Journal of Communications Technology, Electronics and Computer Science, Issue 2, 2015
ISSN 2457-905X

Using manifold structure for automatic image annotation by fusion of multiple feature spaces

Mohammad Ali Zare Chahooki, Hamid Kargar-Shooroki
Electrical and Computer Engineering Department, Yazd University,
Yazd, Iran
chahooki@yazd.ac.ir, hamidkargar@stu.yazd.ac.ir

***Abstract*— Automatic image annotation has been an active research topic in recent years. Low level features like as color, texture, shape as well as object spatial relations are extracted to represent images in general. These syntaxes are further used to retrieve images from large image data sets. However, the similarity of images could not be found correctly by similarity measures such as Euclidean distance in many situations. On the other hand, graph models have been shown powerful in solving many machine learning problems in recent years. In this paper, we propose a graph-based learning approach, named Conceptual Manifold Structure (CMS), based on transition from conceptual to observation space. In the proposed method, a graph including both the trained and tested samples is constructed by fusion of multiple feature spaces. Conceptual transition in graph structure is found by altering the edge values in an innovative manner. This is caused to learn the manifold structure where the samples dissimilarity is closer to the conceptual distance. Furthermore, the continuity between the instances of a semantic in the conceptual space is kept in feature space. Keeping the continuity in manifold structure is the main idea to decrease the semantic gap in this study. The experiments on different image data sets indicated that the geometrical distances between the samples on the manifold space are closer to their conceptual distance. The proposed method has been compared to other well-known approaches. The results confirmed the effectiveness and validity of the proposed method.**

Keywords*— *Automatic image annotation; Manifold structure; Graph learning

I. Introduction

Today, with the advent of digital imagery, the number of images has been growing exponentially. Automatic Image Annotation (AIA) is the task of automatically labeling of images with a set of predefined conceptual keywords. It is mainly used for image retrieval and can be employed in a variety of domains such as image mining, web image classification and conceptual web filtering. In particular, image annotation facilitates Content Based Image Retrieval (CBIR) since annotated keywords narrow the semantic gap between low-level features and high-level semantics.

Over the recent years, the availability of large data collections, which only have limited human annotation, has attracted the attention of a growing community of researchers to the problem of Semi Supervised Learning (SSL) [1]. By leveraging unlabeled data with certain assumptions, semi-supervised learning methods are promising to build more accurate models than those that are achieved by purely supervised learning methods [2].

In this paper, a semi supervised approach for estimation of image manifold structure in automatic image annotation is proposed. Different feature spaces are employed to represent an image. Furthermore, fusion of multiple feature spaces is done on semi-supervised learning method on a graph. Transition of semantics from conceptual space to fused dissimilarity graph is the main innovation of the proposed method.

To perform the semantic transition on the graph, the edge values are changed based on the labels of human annotation and the fused dissimilarities. In this way, the shortest path between unlabeled and labeled samples on the graph would indicate their common concepts. For this purpose, an innovative way for transition from conceptual space to fused dissimilarity graph is proposed. In the proposed method, with the variation that done in the edges values, the distance from unlabeled samples to labeled ones are changed so that the measured distances indicate the shared image concepts.

To evaluate the performance of the proposed method, several experiments on benchmark data of Corel images have been carried out. Besides, a systematic comparison among some related works are given.

This paper is organized as follows: Section II provides our proposed method for annotation by learning the manifold structure. The experimental results are described in section III. The final section is concluding and future work remarks.

II. Learning the Manifold Structure

In this section, we propose our graph-based semi-supervised learning method named Annotation based on Manifold Structure (AMS). Section II-A states some methods for image description by global and local attributes. The approach for learning the manifold structure is described in section II-B.

A. Image description

There are many image description methods that are mainly categorized into global and local-based ones. The global visual features which extracted from the entire images may not be able to characterize the local visual properties [3]. Therefore, in the proposed method, to achieve more sufficient representation of various visual properties of the images, both the global and the local visual features are fused to improve the total recognition rate.

