International Journal of Business Quantitative Economics and Applied Management Research
ISSN: 2349-5677
Volume 1, Issue 4, September 2014

Automated image annotation process: A general overview

Gabriel Mihai, Ph.D.
University of Craiova, Faculty of Automation, Computers and Electronics
Bvd. Decebal, No.107, Craiova, Romania
mihai_gabriel@software.ucv.ro

Abstract
Automated image annotation process consists in assigning meaningful keywords to an image by taking into account its visual content. This process has received a great interest because it allows indexing, retrieving, and understanding of large collections of image data. The results of the annotation process are directly influenced by the quality of the regions obtained during the segmentation process. Image segmentation is a difficult and challenging task in image processing, consisting in dividing an image into different and homogeneous regions. Since there are multiple segmentation algorithms in the literature, numerical evaluations are needed in order to quantify the consistency between them. Error measures can be used for consistency quantification because are allowing a principled comparison between segmentation results on different images. This paper presents the steps involved by the annotation process: choosing a segmentation algorithm based on an evaluation performed with segmentation error measures, choosing a comprehensive data set representing the ground truth, choosing an annotation model that is capable of producing good annotation results, evaluating the annotation results.

Key words : image annotation; annotation model; image segmentation; segmentation evaluation; segmentation comparison; ontology;

1. Introduction

Image annotation is a difficult task for two main reasons: *semantic gap* problem - it is hard to extract semantically meaningful entities using just low level image features and *the lack of correspondence* between the keywords and image regions in the training data. There are multiple annotation models like Co-occurrence Model [3], Translation Model [4], Cross Media Relevance Model (CMRM) [5], Continuous Cross-Media Relevance Model (CRM) [9], Multiple Bernoulli Relevance Model (MBRM) [10], Coherent Language Model (CLM) [11], each of them trying to improve a previous model.

The annotation process used for our experiments system is based on CMRM. This model is much more efficient in implementation than other parametric models because it doesn't have a training stage to estimate model parameters. Using a training set of annotated images, the

International Journal of Business Quantitative Economics and Applied Management Research
ISSN: 2349-5677
Volume 1, Issue 4, September 2014

annotation model learns the joint distribution of the blobs and keywords. The SAIAPR TC-12 Dataset [6] is a set of annotated images having a vocabulary with a hierarchical structure. In the annotation context, the blobs are clusters of image regions obtained using the K-means algorithm on image features. Having the set of blobs, each image from the training set is represented using a discrete sequence of blobs identifiers. This distribution is used to generate a set of keywords for a new image. Each image is segmented using a segmentation algorithm [12] which integrates pixels into a grid-graph.

Image segmentation is one of the most difficult and challenging tasks in image processing and can be defined as the process of dividing an image into different regions such that each region is homogeneous while not the union of any two adjacent regions. The consistency between segmentations must be evaluated because no unique segmentation of an image can exist. If two different segmentations arise from different perceptual organizations of the scene, then these segmentations can be considered as being inconsistent.

The remainder of the paper is organized as follows: segmentation error measures are presented in Section 2, Section 3 contains a description of the dataset used for the annotation process, Section 4 presents the annotation model and the evaluation process and Section 5 concludes the paper.

2. Segmentation error measures

In order to evaluate a segmentation algorithm for the annotation process, it is needed to measure the accuracy, the precision and the performance. From performance point of view the algorithms are to be evaluated by objective comparison of their segmentation results with manual segmentations.

Any error measure should have the following characteristics [1]: tolerant to refinement, independent of the coarseness of pixilation, robust to noise along region boundaries, tolerant of different segment counts between the two segmentations due to the complexity of the images.

When multiple segmentation algorithms are evaluated, some metrics are needed to establish which algorithm produce better results. A segmentation error measure takes two segmentations S_1 and S_2 as input, and produces a real valued output in the range $[0..1]$ where zero signifies no error. For a given pixel p_i two segments S_1 and S_2, containing that pixel, are considered. If one segment is a proper subset of the other, then the pixel lies in an area of refinement, and the local error should be zero. If there is no subset relationship, then the two regions overlap in an inconsistent manner. In this case, the local error should be non-zero. Let \ denote set difference, and $|x|$ the cardinality of set x. If $R(S, p_i)$ is the set of pixels corresponding to the region in segmentation *S* that contains pixel p_i, the local refinement error is defined as:

$$E(S_1, S_2, p_i) = \frac{|R(S_1, p_i) \setminus R(S_2, p_i)|}{R(S_1, p_i)} \quad (1)$$

International Journal of Business Quantitative Economics and Applied Management Research
ISSN: 2349-5677
Volume 1, Issue 4, September 2014

This local error measure is not symmetric and it encodes a measure of refinement in one direction only. Given this local refinement error in each direction at each pixel, there are two natural ways to combine the values into an error measure for the entire image.

