

Image Semantic Distance Metric Learning Approach for Large-scale Automatic Image Annotation

Cong Jin¹ and Shu-Wei Jin²

¹*School of Computer, Central China Normal University, Wuhan, 430079, P. R. China*

²*Département de Physique, École Normale Supérieure, 24, rue Lhomond 75231, Paris, Cedex 5, France*

Keywords: Large-scale Automatic Image Annotation, Image Semantic Distance Metric Learning, Improve Performance, Semantic Similarity.

Abstract: Learning an effective semantic distance measure is very important for the practical application of image analysis and pattern recognition. Automatic image annotation (AIA) is a task of assigning one or more semantic concepts to a given image and a promising way to achieve more effective image retrieval and analysis. Due to the semantic gap between low-level visual features and high-level image semantic, the performances of some image distance metric learning (IDML) algorithms only using low-level visual features is not satisfactory. Since there is the diversity and complexity of large-scale image dataset, only using visual similarity to learn image distance is not enough. To solve this problem, in this paper, the semantic labels of the training image set participate into the image distance measure learning. The experimental results confirm that the proposed image semantic distance metric learning (ISDML) can improve the efficiency of large-scale AIA approach and achieve better annotation performance than the other state-of-the art AIA approaches.

1 INTRODUCTION

Automatic image annotation (AIA) is to automatically annotate an image with appropriate keywords, called labels, and reflect its visual content. Systems managing and analyzing images on image sites heavily depend on textual annotations of images. Various approaches of AIA have two types, i.e., classification-based and probabilistic modeling-based approaches.

In first type, image annotation can be viewed as a classification problem (Zhuang et al. 1999), which can be solved by using a classifier. For annotating an image without caption, first, represent image into a low-level features vector. Then, classify the image into a category. Finally, propagate the semantic of the corresponding category to the image. The unlabeled image may be automatically annotated.

In second type, probabilistic model (Stathopoulos et al. 2009) attempts to infer the joint probabilities between images and semantic concepts. Images given class can be regarded as instances of stochastic process that characterizes the class. Then, statistical models, such as Markov, Gaussian, and Bayes and etc. are trained and images are classified based on probability computation.

An effective image annotation approach can deal with a large number of images, allowing users to query interest images efficiently and effectively. AIA approach also has potential applications in image retrieval (Nguyen and Kaothanthong et al, 2013; Watcharapinchai et al. 2011) and image description (Lasmar et al. 2014) etc.

Currently, many AIA approaches have been proposed (Jin et al. 2015; Wang et al. 2008), among these, the image similarity was determined only using low-level visual features such as colors, textures and shapes (Jin and Guo 2014). The problem is that visual similarity does not equal semantic similarity. Therefore, the performances of some image annotation algorithms were not so satisfactory. The some AIA approaches have the following disadvantages.

(1) They heavily rely on visual similarity for judging semantic similarity.

(2) The image distance is usually measured according to some traditional methods, e.g., Euclidean distance, Mahalanobis distance, Hamming distance, Cosine distance, Histogram distance and so on. Although these traditional distances are simple and convenient, it can not accurately measure the

semantic similarity between two images in many cases.

In this paper, we propose a novel ISDML algorithm based on semantic similarity for large-scale AIA, named AIAISDML, which utilizes ISDML for improving the performance of large-scale AIA.

2 IMAGE DISTANCE METRIC LEARNING

How to more effectively measure the image distance has become a key problem in the field of image recognition. For convenience, some necessary notations and definitions are first introduced. Let $Tr = \{I_1, I_2, \dots, I_N\}$ be the training set with labels, and an image is represented as a M -dimensions vector $I = \{x^1, x^2, \dots, x^M\}$, where $x^i \in I$ is i th visual feature. Let $L = \{l_1, l_2, \dots, l_m\}$ be the set of possible annotated labels, and each image $I \in Tr$ is associated with a subset $Y \subseteq L$. Where, Y may be represented as an m -dimensional vector, i.e., $Y = (y^1, y^2, \dots, y^m)$, which $y^j = 1$ only if image I has label l_j and 0 otherwise. And M and m is the total number of all visual features of the image and labels respectively. So, Tr can be recorded as $Tr = \{(I_i, Y_i) \mid i=1,2,\dots,N\}$, where $Y_i = (y_i^1, y_i^2, \dots, y_i^m)$, y_i^j represents that the j th label l_j belongs to the image I_i .

