AUTOMATIC GROUND-TRUTH IMAGE GENERATION FROM USER TAGS

Triantafillos Tsirelis and Anastasios Delopoulos

Department of Electrical and Computer Engineering
Aristotle University of Thessaloniki - Greece
30fyllos.t@gmail.com, adelo@eng.auth.gr

ABSTRACT

Automatic selection of ground-truth images is very important for the training of image classifiers, like the ones used in concept-based image retrieval. For this purpose, we propose a method which collects a sufficient number of ground-truth images based on their user-assigned tags. The semantic similarity between the tags of the images and the concept is used as a relevancy metric to classify the images in ranked lists. The system is comprised of parts of data preprocessing, WordNet-based synonym retrieval, Natural Language Processing and corpus-based semantic similarity calculations. Experimental results indicate that the proposed method is effective in collecting groundtruth data and that the training of concept classifiers based on this groundtruth leads to effective image retrieval.

1. INTRODUCTION

Large numbers of labeled images are necessary when training image classifiers to detect concepts on the basis of feature vectors extracted from the examined images. However, manually labeling a large number of images is not a trivial task.

A number of methods addressing this problem make use of the tags associated with images. Datta et al. in [1] describe a method that tags untagged images or retags already tagged images and uses the auto-generated tags to improve retrieval. In [2], text, metadata and visual features are used to gather a large number of web images for a specified class. Image tags in [3] are used as additional features of labeled data for the classification of touristic landmark images. The method described in [4] uses the tags associated with labeled and unlabeled images to improve a classifier using semi-supervised learning.

Our approach considers a scenario where we have a set of unlabeled images with their associated tags, such as found on photo sharing websites like Flickr. Based on the tags of these images, the goal is to retrieve a large number of ground-truth images for a given concept and use them to train a classifier and improve its retrieval performance.

The next section is devoted to the detailed description of the proposed method. Section 3 presents experiments conducted on real data and section 4 concludes and makes recommendations for future work.

2. GROUND-TRUTH RETRIEVAL

The proposed method of automatic ground-truth image retrieval can be divided in three main parts: the definition of the concept based on which the images will be classified, the calculation of the relevancy between the given concept and the images and the selection of the ground-truth images.

2.1. Concept definition

Suppose that given a specific concept, an expert provides a set W_1 of keywords that are tightly associated to the concept. Then, the synonyms of each of these words are found with the use of WordNet [5, 6]. Thus, we end up defining the given concept C with an expaned set of keywords $W_2 = \{w_{21}, w_{22}, \ldots, w_{2j}, \ldots, w_{2n}\}$ (Fig. 1).

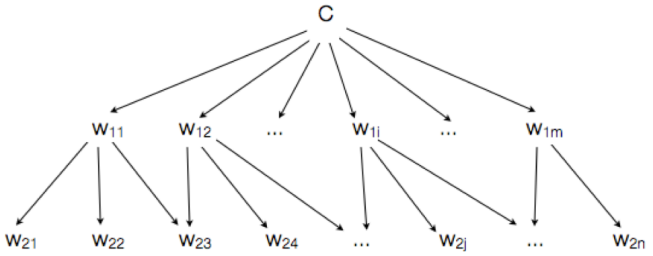

Fig. 1. Defining a concept with keywords.

2.2. Concept-image relevancy

The next step is to find a way to measure the relevancy between a given concept and an image, with a set of associated tags, $T = \{t_1, t_2, \ldots, t_k, \ldots, t_l\}$. To accomplish this, we use the semantic similarity between the set of keywords that define the concept and the tags associated to the image as a metric of the relevancy between the concept and the image.

In order to calculate the semantic similarity between the two word-sets W_2 and T we use the Average Aggregated

Minimum Distance as described in [7]. We are calculating similarity, so we name it Average Aggregated Maximum Similarity (AAMS). If we denote the between word semantic similarity by $s_w(\cdot,\cdot)$ and the between word and set similarity by $\bar{s_w}(\cdot,\cdot)$, then the similarity between a keyword from W_2 and the tag-set T of an image is:

$$\bar{s_w}(w_{2j}, T) = \max_{t_k \in T} \{s_w(w_{2j}, t_k)\}$$

In the same way, the similarity between a tag from T and the keyword set W_2 that defines the concept is:

$$\bar{s_w}(t_k, W_2) = \max_{w_{2j} \in W_2} \{s_w(t_k, w_{2j})\}$$

The similarity between the two sets is computed by averaging over $\bar{s_w}(w_{2j}, T)$, $j = 1, \ldots, n$ and $\bar{s_w}(t_k, W_2)$, $k = 1, \ldots, l$, as shown in (1):

$$sim(W_2, T) = \frac{1}{n} \sum_{j=1}^{n} \bar{s_w}(w_{2j}, T) + \frac{1}{l} \sum_{k=1}^{l} \bar{s_w}(t_k, W_2) \quad (1)$$