There are two metrics that can be used to evaluate the consistency of a pair of segmentations: Global Consistency Error (GCE) and Local Consistency Error (LCE) and are defined as:

$$GCE(S_1, S_2) = \frac{1}{n}\min\left\{\sum_i E(S_1, S_2, p_i), \sum_i E(S_2, S_1, p_i)\right\} \quad (2)$$

$$LCE(S_1, S_2) = \frac{1}{n}\sum_i \min\{E(S_1, S_2, p_i), E(S_2, S_1, p_i)\} \quad (3)$$

GCE forces all local refinements to be in the same direction and LCE allows refinement in different directions.

$\mathbf{LCE \leq GCE}$ for any two segmentations and it is clear that GCE is a tougher measure than LCE. When pairs of human segmentations of the same image are compared, both the GCE and the LCE are low; conversely, when random pairs of human segmentations are compared, the resulting GCE and LCE are high. If the pixel wise minimum is replaced by a maximum it is obtained a new measure named Bidirectional Consistency Error (BCE) that is not tolerating the refinement. This measure is evaluated using

$$BCE(S_1, S_2) = \frac{1}{n}\sum_i \max\{E(S_1, S_2, p_i), E(S_2, S_1, p_i)\} \quad (4)$$

To better understand how the GCE and LCE error metrics work, it's interesting to consider what the metrics report on two extreme cases:

a) a completely under-segmented image where every pixel has the same label; the segmentation contains only one region spanning the whole image
b) a completely over-segmented image in which every pixel has a different label

From the definitions of the GCE and LCE it can be seen that both measures evaluate to 0 on both of these extreme situations regardless of what segmentation they are being compared to.

The reason for this can be found in the tolerance of these measures to refinement. Any segmentation is a refinement of the completely under-segmented image, while the completely over-segmented image is a refinement of any other segmentation. The BCE error measure was introduced to avoid this situation, being non tolerant to refinement.

In [2] we have evaluated three image segmentation algorithms based on the above error measures: color set back-projection algorithm [13], local variation algorithm [14] and hexagonal structure based algorithm [12]. The last algorithm integrates the pixels into a grid-graph. The

International Journal of Business Quantitative Economics and Applied Management Research
ISSN: 2349-5677
Volume 1, Issue 4, September 2014

usage of the hexagonal structure improves the time complexity of the methods used and the quality of the segmentation results. Hence, we have decided to use it for the annotation process.

3. Dataset

We have chosen for our experiments the segmented and annotated SAIAPR TC-12 [6][7] benchmark which is an extension of the IAPR TC-12 [8] collection for the evaluation of automatic image annotation methods and for studying their impact on multimedia information retrieval. SAIAPR TC-12 benchmark contains the pictures from the IAPR TC-12 collection plus: segmentation masks and segmented images for the 20,000 pictures, region-level annotations according an annotation hierarchy, region-level annotations according an annotation hierarchy, spatial relationships information. Each region has associated a segmentation mask and a label from a predefined vocabulary of 275 concepts. This vocabulary is organized according to a hierarchy of concepts having six main branches: Humans, Animals, Food, Landscape-Nature, Man-made and Other. For each pair of regions the following relationships have been calculated in every image: adjacent, disjoint, beside, X-aligned, above, below and Y-aligned. The following features have been extracted from each region: area, boundary/area, width and height of the region, average and standard deviation in x and y, convexity, average, standard deviation and skewness in two color spaces: RGB and CIE-Lab. The dataset contains several folders of images, each folder having the structure presented in figure 1:

