Automatic Image Annotation Using Tag-Related Random Search over Visual Neighbors

Zijia Lin
Tsinghua University
Beijing, 100084, P.R.China
linzijia07@tsinghua.org.cn

Guiguang Ding
Tsinghua University
Beijing, 100084, P.R.China
dinggg@tsinghua.edu.cn

Mingqing Hu
Chinese Academy of Sciences
Beijing, P.R.China
humingqing@ict.ac.cn

Jianmin Wang
Tsinghua University
Beijing, 100084, P.R.China
jimwang@tsinghua.edu.cn

Jiaguang Sun
Tsinghua University
Beijing, 100084, P.R.China
sunjg@tsinghua.edu.cn

ABSTRACT

In this paper, we propose a novel image auto-annotation model using tag-related random search over range-constrained visual neighbors of the to-be-annotated image. The proposed model, termed as TagSearcher, observes that the annotating performances of many previous visual-neighbor-based models are generally sensitive to the quantity setting of visual neighbors, and the probabilities for visual neighbors to be selected is better to be tag-dependent, meaning that each candidate tag can have its own trustworthy part of visual neighbors for score prediction. And thus TagSearcher uses a constrained range rather than an identical and fixed number of visual neighbors for auto-annotation. By performing a novel tag-related random search process over the graphical model made up of range-constrained visual neighbors, TagSearcher can find the trustworthy part for each candidate tag, and further utilize both visual similarities and tag correlations for score prediction. With the range constraint for visual neighbors and the tag-related random search process, TagSearcher can not only achieve satisfactory annotating performances, but also reduce the performance sensitivity. Experiments conducted on benchmark Corel5k well demonstrate its rationality and effectiveness.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Verification.

Keywords

image annotation, TagSearcher, random search

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10 ...$10.00.

1. INTRODUCTION

With the prevalence of social network and digital photography, the number of web images has exploded in recent years, which necessitates effective techniques to manage and retrieve such a large-scale image database. As revealed by recent studies, automatic semantic annotation for unlabelled images can be a potential approach to tackling this problem, which aims to objectively and effectively assign images with tags that can well describe corresponding semantic content.

In this paper, we propose a novel auto-annotation model termed as TagSearcher, which adopts tag-related random search over range-constrained visual neighbors of the to-be-annotated image for predicting tag scores. Experiments conducted on benchmark dataset well demonstrate that TagSearcher is rational and effective, and it presents a promising way to reduce the performance sensitivity. The main motivations of our work are based on the following observations:

The annotating performances of many previously proposed models, which utilize an identical and fixed number of visual neighbors for label propagation, are generally sensitive to the quantity setting of neighbors, since insufficient neighbors cannot provide enough information for mining while redundant neighbors probably introduce much noise, and each image may even has its own optimal quantity setting. To tackle the problem, we propose to utilize both strongly-related and weakly-related ranges of visual neighbors for all to-be-annotated images, i.e. a strong upper bound and a weak upper bound for constraining the number of visual neighbors, as illustrated in Fig. 1. Images in the stronly-related range are supposed to be more reliable for knowledge mining, and those out of the weakly-related range are assumed to be unrelated. Moreover, the proposed tag-related random search over range-constrained visual neighbors can find out the trustworthy part for each candidate tag, and then enhance the robustness of annotating performances.

When predicting tag scores for a to-be-annotated image, previous auto-annotation models usually assume that the probabilities for visual neighbors to be selected remain constant for all candidate tags. In this paper, however, we propose that they are better to be tag-dependent, and denote them as the "trust degrees" of neighbors w.r.t the candidate tag. It is because that in the process of predicting score for a specific candidate tag, visual neighbors associated with it are supposed to be more likely to be selected, which is analogous to using both the target image and the candidate tag to re-select the tag-specific trustworthy visual neighbors. Note that in the same case, the assumption of invariant probabilities for visual neighbors to be selected in previous models will indirectly weakens the positive contributions of trustworthy neighbors and strengthens the negative effects of less trustworthy ones, resulting in "tag-specific" noise.