Each word in each set is matched with the semantically most similar word in the other set. The overall similarity is the average over the similarities between the matched words.

We measure the word-to-word semantic similarity $s_w(\cdot,\cdot)$ using the Pointwise Mutual Information for Information Retrieval (PMI-IR) [8]. PMI-IR is based on word co-occurrence making use of counts collected over large corpora. The PMI-IR between a word from W_2 and a word from T is measured as:

$$PMIIR(w_{2j}, t_k) = \log_2 \frac{Pr(w_{2j}, t_k)}{Pr(w_{2j}) Pr(t_k)}$$

From the four different types suggested in [8] we employ the case which uses the *AND* operator (co-occurrence within a document of the corpus). Furthermore, both word-sets and the corpus are normalized (duplicate and stop-word removal, stemming and lowercasing). If $n_{w_{2j}}$ is the number of corpus documents where a word from W_2 occurs alone, n_{t_k} the document number where a word from T occurs alone, n_c the number of documents where these words co-occur, and n the total number of the corpus documents, then the PMI-IR of those words is approximated as:

$$PMIIR(w_{2j}, t_k) \simeq \log_2 \frac{\frac{n_c}{n}}{\frac{n_{w_{2j}}}{n} \frac{n_{t_k}}{n}} = \log_2 \frac{n_c}{n_{w_{2j}} n_{t_k}} n$$

Since we are looking for the maximum score, we can drop log_2, because it is monotonically increasing and n, because it has the same value for all word couples. As a result, the semantic similarity between the two words is:

$$s_w(w_{2j}, t_k) = \frac{n_c}{n_{w_{2j}} n_{t_k}} \quad (2)$$

Calculating the AAMS between the keyword-set of the concept and the tag-set of each one of the images from the total image-set, we measure the relevancy between the concept and each image.

2.3. Groundtruth image selection

The images are classified based on their relevancy to the concept, forming a ranked list containing the most relevant images at the top and the least relevant at the bottom. Finally, the desired number of the top-ranked images are chosen as positive to the concept and the the desired number of the bottom-ranked images as negative to the concept. In this way, we have retrieved a large number of ground-truth images for the given concept.

3. EXPERIMENTS

We investigate our system's performance on two grounds: how accurately it retrieves ground-truth and how much it improves the performance of an SVM classifier, if the system's results are used to train this classifier.

The dataset used for the experiments is NUS-WIDE [9], a real-world web image dataset created by the Lab for Media Search of the National University of Singapore. It includes: (i) 269,648 Flickr images with their associated tags, (ii) six types of low-level features extracted from these images (64-D color histogram, 144-D color correlogram, 73-D edge direction histogram, 128-D wavelet texture, 225-D block-wise color moments extracted over 5x5 fixed grid partitions, and 500-D bag of words based on SIFT descriptions), (iii) ground-truth for 81 concepts. The Brown Corpus was used for the word-to-word semantic similarity calculation.

The evaluation methods used for the experiments are the Normalized Discounted Cumulated Gain (nDCG) [10] and the Average Precision. The nDCG at a particular rank position i is calculated as follows:

$$nDCG_i = \frac{DCG_i}{IDCG_i} \quad (3)$$

DCG_i is the Discounted Cumulated Gain at the particular rank and $IDCG_i$ the DCG_i of the ideal performance. The DCG_i can be calculated as shown:

$$DCG_i = \begin{cases} CG_i & \text{if } i < b \\ CG_{i-1} + \frac{G_i}{\log_b i} & \text{if } i \geq b \end{cases}$$

CG_i is the Cumulated Gain at the particular rank and is calculated as:

$$CG_i = \sum_{k=1}^{i} G_k$$

G_k is the gain for the particular rank (1 for a relevant image and 0 for an irrelevant). On the other hand, the average precision is:

$$aveP = \frac{\sum_{i=1}^{N} (p_i * rel[i])}{M} \quad (4)$$

where i is the rank, N the number of images recalled, $rel[i]$ a function with value 1 if the image of the given rank is relevant

to the concept or 0 if not, M the number of relevant images and p_i the precision at the given rank i.

