant et al. propose to combine personal tagging history, in the form of tag frequency, and predictions made by a content-based image annotation system [9] in a Naïve Bayes formulation. They conclude that combining tagging history and content analysis is inferior to using the history alone. We argue that their conclusion is true but for their Naïve Bayes model only. In that model the performance of the individual pieces of evidence is not considered. In contrast, we propose a personalization model which is directly optimized in terms of the prior annotation performance. As a consequence, we reach the novel conclusion that combining the personal tagging history and content analysis yields better personalized image annotation.

The rest of the paper is organized as follows. We formulate the personalization problem and elaborate the proposed personalization model in Section 3. We setup experiments in Section 4, with results analyzed in Section 5. We conclude the paper in Section 6.

3. PERSONALIZED IMAGE ANNOTATION

3.1 Problem Formalization

Let u be a user for whom we want to provide personalized image annotation. Let $X_{u,past}$ be a set of images the user has already labeled, and $X_{u,future}$ a set of unlabeled images the user wants to have tags for. We use w to denote a tag and $V = \{w_1, \ldots, w_m\}$ for a large vocabulary. For each image $x \in X_{u,future}$, we aim to annotate it with tags from the vocabulary such that the annotations are relevant with respect to the image from the user's standpoint. To do so, we use both information from the user as well as the social web community. So let X_{comm} indicate images in the community, and $X_{u,past}$ is a subset of X_{comm}. We define $G_u(x, w)$ as a *personalized* image annotation function whose output is a confidence score of the tag w being relevant to the image x. This allows us to rank tags by $G_u(x, w)$ in descending order and preserve the top ranked tags as annotations of unlabeled images for this particular user.

Personalized image annotation is a complex task. It is unlikely that an image annotation function based on a single modality can capture all relevant characteristics. We therefore need to look at the problem from multiple perspectives. Ideally, we exploit varying evidence such as content-based image annotation driven by diverse features [6, 11], tag statistics in personal collections and networks [21], personal daily activities [5], geographic context [15], or camera metadata [23]. In this study, we focus on combining content-based annotation and personal tag statistics, as they are

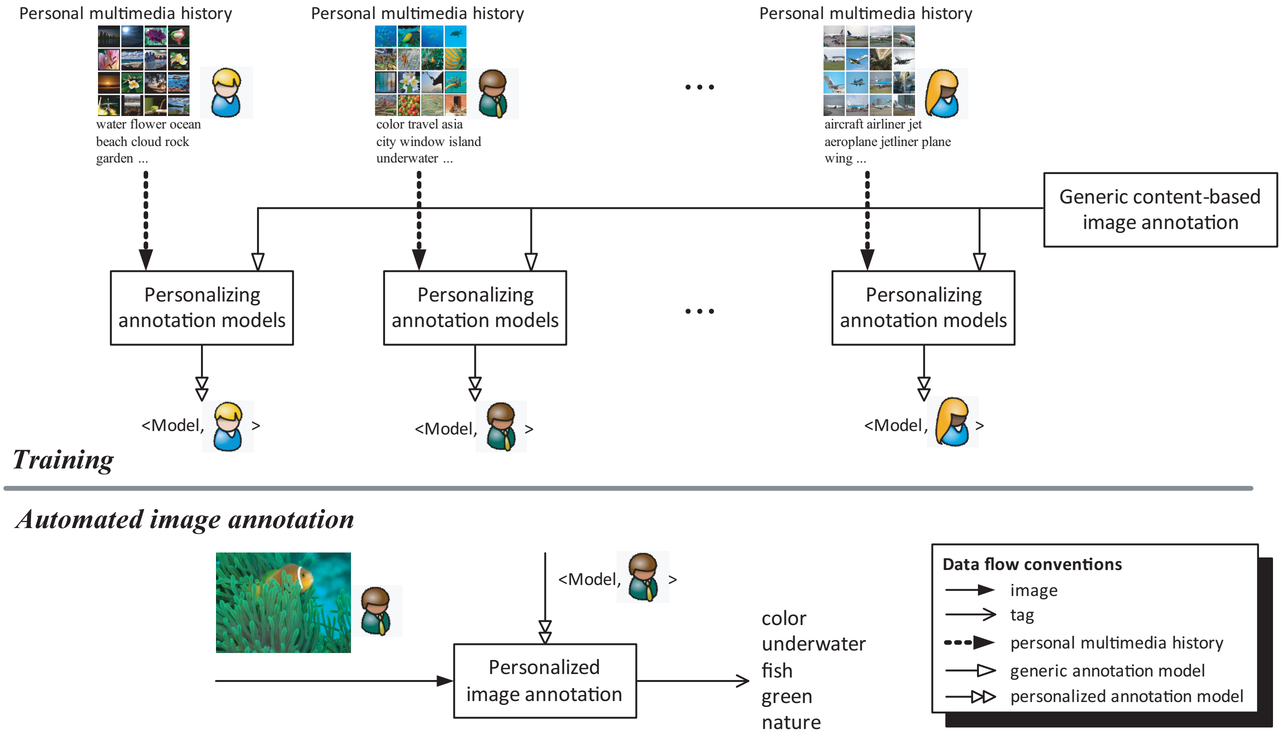

Personal multimedia history
water flower ocean
beach cloud rock
garden ...
Personal multimedia history
color travel asia
city window island
underwater ...
...
Personal multimedia history
aircraft airliner jet
aeroplane jetliner plane
wing ...
Generic content-based image annotation
Personalizing annotation models
Personalizing annotation models
...
Personalizing annotation models
<Model, >
<Model, >
<Model, >
Training
Automated image annotation
<Model, >
Personalized image annotation
color
underwater
fish
green
nature
Data flow conventions
image
tag
personal multimedia history
generic annotation model
personalized annotation model