To encode the spatial layout of the image, histograms are computed over three horizontal regions of the image, reflecting the typical layout of landscape photography. The three histograms are then concatenated to form a new global descriptor. These new histograms are used in addition to the image-wide histograms

Global features- Color histogram and Gist [4] are two global features used in this research. Gist is based on Gabor filter [5]. Typical Gabor features consist of the responses calculated by Gabor filters at several different orientations and scales. Using different orientations and scales ensures invariance; that is objects can be recognized at various different orientations, scales and translations. The color histograms [6] are found in three different spaces: RGB, LAB, and HSV. Therefore, one Gist feature vector and six color histograms (three color spaces for each of two layouts) are extracted to describe image in global approach.

Local features- SIFT [7] and hue [8] are local feature descriptors used in this research. Our approach in detecting the interesting points for extracting the local features [6] is based on two methods; regions on a dense multi-scale grid, and regions found using a Harris-Laplacian detector. Each local feature descriptor is quantized using k-means on samples from the training set, and then images are represented as a bag-of-words histogram. Therefore, four SIFT histograms and four hue histograms (two layouts for each of two detecting methods) are extracted to describe an image in the local approach.

B. Annotation based on manifold structure

Those feature vectors discussed in section II-A are used to describe the semantic of each image. The performance of a graph-based semi-supervised learning method heavily depends on the quality of the constructed graph [9].

AMS is done in three phases. (I) Construction of graph of original structure, (II) Construction of graph of manifold structure, and (III) Graph based kNN image annotation.

Phase I) Constriction of graph of original structure (G) An image data set $I = \{x_1, x_2, \dots, x_{N_l}, x_{N_{l+1}}, \dots, x_N\}$ is initially assumed. N is the images number and N_l is the number of the labeled images. Suppose we have described each image by P description vectors. Therefore, we define $S = \{s_1, s_2, \dots, s_P\}$ as a set of P observation spaces. The dimensionality in each observation space is defined as D_i.$M_{N\times D_i} \subset s_i$ is defined as the set of N feature vectors in each observation space, s_i. Fusion of different feature spaces is done with the method introduced in Algorithm 1.

In the proposed method, the images, x_i, are presented as a graph G=(V, E), where V is the set of N samples and $E = V \times V$. x_i and x_j are two data samples which would be assigned to $(x_i, x_j) \in E$ and be referred by $edge_{ij}$ that is equal to $D(i,j)$. As D is symmetry, $edge_{ij} = edge_{ji}$.

Phase II) Constriction of graph of manifold structure ($\widetilde{GD}$) In the proposed method, $\widetilde{\times}$ defines the operation of mapping from conceptual to dissimilarity space. This operation is illustrated in Fig. 1.

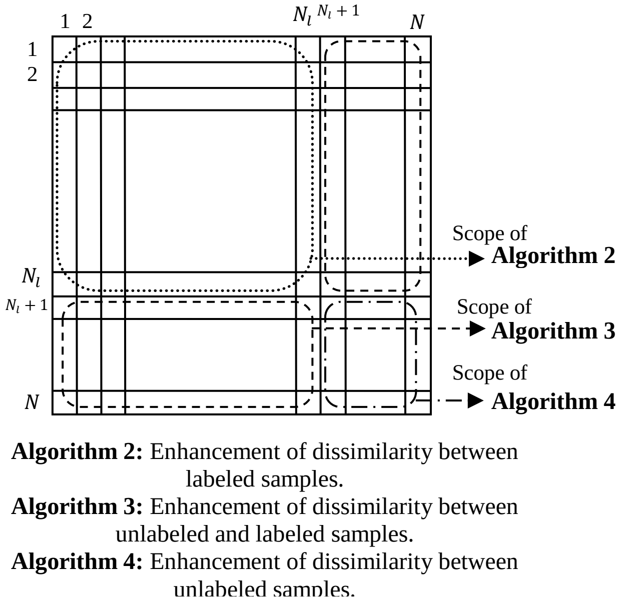

Fig. 1. Illustration of $\widetilde{\times}$, Operation of mapping from conceptual to dissimilarity space.

The operation modifies the edges of G which are valued by *D*. So, the dissimilarity matrices are modified in following algorithms.

$\widetilde{\times}$ constructs an approximation of images structure based on nearest neighbors. Image annotation based on the output structure ($\tilde{G}$) is performed by the nearest neighbors and the evaluation of the shortest paths from unlabeled to labeled samples. In the experimental results, the effect of each of the mentioned values in $\widetilde{\times}$ is reported.