The contributions of our work can be summarized as follows: (1) For reducing performance sensitivity, we propose to utilize a constrained range rather than an identical and fixed number of visual neighbors. (2) An effective and robust image auto-annotation model is proposed, which considers weights of visual neighbors, votes for candidate tags and tag-specific trust degrees of visual neighbors for predicting tag scores. (3) To estimate the trust degrees of visual neighbors w.r.t a candidate tag, we propose a novel optimization algorithm for graphical model named tag-related random search.

The remainder of this paper is organized as follows: Section 2 gives an overview of related work. Section 3 elaborates on the proposed model. Section 4 presents the details of our experiments, including experimental settings, results and analyses. Finally we conclude the paper in Section 5.

2. RELATED WORK

Automatic image annotation has been studied for years, resulting in various models. Among them, a large part adopts the strategy of propagating tags from nearest visual neighbors [1, 2, 3, 4, 5, 6], categorized as visual-neighbor-based (VNB) models. Such models generally assume that visually similar images probably share common tags. In recent years, due to the rapidly increasing image data, VNB models tend to be more preferable. In [1, 3, 4, 5], researchers utilized the visual neighbors of target images to build up real-time annotating frameworks and exploit helpful knowledge from large-scale web images for performance improvements. In general, VNB models will derive weight distribution among visual neighbors and then utilize them for predicting tag scores. F. Wang et al. [2] proposed a graph-based semi-supervised approach for label propagation that estimates weights of neighbors through linear reconstruction. M. Guillaumin et al. [6] adopted metric learning methods for better weight distribution and proposed the sophisticated TagProp, which maintains the state-of-the-art annotating performance.

By surveying previous researches, we conclude that most VNB models utilize an identical and fixed number of visual neighbors for all to-be-annotated images, and the quantity setting of visual neighbors is vital to the annotating performance, thus keeping high the performance sensitivity, especially for models relying heavily on the whole range of visual neighbors.

3. TAGSEARCHER MODEL

3.1 Model Overview

Following previous VNB models, we target at estimating the conditional probability of a candidate tag t_i given the to-be-annotated image I, i.e. $P(t_i|I)$. With the conditional independence assumption between t_i and I, we can derive

$$\begin{aligned} P(t_i|I) &\propto P(I, t_i) \\ &\sim \textstyle\sum_{I_j \in \mathbb{VN}(I)} P(I_j)\, P(t_i|I_j)\, P(I|I_j) \end{aligned} \qquad (1)$$

where $\mathbb{VN}(I)$ is the set of visual neighbors and $P(I_j)$ represents the probability for I_j to be selected. The conditional probability $P(I|I_j)$ and $P(t_i|I_j)$ are also respectively denoted as the weight of visual neighbor I_j and the vote for candidate tag t_i. In previous VNB models, $P(I_j)$ is generally assumed to be a uniform prior probability and assigned with a constant value. Then formula (1) is further simplified as $P(t_i|I) \sim \sum_{I_j \in \mathbb{VN}(I)} P(t_i|I_j)\, P(I|I_j)$. However, according to the observations described formerly, it can be more appropriate to make $P(I_j)$ non-constant and tag-dependent, which in this paper is denoted as the trust degree of I_j w.r.t t_i. Therefore, the proposed TagSearcher predicts tag scores (i.e. estimated $P(t_i|I)$) by considering and estimating three factors: weights of visual neighbors, votes for candidate tags and tag-specific trust degrees of visual neighbors. Specifically, we rewrite formula (1) as:

$$s(I, t_i) = \sum_{I_j \in \mathbb{U}(I)} w(I, I_j)\, v(I_j, t_i)\, c(I_j, t_i) \qquad (2)$$

where $s(I, t_i)$ is the predicted score, $\mathbb{U}(I)$ is the weakly-related range of visual neighbors, $w(I, I_j)$ represents the estimated weight of I_j (i.e. estimated $P(I|I_j)$), $v(I_j, t_i)$ is the estimated vote for t_i from I_j (i.e. estimated $P(t_i|I_j)$), and $c(I_j, t_i)$ means the trust degree of I_j w.r.t t_i (i.e. tag-specific $P(I_j)$). Since the former two factors have been extensively studied in various heuristic ways, we will focus much on the last one in this paper.