Figure 1: The proposed framework for *personalizing* automated image annotation. For a given user, we deliver a personalized image annotation model, by jointly exploiting content-based image annotation and the user's multimedia tagging history.

two fundamental elements related to the problem and are often more accessible than other elements. Nevertheless, to make our discussion general, we consider combining annotation functions driven by multiple sources of evidence. To that end, let $\{g_1(x,w),\ldots,g_t(x,w)\}$, with $g_j(x,w) \in [0,1]$, $j = 1,\ldots,t$, be a set of such image annotation functions. As combining these functions can be viewed as a multi-modal fusion problem, we choose to use a linear weighted sum, an effective strategy for multi-modal fusion according to [1]. The importance of individual tags varies per user, so tag-dependent weights are necessary. To formalize the above notion, for each tag w_i, $i = 1,\ldots,m$, we express a parameterized version of $G_u(x,w)$ as

$$G_{u,\Lambda}(x,w_i) = \sum_{j=1}^{t} \lambda_{i,j} g_j(x,w_i), \tag{1}$$

where $\{\lambda_{i,j}\}$ are non-negative weighting parameters, and $\Lambda = [\lambda_{i,j}]_{m\times t}$ is the parameter matrix. While $\lambda_{i,j}$ indicates the importance of $g_j(x,w_i)$ for predicting w_i, their summation, namely $\sum_{j=1}^{t}\lambda_{i,j}$, reflects the importance of w_i for annotating the personal collection. By optimizing the weights per user, we obtain personalized image annotation models.

To find the optimal weights for a given user, we need information about what tags the user is likely to use for tagging her/his personal collections. We assume that a user's tagging preference is relatively consistent within a certain period. Therefore, the images the user has already labeled $X_{u,past}$ are the prime candidates. To make the above notion operational, we need to formulate an optimization goal per user. Let $rank(V|x,G_{u,\Lambda})$ be a ranking of the vocabulary V for an image $x \in X_{u,past}$, obtained by sorting tags in descending order by $G_{u,\Lambda}(x,w)$. Let $\mathbf{w}_x$ be the set of tags assigned to x by its user, serving as ground truth to assess $rank(V|x,G_{u,\Lambda})$. We define

$$E(rank(V|x,G_{u,\Lambda}),\mathbf{w}_x)$$

as a performance measure function which produces a real-valued score indicating ranking quality. As we learn from a set of images, rather than from a single image, we define a set-level performance measure as

$$S(X,\Lambda) = \frac{1}{|X|}\sum_{x\in X} E(rank(V|x,G_{u,\Lambda}),\mathbf{w}_x), \tag{2}$$

where $|\cdot|$ is the cardinality of a set. Putting everything together, we formulate the problem of personalizing automated image annotation for a given user as solving the following optimization problem:

$$\Lambda^* = \operatorname*{argmax}_{\Lambda} S(X_{u,past},\Lambda), \tag{3}$$

subject to

$$0 \leq \lambda_{i,j} \leq 1. \tag{4}$$

Solving Eq. 3 yields the optimal parameters for the personalized model defined in Eq. 1, which are then used to annotate new, yet unlabeled, images of the user. We illustrate the proposed framework in Fig. 1.

3.2 Personalization using Cross-Entropy

Finding a solution for the optimization problem in Eq. 3 is nontrivial. As the performance measure function E is often not differentiable, a standard gradient-ascent based algorithm is inapplicable. A common approach to such a problem is Monte Carlo simulation. But, when the parameter space is large, as in our case, finding Λ^* or its good approximation by random sampling is a rare event. A crude Monte Carlo approach would imply an unfeasibly large simulation effort. A solution for rare event search is offered by the cross-entropy method [19], which iteratively optimizes an arbitrary function by importance sampling. We first describe the cross-entropy method in general, and then present a cross-entropy based learning algorithm for solving Eq. 3.

3.2.1 The Cross-Entropy Method

Imagine that our goal is to maximize an objective function $S(\Lambda)$, and its maximum is found at Λ^*. The cross-entropy method [19] assumes that Λ is a random variable following a parametric distribution $p(\Lambda; \Theta)$ which is specified by an (unknown) hyper parameter Θ. In a nutshell, the method consists of the following two steps executed iteratively:

Step 1. Randomly generate n samples using $p(\Lambda; \Theta)$; We use $\{\Lambda^{(1)}, \ldots, \Lambda^{(n)}\}$ to denote the n samples.

Step 2. Select the top s samples $\{\widehat{\Lambda}^{(1)}, \ldots, \widehat{\Lambda}^{(s)}\}$ by sorting the n samples in descending order by $S(\Lambda)$, where the selected samples are called *elite* samples. Re-estimate Θ by maximum likelihood estimation on the s elite samples.

From the second step we see that the hyper parameter Θ is updated in terms of the elite samples. As a consequence, the probability of generating good samples is progressively increased, making Λ converge towards its optimal value. The procedure repeats until it hits certain stop criteria. For instance, the performance does not improve or the number of iterations exceed a given threshold. We compute Λ^* as the expectation of $p(\Lambda; \Theta)$.