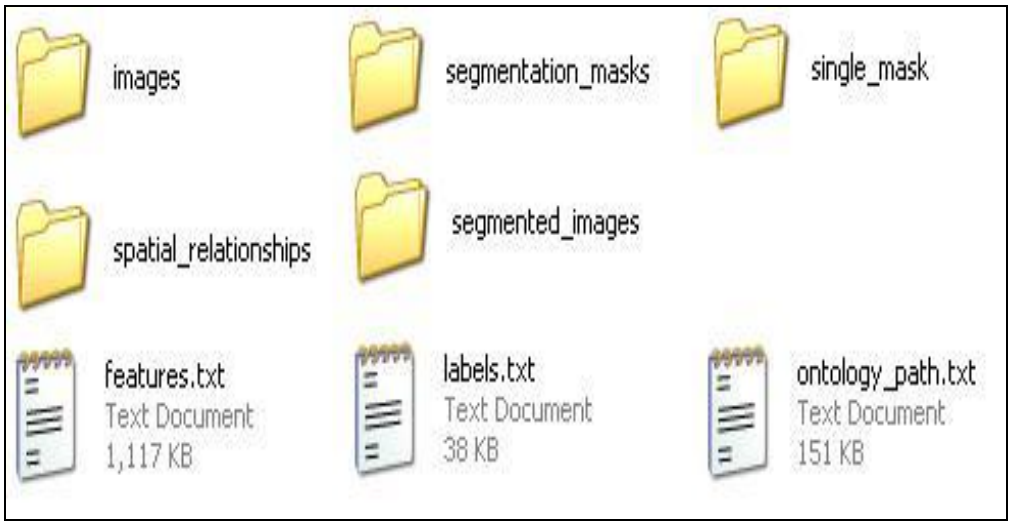

Fig. 1. The structure of images folder

The images associated with each concept can be explored using the viewer presented in figure 2:

International Journal of Business Quantitative Economics and Applied Management Research
ISSN: 2349-5677
Volume 1, Issue 4, September 2014