The values of edges are modified in three steps. Referring to Algorithm 2, the dissimilarities between the labeled images are initially modified. In this step, the nodes of same labeled images would become closer. In Algorithm 2, $D(1{:}N_l, 1{:}N_l)$ indicates the scope of this algorithm as illustrated in Fig. 1. Also, $C(:, i)$ refers to the labeled images which are annotated with v_i. $\tilde{D}$ Indicates as the dissimilarity between nodes in $\tilde{G}$. The structure of $\tilde{G}$ will be incrementally formed by Algorithm 2, 3, and 4.

In the second stage, the dissimilarities between the unlabeled and labeled images are modified. As shown in Algorithm 3, $D(N_l + 1{:}N, 1{:}N_l)$ indicates the scope of this algorithm as illustrated in Fig. 1. In this step, the tags related to k nearest labeled neighbors of each unlabeled image are identified. Then, the labeled images which are labeled with the identified tags are determined. Afterward, the dissimilarity of the unlabeled image with the determined labeled images will be less and with the other unlabeled images will be more. In this algorithm, ε is a real value less than one, and φ is another real value greater than one.

Input: P feature matrices $M^1_{N\times D_1}, \ldots, M^P_{N\times D_P}$. Rows of a feature matrix ($M_{N\times D_i}$), describe the corresponding image in s_i.

Objective: Construction of D, the fused dissimilarity matrix.

% construction of dissimilarities in each observation space

% $M^p(j,:)$ is the description vector of x_j in s_p.

$\forall\ p = 1 \ldots P$ *and* $\forall\ x_j, x_k \in I$ *set*

$$Diss_p(i,j) = \|M^j(j,:) - M^k(j,:)\|;$$

% Normalization of dissimilarities in each observation space

$$\forall\ p = 1 \ldots P\ set\ Diss_p(i,j) = \frac{Diss_p(i,j)}{\sum_{j=1}^{N} Diss_p(i,j)}$$

% Construction of fused dissimilarity matrix. α_p is the effectiveness of s_p in fusion phase.

$$D_{N\times N} = \sum_{p=1}^{P} \alpha_p Diss_p$$

Algorithm 1. Construction of fused dissimilarity matrix.

Input: The fused dissimilarity matrix (D), and the conceptual space (C)
Objective: Enhancement of dissimilarity between labeled samples.
% specifies the scope of this algorithm
$D_1 = D(1{:}N_l, 1{:}N_l)$
$for\ i = 1\ to\ M$ {
 % idx is the set of labeled samples indexes are annotated with v_i.
 $idx = C(:,i);$
 $\forall\ j,k \in idx\ set\ D_1(i,j) = \varepsilon \times D_1(i,j);$
}
% $\widetilde{D}$ is constructed in first step.
$\widetilde{D}(1{:}N_l, 1{:}N_l) = D_1;$

Algorithm 2. The Enhancement of the dissimilarity between labeled samples

In the third stage, the dissimilarities between the unlabeled images are modified (Algorithm 4). In Algorithm 4, $D(N_l + 1{:}N, N_l + 1{:}N)$ indicates the scope of this algorithm as illustrated in Fig. 1. For each unlabeled image, the related dissimilarities are sorted. Then, the values of $(\mu \times N_u)$-closets neighbors will be reduced by multiplying them to ε.

When three algorithms are employed respectively, the neighborhood graph, $\widetilde{NG}$ of $\tilde{G}$, is constructed. This is done by connecting nodes *i* and *j* if they are closer than *e* or *j* is one of the E-nearest neighbors of *i*. The $\widetilde{NG}$ can also be defined as a geodesic distance graph. The shortest path of pair wise nodes of $\widetilde{NG}$ defines $\widetilde{GD}$, called as the geodesic distance matrix.

Phase III) Graph based K-NN image annotation Algorithm 5 shows the annotation of unlabeled images based on $\widetilde{GD}$, the geodesic distance matrix.