In practical applications, a lot of the image distances were measured according to some traditional distances, however these traditional distances are not always appropriate. In this paper, we introduce semantic similarity metric into neighborhood component analysis (NCA) (Goldberger et al. 2005) to learn image semantic distance metric.

Distance metric learning (DML) uses the training images to learn a metric function so that the closer distance between similar images, otherwise the farther is. For given two feature vectors I_i and I_j , the squared Mahalanobis distance is calculated as follows

$$d^2(I_i, I_j) = (I_i - I_j)^T M (I_i - I_j) \quad (1)$$

where, M is a positive semi-definite matrix. Let $M = A^T A$, then A is considered to be a transformation matrix of feature vectors I_i and I_j .

Eq.(1) can be also rewritten as follows

$$d^2(I_i, I_j) = (I_i - I_j)^T A^T A (I_i - I_j) \quad (2)$$

In eq.(2), the distance between two feature vectors I_i and I_j is calculated as the Euclidean distance between AI_i and AI_j . Therefore, the Mahalanobis distance is transformed into Euclidean distance. According to eq.(2), we have

$$d^2(I_i, I_j) = (AI_i - AI_j)^T (AI_i - AI_j) \quad (3)$$

3 SOLVING MATRIX A WITH SEMANTIC SIMILARITY

We notice that the existing research mostly focused on the relation between $d^2(I_i, I_j)$ and visual features, and not the relation between $d^2(I_i, I_j)$ and the semantic labels of images. We know that, the smaller value of $d^2(I_i, I_j)$, the better similarity about visual features, which is not related to the image semantic. Therefore, for training images (I_i, Y_i) and (I_j, Y_j) , we let the semantic similarity $s(I_i, I_j)$ of feature vectors I_i and I_j be

$$s(I_i, I_j) = \frac{1}{a + H(Y_i, Y_j)} \quad (4)$$

where, $H(Y_i, Y_j)$ is Hamming distance between Y_i and Y_j . Thus, the smaller value of $H(Y_i, Y_j)$, the better similarity about image semantics. a is an arbitrary positive constant.

After obtaining the transformation matrix A , it maps image I from feature and semantic spaces to metric space. In the metric space, the distance between the similar images is small, and the distance between the dissimilar images becomes large, which contains both feature similarity and semantic similarity.

3.1 Calculate Probability of Neighbors

Our goal is to learn a matrix A for improving image annotation performance of the images without caption. In this paper, for simplicity and convenience, we use the probability method for finding the neighbor of given image.

For any image $I_i \in Tr$, assume that the probability of image I_i selecting another image I_j in Tr as its neighbor is P_{ij} , then P_{ij} is calculated as follows

$$P_{ij} = \begin{cases} \frac{s(I_i, I_j) \exp(-\|AI_i - AI_j\|^2)}{\sum_{k \neq i} s(I_i, I_k) \exp(-\|AI_i - AI_k\|^2)}, & i \neq j \\ 0, & i = j \end{cases} \quad (5)$$

3.2 Solving Matrix a

In eq.(5), we propose the calculation probability formula of image I_i selecting another image I_j in Tr as its neighbor, which takes into account both the visual similarity and semantic similarity. According to Goldberger et al. 2005, the leave-one-out method was used to obtain A . For simplicity, we use gradient descent to solve matrix A .