3.1. Performance evaluation of the groundtruth selection

To evaluate system performance in the selection of groundtruth datasets, and particularly their positive subset, first we choose the concepts for which the proposed method will be applied. The concepts are carefully chosen in such a way that they contain both general and specific concepts and they belong to different categories including scene, object, event and people. We apply our method for the chosen concepts and keep only the 200 most relevant images for each one. Then, we check them in two ways:

1. With manual annotation: if the examined image contains the concept for which is being examined, it is labeled as positive, else it is labeled as negative.

2. Using the NUS-WIDE ground-truth.

Finally, we calculate precision at rank 200 for both checks. In Table 1 we present for each concept the p_{200} for the manual annotation, the p_{200} based on the NUS-WIDE groundtruth. The last column contains estimates of the prior of each concepts obtained as the percentage of positive labels in the NUS-WIDE groundtruth. For the concepts not included in NUS-WIDE we put a (-) at the associated columns.

We can see that the ground-truth image set retrieved by our method are reliable. The precision values, p_{200} %, for the chosen concepts are high, meaning that most of the 200 images retrieved as positive for each concept, are in fact positive. On the other hand precision values computed on the basis of the NUS-WIDE labels, p_{200} % (NW), are not as high as the ones calculated using the manual annotation. This means that images which are indeed positive, as verified by the annotation, are not labeled as positive by the NUS-WIDE. Ignoring these limitations regarding the accuracy of NUS-WIDE labels, we may observe taking into account the fourth column of Table 1 that the accuracy of the obtained ground truth data set is high even for concepts with very low prior.

concept	p_{200} %	p_{200} % (NW)	$prior$ % (NW)
animal	98.5 %	97 %	12.56 %
beach	96 %	57.5 %	1.94 %
bicycle	90 %	-	-
boat	95.5 %	67 %	1.49 %
building	93.5 %	49 %	6.61 %
butterfly	73.5 %	-	-
car	94.5 %	26 %	0.59 %
cloud	90 %	79.5 %	20.06 %
computer	79 %	44.5 %	0.22 %
cow	91 %	52.5 %	0.31 %
desert	89.5 %	-	-
elephant	87.5 %	-	-
face	89 %	-	-
flower	99 %	87 %	3.19 %
food	93 %	70.5 %	0.99 %
football	87 %	-	-
frost	96.5 %	44 %	0.43 %
hand	94 %	-	-
house	89.5 %	20.5	1.45 %
moon	88.5 %	73 %	0.35 %
mountain	92 %	60 %	1.89 %
night	92.5 %	23.5 %	1.46 %
person	91 %	73 %	19.13 %
plane	98.5 %	90.5 %	0.99 %
police	78.5 %	68.5 %	0.52 %
pyramid	83.5 %	-	-
reflection	91.5 %	55 %	2.92 %
sky	95.5 %	90 %	27.51 %
snow	96 %	85.5 %	2.01 %
sport	97 %	53.5 %	0.67 %
sun	73.5 %	40.5 %	1.35 %
temple	87 %	23 %	0.62 %
train	93 %	40 %	0.35 %
tree	94.5 %	37 %	1.99 %
waterfall	84 %	36 %	0.23 %
window	90 %	58 %	5.58 %
woman	96 %	-	-

Table 1. The system performance results

3.2. SVM classifier enhancement

The next experiment aims to evaluate the appropriateness of the obtained groundtruth for training concept classifiers that will be used in image retrieval. We choose the concepts for which we will train the classifier. The concepts are chosen based on their frequency in the tags: the two least frequent concepts (frost, computer), two of medium frequency (moon, plane) and three more frequent (boat, flower, tree).

The data used for the training of the SVM are the low-level features provided by NUS-WIDE. For each of the chosen concepts, the training data selection is based on the following three scenarios:

1. According to the ground-truth labels provided by NUS-WIDE, we choose the 500 positive and 500 negative images.

2. According to our system, we choose the 500 most relevant and the 500 most irrelevant images.

3. For each concept, the positive part of our training data is formed by choosing only the images labeled as positive from the manual annotation of Section 3.1. The negative images are the rest up to 1000 most irrelevant to each concept, according to our system.

