Visually Weighted Neighbor Voting for Image Tag Relevance Learning

Sihyoung Lee, Wesley De Neve, and Yong Man Ro

In: Multimedia Tools and Applications.

Noname manuscript No.
(will be inserted by the editor)

Visually Weighted Neighbor Voting for Image Tag Relevance Learning

Sihyoung Lee · Wesley De Neve · Yong Man Ro

Received: date / Accepted: date

Abstract The presence of non-relevant tags in image folksonomies hampers the effective organization and retrieval of user-contributed images. In this paper, we propose to learn the relevance of user-supplied tags by means of visually weighted neighbor voting, a variant of the popular baseline neighbor voting algorithm proposed by Xirong Li *et al.* in 2009. To gain insight into the effectiveness of baseline and visually weighted neighbor voting, we qualitatively analyze the difference in tag relevance when using a different number of neighbors, for both tags relevant and tags not relevant to the content of a seed image. Our qualitative analysis shows that tag relevance values computed by means of visually weighted neighbor voting are more stable and representative than tag relevance values computed by means of baseline neighbor voting. This is quantitatively confirmed through extensive experimentation with MIRFLICKR-25000, studying the variation of tag relevance values as a function of the number of neighbors used (for both tags relevant and tags not relevant with respect to the content of a seed image), as well as the influence of tag relevance learning on the effectiveness of image tag refinement, tag-based image retrieval, and image tag recommendation.

Keywords Folksonomy · Neighbor voting · Tag relevance learning

Sihyoung Lee
Image and Video Systems Lab, Korea Advanced Institute of Science and Technology
E-mail: ijiat@kaist.ac.kr

Wesley De Neve
Image and Video Systems Lab, Korea Advanced Institute of Science and Technology
Multimedia Lab, Ghent University - iMinds
E-mail: wesley.deneve@kaist.ac.kr

Yong Man Ro
Image and Video Systems Lab, Korea Advanced Institute of Science and Technology
E-mail: ymro@ee.kaist.ac.kr

1 Introduction

Thanks to the popularity of easy-to-use multimedia devices and services, the availability of cheap storage and bandwidth, and more and more people going online, the number of user-generated images is increasing rapidly [1]. These images are frequently shared on online social network sites such as Flickr[1] and Facebook[2]. For example, as of June 2012, Flickr is known to host more than 7 billion images, with over 2,500 new images uploaded every minute [2]. Similarly, each day, more than 300 million photos are uploaded to Facebook on average [3]. As the number of images contributed by users to online social network sites is increasing at a high rate, the problem of organizing and finding relevant images becomes more apparent.

Current techniques for organizing and retrieving user-contributed images strongly rely on freely-chosen textual descriptors, so-called user-defined labels or tags. Sets of user-contributed images and user-supplied tags are also known as image folksonomies [4]. In general, tags allow providing context for images, facilitating an intuitive understanding of different aspects of the image content. Moreover, tags allow reusing already existing text-based search techniques. However, as for instance pointed out in [5] and [6], the presence of non-relevant tags hampers the effective organization and retrieval of user-contributed images, motivating the design of techniques that allow differentiating relevant tags from non-relevant tags. In this paper, we consider a tag to be non-relevant when users with common knowledge are not able to easily and consistently relate the tag to the image content, a definition also used by the authors of [7], [8], and [9].

2 Rationale, Contributions, and Organization

In this paper, we aim at analyzing and improving the effectiveness of the tag relevance learning technique that has been proposed in [8]. This popular technique estimates the relevance of image tags by means of neighbor voting, assuming that tags are likely to reflect objective aspects of an image when different persons have labeled visually similar images using the same tags. Therefore, given a seed image annotated with a tag, neighbor voting estimates the relevance of the tag with respect to the content of the seed image by accumulating votes from visual neighbors that have also been annotated with the tag under consideration. Note that although this paper explains the basic concepts behind neighbor voting, we assume that the reader has some awareness of the details of [8].

Our rationale to focus on analyzing and improving the effectiveness of neighbor voting is as follows: 1) neighbor voting is straightforward in use, relying on two parameters that are tag-independent (this is, the number of neighbors and a tag relevance threshold); 2) neighbor voting comes with a

[1] http://www.flickr.com/
[2] http://www.facebook.com/

simple yet effective mathematical model; 3) neighbor voting offers support for learning the relevance of an unlimited vocabulary of tags; 4) neighbor voting comes with a low computational complexity; and 5) neighbor voting has recently attracted substantial research attention.

The effectiveness of neighbor voting depends on the number of neighbors used. Thus far, no technique has been made available that allows selecting an optimal number of neighbors, given a particular image folksonomy, seed image, and seed tag (here, optimal refers to the case where image tag relevance learning allows separating non-relevant tags from relevant tags in the most effective way). As a result, in practice, it is common to overestimate the number of neighbors. However, given that neighbor voting assigns a uniform importance to each vote, tags associated with images that are not related to the seed image may negatively affect the effectiveness of tag relevance learning, either underestimating or overestimating the relevance of tags assigned to the seed image. This observation motivated us to enhance the effectiveness of neighbor voting by assigning a weight to each vote that is proportional to the visual similarity between the seed image and the neighbor casting the vote. To that end, we reuse the visual information already computed by neighbor voting. That way, as shown by both a qualitative and quantitative analysis, we are able to compute tag relevance values that are more stable and representative than the tag relevance values computed by neighbor voting (this is, the tag relevance values computed are more robust against overestimating the number of neighbors), with a computational complexity that is of the same order as the computational complexity of neighbor voting.

The remainder of this paper is structured as follows. In Section 3, we discuss related work. In Section 4, we briefly review the neighbor voting algorithm of [8], which is further referred to as baseline neighbor voting. Next, we detail the proposed algorithm for visually weighted neighbor voting in Section 5. This algorithm is the first contribution of this paper. Both Section 4 and Section 5 qualitatively analyze the difference in tag relevance when making use of a different number of neighbors, for both tags relevant and tags not relevant with respect to the content of a seed image. This in-depth qualitative analysis is the second contribution of this paper. In Section 6, we present a quantitative analysis, reporting and discussing experimental results. This extensive quantitative analysis is the third contribution of this paper. In this context, we would like to make note that related research efforts typically only focus on providing a quantitative analysis, foregoing the presentation of a qualitative analysis. Finally, in Section 7, we draw conclusions and we identify a number of directions for future research.

3 Related Work

The scientific literature describes several techniques that aim at estimating the relevance of image tags. In what follows, we discuss a number of representative research efforts. Note that we review baseline neighbor voting in a separate

section (this is, Section 4), given that the research effort presented in this paper builds on top of baseline neighbor voting.

The authors of [10] make use of WordNet in order to measure the semantic correlation among tags assigned to a seed image. Strongly correlated tags are considered to be relevant to the content of the seed image, whereas weakly correlated tags are considered to be non-relevant. It should be clear that this approach can only deal with tags that are present in (the English-language version of) WordNet, which is a subset of the set of tags used in an image folksonomy.

The authors of [11] find reliable textual descriptors by mining the tags assigned by photographers to images and by seeking inter-subject agreement for pairs of images that are judged to be highly similar, assuming that the expertise and reliability of photographers is higher than the expertise and reliability of random human annotators, essentially applying a time-shifted version of the ESP image annotation game explained in [12]. The authors of [13] propose two cluster-inspired metrics to quantify the visual representativeness of a given tag, namely cohesion and separation. The cohesion metric measures the visual consistency among the images tagged with the given tag, whereas the separation metric measures the distinctiveness of the common visual content with respect to the entire image collection. Both [11] and [13] are highly similar in spirit to baseline neighbor voting.

The authors of [14] automatically rank image tags according to their relevance to the image content. To that end, initial relevance scores are first computed by means of probability density estimation, a step that is computationally expensive. Next, a random walk is performed over a tag similarity graph in order to refine the relevance scores.

The authors of [15] formulate the problem of tag relevance estimation as a maximum a posteriori (MAP) problem. Given a seed image, the proposed approach computes a posteriori probability for each tag associated with a seed image, taking advantage of the observation that the Euclidean distance between folksonomy images that have been annotated with the same tag follows a Gaussian distribution in feature space.

