Image Annotation by Learning Label-Specific Distance Metrics⋆

Xing Xu, Atsushi Shimada, and Rin-ichiro Taniguchi

Department of Advanced Information Technology, Kyushu University, Japan
{xing,atsushi}@limu.ait.kyushu-u.ac.jp, ring@ait.kyushu-u.ac.jp

Abstract. Recently, weighted k nearest neighbor based label prediction model combined with distance metric learning (KNN+ML) [10,14,17], has become more attractive and showed exciting results on image annotation task. Usually, in KNN+ML framework, a uniform distance metric is learned given a collection of similar/dissimilar image pairs from training data. Thus, for a couple of images, their distance is globally unique. However, this might not be sufficient for label prediction on annotation task because it is impossible to distinguish the multiple labels attached to each image. In this paper, we are motivated to learn multiple label-specific distance metrics, and measure the distance of an image pair under different labels' distance metrics. We also propose a novel label specific prediction model, in which the weight of each label is determined by its specific distance value rather than previous global distance value. Compared with previous KNN+ML methods, our proposed method is able to exactly discriminate each label in each neighbor, and efficiently reduce the prediction of *false positive* and *false negative* labels. Extensive experimental results on three benchmark datasets demonstrate that proposed method achieves more accurate annotation results and competitive overall performance.

1 Introduction

The task of image annotation is to automatically assign keywords to an image, and it has become an active topic in computer vision and machine learning areas due to its potential useful applications, including image search and photo management. Recently, k nearest neighbor (KNN) based methods has been successfully applied to image annotation problem, as this kind of local learning technique has potentiality to capture the similarity graph of labeled and unlabeled images.

In order to extend KNN based methods to image annotation task, two primary issues need to be considered. The first is how to select appropriate neighbors for an unlabeled image. Metric learning (ML) methods [11,13,19] are often imported to find optimal metric over feature space of provided pairs of labeled images, and linear combination of base metrics for multiple high-dimensional features

is alternative to traditional low-dimensional Mahalanobis metric. The second is how to design efficient label transfer mechanism through learned distance metric. Some well known methodologies including greedy diffusion [14, 23], weighted nearest neighbor label prediction [10, 17, 21], are usually adopted.

Fig. 1. For a test image from ESP Game dataset (first column), the first and second rows on the right section show its 4 nearest images and distance value of each label from proposed method (uses label-specific distance metrics) and JEC [17] (uses global distance metric). Predicted labels (with smallest distance values) in two methods can be compared with ground truth {couple, eye, hat, man, smile, woman}.

The main shortcomings of existing works based on KNN+ML are two folds. First, these works incline to use a single global distance metric to measure the similarity of an image pair, which is convincing to address traditional classification problem. However, in multi-label annotation condition, the degree of similarity ought to vary upon different label affiliated to the image pair. Second, in the celebrated weighted nearest neighbor label prediction model [10, 17], each label of one neighbor has identical weight ($\exp(-D(\cdot))$ in Equ. 1) since the distance is uniquely determined by global distance metric. Thus, during the final label prediction, some labels would get equal weight, such as {*asian*, *hat*} in JEC in Fig. 1, where we can only select these two labels {*asian*, *hat*} randomly and the accuracy of final annotation would be disturbed.

As the proverb goes, "there are a thousand Hamlets in a thousand people's eyes." Given an image pair, we are motivated to measure discriminative distance values ("*Hamlets*") by different label specific distance metrics ("*eyes*"). Then we can distinguish labels according to their specific distance values. In this paper, we propose a new weighted nearest neighbor type model that predicts each label's weight ($\exp(-D_{y_l}(\cdot))$ in Equ. 2) according to its specific distance value. As shown in Fig.1, unlike that different labels of one neighbor share same distance value in JEC, each label of one neighbor has its specific distance value in proposed method. This ensures proposed method to discriminate labels in one neighbor, select proper labels and avoid irrelevant labels simultaneously.

Our contributions are: 1) We propose a label-specific prediction model for annotation task, where the weight of each label in a neighbor is measured by its specific distance metric. 2) We extend [11] in a high dimensional multi-feature fusion setting to learn distance metric for each label. 3) We design a complete annotation framework of training and testing procedures, including learning label-specific distance metrics and predicting labels for new image.