The theoretical foundation of the two-step procedure is that, the original optimization problem can be tackled by iteratively solving the following problem,

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \int_{\Lambda} I(S(\Lambda) \geq \gamma) p(\Lambda; \Theta) d\Lambda, \tag{5}$$

where I is an indicator function, and γ is a given level. The rationale for Eq. 5 is that a good choice of Θ shall generate more elite Λ with $I(S(\Lambda \geq r)) = 1$. We use T to indicate the number of iterations in total and $q = 1, \ldots, T$ to index a specific iteration. For a given level $\gamma^{(q)}$, let $\Theta^{(q)}$ be the solution to Eq. 5. By constructing an increasing sequence of levels $\{\gamma^{(q)}\}$, and correspondingly finding a sequence of hyper parameters $\{\Theta^{(q)}\}$ by solving Eq. 5, the optimal solution is progressively approached. When $\gamma^{(T)}$ is close to $S(\Lambda^*)$, the expectation of $p(\Lambda; \Theta^{(T)})$ will be close to Λ^*.

So in each iteration, our goal is to find Θ such that the cross entropy between $p(\Lambda; \Theta^{(q)})$ and $p(\Lambda; \Theta)$ is minimized. According to [19], minimizing the cross entropy turns out to be maximum likelihood estimation on those elite samples with $S(\Lambda) \geq \gamma^{(q)}$. To see how this conclusion results in the second step of the cross-entropy method, let $\{\widehat{\Lambda}^{(1,q)}, \ldots, \widehat{\Lambda}^{(s,q)}\}$ be the s elite samples found in the q-th iteration. By setting $\gamma^{(q)} = S(\widehat{\Lambda}^{(s,q)})$, $\Theta^{(q)}$ is the result of maximum likelihood estimation on $\{\widehat{\Lambda}^{(1,q)}, \ldots, \widehat{\Lambda}^{(s,q)}\}$.

3.2.2 The Cross-Entropy based Learning Algorithm

We now present an algorithm for image annotation personalization, on the basis of the cross-entropy method. For reasons of simplicity, we assume that the weighting parameters $\{\lambda_{i,j}\}$ are independent of each other. Each parameter $\lambda_{i,j}$ follows a distribution $p(\lambda_{i,j}; \theta_{i,j})$, and $\Theta = [\theta_{i,j}]_{m \times t}$ is the hyper parameter matrix. Moreover, we choose binomial distributions to be the distribution family, because the parameter of a binomial distribution, namely $\theta_{i,j}$, directly measures the impact of $\lambda_{i,j}$ on the personalization process. If a larger (smaller) value of $\lambda_{i,j}$ contributes more to the objective function $S(X_{u,past}, \Lambda)$ in the current learning round, $\theta_{i,j}$ increases (decreases) such that a larger (smaller) value is more likely to be assigned to $\lambda_{i,j}$ in next rounds. Concretely, in the q-th iteration, we first randomly generate a sequence of n samples, $\{\Lambda^{(1,q)}, \ldots, \Lambda^{(n,q)}\}$, where

$$\lambda_{i,j}^{(l,q)} \leftarrow \frac{1}{N} Binomial(N, \theta_{i,j}^{(q-1)}), \text{ for } l = 1, \ldots, n. \tag{6}$$

Note that to satisfy the constraints that $0 \leq \lambda_{i,j} \leq 1$, we divide the output of the Binomial function by the number of trials. Subsequently, we find s elite samples from the n samples by sorting them in descending order according to our objective function $S(X_{u,past}, \Lambda)$. As we have mentioned in Section 3.2.1, the optimal $\Theta^{(q)}$ is found by maximum likelihood estimation on the s elite samples. For a Binomial distribution, this amounts to averaging over the elite samples, namely

$$\theta_{i,j}^{(q)} = \frac{1}{s} \sum_{l=1}^{s} \widehat{\lambda}_{i,j}^{(l,q)}. \tag{7}$$

Since the expectation of $\frac{1}{N} Binomial(N, \theta)$ is θ, the optimal set of weights Λ^* found by the proposed algorithm is $\Theta^{(T)}$.

We summarize our algorithm in Table 1. As there is no need to compute gradients for the objective function, the proposed algorithm can optimize the personalization process in terms of a performance measure which can be arbitrarily chosen. Moreover, its convergence is theoretically guaranteed by the underlying cross-entropy method [19].

Concerning the complexity of our algorithm, the main computational effort is spent on evaluating $S(X_{u,past}, \Lambda)$. We assume that the generic annotation functions $\{g_j(x, w)\}$ are precomputed. For a given Λ, the complexity of constructing a tag rank for an image is $O(m \cdot t + m^2)$, and con-

Table 1: The proposed cross-entropy based learning algorithm for optimizing automated image annotation per user.

```
INPUT: A user's multimedia tagging history $X_{u,past}$,
base image annotation functions $\{g_1, \ldots, g_t\}$,
OUTPUT: optimized weights $\Lambda^*$ for the personalized
image annotation function $G_{u,\Lambda}(x, w)$.
1. Initialize $\Theta^{(0)}$
2. for $q = 1, \ldots, T$
3.     Randomly generate $\{\Lambda^{(1,q)}, \ldots, \Lambda^{(n,q)}\}$ using Eq. 6
4.     Evaluate the generated samples using Eq. 2,
       and select the $s$ elite samples
5.     Obtain $\Theta^{(q)}$ by maximum likelihood estimation
       on the $s$ samples using Eq. 7
6. $\Lambda^* \leftarrow \Theta^{(T)}$
```

sequently $O(|X_{u,past}| \cdot (m \cdot t + m^2))$ for the entire training set. Notice that the evaluations of the n parameters $\{\Lambda^{(l,q)}\}$ are independent of each other. The computation associated with each image is also independent of other training images. Therefore, the algorithm can be easily parallelized.

