An evaluation of nearest-neighbor methods for tag refinement

Conference Paper · July 2013

DOI: 10.1109/ICME.2013.6607547

4 authors, including:

Lamberto Ballan
Stanford University
46 PUBLICATIONS **1,097** CITATIONS

SEE PROFILE

Marco Bertini
University of Florence
168 PUBLICATIONS **1,996** CITATIONS

SEE PROFILE

AN EVALUATION OF NEAREST-NEIGHBOR METHODS FOR TAG REFINEMENT

Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Alberto Del Bimbo

Media Integration and Communication Center (MICC) - Università degli Studi di Firenze, Italy

ABSTRACT

The success of media sharing and social networks has led to the availability of extremely large quantities of images that are tagged by users. The need of methods to manage efficiently and effectively the combination of media and metadata poses significant challenges. In particular, automatic image annotation of social images has become an important research topic for the multimedia community. In this paper we propose and thoroughly evaluate the use of nearest-neighbor methods for tag refinement. Extensive and rigorous evaluation using two standard large-scale datasets shows that the performance of these methods is comparable with that of more complex and computationally intensive approaches and that, differently from these latter approaches, nearest-neighbor methods can be applied to 'web-scale' data.

***Index Terms*—** Tag refinement, tag relevance learning, tag suggestion, image annotation, social media.

1. INTRODUCTION

Social image analysis, annotation and retrieval have become important research topics for the multimedia community. This is due to the success of online social platforms that let users share, rate, comment and tag media. On the one hand the availability of huge quantities of user-generated media and associated metadata are considered valuable resources for improving the results of tasks such as semantic indexing and retrieval. On the other hand there is need to cope with the relatively low quality of these metadata – i.e. tags and annotations are known to be ambiguous, overly personalized, and limited (typically an image is associated with only one-three tags) [1, 2] – and with the 'web-scale' quantity of media. In addition, in a real world social network, users continuously add images and create new terms given the freedom of tagging. Several problems related to content-based tag processing of social images have been recently addressed:

image auto-annotation new tags are added to an image that has not been tagged;

tag-to-region assignment regions of the image are associated with the tags;

tag ranking/relevance existing image tags are ordered according to their relevance with respect to visual content;

Fig. 1. Example of tag refinement: some tags are not relevant with tag content (strike-through), some tags are missing and should be added (bold). Tags may refer to the context or to content depicted in the image.

tag suggestion/recommendation new tags are added (or recommended as candidate tags to a user) to an already tagged image. Existing tags, are considered to be correct;

tag refinement new tags are added to an already tagged image. Existing tags may be eliminated if are not evaluated as relevant for the image content. This is the task addressed by this paper (see the example in Fig. 1).

The methods proposed to address these problems can be divided in those based on statistical modeling techniques and data-driven approaches [3]. Considering the problem of tag refinement, the current state-of-the-art methods [4–6] – often based on matrix factorization approaches – require costly training procedures, that have to be redone periodically if a new set of images or terms are added to the system, thus making the approach impractical for large-scale processing or in social networks undergoing continuous evolution. Recently, data-driven approaches have shown to be able to deal with these latter issues, and have been applied to tag ranking for social image retrieval, tag suggestion for social image annotation (considering also the case in which no tag is associated to an image) [7–9] and tag suggestion and localization in web videos [10]. In order to address the problem of large-scale collections, inherent with social media, we propose to use a data-driven approach also for tag refinement.

To the best of our knowledge a thorough comparison and evaluation of state-of-the-art methods is missing: *i)* not all methods have been applied to the same task on the same

datasets - sometimes proprietary or ad-hoc subsets of standard datasets are used; *ii)* details on description of the methods are sometimes missing - pre-processing steps are not thoroughly described and are hard to be replicated; *iii)* the same procedure for performance evaluation and experimental setup has not been always used and even some parts of standard datasets are not anymore available because refer to web resources that have become inaccessible.