In the next section, we review previous notable works in image annotation field. In Sect.3, we describe our label-specific prediction model, distance metric learning algorithm, and entire training and testing procedures. Experimental results compared with previous KNN+ML based methods are presented in Sect.4. Finally, we make conclusion in Sect.5.

2   Related Work

Large quantities of approaches have been proposed to address image annotation problem. One pipeline of research focuses on modeling medium sized image databases with fixed vocabularies. A common consensus has been reached from [3,8,10,14,17,23] that three main groups exist: 1) Generative models [2,5,22] aim to learn the joint probability of labels and image features, various relationships between semantic and visuality have been imagined, e.g. mixture of topics, and different hypotheses of probability distribution of labels and image features have been assumed, such as multinomials, separate Bernoullis, mixture of Gaussian. 2) Discriminative models [3, 12] treat each label as a semantic class of multi-class multi-label problem, and learn a separate classifier for each label, where balanced training data is required and correlation among the labels may be ignored. 3) Nearest neighbor based models [3,10,14,17,23] have become more attractive recently and shown state-of-the-art annotation performance. As visual close similar images possess certain semantic similarity, after selecting proper neighbors for unlabeled image, labels are then transferred from these neighbors.

On the other line of research, data-driven approaches [4,16,18] have demonstrated their capacity on large-scale web-based image databases with open vocabularies. These approaches usually firstly search a group of visually closely similar candidates for the query image, and then mine relevant tags from associated clues (such as image filename, URL and surrounding texts) available on the web. These approaches can be regarded as hybrid models which combine the generative/discriminative/nearest neighbor models in more practical circumstance.

3   Proposed Method

In this section, we first propose the label-specific prediction model, then describe the label-specific distance metric learning algorithm, finally depict our annotation framework of training and testing procedures.

3.1   Label-Specific Prediction Model

Consider a collection of labeled images $\mathcal{C} = \{\{I_1, Y_1\}, ..., \{I_D, Y_D\}\}$ with a fixed vocabulary of L labels $\mathcal{Y} = \{y_1, ...y_L\}$, where each $Y_i \subseteq \mathcal{Y}$ contains multiple labels. In weighted nearest neighbor based label prediction model [10, 17], each label's presence/absence of an unseen image J is a weighted sum over its K nearest neighbors $N_J = \{I_1, ...I_K\}$ in training set. We denote the weight of k-th neighbor I_k to image J as π_{J,I_k}, where $\pi_{J,I_k} \geq 0$ and $\sum_k \pi_{J,I_k} = 1$. In previous KNN+ML works, π_{J,I_k} is usually represented by a smooth exponential function over distance $D(J, I_k)$, as $\pi_{J,I_k} = \exp\left(-D(J, I_k)\right)/\sum_{k'=1}^{K} \exp\left(-D(J, I_{k'})\right)$. Thus, the presence probability of l-th label y_l in J can be formulated as,

$$
\begin{aligned}
P(y_l = +1|J) &= \sum_{(I_k, Y_k) \in N_J} \pi_{J,I_k} \cdot \delta(y_l \in Y_k | I_k) \\
&= \sum_{(I_k, Y_k) \in N_J} \exp\left(-D(J, I_k)\right) \cdot \delta(y_l \in Y_k | I_k), \quad (1)
\end{aligned}
$$

where $\delta(y_l | I_k)$ is an indicator function that denotes the presence/absence of label y_l in I_k, with $\delta(\cdot) = 1$ when $y_l \in Y_k$ and $\delta(\cdot) = 0$ otherwise. Through ranking probabilities of all the labels $y_1 \to y_L$ based on Equ. 1, top-ranked labels can be assigned to the unseen image J.

Note that in Equ. 1, all labels in Y_k share the same weight value $\exp\left(-D(J, I_k)\right)$ since the distance value $D(J, I_k)$ is unique given the learned global distance metric. This may lead to significant potential risk that, although we can correctly predict *true positive* labels of one neighbor for image J, it is still ambiguous to reject the *false positive* or *false negative* labels in that neighbor. Inspired by [7, 20] which learns local distance functions for every training image or image clusters in visual classification task, here we aims to learn local distance metric for each label. For one image pair, under distance metrics of different labels, the distances are various. Thus, based on these local distance metrics, our label-specific prediction model can be formulated as:

$$
P(y_l = +1|J) = \sum_{(I_k, Y_k) \in N_J} \exp\left(-D_{y_l}(J, I_k)\right) \cdot \delta(y_l \in Y_k | I_k). \quad (2)
$$

Different from traditional prediction model in Equ. 1, here weight $\exp\left(-D_{y_l}(J, I_k)\right)$ includes multiple values involved in different $y_k \in Y_k$. Intuitively, we can measure the distance between neighbor image I_k and J using the specific distance metrics of labels in I_k. This allows us to distinguish the importance of each label $y_k \in Y_k$ of I_k, reduce the prediction of irrelevant labels and preserve relevant labels from Y_k.