3.1.1 *Weights of Visual Neighbors*

Assuming that images ranked after the weak upper bound are unrelated, when estimating weights of visual neighbors, only images within the weakly-related range are considered and others will be directly assigned with zero weights. In the proposed model, weights of visual neighbors are estimated via the following formula:

$$w(I, I_j) = \frac{1}{d(I, I_j)} \log\left(\frac{U+1}{j}\right) \qquad (3)$$

where U and j are respectively the weak upper bound and the rank position of the visual neighbor I_j, and $d(I, I_j)$ is the visual distance between I_j and I. Apparently the formula is both distance-based and rank-based, which assigns larger weights to neighbors ranked in the front.

3.1.2 *Votes for Candidate Tags*

In TagSearcher, when estimating vote for a candidate tag, visual neighbors containing the tag are to return 1, and others will take tag correlations into consideration and give a soft vote. Here we adopt a conditional probability model as following to estimate the soft vote.

$$v(I_j, t_i) \sim P(t_i|\{t_{j1},t_{j2},...,t_{jn}\}) \;\; s.t. \;\; t_i \notin \{t_{j1},t_{j2},...,t_{jn}\} \qquad (4)$$

where $\{t_{j1}, t_{j2}, \ldots, t_{jn}\}$ is the associated tag set of I_j. With the assumption of tag correlations, and the observation that both t_i and $\{t_{j1}, t_{j2}, \ldots, t_{jn}\}$ rarely appear together, we then resort to an approximation scheme as following:

$$v(I_j, t_i) \sim \frac{1}{n} \sum_{k=1}^{n} P(t_i \mid t_{jk}) \sim \frac{1}{n} \sum_{k=1}^{n} \frac{f(\{t_i, t_{jk}\})}{f(t_{jk})} \qquad (5)$$

where n is the number of tags associated with I_j, and $P(t_i|t_{jk})$ is the conditional probability between tags, which is approximated with tag frequencies (i.e. $f(\{t_i, t_{jk}\})$ and $f(t_{jk})$).

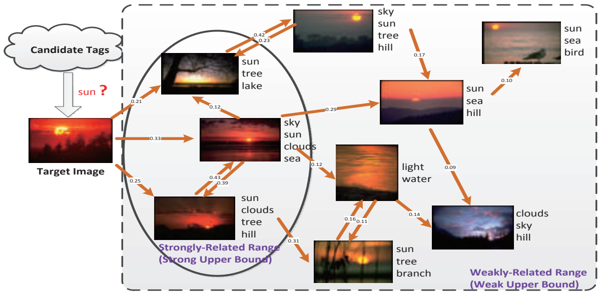

Figure 1: Illustration of the proposed tag-related random search process.

3.1.3 Trust Degrees of Visual Neighbors

As illustrated in Fig. 1, in TagSearcher, for each candidate tag a random search process starting from the to-be-annotated image is performed over its weakly-related range of visual neighbors (i.e. the dashed box) for finding the trustworthy part, and estimating corresponding trust degrees w.r.t the tag. Note that after the first step, the to-be-annotated image is left out in following random search steps. At each step of subsequent random search, the process will determine the probability for each vertex of the graphical model to move forward, which is tag-dependent and varies with the depth of search step. For moving forward, each vertex will choose one of its strongly-related neighbors as a successive vertex and continue. It should be noticed that each visual neighbor also has its own strongly-related neighbors that is constrained within the weakly-related range of the to-be-annotated image. As the random search process moves on, it will finally converge and the probability for the process to stay at each vertex can be utilized to estimate the corresponding trust degree.