The authors of [5] propose a tag quality improvement technique that (1) eliminates non-relevant tags and that (2) recommends additional tags for the given input image and its associated tags. To that end, the proposed technique makes use of both semantic and visual similarity. The authors of [16] address the problem of tag relevance learning by constructing a nonparametric tag weight matrix that encodes the relevance relationship between images and tags. In order to construct the tag weight matrix, an algorithm is presented that takes advantage of both the local visual geometry in image space and the local textual geometry in tag space. Both [5] and [16] solve the problem of tag relevance learning by means of an iterative approach, which is effective but costly from a computational point-of-view.

The authors of [17] discuss a data-driven approach for ranking the tags assigned to an image, taking into account the size of the objects shown in the image. In order to determine the size of the objects shown, image segmenta-

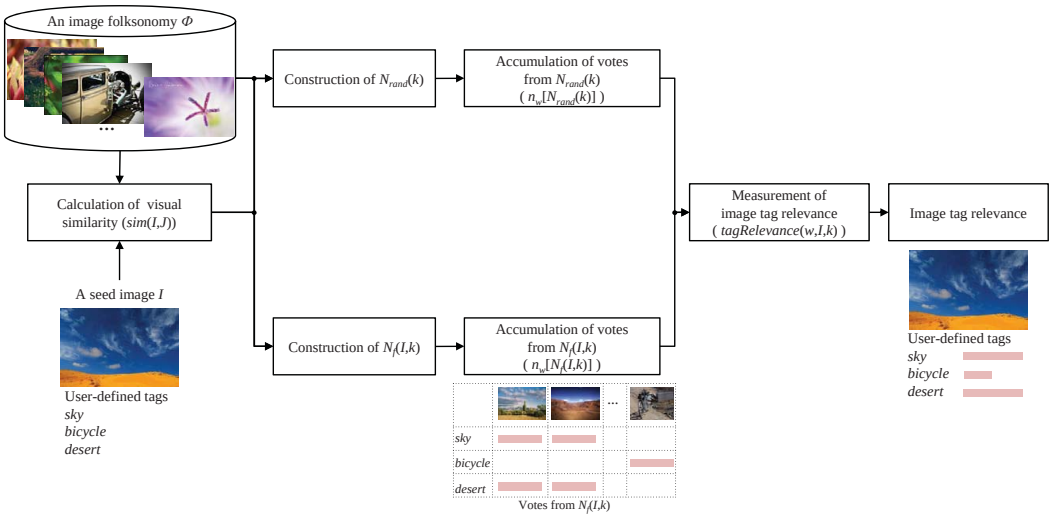


Fig. 1 Visualization of baseline neighbor voting. All votes have a uniform importance.

tion is used. The authors of [18] present a tag ranking method that combines a visual attention model with multi-instance learning, following a three-step procedure: 1) use of multi-instance learning to propagate global image tags to local image regions; 2) use of visual attention modeling to estimate the importance of the different image regions; and 3) ranking of the tags according to the saliency values of the corresponding image regions. Both [17] and [18] make use of image segmentation, a process that is still highly inaccurate. In addition, [18] needs a saliency map, which adds to the computational complexity.

4 Baseline Neighbor Voting

Although the research efforts reviewed in Section 3 have their own distinct merits and demerits, we decided to focus on improving the effectiveness of baseline neighbor voting for the five reasons outlined in Section 2. As such, in this section, we discuss the basic ideas behind baseline neighbor voting, paying particular attention to (1) the difference in accuracy of visual search and random sampling and (2) the difference in tag relevance when making use of a varying number of neighbors, for both tags relevant and tags not relevant with respect to the content of a seed image. Note that Fig. 1 visualizes the way baseline neighbor voting works.

4.1 Background

Given an image folksonomy Φ, baseline neighbor voting estimates the relevance of a tag w with respect to the content of an image I as the difference between 'the number of images annotated with w in a set of k neighbors of I retrieved from Φ by means of visual search' and 'the number of images annotated with w in a set of k neighbors of I retrieved from Φ by means of random sampling'. Following the mathematical notation used in [8], this can be expressed as follows:

$$\begin{aligned} tagRelevance(w, I, k) :=& n_w\left[N_f(I,k)\right] - n_w\left[N_{rand}(k)\right] \\ \approx& \sum_{J \in N_f(I,k)} vote(J,w) - k \cdot \frac{\sum_{J \in \Phi} vote(J,w)}{|\Phi|}, \end{aligned} \tag{1}$$

where $tagRelevance(\cdot)$ denotes the relevance of w with respect to the content of I, computed by means of baseline neighbor voting using k neighbors. The higher the value computed by $tagRelevance(\cdot)$, the higher the relevance of w with respect to the content of I, and vice versa. Further, $n_w[\cdot]$ counts the number of images annotated with w, $N_f(I,k)$ denotes a set of k neighbors of I retrieved from Φ by means of a visual similarity function f (e.g., by means of cosine similarity; please see Section 6.1), and $N_{rand}(k)$ denotes a set of k neighbors retrieved from Φ by means of random sampling. Finally, $vote(J,w)$ represents a voting function, returning one when an image J has been annotated by w, and returning zero otherwise. For the sake of convenience, Table 1 summarizes the mathematical notation used throughout this paper.

4.2 Difference in Accuracy between Visual Search and Random Sampling

When w is relevant to the content of I, it should be clear that the probability that an image from $N_f(I,k)$ is relevant to w is higher than the probability that an image from $N_{rand}(k)$ is relevant to w, given that visual search is supposed to have a higher accuracy than random sampling. To indicate the difference in accuracy of visual search over random sampling, baseline neighbor voting makes use of a variable $\epsilon_{I,w}$. That way, the probability that an image from the set of visual neighbors is relevant to w can be written as $P(R_w) + \epsilon_{I,w}$, where R_w represents the set of all images in Φ relevant to w and where $P(R_w)$ denotes the probability that an image randomly selected from Φ is relevant to w. Given the aforementioned probabilities, $n_w[N_f(I,k)]$ and $n_w[N_{rand}(k)]$ can be determined as follows:

$$\begin{aligned} n_w[N_f(I,k)] =& n_w[N_f(I,k) \cap R_w] + n_w[N_f(I,k) \cap R_w^c] \\ =& k \cdot \{P(R_w) + \epsilon_{I,w}\} \cdot P(w|R_w) \\ &+ k \cdot \{P(R_w^c) - \epsilon_{I,w}\} \cdot P(w|R_w^c), \end{aligned} \tag{2}$$

$$\begin{aligned} n_w[N_{rand}(k)] =& n_w[N_{rand}(k) \cap R_w] + n_w[N_{rand}(k) \cap R_w^c] \\ =& k \cdot P(R_w) \cdot P(w|R_w) + k \cdot P(R_w^c) \cdot P(w|R_w^c), \end{aligned} \tag{3}$$

where R_w^c represents the set of all images in Φ not relevant to w. Further, $P(w|R_w)$ is the probability of correct tagging (*i.e.*, the proportion of images annotated with w in R_w), and $P(w|R_w^c)$ is the probability of incorrect tagging (*i.e.*, the proportion of images annotated with w in R_w^c).

Baseline neighbor voting makes the variable $\epsilon_{I,w}$ dependent on I and w. However, we argue that, in practice, $\epsilon_{I,w}$ is also dependent on k. Indeed, let w denote the tag 'bridge', assigned to an image I that depicts a bridge. Further, for the sake of simplicity, let us assume that visual search is perfect[1]. This implies that, for $k \leq |R_w|$, all images in the set of visual neighbors of I are then relevant with respect to 'bridge'. This implies in turn that $\epsilon_{I,w}$ remains constant, given that the accuracy of visual search is equal to one and that the accuracy of random sampling is constant. However, for $k > |R_w|$, the set of visual neighbors of I will contain $|R_w|$ images that are relevant with respect to 'bridge', as well as $k - |R_w|$ images that are not relevant with respect to 'bridge'. Consequently, for $k > |R_w|$, $\epsilon_{I,w}$ does not remain constant but decreases, and is thus dependent on k.

Given the above example, we subsequently analyze the influence of the value of k on the effectiveness of tag relevance learning, for both a tag w_1 relevant and a tag w_2 not relevant to the content of I.