3.2   Learning Label-Specific Distance Metrics

Suppose we have generated a collection of similar and dissimilar image pairs for the l-th label, in the manner of learning distance metric [11,13,19,20], our goal is to learn a distance metric for the l-th label to ensure distances of similar image

pairs are smaller than dissimilar pairs. Following [11], we model the probability p_n that an pair $n = (A, B)$ is similar (positive), and the pair response r_n is 1, as:

$$p_n = p(r_n = 1|A, B; D_{y_l}(A, B), b) = \sigma(b - D_{y_l}(A, B)), \qquad (3)$$

where $\sigma(z) = (1 + \exp(-z))^{-1}$ is the sigmoid function and b is a bias term. Notably, $D_{y_l}(A, B) = \sum_i u_l(i) \sum_j v_l(j) \cdot d^i_{A,B}(j)$ represents the distance of (A, B) on the view of l-th label y_l. Here in the multiple features fusion setting, $D_{y_l}(A, B)$ is a linear combination of base distances $d^i_{A,B}(\cdot)$ of multiple features [17]. Inter-feature weights $u_l(i)$ and intra-feature weights $v_l(j)$ are specific parameters we need to learn for the l-th label. We use maximum log-likelihood to optimize the parameters of the model. The log-likelihood $\mathcal{L}$ can be written as:

$$\mathcal{L} = \sum_n r_n \ln p_n + (1 - r_n) \ln(1 - p_n), \qquad (4)$$

and the gradient of $\mathcal{L}$ with respect to u_l and v_l equals

$$\frac{\partial \mathcal{L}}{\partial u_l} = \sum_n (r_n - p_n) v_l \cdot d_{A,B}, \qquad \frac{\partial \mathcal{L}}{\partial v_l} = \sum_n (r_n - p_n) u_l \cdot d_{A,B}, \qquad (5)$$

which are smooth and concave. Moreover, non-negative constraints are required to weights $\{u_l, v_l\}$ as the distance value should be non-negative. In practice, we use projected gradient ascend method to optimize $\{u_l, v_l\}$ in an alternating manner.

It is worth saying that our label-specific distance metric learning algorithm is quite different from "*word-specific logistic discriminate model (σML)*" proposed in [10]. Firstly, the target of our algorithm is to learn a metric to identify the similarity of a pair under a label, while σML aims to learn word-specific smooth factors for the weighted nearest neighbor prediction model. Secondly, our algorithm directly impacts the weight of each label (see Equ. 2, $\exp(-D_{y_l}(J, I_k))$), whereas σML does not effect weight factor, which implies similarity of a pair is still considered on a global viewpoint.

3.3 Training and Testing Procedures

To learn distance metric of each label, it's necessary to create a training set of similar/dissimilar image pairs for each label. Our scheme of obtaining label-specific image pairs is a modified version of [17], first we give some key definition:

Semantic cluster. For a label $y_l \in \mathcal{Y}$, its semantic cluster $\mathcal{S}_{y_l} \subseteq \mathcal{C}$ contains all images annotated with label y_l in training set $\mathcal{C}$.

Semantic neighborhood. For an image T, its semantic neighborhood $\mathcal{S}_T$ has L subsets $\{S_{T,y_1} \bigcup ... \bigcup S_{T,y_L}\}$, where the l-th subset $S_{T,y_l} \subseteq \mathcal{S}_{y_l}$ includes K_1 images that are most similar to T in semantic cluster $\mathcal{S}_{y_l}$.

Similar/dissimilar pairs. For an labeled image (T, Y_T) with its semantic neighborhood $\mathcal{S}_T$, its similar samples are images from S_{T,y_p}, $y_p \in Y_T$, the residual S_{T,y_q}, $y_q \in \mathcal{Y} \backslash Y_T$ are dissimilar samples.