Input: The fused dissimilarity matrix (D), and the conceptual space (C)
Objective: Enhancement of dissimilarity between labeled and unlabeled samples.
% specifies the scope of this algorithm

$D_2 = D(N_l + 1{:}N, 1{:}N_l)$

$for\ i = 1\ to\ N_u$ {
 % idx is the set of indexes of labeled samples which have less dissimilarities to (in k nearest neighbors of) x_{N_l+i} .
 $idx = Sort(D_2(i,:),k);$
 % x_j is a labeled image
 $\forall\ j \in idx\ do${
  % r is set of indexes of x_j labels
  $r = C(j,:);$
  $\forall\ m \in r\ do${
   % lidx is set of labeled images which are annotated with v_m.
   lidx=C(:,m);
   $\forall\ t \in lidx\ set\ D_2(i,t) = D_2(i,t) \times \epsilon;$
   $\forall\ t \notin lidx\ set\ D_2(i,t) = D_2(i,t) \times \varphi;$
  }
 }
}
% $\widetilde{D}$ is modified in second step.
$\widetilde{D}(N_l + 1{:}N, 1{:}N_l) = D_2;$
$\widetilde{D}(1{:}N_l, N_l + 1{:}N) = (D_2)^T$;*%due to symmetry*

Algorithm 3. Enhancement the dissimilarity between unlabeled and labeled samples.

Input: The fused dissimilarity matrix (D)
Objective: Enhancement of dissimilarity between unlabeled samples.
% specifies the scope of this algorithm
$D_3 = D(N_l + 1{:}N, N_l + 1{:}N)$
$for\ i = 1\ to\ N_u$ {
 % idx is the set of indexes of unlabeled samples in ascending order in dissimilarity to x_{N_l+i}.
 $idx = Sort(D_3(i,:));$
 % set of μ-closest neighbors
 $idx\mu = idx(1{:}\mu \times N_u);$
 $\forall\ j \in idx\mu\ set\ D_3(i,j) = e \times D_3(i,j)$
}
% $\widetilde{D}$ is modified in third step.
$\widetilde{D}(N_l + 1{:}N, N_l + 1{:}N) = D_3;$

Algorithm 4. The enhancement of dissimilarity between the unlabeled samples.

Input: The geodesic distance matrix ($\widetilde{GD}$).
Objective: Annotation of unlabeled images.
$\forall\, x_i \in I, i \in \{N_l + 1 \dots N\}$ *do*{
  % idx is the set of indexes of labeled samples which have less dissimilarities to (in K nearest neighbors of) x_i .
  $idx = Sort(\widetilde{GD}(i,:), K)$;
  $\forall\, j \in idx$ *do*{
    % r is vector of M elements. The initial values are zero. Most values show the highest annotation probability of the corresponding v_i.
    $r = r + C(j,:)$;
  }
}

Algorithm 5. Annotation of unlabeled images.

III. Experimental results

To prepare a proper comparison with some previous works, we have used the Corel dataset provided by Duygulu et al. [10] without any modification. This data set makes our results comparable to several recent works on this subject. We divided the dataset into 3 categories with 4000 training set images, 500 evaluation set images, and 500 testing set images. The evaluation image set is used to find the optimal system parameters. After setting the parameters, we merged the 4000 training set images and 500 validation set images to make a new training set. This corresponds to the training set of 4500 images and the test set of 500 images used in [10]. Each image is annotated with 1–5 words. The images are of the size of 384×256 pixels. The vocabulary size of the whole data set is 374 and that of the test data set is 263. It is important to mention that since no other words are accessible for the auto annotation process, the crucial vocabulary size is that of the training set. In this case, the vocabulary size of the Corel training set becomes 371.

Similar to the previous works, the quality of automatic image annotation is evaluated through the process of retrieving test images with single keyword. Given a particular keyword *w*, the number of correctly annotated images is denoted as N_c, the number of retrieved images is denoted as N_s, and the number of truly related images in test set is denoted as N_r. Then, precision, recall and F1 measures are computed as follows:

$$Precision(w) = N_c/N_s \tag{1}$$

$$Recall(w) = N_c/N_r \tag{2}$$

$$F1(w) = \frac{2 \times Precision(w) \times Recall(w)}{Precision(w) + Recall(w)} \tag{3}$$

In addition, the values of the precision and recall are averaged respectively over all the words in test data set to evaluate the performance. Finally, the number of words with nonzero recall, which is denoted as “NumWord” for short, is considered as well. The measure provides an indication of the number of words which the system has effectively learned.

In AMS, parameters are estimated by using the validation set. Therefore, for the ongoing experiments in this section, we set $\alpha_1 = \alpha_2 = \dots = \alpha_{15} = 1$, $\varepsilon = 0.93$, $k=1$, $\epsilon = 0.925$, $\varphi = 1.24$, $e=0.99$, $\mu=0.01$ *and* $K=1$.