The test data are all the rest of the NUS-WIDE images that were not used at the above scenarios for each concept.

For each scenario of each concept we use the following procedure: First we scale the training and test data to the

range $[-1, 1]$. Next, considering the RBF kernel for the SVM, we find the best parameters C and γ and use them to train the training set. We test using the test set and classify the predictions based on their probabilities. Then, we evaluate the SVM performance with two checks:

1. we consider the top 200 images with the highest probability as positive and we manually annotate them (as in 3.1).
2. we check the entire set of classified predictions using the ground-truth labels from NUS-WIDE.

For the two checks we calculate the $nDCG_{200}$ with $b = 2$. Additionally, for the second check we calculate the average precision for the entire set of predictions. In Table 2 we present the nDCG values at rank 200 for both checks, the average precision for the second check and the prior of each concept, based on the ground-truth of NUS-WIDE.

concept	scenario	$nDCG_{200}$	$nDCG_{200}$ (NW)	$aveP$ (NW)	$prior\%$ (NW)
boat	1	0.266	0.045	0.046	1.49 %
	2	0.572	0.216	0.07	
	3	0.583	0.296	0.072	
computer	1	0.05	0.02	0.008	0.22 %
	2	0.075	0.025	0.01	
	3	0.11	0.035	0.014	
flower	1	0.186	0.091	0.097	3.19 %
	2	0.613	0.347	0.166	
	3	0.709	0.528	0.174	
frost	1	0.116	0.015	0.006	0.43 %
	2	0.166	0.035	0.007	
	3	0.106	0.03	0.006	
moon	1	0.095	0.0	0.004	0.35 %
	2	0.126	0.005	0.005	
	3	0.201	0.045	0.011	
plane	1	0.116	0.075	0.03	0.99 %
	2	0.226	0.116	0.048	
	3	0.231	0.151	0.054	
tree	1	0.316	0.06	0.049	1.99 %
	2	0.558	0.166	0.05	
	3	0.699	0.176	0.074	

Table 2. The SVM performance results

As it can be seen, our method ground-truth (scenarios 2 and 3) improve the classifier performance over the case where NUS-WIDE ground-truth are used (scenario 1). In most cases scenario 3 performs better, as expected, because the positives it contains are in fact positives. However, in some cases (frost) the number of positives of the third scenario is not enough (less than 200) to effectively train the classifier.

4. CONCLUSIONS AND FUTURE WORK

We have presented a methodology for selecting ground-truth images and employing them for training classifiers. Our approach provides large numbers of labeled images leading to effective classifiers for concept based image retrieval. Experiments with real data have shown that the method is reliable and can be successfully used to enhance the retrieval scores of an SVM classifier.

Of potential interest for future research could be the improvement of the ground-truth retrieval methodology, concerning the part of concept definition, as well as the concept-image relevancy calculation. Experimentation on other image datasets and/or different supervised learning approaches could also be an object of future work.

5. REFERENCES

[1] R. Datta, W. Ge, J. Li, and J. Z. Wang, "Toward bridging the annotation-retrieval gap in image search," *IEEE Multimedia*, vol. 14, no. 3, pp. 24–35, Jul.–Sept. 2007.

[2] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," in *11th Int. Conf. Computer Vision*, Rio de Janeiro, Brasil, 2007, pp. 1–8.

[3] Y. Li, D. J. Crandall, and D. P. Huttenlocher, "Landmark classification in large-scale image collections," in *12th Int. Conf. Computer Vision*, Kyoto, Japan, 2009, pp. 1957–1964.

[4] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *23rd IEEE Conf. Computer Vision and Pattern Recognition*, San Francisco, CA, 2010, pp. 902–909.

[5] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[6] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press, 1998.

[7] J. Li, "A mutual semantic endorsement approach to image retrieval and context provision," in *Proc. 7th ACM SIGMM Int. Workshop Multimedia Information Retrieval*, Hilton, Singapore, 2005, pp. 173–182.

[8] P. D. Turney, "Mining the web for synonyms: PMI-IR versus LSA on TOEFL," in *Proc. 12th European Conference Machine Learning*, London, UK, 2001, pp. 491–502.

[9] T. S Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. T. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Conf. Image and Video Retrieval*, Santorini, Greece, 2009.

[10] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, Oct. 2002.