Let the class of I_j be ω_j and Ω_i an image set, where Ω_i 's elements belong to the same class with I_i , i.e., $\Omega_i = \{I_j | \omega_i = \omega_j\}$, then the probability P_i of Ω_i 's all elements to be neighbors of I_i is

$$P_i = \sum_{I_j \in \Omega_i} P_{ij} \quad (6)$$

So, its logarithmic weighted average is

$$f(A) = \sum_{i=1}^N s(I_i, I_j) \log \sum_{I_j \in \Omega_i} P_{ij} \quad (7)$$

Differentiating $f(A)$ with respect to A can be expressed as follows

$$\frac{\partial f(A)}{\partial A} = 2A \sum_{i=1}^N s(I_i, I_j) \left(\sum_{k \neq i} P_{ik} (I_i - I_k)(I_i - I_k)^T \right. \\ \left. \frac{\sum_{I_j \in \Omega_i} P_{ij} (I_i - I_j)(I_i - I_j)^T}{\sum_{I_j \in \Omega_i} P_{ij}} \right) \quad (8)$$

According to the gradient descent method, we can obtain the matrix A . Furthermore, we can obtain ISDML.

4 LARGE-SCALE AIA APPROACH BASED ON ISDML

4.1 Image Low-level Features Vector

The image low-level feature extraction is a fundamental step in image annotation. The standard feature extraction methods extract the texture, color and shape features of each image respectively, and are used for similarity calculations.

In this paper, we represent an image by several regions. Each region is characterized with the same number of the features. Region distances are defined for these features to use ISDML between the feature representations. The features roughly capture different aspects such as shape, texture, color and location for an image region. The details are listed in Table 1.

Table 1: The 14 region-based features.

Type	Name	Dimension
Shape	BB extent	2
	Centered mask	1024
	Pixel area	1
Texture	Bottom boundary tex-hist	100
	Interior tex-hist	100
	Left boundary tex-hist	100
	Right boundary tex-hist	100
	Top boundary tex-hist	100
Color	Color histogram	33
	Color std	3
	Mean color	3
Location	Absolute mask	64
	Bot height	1
	Top height	1

To capture information of shape, the centered mask, the size of the region and the size of region's bounding box will be computed. To capture information of texture, the normalized texton histograms, and separately, along the boundaries also will be computed. To capture information of color, the mean RGB-value, its standard deviation and a color histogram still are computed. Finally, to capture information of the position of the region in an image, a coarse (blurred) 8×8 absolute mask as well as the height of the top-most and bottom-most pixel in the region are computed.

4.2 Image Annotation Approach

In this section, we discuss large-scale AIA approach based on ISDML, which block diagram is shown in Figure 1.

After dividing a training image into several regions, there is at least one of the labels in each region according to labels of training image. In other words, for each image of the training set, there is at least one label for any region of given image.

In AIA, some important issues will be carefully considered as follows.

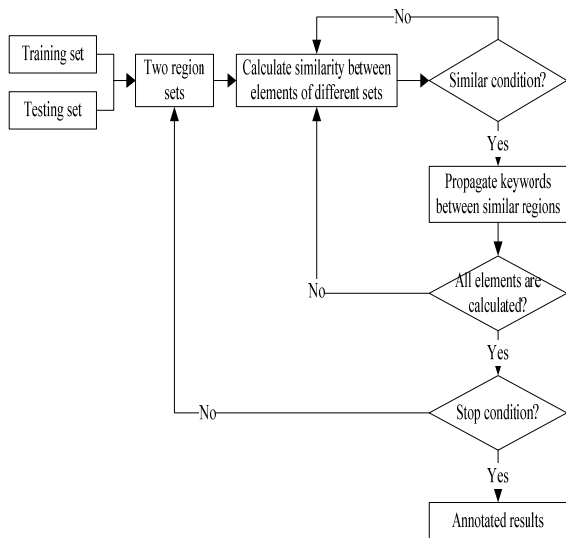


Figure 1: Image annotation scheme.

4.2.1 Calculate Similarity between Different Regions

Each region can be expressed as a feature vector according to subsection 4.1. For the feature vectors I_i and I_j of two regions, calculate its distance using ISDML. The small distance $d^2(I_i, I_j)$ between two regions, the similar between them.

4.2.2 Similar Condition

Similar condition is a criterion to judge whether the similarity of two regions. There are many approaches as similar conditions. In this paper, we simply use a threshold ϵ as criterion condition, namely when $d^2(I_i, I_j) \leq \epsilon$, we can obtain conclusion: I_i and I_j are similar.

4.2.3 Propagate Labels

After judging that two regions of training and test sets is similar, all labels of the region of training set are propagated to the region of testing set. So, this region of testing set is successfully annotated.

