Research Article

Qi Yao, Dayang Jiang*, and Xiancheng Ding

Image retrieval based on weighted nearest neighbor tag prediction

https://doi.org/10.1515/jisys-2022-0045
received August 29, 2021; accepted February 26, 2022

Abstract: With the development of communication and computer technology, the application of big data technology has become increasingly widespread. Reasonable, effective, and fast retrieval methods for querying information from massive data have become an important content of current research. This article provides an image retrieval method based on the weighted nearest neighbor label prediction for the problem of automatic image annotation and keyword image retrieval. In order to improve the performance of the test method, scientific experimental verification was implemented. The nearest neighbor weights are determined by maximizing the training image annotation, and experiments are carried out from multiple angles based on the Mahalanobis metric learning integration model. The experimental results show that the proposed tag correlation prediction propagation model has obvious improvements in accuracy, recall rate, break-even point, and overall average accuracy performance compared with other widely used algorithm models.

Keywords: automatic annotation of images, tag, nearest neighbor weight, relevance prediction, logistic discrimination, precision, recall rate

1 Introduction

Automatic image annotation is a very important and very active research topic in computer vision research [1–3], and its goal is to obtain relevant keywords that can predict new images from the annotated vocabulary. The prediction of these keywords can be used to provide tags for images or to provide images for tags or tag combinations. Retrieval models based on image annotations and keywords mainly use four methods: the method based on the topic model or the mixed model, the discriminative training methods, and the nearest neighbor-type model methods. The methods are based on topic models, such as latent Dirichlet assignment, probabilistic latent semantic analysis, and hierarchical Dirichlet process [4,5]. Inspired by machine translation [6], the translation method from discrete visual features to annotated vocabulary can also be understood as a topic model, which uses a topic for each visual descriptor type. The hybrid model approach is to use a hybrid model determine the joint distribution of image features and annotation labels. Other models use training images as elements to define a hybrid model of visual features and labels [7,8]. The aforementioned two types of generative models are more or less imperfect, so the label prediction discriminant model is proposed [9,10]. This type of method learns a separate classifier for each label and uses these classifiers to predict whether each test image belongs to the image category annotated with each specific label.

* **Corresponding author: Dayang Jiang,** Changzhou College of Information Technology, Changzhou, Jiangsu, 213164, China, e-mail: jiangdayang@ccit.js.cn
Qi Yao: Changzhou College of Information Technology, Changzhou, Jiangsu, 213164, China, e-mail: yaoqi@ccit.js.cn
Xiancheng Ding: Changzhou University, Changzhou, Jiangsu, 213164, China, e-mail: dxc@cczu.edu.cn

As the amount of available training data are increasing, local learning techniques are becoming more and more attractive as a simple and effective alternative to parameterized models. Such techniques include label diffusion methods based on the similarity map of labeled and unlabeled images or learning a discriminant model of the neighborhood of test images [11]. Johnson et al. [12] proposed a simple specific nearest-neighbor label transfer mechanism, which also combines images with multiple common labels and images that do not share any labels by learning a binary classifier, but this linear distance combination does not get a better result than equal weight combination; Uricchio et al. [13] proposed a label delivery framework based on nuclear canonical correlation analysis, which preserves the correlation between visual features and text features as semantic embedding. This method can work when the training set can be well annotated and when the training set is noisy.

The above-mentioned research mainly has two shortcomings: first, the model is usually estimated to maximize the generation possibility of image features and annotations, which may not be optimal for label prediction; second, many parameterized models are not enough to accurately capture the complex dependencies between image content and annotations.