Training
Input: A set of annotated training images $\mathcal{C} = \{\{I_1, Y_1\}, ..., \{I_D, Y_D\}\}$, L semantic cluster $\mathcal{S}_{Y_1}, ..., \mathcal{S}_{y_L}$. For each label $l = \{1, ..., L\}$, do

1. Generate semantic neighborhood using base distance measure for each sample.
2. Generate similar/dissimilar pairs for each sample in $\mathcal{S}_{y_l}$.
3. Learn parameters $\{b_l, u_l, v_l\}$ for l-th distance metric following Sect.3.2.

Output: L label-specific parameters $\{b_1, u_1, v_1\}, ..., \{b_L, u_L, v_L\}$.

Testing
Input: L label-specific parameters $\{b_1, u_1, v_1\}, ..., \{b_L, u_L, v_L\}$, L semantic clusters $\mathcal{S} = \{\mathcal{S}_{y_1}, ..., \mathcal{S}_{Y_L}\}$ from training data, a test image J. For test image J, do

1. Generate its semantic neighborhood $\mathcal{S}_J = \{S_{J,y_1} \bigcup ... \bigcup S_{J,y_L}\}$, where the l-th S_{J,y_1} contains K_1 training samples that are most similar to J measured by the l-th distance metric.
2. For each image $\forall \{I_t, Y_t\} \in \mathcal{S}_J$, do
   • Calculate the presence probability (Equ. 2) of label $y_t \in Y_t$ by its label specific distance $D_{y_t}(J, I_t)$, using l-th distance metric.
   • Accumulate probability score of label y_t.
3. From probability scores of all labels in $\mathcal{S}_J$, select top-5 labels with highest probability scores.

Output: predicted top-5 labels for test image J.

The training and testing procedures are incorporated into the complete annotation framework depicted above. Distance metrics are learned in training stage through the well organized semantic neighborhood of each training sample, which fully leverages image-image, image-label, label-label similarities. In testing stage, similarities between the test image and its neighbors are calculated according to different label's metric, finally labels that are most semantically similar to test image are dug out.

4   Experiments

In this section, we first present the experimental configuration: datasets, multiple features, evaluation measures and details, then we compare our proposed method with previous methods from different aspects.

4.1   Configuration

Datasets and Features. To keep coherence with previous works [10, 14, 17], we also consider three well-explored data sets: **Corel 5K** [6], **ESP Game** [1], **IAPR TC12** [9]. Table 1 summarizes the general statistics of the images and fused multiple features we use in our experiments.

Evaluation Measures. Following [10, 17], we choose top-5 most relevant labels for each test image. Then we compute the mean precision **P**, mean recall **R** and

Table 1. General statistics for the three datasets and multiple features. In column 6 and 7, the items are in the format "mean, maximum." Total dimension of multiple features is 13,900.

Dataset	Num. of images	Num. of labels	Training images	Test images	Labels per image	Images per label
Corel 5K	5000	260	4500	500	3.4, 5	58.6, 1004
ESP Game	20,770	268	18,689	2,081	4.7, 15	326.7, 4553
IAPR TC12	19,627	291	17,665	1,962	5.7, 23	34.7, 4999
Feature	RGB	LAB	HSV	Gist	SIFT	hue
Dimension	4,096	4,096	4,096	512	1,000	100
Base metric	L_1	L_1	L_1	L_2	χ^2	χ^2

their trade-off **F1** score ($F1 = 2.P.R/(P+R)$). Moreover, number of words with non-zero recall **N+** is also taken into account.

Details. In Sect.3.3, to extract similar/dissimilar image pairs, we use entire training images to form semantic cluster of each label on entire Corel 5K dataset, subsets of 30% random training samples of both ESP Game and IAPR TC12 datasets. K_1 is set as 4 for Corel 5K, 3 for both ESP Game and IAPR TC12 datasets. Since for a labeled image, its dissimilar pairs is far more than similar pairs (($L - K_1) \gg K_1$), following the advice in [23], we find that randomly selecting partial dissimilar pairs $4 \sim 10$ times larger than similar pairs (there are average $500 \sim 6000$ pairs (N) for each label in all three datasets) is sufficient to learn stable distance metrics.

The training and testing procedures are repeated three times on each dataset, and we choose models that perform best on testing sets for comparison. All experiments are executed using MATLAB 7.11 on a 3.4 GHz, 8GB RAM PC.