4.3 Tags Relevant to the Image Content

For a tag w_1 relevant to the content of I, we define the difference in accuracy of visual search over random sampling as follows:

$$\epsilon_{I,w_1,k} = \begin{cases} \frac{|R_{w_1}|}{k'} - P(R_{w_1}), k \leq k' \\ \frac{|R_{w_1}|}{k} - P(R_{w_1}), k > k', \end{cases} \tag{4}$$

where k' denotes the value of k for which all images of R_{w_1} are in the set of visual neighbors of I. For more details regarding the derivation of $\epsilon_{I,w_1,k}$, we would like to refer the interested reader to the appendix.

Given Eq. 4, we are able to qualitatively analyze the difference in tag relevance when baseline neighbor voting uses the following two values for k: 1) $k = k'$ (maximum accuracy of visual search) and 2) $k > k'$ (decreasing accuracy of visual search). Denoting k as k_1 in the case of a maximum accuracy of visual search and denoting k as k_2 in the case of a decreasing accuracy of visual search, the difference in tag relevance can then be derived as follows:

[1] The subsequent qualitative analysis does not assume that visual search is perfect.

$$
\begin{aligned}
&tagRelevance(w_1, I, k_1) - tagRelevance(w_1, I, k_2) \\
&\quad = (k_1 \cdot \epsilon_{I,w_1,k_1} - k_2 \cdot \epsilon_{I,w_1,k_2}) \cdot \left\{P(w_1|R_{w_1}) - P(w_1|R^c_{w_1})\right\} \\
&\quad = \left\{k_1 \cdot \left(\frac{|R_{w_1}|}{k'} - P(R_{w_1})\right) - k_2 \cdot \left(\frac{|R_{w_1}|}{k_2} - P(R_{w_1})\right)\right\} \\
&\qquad \cdot \left\{P(w_1|R_{w_1}) - P(w_1|R^c_{w_1})\right\} \\
&\quad = (k_2 - k_1) \cdot P(R_{w_1}) \cdot \left\{P(w_1|R_{w_1}) - P(w_1|R^c_{w_1})\right\}.
\end{aligned} \tag{5}
$$

In line with [8], assuming that the probability of correct tagging is higher than the probability of incorrect tagging, we can observe that $P(w_1|R_{w_1}) - P(w_1|R^c_{w_1})$ is always positive. Consequently, the difference in tag relevance is positive. In addition, we can observe that the larger the value of k_2, the larger the difference in tag relevance. As a result, we can conclude that baseline neighbor voting linearly underestimates the relevance of w_1 with respect to the content of I when selecting values of k higher than k' (see also Fig. 3(a) and Fig. 3(c) in Section 6.2).

4.4 Tags not Relevant to the Image Content

For a tag w_2 not relevant to the content of I, we define the difference in accuracy of visual search over random sampling as follows (please see the appendix for more details):

$$
\epsilon_{I,w_2,k} = \begin{cases} -P(R_{w_2}), & k \leq k' \\ \frac{|R_{w_2}| - |R_{w_2}| \cdot \frac{k'}{k}}{|\Phi| - k'} - P(R_{w_2}), & k > k'. \end{cases} \tag{6}
$$

Given Eq. 6, the difference in tag relevance can then be derived as follows:

$$
\begin{aligned}
&tagRelevance(w_2, I, k_1) - tagRelevance(w_2, I, k_2) \\
&\quad = (k_1 \cdot \epsilon_{I,w_2,k_1} - k_2 \cdot \epsilon_{I,w_2,k_2}) \cdot \left\{P(w_2|R_{w_2}) - P(w_2|R^c_{w_2})\right\} \\
&\quad = \left\{k_1 \cdot (-P(R_{w_2})) - k_2 \cdot \left(\frac{|R_{w_2}| - |R_{w_2}| \cdot \frac{k'}{k_2}}{|\Phi| - k'} - P(R_{w_2})\right)\right\} \\
&\qquad \cdot \left\{P(w_2|R_{w_2}) - P(w_2|R^c_{w_2})\right\} \\
&\quad = (k_2 - k_1) \cdot \left\{\frac{-|R_{w_2}| \cdot k_1}{|\Phi| \cdot (|\Phi| - k_1)}\right\} \cdot \left\{P(w_2|R_{w_2}) - P(w_2|R^c_{w_2})\right\}.
\end{aligned} \tag{7}
$$

In line with [8], assuming that the probability of correct tagging is higher than the probability of incorrect tagging, we can observe that $P(w_2|R_{w_2}) - P(w_2|R^c_{w_2})$ is always positive. Consequently, the difference in tag relevance is negative. In addition, we can observe that the larger the value of k_2, the larger the difference in tag relevance. As a result, we can conclude that baseline neighbor voting linearly overestimates the relevance of w_2 with respect to the

content of I when selecting values of k higher than k' (see also Fig. 3(b) and Fig. 3(d) in Section 6.2).

4.5 Note

For $k > k'$, we found that neighbor voting linearly underestimates and overestimates the tag relevance of w_1 and w_2, respectively. This can be attributed to the selection of $k_2 - k_1$ additional images as neighbors. Indeed, given that neighbor voting assigns a uniform importance to all votes, tags w_1 and w_2 assigned to the $k_2 - k_1$ additional images have the same importance as tags w_1 and w_2 assigned to the first k_1 images, although the $k_2 - k_1$ additional images are not relevant to the seed image I, whereas most of the first k_1 images are.

As such, the selection of $k_2 - k_1$ additional images as neighbors overestimates the importance of the votes cast by the first term in Eq. 1. In addition, the selection of $k_2 - k_1$ additional images as neighbors overestimates the importance of the votes cast by the second term in Eq. 1 (given the use of k as a multiplier).

5 Visually Weighted Neighbor Voting

This section presents visually weighted neighbor voting, a newly developed variant of the neighbor voting algorithm presented in [8]. Fig. 2 visualizes the way visually weighted neighbor voting works. Further, Algorithm 1 provides a formal description of visually weighted neighbor voting.

Algorithm 1 Visually weighted neighbor voting for tag relevance learning.

input: I (an image annotated with w), w (a tag whose relevance to I needs to be learned), k (the number of neighbors of I), Φ (an image folksonomy)
output: $tagRelevance_{visual}(w, I, k)$ (the relevance of w to I)
$tagRelevance_{visual}(w, I, k) = 0$, $v_w\left[N_f(I, k)\right] = 0$, $v_w\left[N_{rand}(k)\right] = 0$
for all $J \in \Phi$ **do**
  compute $sim(I, J)$
end for
construct $N_f(I, k)$
construct $N_{rand}(k)$
for all $J \in N_f(I, k)$ **do**
  if J is annotated with w **then**
    $v_w\left[N_f(I, k)\right] = v_w\left[N_f(I, k)\right] + sim(I, J)$
  end if
end for
for all $J \in N_{rand}(k)$ **do**
  if J is annotated with w **then**
    $v_w\left[N_{rand}(k)\right] = v_w\left[N_{rand}(k)\right] + sim(I, J)$
  end if
end for
$tagRelevance_{visual}(w, I, k) = v_w\left[N_f(I, k)\right]$ - $v_w\left[N_{rand}(k)\right]$
return $tagRelevance_{visual}(w, I, k)$

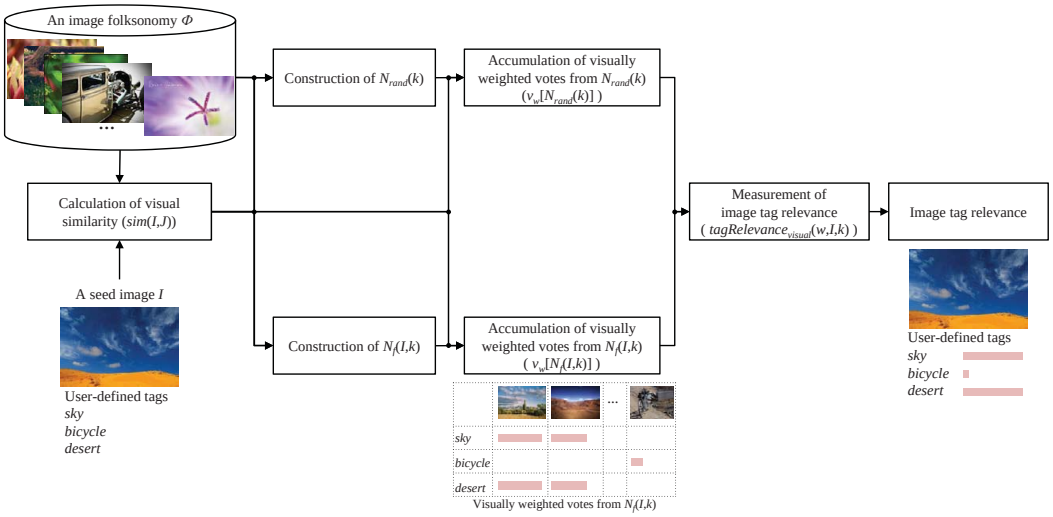


Fig. 2 Visualization of visually weighted neighbor voting. The importance of votes is dependent on the visual similarity between the neighbors selected and the seed image used.