Based on the shortcomings of the above research, this article proposes a new improved model of label correlation prediction – label propagation model. The method is based on the weighted nearest neighbor method, which predicts labels through the weighted combination of non-appearance/appearance of labels between neighborhoods. First, the neighbor's weight is determined according to the neighbor's rank or distance, and it is automatically set by maximizing the possibility of annotation in the training image set. For ranking-based weights, the kth neighbor always receives a fixed weight, while the distance-based weight decays exponentially with distance; second, the model allows the integration of metric learning, so that the Mahalanobis metric between image features can be optimized or the cost is less (a combination of several distance measurements) to determine the neighbor weights of the label prediction task; third, tag propagation includes a logical discriminant model of specific words. The model uses the label prediction of the word invariant model as input. It can also increase the probability of label appearance about rare words or suppress the labels of very frequent words by using exactly two parameters for each word, thereby significantly increasing the number of recall words (i.e., assigned to at least one test image). In order to evaluate the model in this article, three data sets are used-Corel 5k, IAPR TC12, and ESP Game, and standard metrics including accuracy, recall rate, break-even point (BEP), and total average accuracy is used. The label correlation prediction propagation model proposed in this article is algorithmically optimized. The target correlation maximization weight is established by calculating distance-based weights, and the integrated algorithm of metric learning is used to optimize the feature extraction method. This algorithm model has significantly improved prediction accuracy and recall rate, and the improvement of accuracy and recall rate has a direct impact on the performance of the BEP and the total average accuracy.

2 Tag correlation prediction-tag propagation model

The goal of this stage is to predict the relevance of image annotation tags and, then based on these correlation predictions, annotate images by ranking the tags of a given image, or achieve keyword-based retrieval by ranking images with a given tag. The model proposed in this article is based on the weighted nearest neighbor method; that is, the annotation of the training image is propagated to the new image. The proposed model learns in a discriminative manner, rather than adopting neighbors in a specific way to assume certain visual similarities between images [12] the measured value of the performance or distance is given.

2.1 Weighted nearest neighbor label prediction

In the predictive model design process, in order to facilitate the modeling of image annotations, a Bernoulli model is used for each keyword. This model is used because keywords are different from natural texts, only

appearing or not appearing. Here, $y_{iw} \in \{-1, +1\}$ is used to represent the non-appearance and appearance of the keyword w of the image i, that is, to realize the encoding of the image annotation. The label appearance prediction $p(y_{iw} = +1)$ of the image i is the weighted sum of the training image, and the equation is as follows:

$$p(y_{iw} = +1) = \sum_j \pi_{ij} p(y_{iw} = +1|j), \tag{1}$$

$$p(y_{iw} = +1|j) = \begin{cases} 1 - \varepsilon & y_{jw} = +1 \\ \varepsilon & y_{jw} = -1, \end{cases} \tag{2}$$

where π_{ij} represents the weight value of image j used to predict the image i tag, $\pi_{ij} \geq 0$, and $\sum_j \pi_{ij} = 1$. The constant ε of equation (2) is used to avoid zero prediction probability. In practice, set $\varepsilon = 10^{-5}$.

In order to estimate the parameters that control the weight π_{ij}, we maximize the log-likelihood of the training annotation prediction (note that the weight of the training image itself is set to zero, i.e., $\pi_{ij} = 0$); that is, the goal is to maximize, and the equation is as follows:

$$L = \sum_{i,w} c_{iw} \ln p(y_{iw}). \tag{3}$$

In equation (3), c_{iw} is used to represent the unbalanced cost between the appearance and non-appearance of keywords. In practical applications, there are more tags that do not appear than tags that appear, and there is more noise of tags that do not appear than tags that appear. This is because most of the tags in the annotation are related, but the annotation usually does not include all related tags. Suppose $y_{iw} = +1$, $c_{iw} = 1/n^+$, where n^+ is the total number of positive tags. Similarly, when $y_{iw} = -1$, $c_{iw} = 1/n^-$, where n^- is the total number of negative tags.

2.1.1 Weights based on ranking

In the case of ranking-based weights for K neighbors, if j is the kth nearest neighbor of i, set $\pi_{ij} = \gamma_k$. The log-likelihood of the data in equation (3) is concave with respect to the parameter γ_k, and it can be estimated using the EM algorithm or the gradient projection algorithm. The equation for the derivation of equation (3) with respect to γ_k is as follows:

$$\frac{\partial L}{\partial \gamma_k} = \sum_{i,w} \frac{c_{iw} p(y_{iw}|n_{ik})}{p(y_{iw})}, \tag{4}$$

where n_{ik} represents the index of the kth neighbor of image i, the number of parameters is equal to the neighbor size K, and this ranking-based weight model is called RK.