4.2 Comparison

Global *vs.* Label-Specific Distance Metric. First we compare our method with celebrated work of JEC [14] which uses global distance metric, and Fig. 2 (top row) shows performance in terms of **P**, **R** and **N+** with respect to changing neighborhood size. It is remarkable that our method makes significant improvement on these measures compared with global distance metric based method on all three datasets. In addition, unlike JEC which needs large numbers of neighbors (nearly 200) to improve performance, our proposed method achieves best performance using less neighbors (30, 100 and 80) on three datasets correspondingly. This is because in our testing procedure (in Sect.3.3), more semantically related neighbors are pulled nearer when generating semantic neighborhood.

Secondly, we follow σML in [10] and group labels according to their frequency in each dataset and explore which labels benefit most by using specific distance metrics. From Fig. 2 (bottom row), it is illustrated that our method could further care for rare labels and achieve significant improvements for these labels

Table 2. Comparison of annotation performance between proposed label-specific *vs.* previous global distance metric models

Method	Corel 5K				ESP Game				IAPR TC12			
	P	R	F1	N+	P	R	F1	N+	P	R	F1	N+
JEC [14]	27	32	29.3	139	22	25	23.4	224	28	29	28.5	250
GS [23]	30	33	31.4	146	-	-	-	-	32	29	30.4	252
TagProp (ML) [10]	31	37	33.7	146	49	20	28.4	213	48	25	32.9	227
TagProp (σML) [10]	33	42	37.0	160	39	27	31.9	239	46	35	39.8	266
2PKNN [17]	39	40	39.5	177	51	23	31.7	245	49	32	38.7	274
2PKNN+ML [17]	44	46	45.0	191	53	27	35.7	252	54	37	43.9	278
Proposed	40.5	44.7	42.5	185	44.1	26.2	32.9	247	47.6	36.1	41.1	264
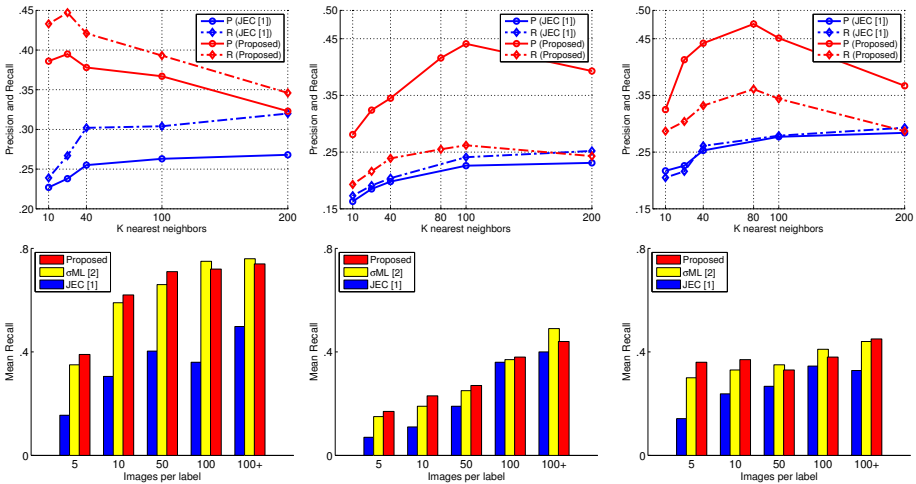
Fig. 2. Label-specific *vs.* global distance metric: performance in terms of **P**, **R** and **N+** with respect to neighborhood size changing (top row) and mean recall of labels (bottom row) on three datasets (from left to right: Corel 5K, ESP Game and IAPR TC12)

compared with σML. The reason is that smooth factor learned in σML could only change the weight for a rare label slightly, whereas the weight is directly decided and promoted by its specific distance metric in proposed method.

Comparison by Annotation Measures. Table 2 summarizes the overall evaluation from our results as well as those reported by previous KNN+ML methods on three datasets. It shows that our method outperforms previous global metric based methods, such as JEC, TagProp, and 2PKNN, but is worse than the prominent 2PKNN+ML. As for 2PKNN+ML, it utilizes sophisticated metric learning algorithm (LMNN [19]) to learn a global large marginalized distance metric, and it requires large quantities of seriously unbalanced similar/dissimilar pairs, e.g. for Corel 5K dataset, there are total 2 million training pairs, and the proportion