4. EXPERIMENTAL SETUP

To verify our proposal of personalized image annotation, we conduct a series of experiments on realistic personal image sets collected from the social web.

4.1 Data Sets

Community Image Set X_{comm} for building a content-based image annotation system. We use a set of 3.5 million images randomly sampled from Flickr by our earlier work [10]. Because batch-tagged images are often (nearly) duplicate and of low tagging accuracy, such images are not helpful for content-based image annotation. Also, we want tags to be meaningful. With these two considerations, we remove batch-tagged images and tags not defined in WordNet [17]. We use the remaining 800K images as X_{comm}. Since tags with very low frequency are unlikely to be well predicted, we preserve tags assigned to at least 100 images, and thus obtain a vocabulary V with $m = 5,073$ tags.

Personal Image Sets X_u for testing personalized image annotation. As this work studies how to personalize automated image annotation, the personal image sets for evaluation should be independent of the 800K community image set. To that end, we choose NUS-WIDE [2], which consists of 20K Flickr images after the same preprocess as we used for the community set. We aim to learn from a user's multimedia tagging history. Therefore, for each user, instead of splitting her/his image set at random, we divide the set into two distinct subsets, namely *Past* and *Future*, such that images from *Past* were uploaded before images from *Future*. The *Past* and *Future* sets are instantiations of $X_{u,past}$ and $X_{u,future}$ defined in Section 3.1. To reveal how much personal tagging history is required for the history information to be useful, we conduct a study on 5,315 users with varying amounts of personal tagging history, as shown in Table 2. The number of images in the *Past* sets ranges from 1 to 205, with an average value of 7.9. The *Future* set has similar statistics. For each user, we use $X_{u,past}$ for training and $X_{u,future}$ for evaluation. Note that we treat each test image as unlabeled. Its user tags are merely used for ground-truth purposes.

Table 2: We build personalized image annotation models for 5,315 users with varying amounts of personal tagging history. The amount of tagging history per user is measured by $|X_{u,past}|$.

$\lvert X_{u,past}\rvert$	Number of users
1	1,422
$2 \sim 9$	2,554
$10 \sim 49$	1,221
≥ 50	118

4.2 Base Image Annotation Functions

We choose two state-of-the-art models, *PersonalPreference* [21] and *Visual* [11], which predict tags using tag statistics and visual content, respectively.

PersonalPreference. We choose this function for its good performance for personalized image annotation, as suggested in [21]. Given an unlabeled image x from a user u, the PersonalPreference model simply annotates x with the most frequent tags in $X_{u,past}$. Let $P(w|X)$ be the tag distribution in a social-tagged image set X, computed as

$$P(w|X) \approx \frac{\text{freq}(w|X) + \epsilon}{\sum_{j=1}^{m} \text{freq}(w_j|x) + \epsilon \cdot m}, \tag{8}$$

where freq$(w|X)$ is the number of images labeled with w in X, and ϵ is a small positive constant for smoothing. We express the PersonalPreference version of $g(x,w)$ as

$$g_{pp}(x,w) = P(w|X_{u,past}). \tag{9}$$

Visual. This model as introduced in our previous work [11] predicts tags purely based on image content. Our experiments show that it outperforms ALIPR [9], the content-based model used in [21]. Given an image x represented by a visual feature f, the Visual model first finds k neighbor images visually close to x from the community image set X_{comm}, and then selects the most frequent tags in the neighbor set as annotations of x. To overcome the limitation of single features in describing image content, predictions made based on individual features are uniformly combined. We express the Visual version of $g(x,w)$ as

$$g_v(x,w) = \frac{1}{|F|}\left(\sum_{f \in F} \frac{\text{freq}(w|X_{x,f,k})}{k} - \frac{\text{freq}(w|X_{comm})}{|X_{comm}|}\right), \tag{10}$$

where F is a set of features, and $X_{x,f,k}$ are the k visual neighbors of x with the visual similarity defined by f.

To implement Eq. 10, we choose three decent visual features as follows: COLOR, CSLBP, and GIST. The COLOR feature is a 64-d global feature, combining the 44-d color correlogram [8], the 14-d texture moments [28], and the 6-d RGB color comments. The CSLBP feature is a 80-d center-symmetric local binary pattern histogram [7], capturing local texture distributions. The GIST feature is a 960-d feature describing dominant spatial structures of a scene by a set of perceptual measures such as naturalness, openness, and roughness [18]. The parameter k is set to 500.

By incorporating the two complementary base functions, $g_{pp}(x,w)$ and $g_v(x,w)$, into our unified framework, we aim for good personalized image annotation.

4.3 Implementation

Parameters of the Proposed Model. There are $m = 5,073$ tags and $t = 2$ base image annotation functions. We empirically set the parameters of the algorithm described in Table 1 as follows: $n = 10$, $s = 2$, and $T = 200$. The computational time of the algorithm is linearly proportional to the size of the training data. For a user with 50 tagged images for training, each learning round costs approximately 42 seconds in our prototype system.

Evaluation Criteria. We use precision at top 1 (P@1) and precision at top 5 (P@5) to evaluate the accuracy of the top predicted tags. To evaluate entire tag rankings, we use average precision (AP), a good combination of precision and recall. The personalization process is optimized in terms of AP. The performance for a given user is averaged over all test images of this user.