2.1.2 Distance-based weights

Of course, the weight can also be directly defined as a function of distance. The advantage of this is that the weight will depend on the distance. Redefine the weight of the training image j used to predict the image i as follows:

$$\pi_{ij} = \frac{\exp(-d_\theta(i, j))}{\sum_{j'} \exp(-d_\theta(i, j))}. \tag{5}$$

In equation (5), d_θ is the distance metric using parameters θ, which is the object we want to optimize. Note that the weight π_{ij} decays exponentially with the distance d from the image i. The selection of d can be the Mahalanobis distance dM parameterized by the positive semi-definite matrix M, such as $d_{\boldsymbol{W}}(i,j) = \boldsymbol{W}^T \boldsymbol{d}_i$. Here, $\boldsymbol{d}_{ij}$ is the base distance vector between the images i and j, W is the positive coefficient including the

linear distance combination, and the number of parameters is equal to the number of base distances of the combination. When a single distance model is used, it is called SD. At this time, W is a scalar, which controls the attenuation of the weight with distance, and it is the only parameter of the model. When multiple distance models are used, call it ML and use it for metric learning.

For the new weight equation (5) and the projected gradient algorithm, under the positive constraint of the elements of W, the gradient of the log-likelihood equation (3) with respect to W is calculated as follows:

$$\frac{\partial L}{\partial \boldsymbol{W}} = \sum_{i,j} W_i(\pi_{ij} - \rho_{ij})\boldsymbol{d}_{ij}, \tag{6}$$

$$\rho_{ij} = \sum_{w} \frac{c_{iw}}{W_i} p(j|y_{iw}), \tag{7}$$

where $W_i = \Sigma_w c_{iw,}$ ρ_{ij} represents the weighted average of all words w of the posterior probability of the neighbor j of the given annotation image i. In order to reduce the computational cost of training the model, equation (7) does not calculate all pairs of π_{ij} and ρ_{ij}. For each i, calculate them only on a large set, assuming that the remaining π_{ij} and ρ_{ij} are zero.

For each i, select K neighbors such that $k^* = \min\{kd\}$ is maximized, and k_d is the largest neighbor ranking. For this ranking, neighbors 1 to k whose distance is d are included in the selected neighbors. In this way, it is possible to include all images with greater than π_{ij}, without considering the learned distance combination W. Therefore, after determining these neighbors, the algorithm has a linear relationship with the number of training images.

2.2 Logical discriminant model of specific words

The weighted nearest neighbor label prediction method described in Section 2.1 often has a relatively low recall rate, because in order to obtain a high probability of label appearance, it needs to appear in most neighbors with important weights. However, this situation is unlikely to appear on rare tags. Therefore, in order to overcome this shortcoming, a specific word logical discriminant model is introduced to increase the probability of rare tags and reduce the probability of very frequent tags. The model uses weighted neighbor prediction, and the equation is as follows:

$$p(y_{iw} = +1) = \sigma(\alpha_w x_{iw} + \beta_w), \tag{8}$$

$$x_{iw} = \sum_{j} \pi_{ij} y_{jw}. \tag{9}$$

In equation (8), x_{iw} is the weighted average of the annotations of the label w between i's neighbors. For a fixed π_{ij}, the model is a logical discriminant model, and the log-likelihood is concave in $\{\alpha_w,\beta_w\}$, and each keyword can be trained. When a logical discriminant model is used, the log-likelihood of training annotations is used. The gradient calculation of the parameters of the control weight is equation (10), and the equations for the model based on ranking and distance are (11) and (12) respectively, as follows:

$$\frac{\partial L}{\partial \theta} = \sum_{i,w} c_{iw}\alpha_w p(-y_{iw}) y_{iw} \frac{\partial x_{iw}}{\partial \theta}, \tag{10}$$

$$\frac{\partial x_{iw}}{\partial y_k} = y_{n_{ik}w}, \tag{11}$$

$$\frac{\partial x_{iw}}{\partial \boldsymbol{W}} = \sum_{j} \pi_{ij}(x_{iw} - y_{jw})\boldsymbol{d}_{ij}. \tag{12}$$

In practice, the parameters θ and $\{\alpha_w,\beta_w\}$ are usually estimated in an alternating manner, and fast convergence is observed after the alternating maximization three times.