ESP Game   IAPR TC12

tree, grass, house, green, road	couple, glasses, people, smile, car	car, dirt, tree, water, wheel	bed, room, lamp, table, window	mountain, range, wall, terrace, front	sky, sea, hill, view, city
home, car, street, tree, people	face, man, woman, teeth, smile	truck, road, tree, green, water	bed, table, shelf, bag, night	landscape, mountain, sky, people, wall	bank, house, sky, boat, tree

Fig. 3. Annotations of example images from ESP game (left three images) and IAPR TC12 (right three images). Since these exemplars have ground truth labels more than 5 words, we explicitly compare proposed method (second row) with JEC [14] (third row) on accuracy of predicted top-5 labels.

of similar/dissimilar pairs is about 1:50. The training procedure is complex and needs to be well designed for unbalanced setting. Our method requires much less (e.g. thousands pairs per label for Corel 5K) and fairly balanced pairs for each label and is scalable to larger vocabulary. On this point, we think our method is promising and competitive to the state-of-the-art method. Moreover, in Fig.3 we present some qualitative annotation results from our method compared with results from global metric based method JEC. It shows that proposed method is able to assign more accurate labels related to image content, whereas JEC might be ambiguous to distinguish the relevant/irrelevant labels, since some equal weighted labels are selected randomly.

5 Conclusion

In this paper, we have proposed a novel label prediction model for image annotation task. In proposed model, labels of one neighbor have different weights depending on their label-specific distance values. And we have extended [11] to high dimensional multiple-feature fusion setting, to learn the specific distance metric for each label. Moreover, we have also designed complete annotation framework of training and testing procedures. To further explore our annotation framework, it is feasible to import more sophisticated metric learning algorithms in high dimensional feature (distance) space, such as LMNN based methodology used in [17] [15]. This would be a primary issue to be tackled in our future work.

References

1. von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2004)
2. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: ACM SIGIR 2003 (2003)

3. Carneiro, G., Chan, A., Moreno, P., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Transactions on PAMI (2007)
4. Dai, L., Wang, X.J., Zhang, L., Yu, N.: Efficient tag mining via mixture modeling for real-time search-based image annotation. In: ICME (2012)
5. Putthividhya, D., Attias, H.T., Nagarajan, S.S.: Topic regression multi-modal latent dirichlet allocation for image annotation. In: CVPR (2010)
6. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
7. Frome, A., Singer, Y., Sha, F., Malik, J.: Learning globally-consistent local distance functions for shape-based image retrieval and classification. In: ICCV (2007)
8. Fu, H., Zhang, Q., Qiu, G.: Random forest for image annotation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 86–99. Springer, Heidelberg (2012)
9. Grubinger, M.: Analysis and Evaluation of Visual Information Systems Performance. Ph.D. thesis, Victoria University (2007)
10. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: ICCV (2009)
11. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? metric learning approaches for face identification. In: ICCV (2009)
12. Huang, S.J., Zhou, Z.H.: Multi-label learning by exploiting label correlations locally. In: AAAI 2012 (2012)
13. Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P., Bischof, H.: Large scale metric learning from equivalence constraints. In: CVPR (2012)
14. Makadia, A., Pavlovic, V., Kumar, S.: A new baseline for image annotation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 316–329. Springer, Heidelberg (2008)
15. Mensink, T., Verbeek, J., Perronnin, F., Csurka, G.: Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 488–501. Springer, Heidelberg (2012)
16. Torralba, A., Fergus, R., Freeman, W.: 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Transactions on PAMI (2008)
17. Verma, Y., Jawahar, C.V.: Image annotation using metric learning in semantic neighbourhoods. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 836–849. Springer, Heidelberg (2012)
18. Wang, X.J., Zhang, L., Liu, M., Li, Y., Ma, W.Y.: Arista - image search to annotation on billions of web photos. In: CVPR (2010)
19. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: NIPS (2006)
20. Weinberger, K., Saul, L.: Fast solvers and efficient implementations for distance metric learning. In: ICML (2008)
21. Wu, P., Hoi, S.C.H., Zhao, P., He, Y.: Mining social images with distance metric learning for automated image tagging. In: WSDM (2011)
22. Xiang, Y., Zhou, X., Chua, T.S., Ngo, C.W.: A revisit of generative model for automatic image annotation using markov random fields. In: CVPR (2009)
23. Zhang, S., Huang, J., Huang, Y., Yu, Y., Li, H., Metaxas, D.: Automatic image annotation using group sparsity. In: CVPR (2010)