5.1 Background

Visually weighted neighbor voting estimates the relevance of a tag w with respect to the content of a seed image I as the difference between 'the visually weighted number of images annotated with w in a set of k neighbors of I retrieved from Φ by means of visual search' and 'the visually weighted number of images annotated with w in a set of k neighbors of I retrieved from Φ by means of random sampling', and where weights are computed by making use of the visual similarity between I and a particular neighbor image. This can be expressed as follows:

$$\begin{aligned} &tagRelevance_{visual}(w,I,k) := v_w\left[N_f(I,k)\right] - v_w\left[N_{rand}(k)\right] \\ &\approx \sum_{J \in N_f(I,k)} sim(I,J)\cdot vote(J,w) - k\cdot\frac{\sum_{J\in\Phi} sim(I,J)\cdot vote(J,w)}{|\Phi|}, \end{aligned} \tag{8}$$

where $tagRelevance_{visual}(\cdot)$ denotes the relevance of w with respect to the content of I, computed by means of visually weighted neighbor voting using k neighbors. Further, $v_w[\cdot]$ represents the sum of the visual similarity of all images annotated with w, and $sim(I,J)$ represents the normalized visual similarity between the two images I and J (when I and J are identical, the visual similarity has a value of one). By adopting the visual similarity between the seed image and a neighbor image as a weight value for each vote, the tags of images that are not visually similar to the seed image I have less influence on the effectiveness of tag relevance learning (*i.e.*, their votes are less important).

In what follows, we provide a qualitative analysis of the difference in tag relevance when visually weighted neighbor voting uses a different number of neighbors. For brevity, we first introduce the following notation:

$$P_{sim}(R_{w_1}) := \frac{\sum_{J \in \Phi} sim(I,J) \cdot vote(J,w_1)}{|\Phi|} = \frac{Q_1}{|\Phi|}. \tag{9}$$

Similar to $P_{sim}(R_{w_1})$, we define $P_{sim}(R_{w_2})$ as follows:

$$P_{sim}(R_{w_2}) := \frac{\sum_{J \in \Phi} sim(I,J) \cdot vote(J,w_2)}{|\Phi|} = \frac{Q_2}{|\Phi|}. \tag{10}$$

Given Eq. 9, $P(R_{w_1})$ can be seen as a special case of $P_{sim}(R_{w_1})$. Indeed, when the visual similarity between the seed image I and all images in R_{w_1} is one, $P(R_{w_1}) = P_{sim}(R_{w_1})$. However, images in the set of visual neighbors are typically not identical to the seed image I. Consequently, we can safely assume that $P(R_{w_1}) > P_{sim}(R_{w_1})$, given that the visual similarity has a maximum value of one when two images are identical.

5.2 Tags Relevant to the Image Content

Similar to the definition of $\epsilon_{I,w_1,k}$ in Section 4, we define the difference in accuracy of visual search over random sampling for a tag w_1 relevant to the content of I as follows:

$$\epsilon'_{I,w_1,k} = \begin{cases} \frac{Q_1}{k'} - P_{sim}(R_{w_1}), & k \leq k' \\ \frac{Q_1}{k} - P_{sim}(R_{w_1}), & k > k'. \end{cases} \tag{11}$$

Given Eq. 11, the difference in tag relevance can then be derived as follows (please see Section 4.3 for the definition of k_1 and k_2):

$$\begin{aligned} &tagRelevance_{visual}(w_1, I, k_1) - tagRelevance_{visual}(w_1, I, k_2) \\ &\quad = \left(k_1 \cdot \epsilon'_{I,w_1,k_1} - k_2 \cdot \epsilon'_{I,w_1,k_2}\right) \cdot \left\{P(w_1|R_{w_1}) - P(w_1|R^c_{w_1})\right\} \\ &\quad = \left\{k_1 \cdot \left(\frac{Q_1}{k'} - P_{sim}(R_{w_1})\right) - k_2 \cdot \left(\frac{Q_1}{k_2} - P_{sim}(R_{w_1})\right)\right\} \\ &\qquad \cdot \left\{P(w_1|R_{w_1}) - P(w_1|R^c_{w_1})\right\} \\ &\quad = (k_2 - k_1) \cdot P_{sim}(R_{w_1}) \cdot \left\{P(w_1|R_{w_1}) - P(w_1|R^c_{w_1})\right\}. \end{aligned} \tag{12}$$

From Eq. 5 and Eq. 12, we can observe that, compared to baseline neighbor voting, the difference in tag relevance increases slower when making use of visually weighted neighbor voting. Indeed, $P_{sim}(R_{w_1})$ is smaller than $P(R_{w_1})$.

5.3 Tags not Relevant to the Image Content

We also analyze the influence of the number of neighbors k on the effectiveness of tag relevance learning by means of visually weighted neighbor voting for a tag w_2 not relevant to the content of I. Similar to the definition of $\epsilon_{I,w_2,k}$ in Section 4, we define the difference in accuracy of visual search over random sampling as follows:

$$\epsilon'_{I,w_2,k} = \begin{cases} -P_{sim}(R_{w_2}), & k \leq k' \\ \frac{Q_2 - Q_2 \cdot \frac{k'}{k}}{|\Phi| - k'} - P_{sim}(R_{w_2}), & k > k'. \end{cases} \tag{13}$$

Given Eq. 13, the difference in tag relevance can then be derived as follows:

$$\begin{aligned} &tagRelevance_{visual}(w_2, I, k_1) - tagRelevance_{visual}(w_2, I, k_2) \\ &= \left(k_1 \cdot \epsilon'_{I,w_2,k_1} - k_2 \cdot \epsilon'_{I,w_2,k_2}\right) \cdot \left\{P(w_2|R_{w_2}) - P(w_2|R^c_{w_2})\right\} \\ &= \left\{k_1 \cdot (-P_{sim}(R_{w_2})) - k_2 \cdot \left(\frac{Q_2 - Q_2 \cdot \frac{k'}{k_2}}{|\Phi| - k'} - P_{sim}(R_{w_2})\right)\right\} \\ &\quad \cdot \left\{P(w_2|R_{w_2}) - P(w_2|R^c_{w_2})\right\} \\ &= (k_2 - k_1) \cdot \left\{\frac{-Q_2 \cdot k_1}{|\Phi| \cdot (|\Phi| - k_1)}\right\} \cdot \left\{P(w_2|R_{w_2}) - P(w_2|R^c_{w_2})\right\}. \end{aligned} \tag{14}$$

Similar to Eq. 7, we can observe that the difference in tag relevance is negative. However, compared to baseline neighbor voting, the difference in tag relevance decreases slower when making use of visually weighted neighbor voting. Indeed, $(-|R_{w_2}| \cdot k_1)/\left\{|\Phi| \cdot (|\Phi| - k_1)\right\}$ in Eq. 7 is smaller than $(-Q_2 \cdot k_1)/\left\{|\Phi| \cdot (|\Phi| - k_1)\right\}$ in Eq. 14.