4.4 Experiments

Experiment 1: User Tagging Consistency. We aim to verify to what extent our conjecture about user tagging consistency made in Section 3.1 is valid. Due to the lack of golden criteria for judging consistency, we compare the divergence between tag distribution of the same user and from different users. Given two users u_i and u_j, we compute the Jensen-Shannon divergence between $P(w|X_{u_i,past})$ and $P(w|X_{u_j,future})$, where the probability masses are computed using Eq. 8. So $i = j$ indicates intra-user divergences, while $i \neq j$ indicates inter-user divergences.

Experiment 2: Comparing Models. We compare the proposed model with the following three baselines: two generic models, namely CommunityPreference and Visual, and one personalized model, PersonalPreference. CommunityPreference annotates an image by simply predicting the most frequent tags within the community set. Visual and PersonalPreference, as mentioned in Section 4.2, are two ingredients in the proposed model.

5. RESULTS

5.1 Experiment 1: User Tagging Consistency

The intra-user and inter-user divergence matrix is shown in Fig. 2, where the diagonal line denotes the intra-user divergences. For a better view of the four user groups in Table 2, we randomly select 100 users from each group, and arrange the matrix in ascending order in terms of $|X_{u,past}|$. For users with a very short tagging history as shown in the top left corner of Fig. 2, the intra-user divergences are smaller than their inter-user counterparts within the same group. But the difference is relatively small, largely due to the fact that the lack of tagging history makes the estimated tag distributions less distinguishable. As the amount of the tagging history increases, we observe a more clear difference between intra-user divergences and inter-user divergences. See for instance $|X_{u,past}| \geq 50$ as shown in the bottom right corner of Fig. 2. Viewing the inter-user divergences as a baseline, we conclude that the tagging preferences of the same user is relatively consistent.

5.2 Experiment 2: Comparing Models

Personalized Models versus Generic Models. As shown in Table 3, when a user's personal tagging preference is unknown, content-based prediction, i.e, the *Visual* model, with an AP score of 0.091, is much better than *CommunityPreference* with an AP score of 0.044. Once a user's tagging history is available, the simple *PersonalPreference* model, with an AP score of 0.232, clearly outperforms content-based prediction. The statement is valid even for $|X_{u,past}| = 1$. The result is consistent with the observation made by [21] that a user's previously used tags are important for predicting her/his future uploads.

Comparing Two Personalized Models. As shown in Table 3, the proposed model compares favorably to *PersonalPreference* under all evaluation criteria. In contrast to [21] where *PersonalPreference* is considered as an upper bound on image annotation performance, our model surpasses the "upper bound".

We compare the two personalized models, given users with varying amounts of personal tagging history. In the extreme case, there are 1,422 users, each having only one tagged images available for training. For 67% of the 1,422 users, we observe improvements, with a relative gain of 8% in terms of AP. Richer tagging history results in better personalized models in general.

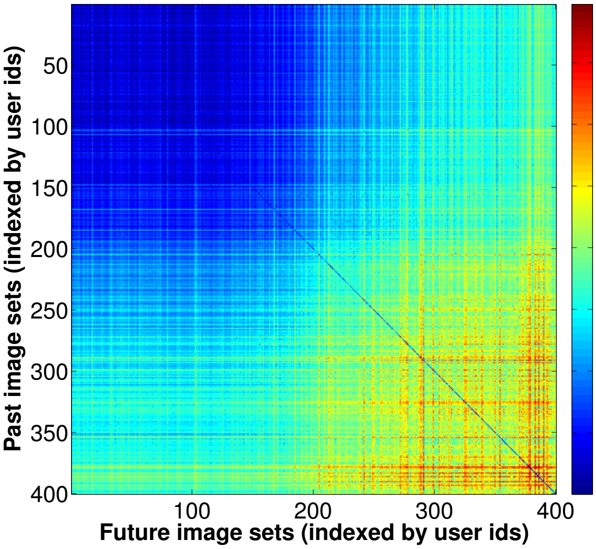


Figure 2: Experiment 1. User tagging consistency. The axes represent the *Past* and *Future* sets of 400 users with $|X_{u,past}|$ ranging from 1 to 168. Each entry in the matrix is the Jensen-Shannon divergence between the tag distribution in a *Past* set and in a *Future* set. The matrix is asymmetrical due to inter-user tagging divergences. The diagonal line indicates intra-user divergences. Best viewed in color.

For a comprehensive study, we make a per-user comparison between our model and *PersonalPreference*. The **absolute** improvement in terms of AP is shown in Fig. 3. For 4,442 out of the 5,315 users in our experiments, the proposed model is better than *PersonalPreference*. For 1,088 users, we obtain an absolute improvement of at least 0.05 in terms of AP. We provide in Table 4 a close-up view of the two extremes of the performance curve. In the worst case (Bottom 1), the two ground truth tags 'peninsula' and 'winchester' correspond to abstract notions with rare frequency in the community set. As a consequence, the *Visual* model fails to predict these two tags, resulting in a worse personalized model compared to *PersonalPreference*. In the best case (Top 1), our model, by ranking 'balloon' at the top, improves AP from 0.333 to 1. Overall the proposed algorithm strikes a proper balance when combining *PersonalPreference* and *Visual* in the process of model personalization.