The effectiveness of some AMS steps in phase II are shown in Table I. $\widetilde{D}$ is initialized with D in all experiments. In the evaluation of experimental results, annotation is performed by G, which is the primary fused dissimilarity graph in the first step. For this purpose $\widetilde{\mathrm{GD}}$ is replaced by D in Algorithm 5. The results of the annotation quality are shown in the first row of Table I.

In the second step of experiments, G is employed instead of $\widetilde{\mathrm{G}}$ in construction of neighborhood graph and also in evaluation of geodesic distances. In this regard, the effect of the elimination of dissimilarity enhancement between samples (Algorithm 2, 3, and 4) is studied in annotation context. The results are shown in the second row of Table I.

In the third step, Algorithm 3 and 4 are ignored. Therefore, the dissimilarity between labeled samples are only improved (refer to Algorithm 1). The results are shown in the third row of Table I.

In the fourth step, Algorithm 4 is ignored, and finally, the ASM is employed completely in the fifth step. The results are shown in the fourth and in the fifth rows of Table I respectively.

Let’s the results in the step 1 are considered as basic results. Using shortest path in the neighborhood graph of the original structure not only improves the results, but also decreases the precision. This is because of noise between labeled and not-labeled samples in the fused dissimilarity graph (G). Therefore, the geodesic distance of samples is not well represented by the shortest path. In the third step, although the results become better than the second step, the precision is still less than to that of the first step. So, better approximation of the geodesic distance has been found by mapping the conceptual space to the dissimilarities of labeled samples (Algorithm 2). In the fourth step, in addition to the dissimilarity enhancement between the labeled samples, the dissimilarities between the unlabeled and the labeled ones have been corrected (Algorithm 3). In this regard a better approximation in evaluating geodesic distance which provides better results than the first step has been found. Finally, in the fifth step, ASM is employed. Using the enhancement of the dissimilarities between unlabeled samples in an unsupervised approach (Algorithm 4), the shortest path provides the best approximation of geodesic distance in our proposed annotation method.

A comparison of the performance of the proposed algorithm with other approaches is shown in Table II.

The proposed SML model gives a F-measure of 0.3%, which is comparable to the best approaches in the literature. Moreover, NumWord is 130 which is between 49 [10] and 147 [18], as reported in the literature.

TABLE I. COMPARISON OF ANNOTATION QUALITY IN DIFFERENT LEVELS FROM THE ORIGINAL STRUCTURE TO THE MANIFOLD STRUCTURE.

Graph structure	Precision	Recall	F1	NumWords
Step 1. original (basic) structure	0.3052	0.2677	0.2852	126
Step 2. Manifold structure without dissimilarity enhancing	0.2131	0.1791	0.1946	103
Step 3. Manifold structure with only enhancing the dissimilarity between the labeled samples.	0.2503	0.2133	0.2303	109
Step 4. Manifold structure with ignoring the dissimilarity enhancement between the unlabeled samples.	0.3055	0.2834	0.29403	130
Step 5. Manifold structure	0.3127	0.2838	0.2975	130

TABLE II. PERFORMANCE COMPARISON ON COREL DATASET.

Methods	Precision	Recall	F1
TM [10]	0.06	0.04	0.048
CMRM [11]	0.09	0.1	0.095
CLM [12]	0.19	0.12	0.147
CRM [13]	0.16	0.19	0.174
AGAnn [14]	0.236	0.269	0.251
SML [15]	0.228	0.29	0.255
MBRM [16]	0.22	0.24	0.230
CLP [16]	0.21	0.26	0.232
GBM [16]	0.23	0.28	0.253
TGLM [17]	0.253	0.291	0.271
CKSM [18]	0.287	0.349	0.315
HDGM [19]	0.29	0.3	0.3
SSML [20]	0.4	0.25	0.31
Proposed AMS	**0.313**	**0.284**	**0.298**

IV. CONCLUSIONS AND FUTURE WORKS

In this paper, we introduced a method for image Annotation based on Manifold Structure (AMS). The proposed method bridges the semantic gap between the visual contents and textual tags in an efficient way based on a graph-based semi supervised algorithm. Images were described in fifteen local and global feature spaces. Furthermore, fusion was done in graph level instead of feature and decision levels. The structure of the fused dissimilarity graph was altered by using labeled samples of images. For this purpose, an innovative way for transition from the conceptual space to the fused dissimilarity graph was proposed. Using the calculation of the shortest path in the altered graph, a suitable approximation of the intrinsic dissimilarity was obtained. Transition of semantics from the conceptual space to the fused dissimilarity graph is the main innovation of the proposed method.