4.2.4 Stop Condition

When all regions of testing set are annotated, we say stop condition is satisfied. If stop condition is not satisfied, we will re-divide image, and obtain some new regions.

4.2.5 Obtain Annotation Results

After all regions of testing set are annotated, we

extract labels from all regions of a testing image without caption and generate a label set, which is annotation result of this testing image.

5 EXPERIMENTAL RESULTS

5.1 Dataset

For evaluating the performance of AIAISDML approach, we used three standard benchmark datasets for AIA, namely the ESP Game (Von Ahn, et al., 2004), the IAPR TC-12 (Grubinger, 2007) and the NUS-WIDE (Chua, et al., 2009). ESP Game contains images annotated using an on-line game, where two players are randomly given an image for which they have to predict same keywords to score points (Von Ahn et al, 2004). This way the players are encouraged to provide important and meaningful labels to images. Because many people participate in the manual annotation task, this dataset very challenging and diverse. IAPR-TC12 was originally used in ImageCLEF, and this set of 20.000 images accompanied with descriptions in several languages was initially published for cross-lingual retrieval. It can be transformed into a format comparable to the other sets by extracting common nouns using natural language processing techniques. NUS-WIDE dataset is a comparatively large web image dataset (Chua, et al., 2009) consists of 269648 real-world web images crawled from Flickr. We have downloaded all the 269648 images. All samples are supervised and annotated with 81 concepts, where these ground-truth concepts for all images are provided for evaluation.

Table 2 summarizes the statistics for each dataset.

Table 2: Statistics for the datasets used in the experiments.

Dataset	# of images	# of labels	labels per image	Images per labels
ESP Game	20768	268	4.69/15	363/5059
IAPR TC-12	19627	291	5.72/23	386/5534
NUS-WIDE	269648	81	1.9/12	3722/44255

For every dataset, we randomly select 90% of images as training set and use the remaining 10% for testing set.

5.2 Performance Evaluation

In this paper, each image is divided several regions using normalized cut technology (Shi et al. 2000; Jin

and Liu et al. 2015). Many regions are obtained for three image datasets respectively. For each region, low-level features, such as shape, texture, color and location etc. are considered like subsection 4.1.

In order to estimate the annotation performances of AIAISDML approach, some popular evaluations are used for AIA task. We compute precision and recall of each label in a test dataset. Suppose a label l is present in the ground-truth of R images, and it is predicted for P images during testing out of which Q predictions are correct ($Q \leq P$ and $Q \leq R$), then we have

$$\text{Precision} = \frac{Q}{P}, \quad \text{Recall} = \frac{Q}{R} \quad (9)$$

We average these values over all the labels of a test dataset and get percentage average precision (called, AP) and percentage average recall (called, AR), similar to the previous annotation approaches such as Guillaumin et al., 2009 and Nakayama 2011.

As a tradeoff between the above indicators, the geometric mean of them is adopted widely, namely

$$F = \frac{2(\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}} \quad (10)$$

The larger values of F , the better performance of AIA approach.

In addition, the statistics value, i.e., the number of labels annotated correctly at least, also is used, which reflects the coverage of labels in proposed approach, denoted by N^+ .

The larger values of N^+ , the better performance of AIA task is.

5.3 Comparison with State-of-the-Art Image Annotation Approaches

In this subsection, we will estimate the performance of AIAISDML approach, and also compare with state-of-the art algorithms from the literatures. These algorithms have been shown to be successful and can obtain suitable annotation results. For example, MBRM (Feng, et al., 2004), JEC (Makadia et al. 2008), TagProp (Guillaumin et al. 2008 and Yashaswi et al. 2012), CCD(HLAD) (Nakayama et al. 2011), ML-LGC (Zhou et al. 2004), SMSE (Chen et al. 2008) and MISSL (Rahmani et al. 2006) and WSG.