We also look into the scenario when richer personal tagging history is available. For 94.9% of the 118 users with $X_{u,past}| \geq 50$, we observe improvements when compared to *PersonalPreference*. While in the worse case there is a relative loss of 4%, in a successful case we reach a relative improvement of 60%. For a better understanding of (un)successful cases, we illustrate two of them in Fig. 4. For both cases, due to the divergence between the tag distribution in *Past* and in *Future*, PersonalPreference yields relatively lower performance, with AP scores of 0.285 and 0.195, respectively. For the successful case, however, pictures of flowers can be well annotated by the *Visual* model, with an average precision of 0.343. By cross-entropy based

Table 3: Experiment 2. Comparing the overall performance of generic and personalized image annotation models. The amount of personal tagging history is reflected by $|X_{u,past}|$. Scores are averaged over users. A gray cell indicates the top performer.

	$\|X_{u,past}\| = 1$			$\|X_{u,past}\| = 2 \sim 9$			$\|X_{u,past}\| = 10 \sim 49$			$\|X_{u,past}\| \geq 50$			**MEAN**		
Annotation models	*P@1*	*P@5*	*AP*	*P@1*	*P@5*	*AP*	*P@1*	*P@5*	*AP*	*P@1*	*P@5*	*AP*	*P@1*	*P@5*	*AP*
CommunityPreference	0.033	0.046	0.036	0.044	0.058	0.043	0.058	0.088	0.053	0.077	0.106	0.061	0.045	0.063	0.044
Visual [11]	0.175	0.108	0.079	0.198	0.128	0.090	0.249	0.164	0.105	0.304	0.200	0.118	0.206	0.132	0.091
PersonalPreference [21]	0.271	0.233	0.194	0.403	0.273	0.219	0.571	0.398	0.302	0.610	0.437	0.328	0.411	0.295	0.232
This paper	**0.307**	**0.244**	**0.209**	**0.439**	**0.293**	**0.245**	**0.597**	**0.419**	**0.328**	**0.655**	**0.469**	**0.356**	**0.445**	**0.313**	**0.257**

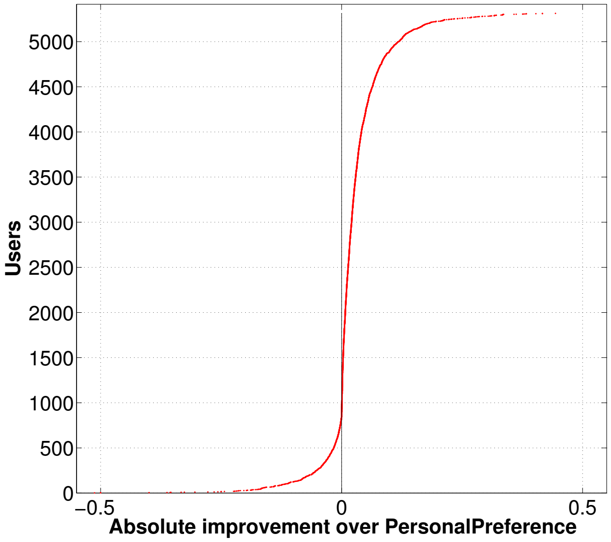


Figure 3: Experiment 2. Comparing two personalized models: The proposed model *versus* PersonalPreference. The performance measure is average precision. For the majority of users in consideration, we obtain personalized image annotation with a higher accuracy.

learning, our model reaches an AP of 0.455. Since images in the worst case are heavily edited, making image content analysis more difficult, *Visual* performs badly, with an average precision of 0.073. We conclude that unless both base image annotation functions fail, the combined model in general yields better or at least comparable performance, compared to the base functions.

Finally, we present some qualitative results in Table 5. The proposed algorithm emphasizes tags which are less frequent, yet more meaningful than the most frequent tags which tend to be general. To summarize, both quantitative and qualitative results verify the effectiveness of the proposed algorithm for personalizing automated image annotation.

6. DISCUSSION AND CONCLUSIONS

Automated image annotation is an important yet challenging research problem. In this paper, we study a novel aspect of the problem: *personalization* – personalizing generic image annotation models with respect to a given user.

We confirm the observation from previous work [21] that personal tagging preference is a strong source of evidence for predicting a user's future annotation. Similar to [21], our experiments also show that a model simply using personal tagging statistics clearly outperforms a content-based model [11] which exploits multi-feature visual content analysis. This nontrivial phenomenon implies that the landscape of personalized image annotation is much different from generic image annotation. Let us now look back at the fundamental challenge in image annotation, namely the semantic gap [24]. The objective aspect of the gap might be ultimately surmounted by machine vision, which aims for an understanding of the visual content independent of the user. Since the human interpretation of the content depends on the specific user in a given situation [24], personalization is essential for solving the subjective aspect of the gap. Thus, to fully bridge the semantic gap, personal information such as tagging history is a factor of major importance. Nevertheless, we challenge the conclusion of [21] that annotating images using personal tagging statistics is a performance upper bound for personalized image annotation.

To personalize generic image annotation models, we propose a linear fusion framework which jointly exploits a user's personal multimedia tagging history and content-based image annotation. The proposed cross-entropy based model enables the personalization process to be optimized in terms of an (arbitrarily) chosen performance measure. It is due to this technical innovation that we can go beyond the performance upper bound defined in [21].

We have conducted an extensive evaluation on 5,315 realistic users with varying amounts of personal tagging history. For the majority of users in consideration, the proposed personalization model surpasses the "upper bound". In an extreme scenario, where a user has only one tagged image available for training, we observe improvements for 67% of these one-training-image users, with a relative gain of 8% in terms of average precision. In general, richer personal tagging history leads to better personalized annotation models. These results clearly verify the effectiveness of the proposed framework for personalized image annotation.