The accuracy improvement based on AMS is due to the altering the fused dissimilarity graph by some semantic information. This is done by three algorithms in constriction of graph of manifold structure.

The goal of automatic image annotation is to automatically annotate a huge amount of images effectively and efficiently while the required labeled information is as less as possible. Hence, in the future, we will work on fusion of different feature spaces in a more effective way to reduce the semantic gap by extracting features with the aim of closing Euclidean and semantic distances.

REFERENCES

[1] Tang J., Qi G. J., M. Wang, and X. S. Hua, "Video semantic analysis based on structure-sensitive anisotropic manifold ranking", Signal Processing, vol. 89 no. 12, pp. 2313-2323, 2009.

[2] Tang J., Hua X. S., Qi G. J., Wang M., Mei T., and Wu X., Structure-Sensitive Manifold Ranking for Video Concept Detection, in: 15th international conference on Multimedia, 2007, pp. 852-861 .

[3] J. Fan, Y. Gao, H. Luo, Integrating Concept Ontology and Multitask Learning to Achieve More Effective Classifier Training for Multilevel Image Annotation, IEEE Transactins on Image Processing 17 (2008) 407-426.

[4] A. Oliva, A. Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, International Journal of Computer Vision 42 (2001) 145-175.

[5] J. Ilonen, J. K. Kamarainen, H. Kalviainen, Y. L. Teknillinen, Efficient computation of Gabor features, Lappeenranta University of Technology, 2005.

[6] J. Verbeek, M. Guillaumin, T. Mensink, C. Schmid, Image annotation with tagprop on the mirflickr set, in: International Conference on Multimedia information retrieval, 2010, pp. 537-546.

[7] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91-110.

[8] J. V. Weijer, C. Schmid, Coloring local feature extraction, Lecture Notes in Computer Vision 3952 (2006) 334-348.

[9] B. Wang, F. Hu, J. C. Paul, Manifold-ranking based retrieval using k-regular nearest neighborgraph, Pattern Recognition 45 (2012) 1569-1577.

[10] P. Duygulu, K. Barnard, N. Freitas, D. Forsyth, Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, in: Seventh European Conference on Computer Vision, 2002, pp. 97-112.

[11] J. Jeon, V. Lavrenko, R. Manmatha, Automatic image annotation and retrieval using cross-media relevance models, in: 26th Annual International ACM SIGIR, 2003, pp. 119-126.

[12] R. Jin, J. Chai, L. Si, Effective automatic image annotation via a coherent language model and active learning, in: 12th Annual ACM Internatational Conference on Multimedia, 2004, pp. 892-899.

[13] V. Lavrenko, R. Manmatha, J. Jeon, A model for learning the semantics of pictures, in: Advance Neutral Information Processing, 2003.

[14] J. Liu, M.J. Li, W. Ma, Q. Liu, H.Q. Lu, An adaptive graph model for automatic image annotation, in: Eighth ACM International Workshop on Multimedia Information Retrieval, 2006, pp. 61-70.

[15] G. Carneiro, A.B. Chan, P.J. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, IEEE Transaction on Pattern Analysis and Machine Intelligence 29 (2007) 394-410.

[16] J. Liu, B. Wang, H. Lu, S. Ma, A graph-based image annotation framework, Pattern Recognition Letters 29 (2008) 407-415

[17] J. Liua, M. Lib, Q. Liua, H. Lua, S. Ma, Image annotation via graph learning, Pattern Recognition 42 (2009) 218-228.

[18] Z. Lu, Horace H. S. Ip, Y. Peng, Contextual Kernel and Spectral Methods for Learning the Semantics of Images, IEEE Transactins on Image Processing 20 (2011) 1739-1750.

[19] X. Ke, Shaozi Li, Donglin Cao, A two-level model for automatic image annotation, Multimedia Tools and Applications (2011).

[20] M. A. Z. Chahooki, N. M. Charkari, "Bridging the semantic gap for automatic image annotation by learning the manifold space", Computer Systems Science and Engineering (2015) In Press.