3 Data set and experimental setup

3.1 Data set

Considering the publicly available data sets often used in this type of research, Table 1 gives some statistics on the three data sets.

The example image is shown in Figure 1. The figure shows five examples of annotated images and the predictions using the model in this article. Next to each image, the real annotations (left) are given, as well as the five highest correlation predictions (correct prediction values underlined) given by the label propagation model (σML variant of $K = 200$) in this article (right). Pay attention to the differences between the data sets. For example, the real annotations do not always contain all relevant tags ("Water" in the second image in Figure 1(a)) and may also include whether they are related (Figure 1(c) the "Lot") label of the second image.

Corel 5k: This data set is used in most image retrieval and image annotations and has become an important benchmark for keyword-based image retrieval and image annotation. It contains about 5,000 images manually annotated with 1–5 keywords, and the vocabulary contains 260 words. A fixed set of 499 images is used for testing, and the rest are used for training.

ESP Game: This data set is obtained from an online game with two players. This article uses 20,000 subsets of the available 60,000 images. This data set is very challenging because it contains a variety of images: logos, drawings, and personal photos.

IAPR TC12: The 20,000 images in this dataset are accompanied by descriptions in multiple languages, which are mainly used for cross-language retrieval [14].

3.2 Feature extraction

In order to extract different types of features for image search and classification, this article adopts two types of global image descriptors: Gist features [15] and color histograms, each color channel has 16 buckets, represented by RGB, LAB, and HSV. Local features include SIFT and robust hue descriptors [16]. Each local feature descriptor is quantized using the k-means from training set samples; all descriptors except Gist are L1-normalized and are spatially calculated in the arrangement. What is computed here is the histogram of the image over three horizontal regions and concatenated to form a new global descriptor.

3.3 Evaluation method

For the evaluation of precision and recall with a fixed number of annotations: in the experiment, each image was annotated with the five most relevant keywords, and then, the average precision and recall of these five keywords were calculated using the model in this article ($K = 200$). (The average precision is represented by P, and the recall is represented by R.) $N+$ represents the number of keywords with non-zero

Table 1: Statistics of the training sets of the three data sets

	Corel 5k	ESP game IAPR	TC12
Vocabulary size	260	268	291
Number of images	4,493	18,689	17,665
Words per image	3.4/5	4.7/15	0.7/23
Image per word	58.6/1,004	362.7/4,553	347.7/4,999