Thus far, we have successfully exploited two heterogenous image annotation functions in the proposed framework. Since our framework is general, other annotation functions driven by varied evidence could be easily added in the future.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback. This work was partially sponsored by the EC-FP6 VIDI-Video, the STW SEARCHER, and the FES COMMIT projects.

Table 4: A close-up of the two extremes in Fig. 3. Images ranging from Top 1 to Top 6 (Bottom 1 to Bottom 6) have the largest (least) absolute improvements, when comparing our model to the *PersonalPreference* model. In the context of personalization image annotation, we consider user tags as the ground truth. The function $g_v(x,w)$ indicates the *Visual* model [11], $g_{pp}(x,w)$ for the *PersonalPreference* model [21], and $G_u(x,w)$ for the proposed model. For each model, the top ranked tags are shown. Correct annotations are marked by an *italic* font.

		Annotation results					Annotation results		
Top 1	**Truth**	$g_v(x,w)$	$g_{pp}(x,w)$	$G_u(x,w)$	**Top 2**	**Truth**	$g_v(x,w)$	$g_{pp}(x,w)$	$G_u(x,w)$
	balloon	flower pink macro cat girl	mickey school *balloon* high gym	*balloon* gym high school mickey		market	beach water car street food	vendor indian *market*	*market* indian vendor bw two
Top 3 	rabbit sit sitting pose portrait	dog cat *pet* *animal* cute	cute coaster summer color sweet	*animal* *pet* *portrait* *bunny* young	**Top 4** 	toy anniversary	flower food nature garden dog	duke *toy* *anniversary* ho lady	*toy* *anniversary* ho lady duke
Top 5 	navy ship	boat bus harbour water river	usa *navy* *ship* alabama museum	*ship* *navy* usa museum alabama	**Top 6** 	beach coast water storm weather	cloud sky sunset *beach* *water*	panorama canon rip ocean *weather*	*beach* *storm* *weather* rock rip
Bottom 1 	peninsula winchester	bridge building house city river	*peninsula* *winchester* water square fountain	fountain water square navy pier	**Bottom 2** 	nj squirrel explore	animal zoo night mountain building	*nj* *squirrel* halloween dog ny	manhattan ny dog parade *nj*
Bottom 3 	microphone	bw blackandwhite street people portrait	*microphone* music	music *microphone* smile vienna subway	**Bottom 4** 	elephant	tree animal nature green bird	*elephant* dog mutt collar	dog *elephant* mutt collar plant
Bottom 5 	toledo windmill	sky night cloud bridge blue	*windmill* holland sky landscape cloud	sky holland landscape cloud *windmill*	**Bottom 6** 	pottery ceramic wheel me	food flower dog macro cat	*pottery* *wheel* disaster big *ceramic*	disaster big bowl *wheel* *pottery*

7. REFERENCES

[1] P. Atrey, M. Hossain, A. El Saddik, and M. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Syst.*, 16(6):345–379, 2010.

[2] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. NUS-WIDE: A real-world web image database from National University of Singapore. In *CIVR*, 2009.

[3] R. Datta, D. Joshi, J. Li, and J. Wang. Tagging over time: real-world image annotation by lightweight meta-learning. In *ACM Multimedia*, 2007.

[4] M. Duan, A. Ulges, T. Breuel, and X.-Q. Wu. Style modeling for tagging personal photo collections. In *CIVR*, 2009.

[5] A. Gallagher, C. Neustaedter, L. Cao, J. Luo, and T. Chen. Image annotation using personal calendars as context. In *ACM Multimedia*, 2008.

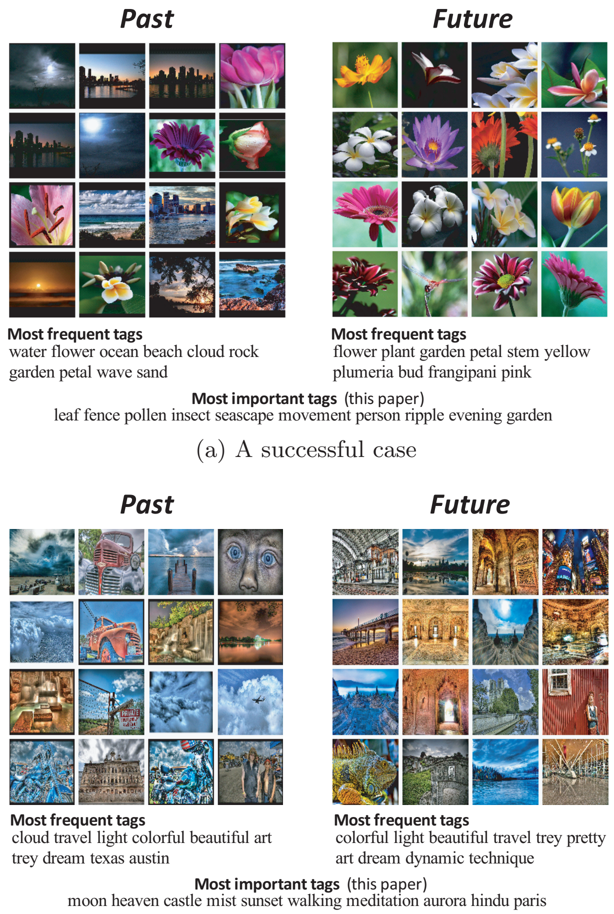
[6] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.