3.2 Tag-Related Random Search

In TagSearcher, the constructed graphical model is directed, where each vertex V_a represents a visual neighbor I_a, and the weight on a directed edge means the probability for one to choose the other as a successive vertex for a further step. Note that in the graphical model, all vertices except for the to-be-annotated image can be the successive vertices of others. Hence the to-be-annotated image can be denoted as the source of the graphical model. Regarding the weight on a directed edge, we estimate it as following:

$$s_{a,b} = \begin{cases} \frac{\eta}{d(I_a,I_b)} \log\left(\frac{U+1}{r(I_a,I_b)}\right), & I_b \in \mathbb{V}(I_a) \\ 0, & I_b \notin \mathbb{V}(I_a) \end{cases} \tag{6}$$

where $s_{a,b}$ is the weight on the directed edge from vertex V_a to V_b, U is the weak upper bound for visual neighbors, $\mathbb{V}(I_a)$ represents the strongly-related visual neighbor set of I_a, $d(I_a, I_b)$ is the visual distance, $r(I_a, I_b)$ is the rank position of I_b among the neighbors of I_a, and η is normalizing factor to ensure all the weights on edges from a vertex to sum up to 1. Here we conservatively utilize the more reliable neighbor range for avoiding too much noise. Then a successive matrix $\mathbf{S}_{U\times U}$ representing the successive relationships between visual neighbors can be derived, of which the element $\mathbf{S}_{ij}$ is the weight on the directed edge from V_i to V_j. Assuming at first only the source of the graphical model is assigned with 1 while others with zero, we can then get the expectation values of other vertices after the first step, denoted as the initial value vector $\mathbf{p}$. Apparently, $\mathbf{p}_i$ equals the weight on the edge from the source to V_i, and $\mathbf{p}$ sums up to 1. Then tag-related random search can be further performed for achieving trust degrees of visual neighbors.

At a specific step of the proposed tag-related random search process, the expectation value of a vertex at the kth step is calculated as following:

$$\mathbf{p}_i^{(k)} = \delta \mathbf{p}_i + (1-\delta) \sum_{j \leqslant U, j \neq i} \mathbf{p}_j^{(k-1)} \mathbf{f}_j^{(k-1)} \mathbf{S}_{ji}^{(k-1)} \tag{7}$$

where $\mathbf{p}_i$ is the initial value of the vertex V_i, $\mathbf{p}_j^{(k-1)}$ is the expectation value of V_j at the $(k-1)$th step, $\mathbf{f}_j^{(k-1)}$ and $\mathbf{S}_{ji}^{(k-1)}$ are respectively the probability of moving forward on V_j and the probability for V_i to be chosen as a successive vertex of V_j at the $(k-1)$th step, and δ is a weighting parameter between 0 and 1. The successive matrix $\mathbf{S}_{ji}^{(k-1)}$ is constructed in nearly the same way as formula (6), while the number of candidate successive vertices decreases by a ratio λ and remains no less than 1 as the step increases in order to avoid reaching too many less related neighbors. In formula (7), the forward probability of each vertex, i.e. $\mathbf{f}_j^{(k-1)}$, is vital to finding trustworthy visual neighbors for a candidate tag, which is estimated as following:

$$\mathbf{f}_j^{(k)} = \begin{cases} 0, & \widehat{t} \in \{t_{j1}, t_{j2}, \ldots, t_{jn}\} \\ 1 - \frac{\alpha_j \exp(k)}{\alpha_j \exp(k) + \beta_j}, & \widehat{t} \notin \{t_{j1}, t_{j2}, \ldots, t_{jn}\} \end{cases} \tag{8}$$

where α_j is the conditional probability of the candidate tag $\widehat{t}$ given image I_j. For simplicity, here we use the vote for $\widehat{t}$ from I_j, i.e. $v(I_j, \widehat{t})$, as an approximation. β_j is the expectation value of the conditional probability at a further step, which can be estimated as $\beta_j = \sum_{m \leqslant U, m \neq j} \mathbf{S}_{jm} \alpha_m$ with the law of total probability. As implied by formulas above, it is more likely for tag-related random search to stay at a vertex which contains the specific tag or strongly-related tags, and the staying probability increases with the depth of search step. By constructing a diagonal matrix $\mathbf{F}^{(k)}$ with forward probabilities of all vertices, formula (7) can be further rewritten with matrix notations as following:

$$\mathbf{p}^{(k)} = \delta \mathbf{p} + (1-\delta) \left(\mathbf{S}^{(k-1)^T} \mathbf{F}^{(k-1)} \right) \mathbf{p}^{(k-1)} \tag{9}$$