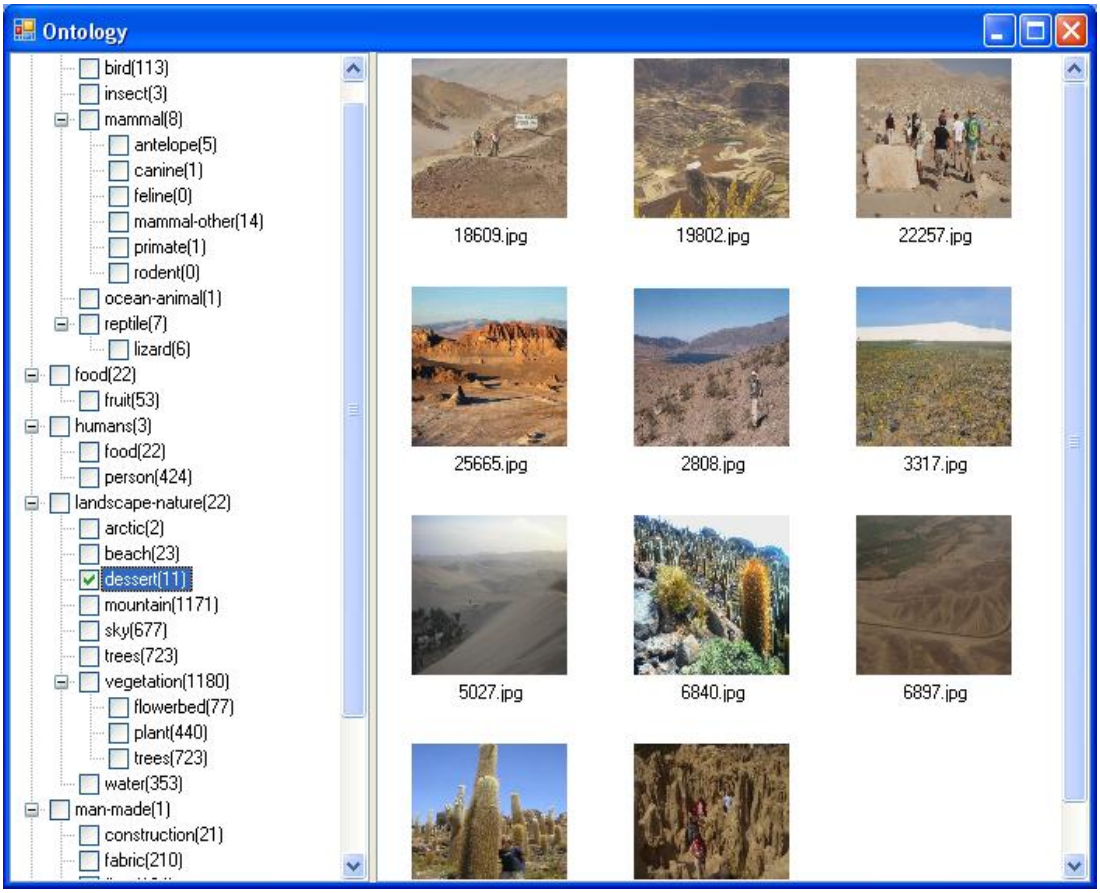

Fig. 2. Ontology viewer

4. The Annotation process

4.1 The Annotation model

The Cross Media Relevance Model [5] is a non-parametric model for image annotation and assigns keywords to the entire image and not to specific blobs – clusters of image regions, because the blob vocabulary can give rise to many errors. Given a training set of images with annotations, this model allows predicting the probability of generating a keyword given the blobs in an image. A test image *I* is annotated by estimating the joint probability of a keyword *w* and a set of blobs (cluster of image regions):

International Journal of Business Quantitative Economics and Applied Management Research
ISSN: 2349-5677
Volume 1, Issue 4, September 2014

$$P(w, b_1, \ldots, b_m) = \sum_{J \in T} [P(J)P(w,]b_1, \ldots, b_m|J).$$ (5)

For the annotation process the following assumptions are made:

a) it is given a collection *C* of un-annotated images
b) each image *I* from *C* to can be represented by a discrete set of blobs $I = \{b_1 \ldots b_m\}$
c) there exists a training collection *T*, of annotated images, where each image *J* from *T* has a dual representation in terms of both words and blobs: $J = \{b_1 \ldots b_m; w_1 \ldots w_n\}$
d) $P(J)$ is kept uniform over all images in *T*
e) the number of blobs m and words in each image (m and n) may be different from image to image.
f) no underlying one to one correspondence is assumed between the set of blobs and the set of words; it is assumed that the set of blobs is related to the set of words.

$P([w, b]_1, \ldots, b_m|J)$ represents the joint probability of keyword *w* and the set of blobs $(b_1, \ldots, b_m)$ conditioned on training image *J*. An intuitive interpretation of this probability is how likely *w* co-occurs with individual blobs given that we have observed an annotated image *J*.

In CMRM it is assumed that, given image *J*, the events of observing a particular keyword *w* and any of the blobs $(b_1, \ldots, b_m)$ are mutually independent, so that the joint probability can be factorized into individual conditional probabilities. This means that $P(b_1, \ldots, b_m|J)$ can be written as:

$$P([w, b]_1, \ldots, b_m|J) = P(w|J)\prod_i (i = 1)^m [P(b_i|J)].$$ (6)

$$P(w|J) = (1 - \alpha_J)\frac{\#(w, J)}{|J|} + \alpha_J \frac{\#(w, T)}{|T|}.$$ (7)

$$P(b|J) = (1 - \beta_J)\frac{\#(b, J)}{|J|} + \beta_J \frac{\#(b, T)}{|T|}.$$ (8)

where:

a) P(w|J) , P(w|J) denote the probabilities of selecting the word w, the blob b from the model of the image J.

b) #(w, J) denotes the actual number of times the word w occurs in the caption of image J.

International Journal of Business Quantitative Economics and Applied Management Research
ISSN: 2349-5677
Volume 1, Issue 4, September 2014

c) #(w, T) is the total number of times w occurs in all captions in the training set T .

d) #(b, J) reflects the actual number of times some region of the image J is labeled with blob b.

e) #(b, T) is the cumulative number of occurrences of blob b in the training set.

f) |J| stands for the count of all words and blobs occurring in image J.

g) |T| denotes the total size of the training set.

h) The prior probabilities P(J) can be kept uniform over all images in T

The smoothing parameters α and β determine the degree of interpolation between the maximum likelihood estimates and the background probabilities for the words and the blobs respectively. The values determined after experiments for the Cross Media Relevance Model were α = 0.1 and β = 0.9.

4.2Evaluation of the annotation task

In order to evaluate the annotation process from two perspectives (annotation, retrieval) we have used the images included in the SAIAPR TC-12 dataset. From annotation perspective, the number of relevant keywords automatically assigned (image 3) by the annotation system was compared against the number of relevant keywords manually assigned by computing a recall value. Using this approach for each image, we have obtained a statistic evaluation having the following structure:

Index	*Image*	*Relevant keywords automatically assigned (RWAA)*	*Keywords manually assigned (WMA)*	*Recall = RWAA/ WMA*
0		sky-blue, sand-beach, ocean	sand-beach, ocean, boat, palm, hut, sky-blue	3/6 = 0.50
1		sky-blue, grass, ocean, cloud	grass, ocean, boat, cloud, sky-blue, branch	4/6 = 0.66
2		sky, mountain, lake	lake, vegetation, mountain, cloud, sky	3/5 = 0.60

International Journal of Business Quantitative Economics and Applied Management Research
ISSN: 2349-5677
Volume 1, Issue 4, September 2014

3		mountain, sky-blue, sand-dessert	mountain, lake, sand-dessert, sky-blue	3/4 0.75	=

After computing the recall value for each image, it was obtained a medium recall value equal to 0.66.