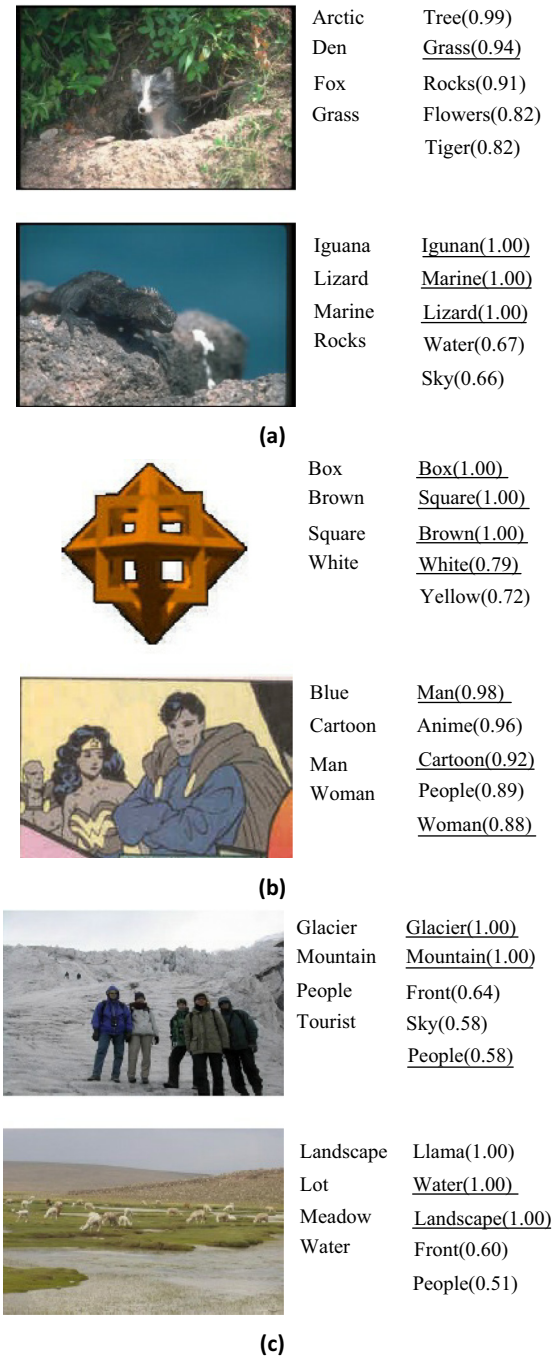

Figure 1: Example test images from the three data sets: (a) Corel 5k, (b) ESP Game, and (c) IAPR TC12.

recall values, with the caveat that each image must be annotated with five keywords. Finally, accurate experimental results are obtained according to the performance comparison of various variants of this model with other algorithmic models of P, R, and $N+$.

For the evaluation of precision with different recall rates: in the experiments, the precision with different recall rates, that is, BEP or R-precision, is also calculated. It is a measure of the accuracy of the top n_w-related images for each keyword w, where n_w is the number of images annotated with this keyword. The mean Average Precision (mAP) [17] is obtained by calculating the average precision of each keyword, which is measured after each relevant image is retrieved. Then, according to the relationship of the distance

model and its variants with respect to the neighbor size K in terms of P, R, BEP, and mAP performance, the relevant data are calculated, and the graph is drawn accordingly to determine the difference between them, and then, the experimental results are obtained.

4 Experimental results

4.1 Corel 5k experimental results

The first set of experiments is to compare the performance of different variants of our algorithmic model. It is mainly compared with the following algorithms: such as the label diffusion method of ref [2], the original results of the specific nearest neighbor label transfer mechanism of ref [12] (denoted as JEC), the results obtained by ref [12] using the features of this article (denoted as JEC-15, that is, a weighted combination of 15 normalized base distances is used to determine the similarity of images), and ref [13] of a label transfer algorithm based on kernel canonical correlation analysis.

Table 2 shows the experimental results and also that, using the features extracted in this article, the label transfer mechanism proposed in refs [12] and [13] can obtain results that are very similar to their original results. Therefore, the other performance difference obtained with the algorithmic model in this article lies in the label prediction method. The model described in this article performs quite well using this combination of fixed distances to determine weights (either directly in SD or in RK). Among these results, the performance results of the σSD model with our distance-based weights are the best.

More importantly, the ensemble metric learning model (ML and σML) adopted in this article is even more improved. In particular, the σML variant has a 5% improvement in precision and a 9% improvement in recall compared to JEC-15 with the same features, and the number of words with positive recall exceeds 20. This indicates that the nearest neighbor-type label prediction in this article benefits from the integration of metric learning in the prediction model.

Figure 2 shows the relationship curve between the P, R, BEP, and mAP performance of the distance model and the number of neighbors, K. As can be seen from Figure 2, for all neighbors, regardless of whether there is σ, the distance combination obtained by metric learning is always better than the equal weight combination. Also, adopting a large number of neighbors (like more than 100) can improve performance, especially for the ML variant. The reason is that in ML, the ranking of adjacent images varies with the learned metric. Therefore, to ensure that all useful training images are included in the initial neighborhood [18], these sets need to be large enough.

Figure 3 shows the average recall of words in buckets [19,20] using ML and its variant σML. The blue and yellow bars represent the average recall of ML and its variant ML, respectively. As analyzed in Section 2.2, the introduction of a word-specific logical discriminant model increases the probability of rare labels and reduces the probability of very frequent labels, making the improvement for rare words higher.

From these experimental results above, it can be seen that the distance-based variants (σRK, σSD, and σML) perform the best. So use them for the other two datasets that follow and take $K = 200$ as the default choice for the number of neighbors.