According to formula (9), we can derive that

$$\begin{aligned} \mathbf{p}_\pi = \lim_{n\to\infty} \delta \left(1 + \sum_{k=1}^{n-1} (1-\delta)^k \prod_{h=1}^{k} \left(\mathbf{S}^{(n-h)^T} \mathbf{F}^{(n-h)} \right) \right) \mathbf{p} \\ + \left(\prod_{h=1}^{n-1} \left((1-\delta) \mathbf{S}^{(n-h)^T} \mathbf{F}^{(n-h)} \right) \right) \mathbf{p}^{(1)} \end{aligned} \tag{10}$$

where $\mathbf{p}_\pi$ is the final value vector as the step of tag-related random search tends to positive infinity (i.e. $\mathbf{p}_\pi = \lim_{n\to\infty} \mathbf{p}^{(n)}$). It can be demonstrated that the proposed tag-related random search process will finally converge, and the second part of formula (10) tends to zero. Then formula (10) can be further simplified as:

$$\mathbf{p}_\pi \sim \lim_{n\to\infty} \left(1 + \sum_{k=1}^{n-1} (1-\delta)^k \prod_{h=1}^{k} \left(\mathbf{S}^{(n-h)^T} \mathbf{F}^{(n-h)} \right) \right) \mathbf{p} \tag{11}$$

By normalizing $\mathbf{p}_\pi$ to make it sum up to 1, the trust degree of visual neighbor I_j w.r.t the candidate tag t_i, i.e. $c(I_j.t_i)$ in formula (2), can be estimated as the jth element of $\mathbf{p}_\pi$.

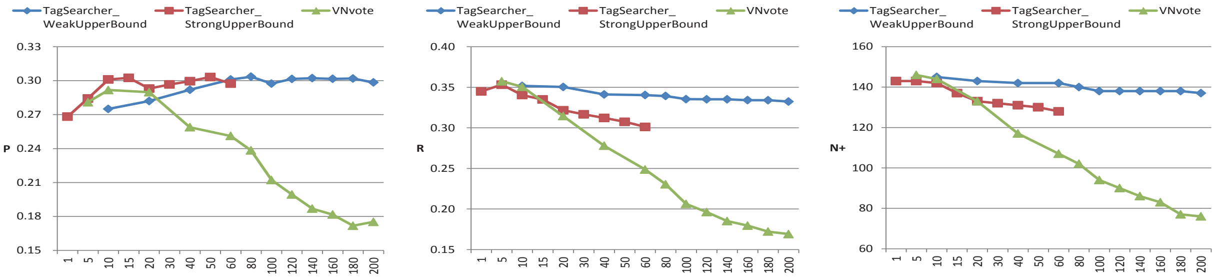

Figure 2: Annotating performances of TagSearcher on Corel5k with strong or weak upper bound varying, compared with VNvote as a baseline, in terms of average precision, recall and $N+$ (from left to right).

3.3 Model Refinement

In the basic TagSearcher presented above, there exist some drawbacks w.r.t non-frequent tags, and thus we further propose refining approaches. Firstly, since non-frequent tags rarely co-occur with other tags in sufficient images, the returned votes for them tend to be smaller, resulting in their lower predicted tag scores. Hence, we resort to the widely-used WordNet for completing the tagging matrix. Specifically, for each non-frequent tag, all the tagging vectors of other tags will be summed up with their corresponding semantic similarities to the specific tag as weights. Thus zero positions in the tagging vector of the non-frequent tag will be changed to 1 if they are above some predefined threshold in the summed vector, and then tag correlations and votes for tags are re-estimated. Secondly, for a to-be-annotated image the weight distribution among visual neighbors in TagSearcher is better to be different between frequent and non-frequent candidate tags, since in most cases a non-frequent tag just appears as an unrelated bundled attachment in visual neighbors. Hence we propose that the weight distribution for non-frequent tags should be more insensitive to the rank positions of neighbors. Specifically, for non-frequent tags, we adjust formula (3) into $w\left(I, I_j\right) = \frac{1}{d\left(I, I_j\right)} \log\left(\frac{U+1}{\lceil \mu j \rceil}\right)$ where μ is a smoothing factor between 0 and 1, and $\lceil\cdot\rceil$ is a ceiling function. Namely, we utilize an echelon decline curve to estimate the weight distribution for any non-frequent tag.