This paper proposes a complete and standardized framework for tag refinement based on nearest-neighbor techniques, presenting a rigorous evaluation on two large-scale standard datasets (MIRFlickr-25K and NUS-WIDE-270K), using the same experimental setup and evaluation metrics. This allows a consistent comparative analysis of these methods. Finally, to fully disclose the implementation details of the proposed methods and to ease future comparisons with other methods, the code of their implementation is provided[1].

The paper is organized as follows: related works on tag refinement are discussed in Sect. 2; descriptions of nearest neighbor methods used for tag ranking and suggestion are presented in Sect. 3, discussing how they can be applied to tag refinement; thorough discussion of the experimental protocols and extensive experimental results are reported in Sect. 4; finally, conclusions are drawn in Sect. 5.

2. RELATED WORK

The first attempt in the literature for image tag refinement is the RWR algorithm presented in [11]. In this work, Wang *et al.* performed belief propagation among tags within the Random Walk with Restart framework, to refine the imprecise original annotations.

The problem of filtering out unreliable tags in social images has been considered also by Kennedy *et al.* in [12], where it is shown that the tags used by different persons to annotate visually similar images are more related to visual content than the others. In the proposed approach 20 nearest neighbors of each processed image are considered and scalability is addressed using a learned low-dimensional image feature, and using the Map/Reduce framework to speed the exhaustive search.

The assumption of consistency between visual and semantic similarity in social images is used by Liu *et al.* in [4] to formulate the tag refinement task as an optimization framework which tries to maximize the consistency while minimize the deviation the tags from initially provided by users. Considering that the consistency assumption is mainly applicable for content-related tags (see Fig. 1), a filtering procedure based on Wordnet is used to constrain the tagging vocabulary within content-related tags. Tag enrichment is done by considering tag synonyms and hypernyms. This method is usually referred in the literature as tag refinement based on visual and semantic consistency (TRVSC).

The method proposed by Zhu *et al.* in [5] is based on the assumptions that visually similar images are similarly tagged, that tags are often correlated and interact at the semantic level, that the semantic space spanned by all the tags can be approximated by a smaller subset of them and that user tags are accurate enough so that it can be assumed a condition of error sparsity for the image tag matrix. The problem of tag refinement is then cast into a decomposition of the user-provided tag matrix into a low-rank refined matrix and a sparse error matrix, and a convergence provable iterative procedure is proposed to accomplish the optimization. This tag refinement approach is referred as Low-Rank approximation (LR).

Recently, Sang *et al.* [6] have proposed to jointly model the ternary relations between users, tags and images employing tensor factorization and using Tucker decomposition for the latent factor inference. Since the traditional factorization models used in recommendation and collaborative filtering systems cannot fully account for missing and noisy tags, the task is cast into a ranking problem to determine which tag is more relevant for a user to describe an image than another tag. To this end is introduced a ternary semantics for tags, that can be positive (those assigned by the users), negative (tags that are dissimilar and that rarely occur together with positive tags) and neutral (all the other tags).

3. TAG REFINEMENT USING NEAREST NEIGHBOR METHODS

The basic idea of the nearest-neighbor methods is to select a set of visually similar images and then to select a set of relevant associated tags based on a tag transfer procedure. This type of methods has been applied to different tasks such as image auto-annotation and tag ranking/relevance.

3.1. Simple Label Transfer: Makadia *et al.* [7]

Makadia *et al.* have proposed a new baseline for image auto-annotation by using a simple method to transfer n tags to a test image from its visual neighborhood. Considering a test image I and a set of K visually similar images $N_k(I, K) = \{I_1, I_2, \ldots, I_K\}$, ordered according to their increasing distance (where I_1 is the nearest image and I_K is the farthest), the procedure used is:

1. Rank the tags of I_1 according to their frequency in the training set. We denote this set as S_1.
2. Transfer the highest n ranking tags of I_1. If I_1 has at least n tags, the algorithm terminates.
3. Rank the tags of neighbors I_2 through I_K (excluding $|S_1|$) according to the co-occurrence in the training set with the tags transferred in step 2 (S_1) and according to the local frequency.