5.4 Complexity Considerations

In this section, we briefly discuss the complexity of visually weighted neighbor voting, relative to the complexity of baseline neighbor voting. Compared to the latter, the proposed approach additionally makes use of the visual similarity between the seed image I and the folksonomy images used in order to compute weights. To that end, the proposed approach reuses the visual similarity values that already had to be computed for the construction of $N_f(I, k)$, the set of visual neighbors of I (note that $N_f(I, k)$ is constructed by first computing the visual similarity between I and the folksonomy images used, and by subsequently selecting the k folksonomy images that are visually most similar to I). As such, it should be clear that the complexity of visually weighted neighbor voting is of the same order as the complexity of baseline neighbor voting, while coming with an effectiveness that is higher than the effectiveness of baseline neighbor voting. Finally, we would like to make note that image tag relevance learning in an image folksonomy, either by making use of baseline neighbor

voting or visually weighted neighbor voting, will typically be executed offline (as a form of preprocessing).

6 Experiments

This section discusses four experiments that compare the effectiveness of visually weighted neighbor voting with the effectiveness of baseline neighbor voting. First, we study how tag relevance values vary as a function of the number of neighbors used. Second, we investigate the ratio of non-relevant to relevant tags in an image folksonomy, before and after executing image tag refinement by means of baseline and visually weighted neighbor voting. In this context, we see image tag refinement as an application of tag relevance learning, removing tags with a relevance value lower than a particular threshold. Third, we analyze the influence of baseline and visually weighted neighbor voting on the effectiveness of tag-based image retrieval. Finally, we study the influence of baseline and visually neighbor voting on the effectiveness of image tag recommendation.

6.1 Experimental Setup

Our experiments made use of the publicly available MIRFLICKR-25000 image set [19], a collection of 25,000 user-contributed Flickr images, annotated with a total of 223,537 tags by 9,862 users (the average number of tags per image is 8.94).

We characterized each image by means of the 256-D MPEG-7 Scalable Color Descriptor (SCD) [20]. We also represented each image by means of Bag-of-Visual-Words (BoVW), relying on a vocabulary of 500 visual words [21]. Similar to [22], we adopted the cosine similarity to compute $sim(I, J)$, for both MPEG-7 SCD and BoVW, and similar to [8], we adopted k-nearest neighbor (k-NN) search to find visual neighbors.

For the first two experiments, we used the MIRFLICKR-25000 collection to create a set of 500 test images, annotated with a total of 14,710 tags (each test image was annotated with at least five tags). We manually classified the 14,710 tags as either relevant or non-relevant by making use of a two-step procedure. In the first step, we made use of three annotators to manually classify the 14,710 tags as either relevant or non-relevant. In the second step, we made use of the following criterion to take a final classification decision: if at least two people agree that a tag is relevant, then the tag in question is considered to be relevant, and vice versa. As a result, we found 3,845 tags to be correct (*i.e.*, relevant) and 10,865 tags to be noisy (*i.e.*, non-relevant).

We evaluated the effectiveness of image tag refinement by adopting the noise level (NL) metric [9] [15], which represents the proportion of noisy tags in the set of user-supplied tags of an image folksonomy. When NL is close to one, the number of noisy tags in a folksonomy is high. Likewise, when NL is

close to zero, the number of noisy tags in a folksonomy is low. We determined the value of the threshold for differentiating relevant tags from non-relevant tags offline, using an empirical approach, varying the value of the threshold till we removed 10% of the relevant tags. Note that we used the same threshold value for all test images.

We tested the effectiveness of tag-based image retrieval by using 24 query tags: 'animals', 'baby', 'bird', 'car', 'clouds', 'dog', 'female', 'flower', 'food', 'indoor', 'lake', 'male', 'night', 'people', 'plant_life', 'portrait', 'river', 'sea', 'sky', 'structures', 'sunset', 'transport', 'tree', and 'water'. In this context, we would like to make note that the founders of MIRFLICKR-25000 created a ground truth for these 24 query tags. Before the execution of tag-based image retrieval, we learned the relevance of the 24 query tags to the MIRFLICKR-25000 images they were assigned to. After the execution of tag-based image retrieval, we ranked the images according to their relevance to the query tag under consideration, with the image at rank 1 considered to be the most relevant. In order to know whether the images retrieved were relevant to a particular query tag, we made use of the aforementioned ground truth. Note that we measured the effectiveness of tag-based image retrieval by averaging the precision at rank n ($P@n$) over the 24 query tags, with $P@n$ representing the proportion of relevant images retrieved. When a high number of relevant images can be found among the images retrieved, $P@n$ is close to one. Likewise, when a low number of relevant images can be found among the images retrieved, $P@n$ is close to zero.

Finally, we measured the effectiveness of image tag recommendation by making use of $P@5$, with $P@5$ representing the ratio of correctly recommended tags to the total number of recommended tags (five). When $P@5$ is close to one, the number of correctly recommended tags is high. Likewise, when $P@5$ is close to zero, the number of correctly recommended tags is low. In order to recommend tags to a seed image, we first estimated the relevance between the seed image and the tags in an image folksonomy by making use of baseline neighbor voting and visually weighted neighbor voting. Next, in order to calculate $P@5$, we propagated the top five relevant tags to the seed image under consideration. Note that we used 500 randomly selected images from MIRFLICKR-25000 as test images, using the remaining images for the purpose of retrieving neighbors.

6.2 Experimental Results

6.2.1 Image Tag Relevance as a Function of k

Fig. 3(a) illustrates that, when adopting BoVW and for the 3,845 tags relevant to the test images, the average tag relevance value starts to decrease when k surpasses a value of 1,000. In particular, for k = 1,000 and k = 3,000, the average tag relevance value decreases with 63% (from 5.99 to 2.22, with a standard deviation of 1.86 and 1.24, respectively) when making use of baseline

neighbor voting and with 52% (from 6.30 to 3.04, with a standard deviation of 1.74 and 1.18, respectively) when making use of visually weighted neighbor voting. We can also observe that the average tag relevance value computed by visually weighted neighbor voting decreases more slowly than the average tag relevance value computed by baseline neighbor voting, thus showing that visually weighted neighbor voting is more resilient against underestimating the relevance of correct tags than baseline neighbor voting. We can observe similar results when making use of MPEG-7 SCD (please see Fig. 3(c)).

Fig. 3(b) illustrates that, when adopting BoVW and for the 10,865 tags not relevant to the test images, the average tag relevance value increases when k increases. In particular, for $k = 1{,}000$ and $k = 3{,}000$, the average tag relevance value increases with 60% (from 1.29 to 2.06, with a standard deviation of 1.12 and 1.21, respectively) when making use of baseline neighbor voting and with 42% (from 1.20 to 1.70, with a standard deviation of 1.11 and 1.16, respectively) when making use of visually weighted neighbor voting. We can also observe that the average tag relevance value computed by visually weighted neighbor voting increases more slowly than the average tag relevance value computed by baseline neighbor voting, thus showing that visually weighted neighbor voting is more robust against overestimating the relevance of noisy tags than baseline neighbor voting. We can observe similar results when making use of MPEG-7 SCD (please see Fig. 3(d)).

In summary, the quantitative results reported above are in line with the outcome of the qualitative analysis presented in Section 4: when overestimating the number of neighbors used, baseline neighbor voting underestimates and overestimates the relevance of correct tags and noisy tags, respectively. In addition, both our quantitative and qualitative results demonstrate that visually weighted neighbor voting is more robust against underestimating and overestimating the relevance of correct tags and noisy tags than baseline neighbor voting, thanks to the use of visual similarity information for the purpose of weighting votes (compared to the use of uniformly weighted votes by baseline neighbor voting).

6.2.2 Image Tag Refinement

Fig. 4 shows the effectiveness of image tag refinement in terms of NL (benefit), for the case where we allowed image tag refinement to remove 10% of the relevant tags (cost). We can observe that image tag refinement by means of visually weighted neighbor voting is consistently more effective than image tag refinement by means of baseline neighbor voting, especially when making use of a high number of neighbors. We can also observe that image tag refinement is most effective when retrieving 1,000 neighbors from MIRFLICKR-25000, for both baseline and visually weighted neighbor voting, and for both BoVW and MPEG-7 SCD. However, when making use of more than 1,000 neighbors, we can observe that the effectiveness of image tag refinement starts to decrease (given the higher NL values), for both baseline and visually weighted neighbor voting. In this context, we can also observe that the difference in effectiveness

of image tag refinement by means of baseline neighbor voting on the one hand, and by means of visually weighted neighbor voting on the other hand, starts to increase when making use of more than 1,000 neighbors, especially when making use of MPEG-7 SCD.