MBRM describe a statistical model for automatic annotation of images and video frames, which proposed a multiple-Bernoulli relevance model for image annotation, to formulate the process of a human annotating images. The results show that it outperforms, especially on the ranked retrieval

task, the (multinomial) continuous relevance model and other models on both the Corel dataset and a more realistic Trec Video dataset. JEC treats image annotation as retrieval. Using multiple global features, a greedy algorithm is used for label transfer from neighbours. They also performed metric learning in the distance space but it could not do any better than using equal weights. TagProp is a weighted K Nearest Neighbor (KNN) that transfers labels by taking a weighted average of keywords' presence among the neighbours, which also address the class-imbalance problem, logistic discriminant models are wrapped over the weighted KNN method with metric learning. CCA is a theoretically optimal distance metric. To use CCD efficiently, image features were embedded in a Euclidean space. Image annotation based on CCD is shown to achieve comparable performance to state-of-the-art works with lower computational costs for learning and recognition. ML-LGC was proposed by Zhou et al. 2004. It aims to design a classifying function which is sufficiently smooth with respect to the intrinsic structures collectively revealed by both annotated and unlabeled points. SMSE was proposed by Chen et al. 2008. Two graphs are first constructed on instance level and category level. Then a regularization framework combining two regular terms for the two graphs is used. MISSL was proposed by Rahmani et al. 2006. It transforms multiple instance problem into an input for a graph-based single instance semi-supervised learning method.

5.4 Results and Discussions

Figure 2 shows that the annotated results of proposed AIAISDML approach, keep rather a high consistent with the ground truth. This fact verifies the effectiveness of proposed AIAISDML approach.



Test image		
Ground truth	chair house landscape pool sun terrace tree	cactus flower lake landscape middle mountain slope
Proposed	chair building, umbrella pool landscape sun terrace tree	cactus bush lake landscape middle mountain slope

Figure 2: Illustrations of annotation results of proposed approach.

Test image			
Ground truth	car dirt sky tree wheel white	black dog grass green guy man run shoes white	brick classroom desk front girl wall
Proposed	car land sky tree wheel white	black dog grass green jerkin man run shoes white	brick classroom desk green girl wall

Figure 2: Illustrations of annotation results of proposed approach (cont.).

We let the threshold ϵ of similarity condition be 1, 3, 5, 7 and 10 respectively. After running AIAISDML approach, we can obtain the average Precision (AP), average Recall (AR), average F (AF), and then these average values are compared with other state-of-the-art algorithms. The results are shown in Table 3-5.

Table 3: Performance comparison on ESP Game.

Approach	AP	AR	AF	N
MBRM	0.18	0.19	0.18	209
JEC	0.24	0.19	0.21	222
TagProp	0.39	0.27	0.32	239
CCD(HLAC)	0.27	0.18	0.22	221
Proposed	0.41	0.29	0.34	241

Table 4: Performance comparison on IAPR-TC12.

Approach	AP	AR	AF	N
MBRM	0.24	0.23	0.23	223
JEC	0.29	0.19	0.23	211
TagProp	0.46	0.35	0.40	266
CCD(HLAC)	0.35	0.26	0.30	249
Proposed	0.47	0.36	0.41	267

Table 5: Performance comparison using NUS-WIDE.

Approach	AP	AR	AF	N
JEC ^a	0.22	0.25	0.23	/
ML-LGC ^a	0.28	0.29	0.29	/
SMSE ^a	0.32	0.32	0.32	/
MISSL ^a	0.27	0.33	0.30	/
Proposed	0.48	0.29	0.35	74

Results of ^a provided by Liu et al. 2012

Table 3-5 show the comparison results of proposed AIAISDML approach with other different state-of-the-art image annotation approaches. We can observe that the AIAISDML approach outperforms all the compared state-of-the-art image annotation approaches on IAPR TC12, ESP Game and NUS-WIDE datasets, which shows that the annotation performance of AIAISDML approach is satisfactory. We also notice that the performance of AIAISDML approach is always higher than other compared approaches, which show that the ISDML is very effective for improving annotation performance.

6 CONCLUSIONS

In this paper, we propose ISDML and investigate its applications to large-scale AIA. The proposed AIAISDML approach based on ISDML can improve annotation performances of AIA task. The main advantages of proposed AIAISDML are as follows:

(1) To improve the annotation performance of AIA, both semantic and low-level visual features knowledge of training set are sufficiently considered, and they are also introduced into ISDML.