[1]Full code, standard datasets with additional metadata, description of experimental protocol and results are available at: `http://www.micc.unifi.it/vim`

4. Transfer the highest n - $|S_1|$ ranking tags from step 3.

The method is comprised of a composite image distance measure (JEC - Joint Equal Contribution - or Lasso) for nearest neighbor ranking, combined with the tag transfer algorithm, and has been tested on Corel5K, IAPR TC-12 and ESP datasets.

In our implementation the distance between images is computed as:

$$d(I_i, I_k) = \frac{e^{||\mathbf{f}_i - \mathbf{f}_k||}}{\sigma^2} \quad (1)$$

where I_i is the visual neighbor in the i position, with N features $\mathbf{f}_i = (f_i^1, \ldots, f_i^N)$, and σ^2 is set as the median value of all the distances.

3.2. Learning Tag Relevance from Visual Neighbors: Li *et al.* [8]

Li *et al.* have proposed a tag relevance measure for image retrieval based on the consideration, originally proposed in [12], that if different persons label visually similar images using the same tags, then these tags are more likely to reflect objective aspects of the visual content. Therefore it can be assumed that the more frequently the tag occurs in the neighbor set, the more relevant it might be. However, some frequently occurring tags are unlikely to be relevant to the majority of images. To account for this fact the proposed tag relevance measurement takes into account both the distribution of a tag t in the neighbor set for an image I and in the entire collection:

$$tagRelevance(t, I, K) := n_t[N_k(I, K)] - Prior(t, K) \quad (2)$$

where n_t is an operator counting the occurrences of t in the neighborhood $N_k(I, K)$ of K similar images, and $Prior(t, K)$ is the occurrence frequency of t in the entire collection. In order to reduce user bias, only one image per different user is considered when computing the visual neighborhood. The method has been tested for image retrieval on a proprietary Flickr dataset with 20,000 manually checked images and for image auto-annotation using a subset of 331 images.

Considering the setup of the auto-annotation experiment, we estimate $tagRelevance$ for each candidate tag and then rank the tags in descending order by $tagRelevance$. Considering a test image I the procedure used for tag refinement is:

1. Estimation of the distribution of each tag t of I in $N_k(I, K)$.
2. Computation of $tagRelevance$ of each tag t subtracting $Prior(t, K)$ from the distribution of t in $N_k(I, K)$.
3. Ranking of the tags according to their $tagRelevance$ score.
4. Transfer the n highest ranking tags.

3.3. TagProp, Discriminative Metric Learning in Nearest Neighbor Models: Guillaumin *et al.* [9]

Guillaumin *et al.* have proposed to learn a weighted nearest neighbor model, to automatically find the optimal combination of feature distances, to solve the task of image auto-annotation and tag relevance. Using $y_{It} \in \{-1, +1\}$ to represent if tag t is relevant or not for the test image I, the probability of being relevant given a neighborhood of K images $N_k(I, K) = \{I_1, I_2, \ldots, I_K\}$ is:

$$p(y_{It} = +1) = \sum_{N_k(I,K)} \pi_{II_i} p(y_{It} = +1 | N_k(I, K)) \quad (3)$$

$$p(y_{It} = +1 | N_k(I, K)) = \begin{cases} 1 - \epsilon & \text{for } y_{It} = +1, \\ \epsilon & \text{otherwise} \end{cases} \quad (4)$$

where π_{II_i} is the weight of a training image I_i of the neighborhood $N_k(I, K)$, $p(y_{It} = +1 | N_k(I, K))$ is the prediction of tag t according to each neighbor in the weighted sum, with $\pi_{II_i} \geq 0$ and $\sum_{N_k(I,K)} \pi_{II_i} = 1$. The objective is to maximize $\sum_{I,t} \ln p(y_{It})$.