6.2.3 Tag-based Image Retrieval

Table 2 and Table 3 summarize the effectiveness of tag-based image retrieval, for learning image tag relevance with 1,000 and 3,000 neighbors, respectively.

Given Table 2, when making use of BoVW and compared to baseline neighbor voting, we can observe that visually weighted neighbor voting allows improving the effectiveness of tag-based image retrieval in terms of Average P@5 with 7% (from 0.61 to 0.65) and in terms of Average P@10 with 9% (from 0.57 to 0.62). We can also observe that the effectiveness of visually weighted neighbor voting is higher than the effectiveness of the rank-based weighting method of [23]. This method computes a weight for each neighbor that is inverse proportional to the rank of the neighbor (i.e., 1/rank), and where the rank of the neighbor is dependent on the visual similarity between the neighbor and the seed image used. Specifically, when making use of BoVW and compared to rank-based weighting, we can observe that visually weighted neighbor voting allows increasing the effectiveness of tag-based image retrieval in terms of Average P@5 with 5% (from 0.62 to 0.65) and in terms of Average P@10 with 7% (from 0.58 to 0.62). We can observe similar results when making use of MPEG-7 SCD.

Given Table 3, when making use of BoVW and compared to baseline neighbor voting, we can observe that visually weighted neighbor voting allows improving the effectiveness of tag-based image retrieval in terms of Average P@5 with 17% (from 0.48 to 0.56) and in terms of Average P@10 with 21% (from 0.43 to 0.52). Compared to rank-based weighting, we can observe that visually weighted neighbor voting allows improving the effectiveness of tag-based image retrieval in terms of Average P@5 with 10% (from 0.51 to 0.56) and in terms of Average P@10 with 13% (from 0.46 to 0.52). We can observe similar results when making use of MPEG-7 SCD.

Further, by analyzing the statistical significance of the improvement in effectiveness of tag-based image retrieval in terms of Average P@5 by means of a paired t-test, we found that the improvement offered by visually weighted neighbor voting over baseline neighbor voting is statistically significant ($p < 0.05$ for both the use of 1,000 neighbors and the use of 3,000 neighbors).

Finally, when comparing the results presented in Table 2 and Table 3, we can observe that the effectiveness of tag-based image retrieval decreases more slowly in the case of visually weighted neighbor voting than in the case of baseline neighbor voting. Specifically, when making use of BoVW, the effectiveness of tag-based image retrieval decreases in terms of Average P@5 with 21% (from 0.61 to 0.48) and in terms of Average P@10 with 25% (from 0.57 to 0.43) in the case of baseline neighbor voting, whereas the effectiveness of tag-based image retrieval decreases in terms of Average P@5 with 14% (from

0.65 to 0.56) and in terms of Average P@10 with 16% (from 0.62 to 0.52) in the case of visually weighted neighbor voting.

6.2.4 Image Tag Recommendation

Table 4 and Table 5 show the effectiveness of image tag recommendation in terms of P@5 for learning image tag relevance with 1,000 and 3,000 neighbors, respectively.

When making use of 1,000 neighbors and compared to baseline neighbor voting, visually weighted neighbor voting allows improving the effectiveness of image tag recommendation in terms of P@5 with 7% (from 0.201 to 0.215) when making use of BoVW and with 6% (from 0.193 to 0.205) when making use of MPEG-7 SCD. Similarly, when making use of 3,000 neighbors and compared to baseline neighbor voting, visually weighted neighbor voting allows improving the effectiveness of image tag recommendation in terms of P@5 with 11% (from 0.183 to 0.203) when making use of BoVW and with 10% (from 0.174 to 0.192) when making use of MPEG-7 SCD.

Compared to rank-based weighting, we can observe that visually weighted neighbor voting allows improving the effectiveness of image tag recommendation in terms of P@5 with 4% (from 0.206 to 0.215) when making use of 1,000 neighbors and with 6% (from 0.191 to 0.203) when making use of 3,000 neighbors. We can observe similar results when making use of MPEG-7 SCD.

Further, by analyzing the statistical significance of the improvement in effectiveness of image tag recommendation in terms of P@5 by means of a paired t-test, we found that the improvement offered by visually weighted neighbor voting over baseline neighbor voting is statistically significant ($p < 0.04$ for both the use of 1,000 neighbors and the use of 3,000 neighbors).

Finally, when comparing the results presented in Table 4 (for 1,000 neighbors) and Table 5 (for 3,000 neighbors), we can observe that the effectiveness of image tag recommendation decreases more slowly when making use of visually weighted neighbor voting than when making use of baseline neighbor voting. Specifically, when making use of BoVW-based baseline neighboring voting, the effectiveness of image tag recommendation decreases with 9% (from 0.201 to 0.183), whereas when making use of BoVW-based visually weighted neighbor voting, the effectiveness of tag-based image retrieval only decreases with 5% (from 0.215 to 0.203).

Fig. 5 shows three example images and their corresponding recommended tags. The recommended tags are sorted according to decreasing tag relevance. Tags related to the image content are underlined. For the example images shown, we can observe that image tag recommendation based on visually weighted neighbor voting is more effective than image tag recommendation based on baseline neighbor voting.

Table 1 Mathematical notation used.

	Notation	Definition
Common	Φ	An image folksonomy
	I	A user-contributed image
	w	A user-defined image tag
	R_w	Images in Φ relevant to w
	R_w^c	Images in Φ not relevant to w
	f	A function that measures the visual similarity between two images
	$\epsilon_{I,w,k}$	Difference in accuracy between visual search and random sampling
	$N_f(I,k)$	A set of k images, selected from Φ by means of f
	$N_{rand}(k)$	A set of k images, selected from Φ by means of random sampling
	$\lvert\cdot\rvert$	The number of elements in a set
Baseline neighbor voting	$n_w[\cdot]$	The number of images annotated with w
	$vote(J,w)$	A voting function, returning one when J has been annotated with w, and returning zero otherwise
	$P(R_w)$	The probability that an image randomly selected from Φ is relevant to w
	$P(R_w^c)$	The probability that an image randomly selected from Φ is not relevant to w
Visually weighted neighbor voting	$v_w[\cdot]$	Sum of the visual similarity of all images annotated with w, given a particular seed image
	$sim(I,J)$	The normalized visual similarity between two images I and J
	$P_{sim}(R_w)$	The visually weighted probability that an image randomly selected from Φ is relevant to w
	$P_{sim}(R_w^c)$	The visually weighted probability that an image randomly selected from Φ is not relevant to w

Table 2 Effectiveness of tag-based image retrieval when making use of 1,000 neighbors.

	Baseline		Rank		Visual	
	BoVW	MPEG-7 SCD	BoVW	MPEG-7 SCD	BoVW	MPEG-7 SCD
Avg.P@5	0.61	0.57	0.62	0.58	0.65	0.62
Avg.P@10	0.57	0.54	0.58	0.56	0.62	0.59

Table 3 Effectiveness of tag-based image retrieval when making use of 3,000 neighbors.

	Baseline		Rank		Visual	
	BoVW	MPEG-7 SCD	BoVW	MPEG-7 SCD	BoVW	MPEG-7 SCD
Avg.P@5	0.48	0.43	0.51	0.46	0.56	0.51
Avg.P@10	0.43	0.37	0.46	0.41	0.52	0.46

Table 4 Effectiveness of image tag recommendation when making use of 1,000 neighbors.

	Baseline		Rank		Visual	
	BoVW	MPEG-7 SCD	BoVW	MPEG-7 SCD	BoVW	MPEG-7 SCD
P@5	0.201	0.193	0.206	0.197	0.215	0.205

7 Conclusions and Directions for Future Work

This paper proposed to learn the relevance of user-defined tags in an image folksonomy by means of a visually weighted variant of the popular neighbor voting algorithm proposed in [8]. To that end, we adopted the visual similarity between a seed image and a neighbor image as a weight value for each vote, reusing the visual similarity information already computed by the aforementioned neighbor voting algorithm.