[7] M. Heikkilä, M. Pietikäinen, and C. Schmid. Description of interest regions with local binary patterns. *Pattern Recogn.*, 42:425–436, 2009.

[8] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image indexing using color correlograms. In *CVPR*, 1997.

[9] J. Li and J. Wang. Real-time computerized annotation of pictures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(6):985–1002, 2008.

[10] X. Li, C. Snoek, and M. Worring. Learning social tag


Past
Future
Most frequent tags
water flower ocean beach cloud rock garden petal wave sand
Most frequent tags
flower plant garden petal stem yellow plumeria bud frangipani pink
Most important tags (this paper)
leaf fence pollen insect seascape movement person ripple evening garden
Past
Future
Most frequent tags
cloud travel light colorful beautiful art trey dream texas austin
Most frequent tags
colorful light beautiful travel trey pretty art dream dynamic technique
Most important tags (this paper)
moon heaven castle mist sunset walking meditation aurora hindu paris

(a) A successful case

(b) An unsuccessful case

Figure 4: Illustrating successful and unsuccessful cases of image annotation personalization when rich personal tagging history ($|X_{u,past}| \geq 50$) are available for training. We personalize the *Visual* model $g_v(x, w)$ in terms of a user's *Past* and apply the personalized model $G_u(x, w)$ to annotate the user's *Future* images. In (a), $g_v(x, w)$ is stressed due to its good performance in *Past*, resulting in a relative improvement of 60% over PersonalPreference $g_{pp}(x, w)$. In (b), both $g_v(x, w)$ and $g_{pp}(x, w)$ perform poorly, resulting in a relative loss of 4% against $g_{pp}(x, w)$. Compared to $g_v(x, w)$ and $g_{pp}(x, w)$, unless both base functions fail, our algorithm generally yields better or at least comparable performance.

Table 5: Important tags *versus* frequent tags. The most important tags within a user's tagging vocabulary are found by sorting tags in descending order using $\sum_{j=1}^{t} \lambda_{i,j}$. Recall that $\lambda_{i,j}$ reflects the importance of an image annotation function $g_j(x, w)$ for predicting tag w_i. As the cross-entropy method was original invented for rare event search [19], our model recognizes tags which are less frequent yet more meaningful. For instance, 'sheep' and 'basket' for User 1, 'stage' and 'racing' for User 2, and 'house' and 'sunlight' for User 3.

	Top ranked tags	
$X_{u,past}$ of User 1	*Frequency*	*Importance*
	washington	sheep
	animal	dock
	wildlife	tent
	vancouver	peanut
	outdoor	basket
	nature	snake
	pet	feather
	female	holiday
	bird	black
	rescue	rooster
$X_{u,past}$ of User 2	*Frequency*	*Importance*
	image	stage
	picture	music
	photo	racing
	animal	aquarium
	zoo	bird
	nature	swim
	austria	angry
	tier	mare
	google	flying
	art	floral
$X_{u,past}$ of User 3	*Frequency*	*Importance*
	cloud	land
	brasil	tunnel
	road	house
	sky	sol
	brazil	sunlight
	car	field
	blue	rural
	highway	ga
	landscape	muscle
	green	pb

relevance by neighbor voting. *IEEE Trans. Multimedia*, 11(7):1310–1322, 2009.

[11] X. Li, C. Snoek, and M. Worring. Unsupervised multi-feature tag relevance learning for social image retrieval. In *CIVR*, 2010.

[12] D. Liu, X.-S. Hua, M. Wang, and H.-J. Zhang. Image retagging. In *ACM Multimedia*, 2010.

[13] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma. Dual cross-media relevance model for image annotation. In *ACM Multimedia*, 2007.

[14] Y. Liu, D. Xu, I. Tsang, and J. Luo. Textual query of personal photos facilitated by large-scale web data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):1022–1036, 2011.

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

[16] A. Makadia, V. Pavlovic, and S. Kumar. Baselines for image annotation. *Int. J. Comput. Vision*, 90(1):88–105, 2010.

[17] G. Miller. WordNet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[18] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, 2001.

[19] R. Rubinstein and D. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning*. Springer-Verlag, New York, 2004.

[20] P. Sandhaus and S. Boll. Semantic analysis and retrieval in personal and social photo collections. *Multimedia Tools Appl.*, 51(1):5–33, 2011.

[21] N. Sawant, R. Datta, J. Li, and J. Wang. Quest for relevant tags using local interaction networks and visual content. In *ACM MIR*, 2010.

[22] Y. Shen and J. Fan. Leveraging loosely-tagged images and inter-object correlations for tag recommendation. In *ACM Multimedia*, 2010.

[23] P. Sinha and R. Jain. Classification and annotation of digital photos using optical context data. In *CIVR*, 2008.

[24] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[25] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain. Image annotation by knn-sparse graph-based label propagation over noisily tagged web images. *ACM Trans. Intell. Syst. Technol.*, 2:14:1–14:15, 2011.

[26] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Mach. Learn.*, 81:21–35, 2010.

[27] L. Wu, S. Hoi, R. Jin, J. Zhu, and N. Yu. Distance metric learning from uncertain side information with application to automated photo tagging. In *ACM Multimedia*, 2009.

[28] H. Yu, M. Li, H.-J. Zhang, and J. Feng. Color texture moment for content-based image retrieval. In *ICIP*, 2002.

[29] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *ACM Multimedia*, 2010.