The model can be used with rank-based or distance-based weighting. Furthermore, to compensate for varying frequencies of tags, a tag-specific sigmoid is used to scale the predictions, to boost the probability for rare tags and decrease that of frequent ones. Image tags have been used for model learning. The method has been initially experimented on Corel5K, IAPR TC-12 and ESP datasets. More recently it has also been tested on MIRFlickr-25K [13], using two sets of manually annotated concepts with different degrees of relevance, and a train/test split of the dataset that is different from the one proposed by its creators.

4. EXPERIMENTS AND DISCUSSION

To demonstrate the effectiveness of nearest neighbor methods for tag refinement in a real large-scale scenario, we performed thorough experiments on two large image datasets: MIRFlickr-25K [14] and NUS-WIDE-270K [15]. Both datasets have been collected from Flickr.

The MIRFlickr-25K dataset contains 25,000 images with 1,386 tags. The NUS-WIDE-270K dataset comprises a total of 269,648 images (provided as URLs) with 5,018 unique tags. In order to implement the method described in [8] (see Section 3.2) we had to download again the original data from Flickr for the NUS-WIDE-270K dataset, in order to obtain the users information that is not contained in the dataset; due to the fact that some of the original images of the NUS-WIDE-270K collection are not anymore available, we have been forced to use a subset of the 238,251 images that are still present on Flickr. Hereafter, we refer to this image collection as NUS-WIDE-240K.

Since the tags in the above two image collections are rather noisy and many of them are meaningless words, a pre-processing step was performed to filter out these tags. To this end we matched each tag with entries in Wordnet and only those tags with a corresponding item in Wordnet were retained, similarly to the approach used in [15]. Moreover, we removed the less frequent tags, whose occurrence numbers are below 50. The result of this pre-processing is that 219 and 684 unique tags were obtained in total for MIRFlickr-25K and NUS-WIDE-240K, respectively.

4.1. Visual Features

For both these datasets, the visual similarity between images has been calculated using some simple visual descriptors. We started from the features provided by the authors of the NUS-WIDE dataset and, as in [5], for each image we have extracted a single 428-dimensional descriptor. This feature vector has been obtained as the early-fusion of a 225-d block-wise color moment features generated from 5-by-5 fixed partition on image, a 128-d wavelet texture features, and a 75-d edge distribution histogram features. These features have been computed for both the MIRFlickr-25K and NUS-WIDE-240K datasets, in order to have comparable results.

4.2. Our Evaluation Framework

In order to measure the effectiveness of different tag refinement approaches, we evaluated the performance on the 18 tags in MIRFlickr-25K and the 81 tags in NUS-WIDE-240K where the ground-truth annotations have been provided by the respective authors of these datasets. Following the most relevant previous works in the field [4–6, 11, 16], we report F-measure figures which have been widely used as evaluation metric of tag refinement. The F-measure is defined by $F = \frac{2RP}{(R+P)}$, where P is precision and R is recall.

The F-measure has been calculated to evaluate the refinement results for each tag, and then the overall results were usually obtained by averaging over the number of ground-truth annotations (i.e. classes) as a *macro-average*. Moreover, since both datasets are highly unbalanced, we show also the F-scores obtained by averaging over all the images as a *micro-average*. We believe that both *micro* and *macro* average F-scores are necessary to evaluate the performance of different tag refinement algorithms. The main reason is that because of the unbalance in the number of images per label, simple algorithms like Makadia *et al.* [7] tend to always predict the most common tags.

As previously done by most of the related works [5, 16], we report the overall results by retaining $m = 5$ tags per image. This is an important aspect since the performance are highly influenced by this number. For this reason, we report for both the datasets also some figures by varying m between 1 and 10. It has to be noticed that, on average, each image of the MIRFlickr-25K dataset contains 1.3 tags, while in the NUS-WIDE-240K dataset there are 4 tags per image.

	UT	SLT [7]	TR [8]
F-score *macro*	0.18	0.26	0.27
F-score *micro*	0.06	0.14	0.13

Table 1. Average performances of different algorithms for tag refinement on MIRFlickr-25K (full dataset).

	UT	SLT [7]	TR [8]	TP [9]
F-score *macro*	0.18	0.20	0.19	0.20
F-score *micro*	0.06	0.11	0.11	0.11

Table 2. Average performances of different algorithms for tag refinement on MIRFlickr-25K (test set).