Table 2: Performance comparison of algorithm model ($K = 200$) and various variants with other algorithm models in P, R, and $N+$

	References [2]	References [13]	JEC	JEC-15	RK	σRK	SD	σSD	ML	σML
P (%)	23	28	27	28	28	26	30	28	31	33
R (%)	29	32	32	33	32	34	33	35	37	42
$N+$	137	138	139	140	136	143	136	145	146	160

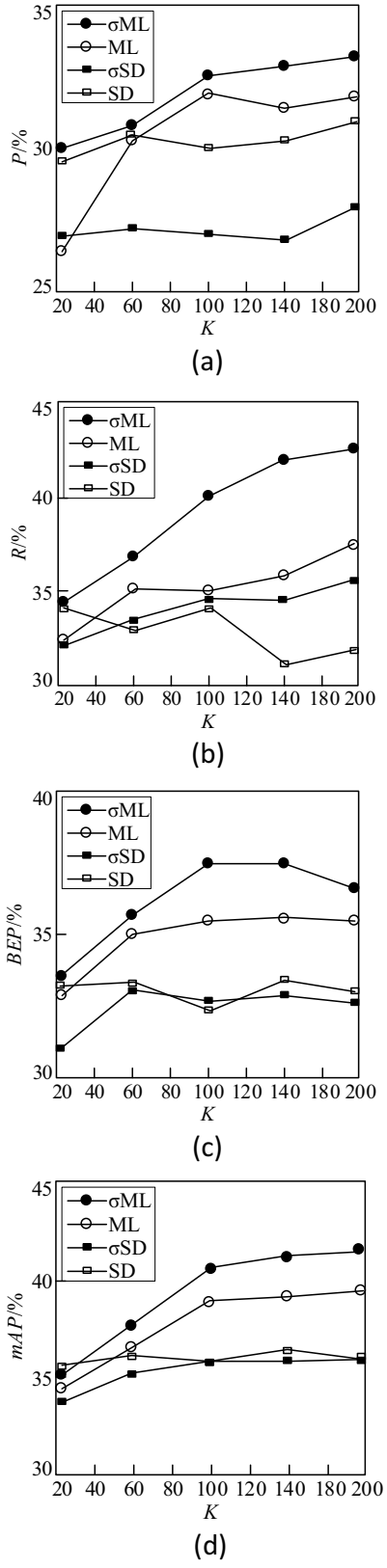

Figure 2: The relation of neighbor size *K* in terms of *P*, *R*, BEP, and mAP performance based on the distance model and its variants: (a) *P* versus the number of neighbors, (b) *R* versus the number of neighbors, (c) BEP versus the number of neighbors, and (d) mAP versus the number of neighbors.

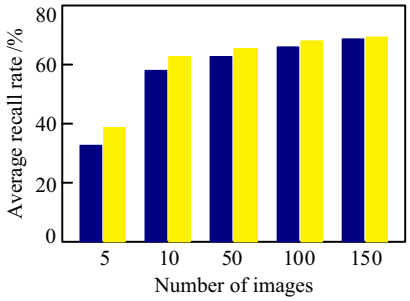

Figure 3: Comparison of average recall rate of words in ML and σML bins.

4.2 Experimental results of ESP Game and IAPR TC 12

The distance-based variant of this article uses a distance combination of equal weight and metric learning, and Table 3 shows the results obtained for these two datasets. It can be seen that compared to the label diffusion method of ref. [2], and the two algorithms of the specific nearest-neighbor label transfer mechanism model of ref. [12], and the performance of the label transfer algorithm based on kernel canonical correlation analysis of ref. [13], the algorithm model used in this article has obvious improvement. In addition, it can be seen from Table 3 that for IAPR TC 12 and ESP Game, compared to the Corel 5k data set, the most significant improvement is the increase in accuracy.

Figure 4 shows for two data sets ESP Game and IAPR TC12, the relationship between the different performance measures P, R, BEP, and mAP of the two variants σSD and σML of the distance-based algorithm model and the neighbor size K in this article. It can be seen that the ensemble metric learning algorithm model has the best performance.