(2) In proposed AIAISDML approach, through calculating similarity between different regions using ISDML, it is possible to more effectively annotate an image, which shows that the proposed AIAISDML can be applied on a large-scale image dataset.

ACKNOWLEDGEMENTS

This work was supported by Natural Social Science Foundation of China (Grant No.13BTQ050).

REFERENCES

Chen, G., Song, Y., Wang, F., Zhang C., 2008. Semi-supervised multilabel learning by solving a sylvester equation. *SIAM International Conference on Data Mining*, 410-419

Chua, T.S., Tang, J., Hong, R., et al., (2009). NUS-WIDE: a real-world web image database from National University of Singapore. *ACM International Conference on Image and Video Retrieval*, 48

Feng, S.L., Manmatha, R., Lavrenko, V., 2004. Multiple bernoulli relevance models for image and video annotation, *IEEE Computer Society Conference on*

- Computer Vision and Pattern Recognition (CVPR 2004)*, II-1002-II-1009, Vol.2, 1002-1009
- Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R., 2005. Neighbourhood components analysis. *Advances in Neural Information Processing Systems*, 17, 103-110
- Grubinger, M., 2007. Analysis and evaluation of visual information systems performance. PhD thesis, Victoria University, Melbourne, Australia
- Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C., 2009. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation, *IEEE 12th International Conference on Computer Vision*. 309-316
- Jin, C., Guo, J.L., 2014. Image semantic annotation approach based on the feature matching. Springer, *Advances in Intelligent Systems and Computing*, Vol.250, 281-288
- Jin, C. Jin, S.W., 2015. Automatic image annotation using feature selection based on improving quantum particle swarm optimization. *Signal Processing*, 109, 172-181
- Jin, C., Liu, J.A., Guo, J.L., 2015. A hybrid model based on mutual information and support vector machine for automatic image annotation. *Artificial Intelligence Perspectives and Applications*. Springer, 347, 29-38
- Lasmar, N.E., Berthoumieu, Y., 2014. Gaussian copula multivariate modeling for texture image retrieval using wavelet transforms. *IEEE Transactions on Image Processing*, 23(5), 2246-2261
- Liu, S., Yan, S.C., Zhang, T.Z., Xu, C.S., Liu, J., Lu, H.Q., 2012. Weakly supervised graph propagation towards collective image parsing, *IEEE Transactions on Multimedia*, 14(2), 361-373
- Makadia, A., Pavlovic, V., Kumar, S., 2008. A new baseline for image annotation. *Computer Vision–ECCV 2008*. Springer Berlin Heidelberg, 316-329
- Nakayama, H., 2011. Linear distance metric learning for large-scale generic image recognition. PhD thesis, *The University of Tokyo*, Japan
- Nguyen, C.T., Kaothanthong, N., Tokuyama, T., Phan, X.H., 2013. A feature-word-topic model for image annotation and retrieval. *ACM Transactions on the Web*, 7(3), 1-12
- Rahmani, R., Goldman, S., 2006. Missl: Multiple-instance semi-supervised learning, *International Conference on Machine Learning*, 705-712
- Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888-905
- Von Ahn, L., Dabbis, L., 2004. Labeling images with a computer game. *SIGCHI Conference on Human Factors in Computing Systems*. ACM, 319-326
- Wang, C., Zhang, L., Zhang, H.J., 2008. Learning to reduce the semantic gap in web image retrieval and annotation. *The 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, Singapore, 355-362
- Watcharapinchai, N., Aramvith, S., Siddhichai, S., 2011. Two-probabilistic latent semantic model for image annotation and retrieval, *Lecture Notes in Computer Science*, vol.6468, 359-369
- Yashaswi, V., Jawahar, C.V., 2012. Image annotation using metric learning in semantic neighbourhoods. *ECCV(3)*, 836-849
- Zhou, D., Bousquet, O., Lal, T.N., et al., 2004. Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 16(16), 321-328
- Zhuang, Y., Liu, X., Pan, Y., 1999. Apply semantic template to support content-based image retrieval. *Lecture Notes in Computer Science*. 3972, 442-449