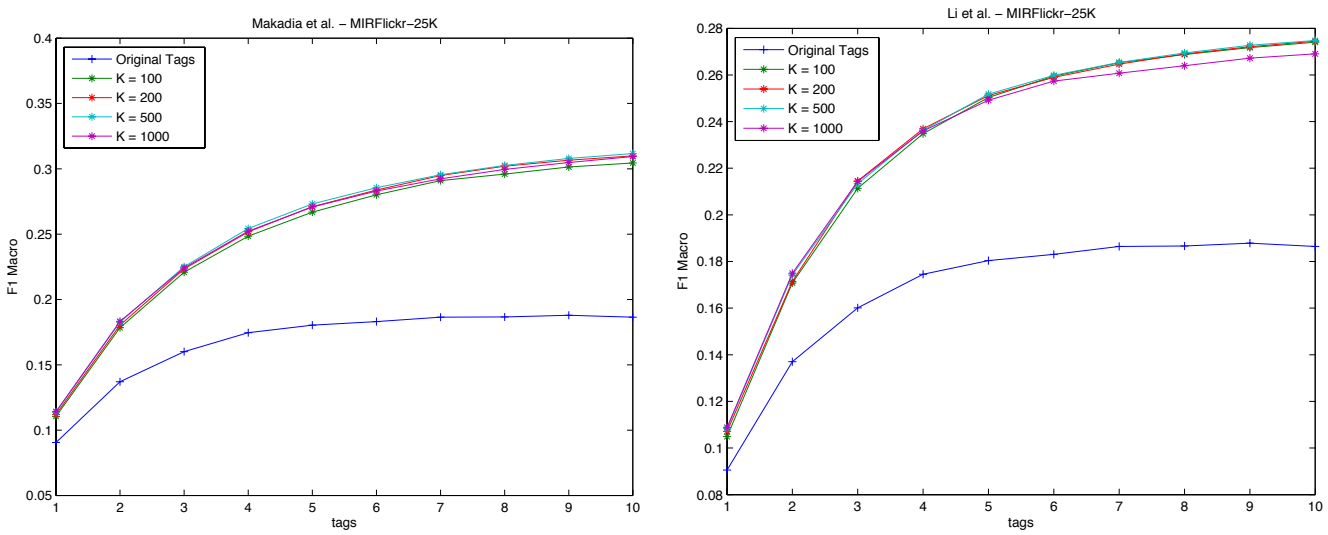
4.3. Evaluation of Tag Refinement on MIRFlickr-25K

To evaluate the effectiveness of the proposed methods, we compare the following four algorithms:

- Baseline, the original tags provided by the users (UT);
- Simple Label Transfer (SLT) [7], described in Sect. 3.1; as shown in Fig. 2 the best results are obtained using $K = 500$ neighbors;
- Learning Tag Relevance from Visual Neighbors (TR) [8], described in Sect. 3.2; again, see Fig. 2, the best results are obtained using $K = 500$ visual neighbors;
- TagProp, Discriminative Metric Learning in Nearest Neighbor Models (TP) [9], described in Sect. 3.3; the best results are obtained by defining the weights of the model directly as a function of the distance.

We performed two sets of experiments. The first one has been conducted on the entire dataset (i.e. 25,000 images) and the results are shown in Table 1. The second one has been conducted using 15,000 images as training set and 10,000 images as test set. Therefore, the results reported in Table 2 refer to the F-scores obtained on the test set (as averages among 10 random train/test splits). It has to be noticed that in this second set of experiments, the performance drop - about 5% for each method - is due to the smaller number of visual neighbors available for the tag propagation.

In general, the Tag Relevance algorithm by Li *et al.* [8] guarantees superior performance with respect to the Simple Label Transfer algorithm by Makadia *et al.* [7] (e.g. 0.27 vs 0.26 on the MIRFlickr-25K full dataset, see Table 1). TagProp shows very similar results (e.g. 0.20 vs 0.19, as reported in Table 2) but it requires a learning phase and more computational costs. The F-score micro-average figures emphasize the better performance given by the method of Li *et al.* [8]. Regarding other methods recently presented in the literature, we report in Table 3 the most relevant previous results.