4.3 Image retrieval in multi-word query

The above experimental results are aimed at image retrieval performance for single-word queries, but any practical image retrieval system should support multi-word queries. This section presents the BEP and mAP performance of our algorithm model for multi-word query on the Corel 5k dataset and compares it with the multi-word query-based image retrieval in ref. [9]. To facilitate direct comparison, the experiments use a subset of 179 words from the 260 annotated words of Corel 5k, and they appear at least twice in the test set. Images were considered relevant to the query when they were annotated with full words, and all 2,241 queries consisting of 1 or more words were considered. This way the test set contains at least one relevant image. Table 4 shows the obtained experimental results.

Table 3: Two variants of the proposed algorithm model ($K = 200$) comparison with other algorithm models in terms of the performance of P, R, and $N+$

	IAPR TC12			ESP game		
	P	***R***	***N+***	***P***	***R***	***N+***
References [2]	24	23	223	18	19	209
JEC	28	29	250	22	25	224
JEC-15	29	19	211	24	19	222
References [13]	28	27	233	23	20	221
SD	50	20	215	48	19	212
σSD	41	30	259	39	24	232
ML	48	25	227	49	20	213
σML	46	35	266	39	27	239

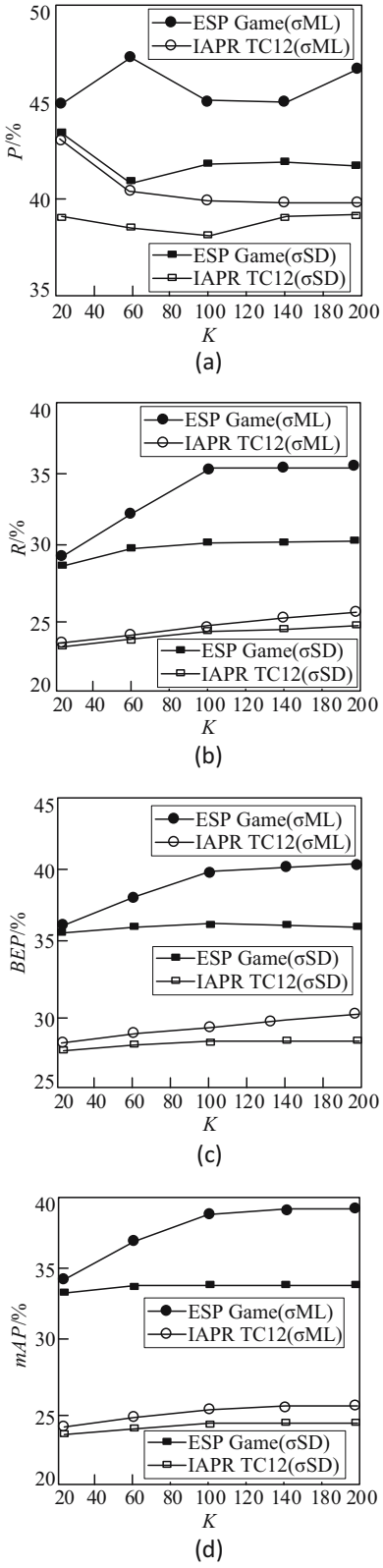

Figure 4: Relationship between neighbor size K and based on the distance model variants of performance of σSD and σML in terms of P, R, BEP, and mAP: (a) P versus the number of neighbors, (b) R versus the number of neighbors, (c) BEP versus the number of neighbors, and (d) mAP versus the number of neighbors.

Table 4: Comparison between the algorithm model in this article (K = 200) and the algorithm in the literature [9] in terms of mAP and BEP performance

	SD	σSD	ML	σML	References [9]
BEP	24	23	27	27	17
mAP	32	31	36	36	26

From Table 4, it can be seen that the algorithm model in this article has an average improvement of about 6–10% in mAP performance compared with the image retrieval based on multi-word query [9] for all query categories and in terms of BEP performance that also got a gain of about 6–10%.

5 Concluding remarks

This article proposes a new model for image retrieval based on image annotations and keywords. These models combine weighted nearest neighbor methods and metric learning capabilities in a discriminative framework. Extensive experimental results based on several performance metrics on three typical data sets show that the ML variant of the label propagation model proposed in this article (i.e., employing distance-based weights and integrating metric learning) has the best performance. On all data sets, it not only has a good recall rate and high precision on the set but also significantly improves the recall rate of rare words and the overall performance.