To gain insight into the effectiveness of both baseline and visually weighted neighbor voting, we qualitatively analyzed the difference in tag relevance when using a different number of neighbors, for both tags relevant and tags not relevant with respect to the content of a given seed image. Our in-depth qualitative analysis, which is one of the main contributions of this paper, demonstrated that tag relevance values computed by means of visually weighted neighbor voting are more stable and representative than tag relevance values computed by means of baseline neighbor voting.

Our qualitative observations are quantitatively confirmed through extensive experimentation with MIRFLICKR-25000. In particular, a first experiment tested the stability of tag relevance values, showing that tag relevance values are less dependent on the number of neighbors retrieved when making use of visually weighted neighbor voting than when making use of baseline neighbor voting. A second experiment tested the representativeness of tag relevance values, showing that tag relevance learning by means of visually weighted neighbor voting allows for more effective image tag refinement than tag relevance learning by means of baseline neighbor voting. A third experiment demonstrated that tag relevance learning by means of visually weighted neighbor voting allows for more effective tag-based image retrieval than tag relevance learning by means of baseline neighbor voting. Finally, a fourth experiment demonstrated that tag recommendation by making use of visually weighted neighbor voting is more effective than tag recommendation by making use of baseline neighbor voting.

We can identify a number of directions for future research. First, given that the effectiveness of the proposed approach is dependent on the use of visual information for exploiting objective image aspects, we plan to further improve its robustness by taking into account information that originates from other image folksonomy modalities (like the tag and user modality of an image folksonomy). Second, based on the observations outlined in this paper, we plan to study techniques that allow automatically computing a proper value for the number of neighbors to use and the tag relevance threshold. Third, we plan to investigate techniques that allow trading off computational complexity with accuracy. We could for instance study how the effectiveness of the proposed approach is influenced by taking advantage of techniques that are computationally less costly, like the use of vocabulary trees to speed up the retrieval of images that are visually similar to the seed image used [24] (the research effort discussed in this paper made use of exhaustive search to construct the set of visual neighbors). On the other hand, we could also study how the effective-

ness of the proposed approach is influenced by taking advantage of techniques that are computationally more costly, such as the simultaneous use of multiple visual features [25].

Acknowledgements This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2012K2A1A2033054).

Appendix

This appendix details the derivation of the difference in accuracy of visual search over random sampling. To that end, given a seed image I, we make a distinction between a tag w_1 relevant to the content of I and a tag w_2 not relevant to the content of I.

Difference in search accuracy for w_1 - We make use of $V_{I,w_1}(k)$ to represent the number of images relevant to w_1 in the set of k visual neighbors of I. We assume that the value of $V_{I,w_1}(k)$ is (1) upper-bounded by the number of images relevant to w_1 when making use of perfectly working visual search and (2) lower-bounded by the number of images relevant to w_1 when making use of random sampling. This is conceptually illustrated by Fig. A-1.

When visual search works perfectly, $V_{I,w_1}(k)$ increases linearly from zero to $|R_{w_1}|$ for k varying from zero to $|R_{w_1}|$. Indeed, all images in the set of visual neighbors belong to $|R_{w_1}|$. For $k > |R_{w_1}|$, $V_{I,w_1}(k) = |R_{w_1}|$ because Φ only contains $|R_{w_1}|$ images related to w_1. This is denoted in Fig. A-1 by "*ideal*". When making use of random sampling, we assume that $V_{I,w_1}(k)$ increases linearly and that all images of R_{w_1} can only be found in the set of visual neighbors when this set is identical to Φ (this is, when k is equal to $|\Phi|$). This is denoted in Fig. A-1 by "*random*". In practice, we also assume that $V_{I,w_1}(k)$ increases linearly until the value of $V_{I,w_1}(k)$ is equal to $|R_{w_1}|$. This is denoted in Fig. A-1 by "*real*". When visual search is effective, the dashed line will be close to "*ideal*". Otherwise, when visual search is not effective, the dashed line will be close to "*random*". In Fig. A-1, k' represents the minimal value of k for which all images of R_{w_1} can be found in the set of visual neighbors of I.

In general, given a tag w_1, the accuracy of visual search $A_{I,w_1,k}$ can be written as $V_{I,w_1}(k)/k$. Given the above observations made for $V_{I,w_1}(k)$, $A_{I,w_1,k}$ can also be expressed as follows:

$$A_{I,w_1,k} = \begin{cases} \frac{|R_{w_1}|}{k'}, \ k \leq k' \\ \frac{|R_{w_1}|}{k}, \ k > k'. \end{cases} \tag{A-1}$$

The difference in accuracy of visual search over random sampling for w_1 can then be expressed as follows:

$$\epsilon_{I,w_1,k} = \begin{cases} \frac{|R_{w_1}|}{k'} - P(R_{w_1}), \ k \leq k' \\ \frac{|R_{w_1}|}{k} - P(R_{w_1}), \ k > k'. \end{cases} \tag{A-2}$$

Difference in search accuracy for w_2 - We make use of $V_{I,w_2}(k)$ to represent the number of images relevant to w_2 in the set of k visual neighbors of I. Further, we assume that the value of $V_{I,w_2}(k)$ is (1) lower-bounded by the number of images relevant to w_2 when visual search works perfectly and (2) upper-bounded by the number of images relevant to w_2 when making use of random sampling. This is conceptually illustrated by Fig. A-2.

When visual search works perfectly (in this case, when visual search finds all images relevant to I in Φ), then the images in R_{w_2} should not be among the visual neighbors of I when $k \leq |R_I|$, where R_I represents the set of images relevant to I. Here, we assume that images are relevant to each other when they have semantic concepts in common (for the sake of simplicity, we also assume that images relevant to I are not relevant to w_2). However, for $k > |R_I|$, the set of visual neighbors of I will start to contain images belonging to R_{w_2}.

This is denoted in Fig. A-2 by "*ideal*". When making use of random sampling, we assume that the number of images of R_{w_2} in the set of visual neighbors increases linearly when k varies from zero to $|\Phi|$. This is denoted in Fig. A-2 by "*random*". In practice, we are able to find a k' for which we can start to see images of R_{w_2} in the set of visual neighbors. This is denoted in Fig. A-2 by means of "*real*". In practice, we also assume that the number of images of R_{w_2} in the set of visual neighbors increases linearly. The accuracy of visual search for w_2, $A_{I,w_2,k}$, is calculated by dividing $V_{I,w_2}(k)$ by k:

$$A_{I,w_2,k} = \begin{cases} 0, & k \leq k' \\ \frac{|R_{w_2}| - |R_{w_2}| \cdot \frac{k'}{k}}{|\Phi| - k'}, & k > k'. \end{cases} \tag{A-3}$$

The difference in accuracy of visual search over random sampling for w_2 can then be expressed as follows:

$$\epsilon_{I,w_2,k} = \begin{cases} -P(R_{w_2}), & k \leq k' \\ \frac{|R_{w_2}| - |R_{w_2}| \cdot \frac{k'}{k}}{|\Phi| - k'} - P(R_{w_2}), & k > k'. \end{cases} \tag{A-4}$$