(a) (b)

Fig. 2. F-score results (y axis) on the MIRFlickr-25K dataset with (a) the Simple Label Transfer algorithm [7], (b) the Tag Relevance Learning algorithm [8]. These results are obtained by varying the number of visual neighbors (K) and the number m of retained tags per image (x axis).

	UT	RWR [11]	TRVSC [4]	LR [5]
Zhu *et al.* [5]	0.22	0.34	0.41	0.42
Liu *et al.* [16]	0.2	0.31	0.37	-

Table 3. F-score performances of other algorithms for tag refinement on MIRFlickr-25K, as reported in the literature.

These results demonstrate that nearest-neighbor methods, when applied to tag refinement, give comparable results to more complex state-of-the-art approaches, despite their simplicity and low computational cost. Complex and computationally intensive algorithms such as TRVSC [4] and LR [5] give an improvement in performance of about 2 percent, but require re-training if the datasets change. The recent results by Liu *et al.* [16], obtained using different visual features (i.e. 500-d BoW of SIFT descriptors), confirm the same trend.

4.4. Evaluation of Tag Refinement on NUS-WIDE-240K

We have done similar experiments on the NUS-WIDE-240K dataset, using the same parameters and the same experimental methodology. Again, we performed two sets of experiments. The first one has been conducted on the entire dataset (i.e. 238,251 images) and the results are shown in Table 4. The second one has been conducted using 158,834 images as training set and the remaining 79,417 as test set. In this case, the results are reported in Table 5. The variation of performance due to changes in the number of visual neighbors K and number of retained tags m per image is similar to that reported in Fig. 2 for MIRFlickr-25K.

The experiments on the NUS-WIDE-240K dataset confirm that the TR algorithm of Li *et al.* [8] gives the best results. It is more difficult to compare our results with the previous works since, in the case of the NUS-WIDE dataset, the previous works often use a subset of the full dataset (often due to the large-scale nature of this dataset) and some undocumented/non-standard experimental procedures. Zhu *et al.* [5] reported in their paper some results on the NUS-WIDE-270K dataset. Their pre-processing step on the tags vocabulary results in 521 tags (instead of our 684 tags). Their results are lower than the others reported by us and by the other works in the literature; their baseline UT is 0.269 while in our case is 0.35 (see Table 4) and so their results are not comparable to us; our results is more similar to those reported by Liu *et al.* [16] (UT=0.45) and Sang *et al.* [6] (UT=0.477). But both [16] and [6] used subsets of the NUS-WIDE-270K dataset, due to the inapplicability of their methods for such a huge number of images. In particular, Liu *et al.* [16] used a subset of only 24,300 images, while Sang *et al.* [6] used a subset of 124,099 images (about half of our NUS-WIDE-240K). Sang *et al.* have used also the same features of us but they have reported results obtained with $m = 10$ tags per image. On their dataset, they have obtained 0.475 with the RWR [11] method, 0.49 with TRVSC [4], 0.523 with LR [5], and 0.571 with their best algorithm.

	UT	SLT [7]	TR [8]
F-score *macro*	0.35	0.37	0.44
F-score *micro*	0.11	0.18	0.23

Table 4. Average performances of different algorithms for tag refinement on NUS-WIDE-240K (full dataset).

Also in the case of a large-scale dataset such as NUS-WIDE-240K, nearest-neighbor based methods show competitive performance. Moreover, an important aspect that is clear

	UT	SLT [7]	TR [8]	TP [9]
F-score *macro*	0.35	0.36	0.45	0.44
F-score *micro*	0.11	0.18	0.22	0.21

Table 5. Average performances of different algorithms for tag refinement on NUS-WIDE-240K (test set).

from the other previous works is that this kind of approaches (i.e. matrix factorization and graph-based methods) suffer in a large-scale scenario. This fact enforces the interest in nearest-neighbor methods for tag refinement.