In future research, we will further consider extending the model and assigning labels to image regions to address tasks such as image region labeling and object detection from image range annotations.

Conflict of interest: Authors state no conflict of interest.

References

[1] Zhang W, Hu H. Training visual-semantic embedding network for boosting automatic image annotation. Neural Process Lett. 2018;48(3):1503–19.
[2] Liang Y, Xin Z, Xiaohai HE, Shuhua X, Linbo Q. Violent image annotation using ensemble learning. J Terahertz Sci Electron Inf Technol. 2020;18(2):306–12.
[3] Mehmood Z, Mahmood T, Javid MA. Content-based image retrieval and semantic automatic image annotation based on the weighted average of triangular histograms using support vector machine. Appl Intell. 2018;48(1):166–81.
[4] Houlin Q, Lei G. KNN text classification algorithm for probabilistic latent semantic analysis. Computer Technol Dev. 2017;27(7):57–61.
[5] Tian DP. Semi-supervised learning based probabilistic latent semantic analysis for automatic image annotation. High Technol Lett. 2017;23(4):367–74.
[6] Ruiying J, Lei C, Jing H, Ming Z, Zhigeng P. Automatic generation of Chinese metrical poetry based on topic model and statistical machine translation method. Acta Computer Sin. 2015;38(12):2426–36.
[7] Yao Z, Xiaojiao M, Yubin Y. Visual object tracking based on multi feature hybrid model. J Nanjing Univ (NAT SCI Ed). 2016;52(4):762–70.
[8] Panlong R. Research on hybrid recommendation system model based on multi-dimensional features. Chengdu: University of Electronic Science and Technology; 2018. p. 56–67.
[9] Huang S, Ye J, Wang T, Jiang L, Xing C, Li Y. Learning a similarity constrained discriminative kernel dictionary from concatenated low-rank features for action recognition. IEICE Trans Inf Syst. 2016;E99.D(2):541–4.
[10] Cong J. Joint significance detection based on generation model and discriminant model. Dalian: Dalian University of Technology; 2015. p. 41–58.
[11] Abu Alfeilat HA, Hassanat A, Lasassmeh O, Tarawneh AS, Alhasanat MB, Eyal Salman HS, et al. effects of distance measure choice on K-nearest neighbor classifier performance: a review. Big Data. 2019;7(4):221–48.

[12] Johnson J, Ballan L, Li FF. Love thy neighbors: image annotation by exploiting image metadata. 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile; 2015. p. 4624–32.
[13] Uricchio T, Ballan L, Seidenari L, Del Bimbo A. Automatic image annotation via label transfer in the semantic space. Pattern Recognit. 2017;71:144–57.
[14] Jia L. Research on cross language retrieval platform based on co-occurrence of words. J Inf. 2015;34(8):195–8.
[15] Zhang YH, Du JX, Wang J, Zhai CM. Reverse training for leaf image set classification. in Proceedings of Advanced Intelligent Computing Theories and Applications: 11th International Conference (ICIC), Fuzhou, China; 2015. p. 233–42.
[16] Shinde SR, Sabale S, Kulkarni S, Bhatia D. Experiments on content based image classification using Color feature extraction. Proceedings of 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India; 2015. p. 1–6.
[17] Meng W. Research on image annotation based on graph learning and generation countermeasure network. Dalian: Dalian University of technology; 2020.
[18] Yanchun M, Yongjian L, Qing X, Shengwu X, Lingli T. Overview of automatic image annotation technology. Computer Res Dev. 2020;57(11):2348–74.
[19] Jianfang C, Aidi Z, Zibang Z. Application of convolution neural network based on fusion threshold optimization in image annotation. Computer Appl. 2020;40(6):1587–92.
[20] Markatopoulou F, Mezaris V, Patras I. Implicit and explicit concept relations in deep neural networks for multi-label video/image annotation. IEEE Trans Circuits Syst Video Technol. 2019;29(6):1631–44.