4. EXPERIMENTS

Our experiments were conducted on the widely-used well-known benchmark dataset Corel5k for making fair comparisons. Corel5k is one of the most important evaluation benchmarks in the community of image auto-annotation, containing around 4,999 images that are manually annotated with 1 to 5 tags. A fixed set of 499 images is split out for test, and the remaining works as the training set (i.e. the labelled database). There are totally 260 tags existing in both training and test sets. With accurate manual annotations, the dataset contains little noise and acts as an ideal evaluation benchmark.

To retrieve visual neighbors of to-be-annotated images, we use the open-source Lire [7] project for feature extraction and visual similarity measurement. In our experiments, eleven kinds of features[1] are extracted for each image, and distances between feature vectors are calculated with Lire, then normalized and merged with equal weights for measuring visual distance. Following previous researches, we annotate each test image with the top 5 tags, and calculate the average precision p and recall r for all tags to measure the annotating performances. Additionally, the number of tags with non-zero recall, denoted as $N+$, is also a widely-used important measurement.

4.1 Performance Sensitivity Analysis

Firstly we investigate whether the range constraint for visual neighbors and the proposed tag-related random search can reduce the performance sensitivity to the quantity setting of visual neighbors. We respectively keep either the strong upper bound (i.e. 10) or weak upper bound (i.e. 60) invariant, and investigate the performance variations as the other bound changes. Here we introduce a simple but typical baseline denoted as VNvote, which predicts tag scores with formula (2) except for $c\left(I_j, t_i\right)$ (i.e. trust degrees of visual neighbors w.r.t the candidate tag), similar to many previous VNB models. Hence the comparisons between both can well reflect the effects of the range constraint of visual neighbors and the tag-related random search. Fig. 2 illustrates the performance variations of TagSearcher with the strong or weak upper bound varying, in terms of average precision, recall and $N+$. From Fig. 2, we can draw the following conclusions. (1) The annotating performances of TagSearcher is much less sensitive to the bound settings of visual neighbors, which is attributed to the range constraint for visual neighbors and the proposed tag-related random search. (2) The annotating performances of TagSearcher remain comparable to or even better than the best performance of the baseline, though the settings of the strong and weak upper bounds vary in quite a large range, which well demonstrates the rationality of introducing trust degrees of visual neighbors for auto-annotation and the effectiveness of the proposed tag-related random search. (3) The strong upper bound for visual neighbors has a relatively more significant effect on annotating performance than the weak one. It is because that in the proposed model we rely much more on the strongly-related range of visual neighbors. (4) Though the performance insensitivity is significantly improved, both strong and weak upper bounds can still slightly affect the annotating performance as a convex curve.

4.2 Annotation Result

Table 1 gives an overview of the annotating performances in terms of average precision, recall and $N+$ of the proposed model and those reported in other remarkable earlier

[1] The features include: Color Correlogram, Color Layout, CEDD, Edge Histogram, FCTH, HSV Color Histogram, JCD, Jpeg Coefficient Histogram, RGB Color Histogram, Scalable Color, SURF with bag-of-words model.