4.5. Image auto-annotation experiments

Finally, we report also some results on image auto-annotation task, using SLT [7] and TR [8] algorithms on the MIRFlickr-25K dataset. This is an harder task than tag refinement since we try to re-tag the image without retaining any tag from the initial list. These results are reported in Table 6. If compared to the tag-refinement results previously presented in Table 1, these numbers are much lower (0.17 vs 0.26 for SLT, and 0.19 vs 0.27 for TR). Anyway these are interesting numbers: they confirm that the tags suggested by only relying on visual similarity are reliable, and this procedure gives at least the same performance with respect to the initial list of tags given by the users. However, a full tag-refinement approach, that is able to filter out only the noisy tags from the original ones, and to suggest new tags using the social knowledge given by the visual neighbors, can give more satisfying results. This is mainly due to the fact that all content analysis algorithms, including our nearest-neighbor methods for tag refinement, can only handle content-related tags. More abstract concepts such as the name of a city (e.g. "Rome") or the name of a season (e.g. "Spring"), are not directly related to the visual aspect of the image and can only be given by the original users who possess knowledge of the image context.

	UT	SLT [7]	TR [8]
F-score *macro*	0.18	0.17	0.19
F-score *micro*	0.06	0.10	0.10

Table 6. Average performances of different algorithms for image auto-annotation on the MIRFlickr-25K dataset.

5. CONCLUSION

In this paper we have proposed the use of nearest-neighbor models for tag refinement of social images, presenting a standardized evaluation framework using two standard large-scale datasets. In particular, we propose the use of *macro* and *micro* F-scores to better understand the influence of imbalances of the datasets on the performance of the methods. The comparison with current state-of-the-art approaches for tag refinement, based on complex models and computationally intensive algorithms, shows that the simpler nearest-neighbor models obtain quite comparable performance but have the advantage of being usable on large-scale datasets such as the NUS-WIDE-240K.

6. REFERENCES

[1] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev, "To search or to label? Predicting the performance of search-based automatic image classifiers," in *Proc. of ACM MIR*, 2006.

[2] B. Sigurbjörnsson and R. van Zwol, "Flickr tag recommendation based on collective knowledge," in *Proc. of WWW*, 2008.

[3] D. Liu, X.-S. Hua, and H.-J. Zhang, "Content-based tag processing for internet social images," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 723–738, 2011.

[4] D Liu, X.-S. Hua, M. Wang, and H.-J. Zhang, "Image retagging," in *Proc. of ACM Multimedia*, 2010.

[5] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proc. of ACM Multimedia*, 2010.

[6] J. Sang, C. Xu, and J. Liu, "User-aware image tag refinement via ternary semantic analysis," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 883–895, 2012.

[7] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *Proc. of ECCV*, 2008.

[8] X. Li, C. G. M. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1310–1322, 2009.

[9] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. of ICCV*, 2009.

[10] L. Ballan, M. Bertini, A. Del Bimbo, M. Meoni, and G. Serra, "Tag suggestion and localization in user-generated videos based on social knowledge," in *Proc. of ACM-MM Workshop on Social Media*, 2010.

[11] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang, "Content-based image annotation refinement," in *Proc. of CVPR*, 2007.

[12] L. S. Kennedy, M. Slaney, and K. Weinberger, "Reliable tags using image similarity: mining specificity and expertise from large-scale multimedia databases," in *Proc. of ACM-MM Workshop on Web-Scale Multimedia Corpus*, 2009.

[13] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid, "Image annotation with tagprop on the mirflickr set," in *Proc. of ACM MIR*, 2010.

[14] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proc. of ACM MIR*, 2008.

[15] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from national university of singapore," in *Proc. of ACM CIVR*, 2009.

[16] D. Liu, S. Yan, X.-S. Hua, and H.-J. Zhang, "Image retagging using collaborative tag propagation," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 702–712, 2011.