References

1. OECD (2007) OECD Study on the Participative Web: User Generated Content http://www.oecd.org/dataoecd/57/14/38393115.pdf. Accessed 24 August 2012
2. Flickr's Photostream (2012) Trend Report - Summer'12 http://www.flickr.com/photos/flickr/. Accessed 24 August 2012
3. PlanetTech (2012) Facebook Reveals Staggering New Stats http://www.planettechnews.com/business/item1094. Accessed 24 August 2012
4. Vander Wal T (2007) Folksonomy Coinage and Definition http://www.vanderwal.net/folksonomy.html. Accessed 24 August 2012
5. Liu D, Wang M, Yang L, Hua X S, Zhang H J (2009) Tag Quality Improvement for Social Images. In: IEEE International Conference on Multimedia & Expo (ICME), pp 350-353
6. Lindstaedt S, Morzinger R, Sorschag R, Pammer V, Thallinger G (2009) Automatic Image Annotation using Visual Content and Folksonomies. Multimedia Tools and Applications 42(1):97-113
7. Wu L, Yang L, Yu N, Hua X S (2009) Learning to Tag. In: 18th International Conference on World Wide Web(WWW), pp 361-370
8. Li X, Snoek C G M, Worring M (2009) Learning Social Tag Relevance by Neighbor Voting. IEEE Transactions on Multimedia 11(7):1310-1322
9. Chua T, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In: ACM International Conference on Image and Video Retrieval (CIVR), pp 1-9
10. Jin Y, Khan L, Wang L, Awad M (2005) Image Annotation by Combining Multiple Evidence & WordNet. In: 13th ACM International Conference on Multimedia, pp 706-715
11. Kennedy L, Slaney M, Weinberger K (2009) Reliable Tags Using Image Similarity: Mining Specificity and Expertise from Large-Scale Multimedia Databases. In: 17th ACM International Conference on Multimedia, pp 17-24
12. Ahn L, Dabbish L (2004) Labeling Images with a Computer Game. In: SIGCHI conference on Human factors in computing systems, pp 319-326
13. Sun A, Bhowmick S S (2010) Quantifying Tag Representativeness of Visual Content of Social Images. In: 18th ACM International Conference on Multimedia, pp 471-480
14. Liu D, Hua XS, Yan L, Wang M, Zhang H J (2009) Tag Ranking. In: 18th International Conference on World Wide Web(WWW), pp 351-360
15. Lee S, De Neve W, Ro YM (2010) Tag Refinement in an Image Folksonomy using Visual Similarity and Tag Co-occurrence Statistics. Signal Processing: Image Communication 25(10):761-773
16. Zhuang J, Hoi S C H (2011) A Two-View Learning Approach for Image Tag Ranking. In: ACM International Conference on Web Search and Data Mining, pp 625-634

17. Agrawal G (2011) Relevancy Tag Ranking. In: International Conference on Computer and Communication Technology, pp 169-173
18. Feng S, Hong B, Lang C, Xu D (2011) Combining Visual Attention Model with Multi-instance Learning for Tag Ranking. Neurocomputing 74(17):3619-3627
19. Huiskes MJ, Lew MS (2008) The MIR Flickr Retrieval Evaluation. In: ACM International Conference on Multimedia Information Retrieval, pp 39-43
20. Manjunath B, Salembier P, Sikora T (2003) Introduction to MPEG-7: Multimedia Content Description Interface. Wiley, New Jersey
21. van de Sande K E A, Gevers T, Snoek C G M (2010) Evaluating Color Descriptors for Object and Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9):1582-1596
22. Singh K, Ma M, Park D, An S (2005) Image Indexing Based On MPEG-7 Scalable Color Descriptor. Key Engineering Materials 277:375-382
23. Ferreira J, Silva A, Delgado J (2004) How to Improve Retrieval Effectiveness on the Web. In: IADIS E-Society Conference, pp 1-9
24. Wang X, Yang M, Cour T, Zhu S, Yu K, and Han TX (2011) Contextual weighting for vocabulary tree based image retrieval. In: IEEE International Conference on Computer Vision, pp 6-13
25. Li X, Snoek C G M, Worring M (2010) Unsupervised multi-feature tag relevance learning for social image retrieval. In: ACM International Conference on Image and Video Retrieval (CIVR), pp 10-17

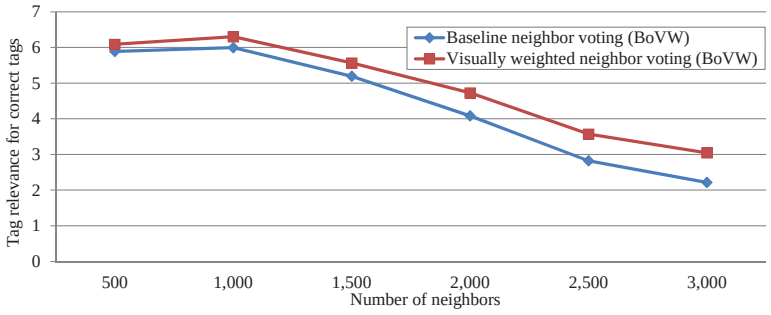
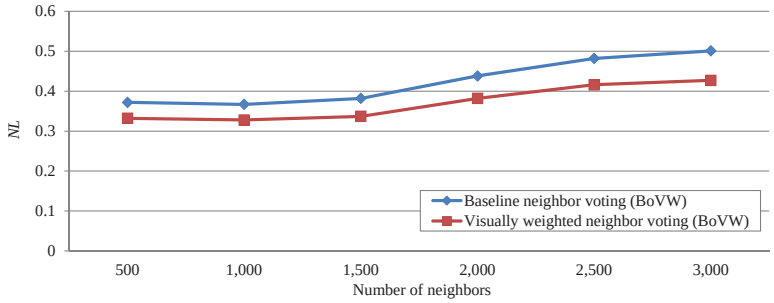
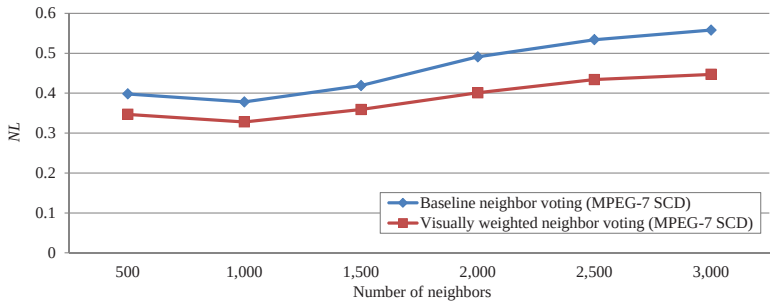


Fig. 3 Average tag relevance as a function of the number of neighbors used: (a) for w_1 using BoVW, (b) for w_2 using BoVW, (c) for w_1 using MPEG-7 SCD, and (d) for w_2 using MPEG-7 SCD.



(a)



(b)

Fig. 4 Effectiveness of image tag refinement for a varying number of neighbors: (a) BoVW and (b) MPEG-7 SCD. The lower NL, the more effective image tag refinement.

Table 5 Effectiveness of image tag recommendation when making use of 3,000 neighbors.

	Baseline		Rank		Visual	
	BoVW	MPEG-7 SCD	BoVW	MPEG-7 SCD	BoVW	MPEG-7 SCD
P@5	0.183	0.174	0.191	0.181	0.203	0.192

Images	Baseline neighbor voting	Visually weighted neighbor voting
	green, cloud, explore, sky, light	green, cloud, sky, explore, landscape
	nature, structure, white, sun, art	structure, nature, white, sun, street
	cloud, reflection, blue, sun, sand	cloud, blue, reflection, sun, sky
	flower, nature, tree, macro, art	flower, nature, tree, macro, petal

Fig. 5 Example images and their recommended tags.

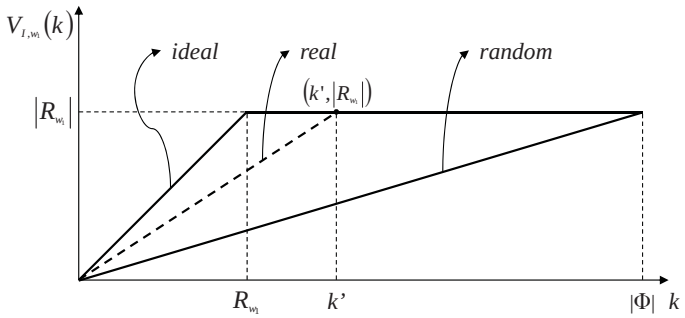


Fig. A-1 The number of images relevant to w_1 in the set of k visual neighbors of I.

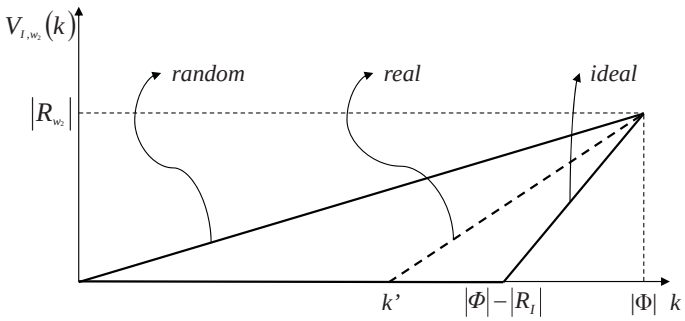


Fig. A-2 The number of images relevant to w_2 in the set of k visual neighbors of I.