	Previously Reported Results											TagSearcher			
	SML[8]	MBRM[9]	TGLM[10]	MSC[11]	JEC[12]	HDGM[13]	GS[14]	En-CRF[15]	TagProp[6]	JEC*	TagProp*	TS	TS+WN	TS+WDA	TS+Both
p	23	24	25	25	27	29	30	32	**33**	29	30	30	31	31	**32**
r	29	25	29	32	32	30	33	33	**42**	33	32	34	33	**36**	35
$N+$	137	122	131	136	139	146	146	148	**160**	139	141	142	142	146	**149**

Table 1: Annotating performances on Corel5k in terms of average precision (p), recall (r) and N+ of the proposed TagSearcher, and those reported in a selection of remarkable earlier researches.

researches on the benchmark Corel5k. JEC* is our implementation of JEC using our eleven kinds of features, and TagProp* refers to the corresponding model with published implementation by M. Guillaumin [6] and our features. Regarding TagSearcher, we empirically set the strong and weak upper bound as 10 and 60 respectively, and decreasing ratio λ as 2, weighting factor δ as 0 for reducing computation complexity. Here we denote the refined variant of TagSearcher with tagging matrix completion using WordNet as TS+WN, and that with weight distribution adjustment for non-frequent tags as TS+WDA ($\mu = 0.5$). Furthermore, we merge both refining strategies and denote it as TS+Both.

From table 1 we can make several observations. (1) The annotating performance of JEC* on Corel5k is similar to those reported in previous researches, making it relatively fair to make comparisons with their reported results. It can be seen that the proposed TagSearcher and its refined variants outperform most previous remarkable models, and achieve comparable annotating performances to the state-of-the-art TagProp [6]. (2) When comparisons are strictly made with the same kinds of features, TagSearcher and its refined variants outperform the state-of-the-art TagProp. Both observations above well demonstrate the effectiveness of TagSearcher for image auto-annotation. (3) The refined variants of TagSearcher achieve more satisfactory annotating performances, especially TS+Both, which demonstrates the rationality and effectiveness of our proposed refining strategies for non-frequent tags.

5. CONCLUSIONS

In this paper, we propose a novel model named TagSearcher for image auto-annotation, using tag-related random search over range-constrained visual neighbors. TagSearcher proposes to use a constrained range of visual neighbors for label propagation, and utilizes tag-related random search processes to find out the trustworthy part for each candidate tag. By merging weights of visual neighbors, votes for candidate tags and tag-specific trust degrees of visual neighbors in score predictions for candidate tags, TagSearcher can not only achieve satisfactory annotating performances but also effectively reduce the performance sensitivity.

6. ACKNOWLEDGMENTS

This research was supported by the National Basic Research Project of China (Grant No. 2011CB707000), the National Natural Science Foundation of China(Grant No. 60972096, 61005045).

7. REFERENCES

[1] X. Li, L. Chen, L. Zhang, F. Lin, and W. Ma. Image annotation by large-scale content-based image retrieval. In *MM '06*.

[2] F. Wang and C. Zhang. Label propagation through linear neighborhoods. In *ICML '06*.

[3] J. Li and J.Z. Wang. Real-time computerized annotation of pictures. *PAMI*, 30(6):985 –1002, 2008.

[4] X. Wang, L. Zhang, X. Li, and W. Ma. Annotating images by mining image search results. *PAMI*, 30(11):1919 –1932, 2008.

[5] X. Li, C.G.M. Snoek, and M. Worring. Annotating images by harnessing worldwide user-tagged photos. In *ICASSP '09*.

[6] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV '09*.

[7] M. Lux and S.A. Chatzichristofis. Lire: lucene image retrieval: an extensible java cbir library. In *MM '08*.

[8] G. Carneiro, A.B. Chan, P.J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *PAMI*, 29(3):394 –410, 2007.

[9] S.L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR '04*.

[10] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma. Image annotation via graph learning. *Pattern Recogn.*, 42(2):218–228, 2009.

[11] C. Wang, S. Yan, L. Zhang, and H. Zhang. Multi-label sparse coding for automatic image annotation. In *CVPR '09*.

[12] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV '08*.

[13] Xiao Ke, Shaozi Li, and Donglin Cao. A two-level model for automatic image annotation. *Multimed Tools Appl.*, pages 1–18, 2011.

[14] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D.N. Metaxas. Automatic image annotation using group sparsity. In *CVPR '10*.

[15] X. Xu, Y. Jiang, L. Peng, X. Xue, and Z. Zhou. Ensemble approach based on conditional random field for multi-label image and video annotation. In *MM '11*.