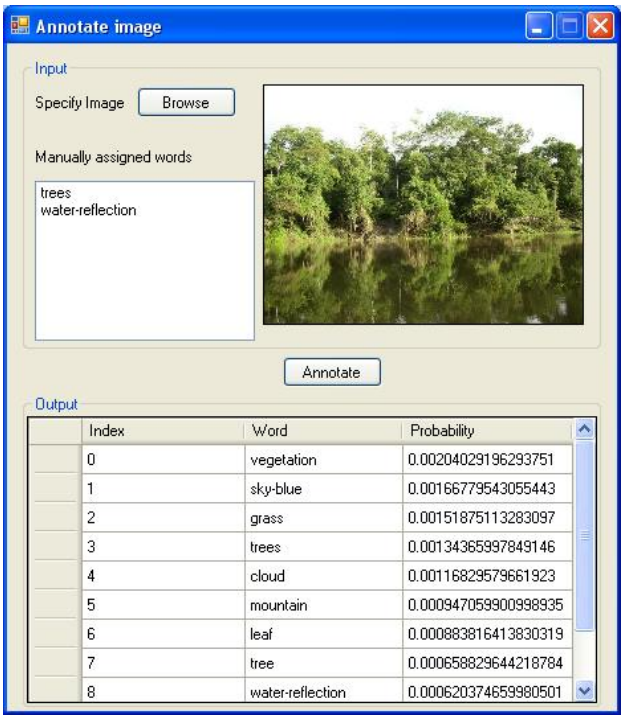

Fig.3. Image annotation

From retrieval perspective, we've computed the precision and recall values for each concept. A sample of this computation can be found in table 1:

Table 1: Evaluation from retrieval perspective

Index	Concept	Precision	Recall
0	blue-sky	0.34	0.40
1	cloud	0.51	0.46
2	vegetation	0.43	0. 57
3	mountain	0.42	0.49
4	ground	0.36	0.56
5	grass	0.54	0.48
6	wall	0.26	0.59
7	ocean	0.52	0.38

International Journal of Business Quantitative Economics and Applied Management Research
ISSN: 2349-5677
Volume 1, Issue 4, September 2014

8	tree	0.47	0.52
9	building	0.69	0.32
10	hill	0.47	0.53
11	city	0.35	0.42
12	palm	0.33	0.42
		Average precision	**Average recall**
		0.43	0.47

5. Conclusions

In this paper we have presented how to identify a segmentation algorithm based on segmentation error measures. This kind of evaluation is required because a good segmentation algorithm will produce better annotation results. These results are influenced also by the annotation model. For this reason, we have used the CMRM annotation model which was proven to be efficient by several studies. Our experimental results confirmed the efficiency of this annotation model. A dataset containing manually segmented and annotated images is used as a reference point for the annotation process. The SAIAPR TC-12 dataset contains a large-size image collection comprising diverse and realistic images, includes an annotation vocabulary having a hierarchical organization, well defined criteria for the objective segmentation and annotation of images. Further extensions of the annotation process evaluation will include the two models of image retrieval provided by CMRM: Annotation-based Retrieval Model and Direct Retrieval Model.

References

1. D. Martin, C. Fowlkes, D. Tal, J. Malik "A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics " In IEEE (ed.), Proceedings of the Eighth International Conference On Computer Vision (ICCV-01), July 7-14, 2001, Vancouver, British Columbia, Canada, vol. 2, pp. 416–425.
2. Gabriel Mihai, Liana Stanescu, Dumitru Burdescu, Alina Doringa,A Comparison of Three Graph-Based Image Segmentation, IPCV'10 - 14th International Conference on Image Processing, Computer Vision, & Pattern Recognition, July 12-15, USA, (2010)
3. Mori, Y., Takahashi, H., Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words. In: MISRM'99 First Intl. Workshop on Multimedia Intelligent Storage and Retrieval Management (1999)

International Journal of Business Quantitative Economics and Applied Management Research
ISSN: 2349-5677
Volume 1, Issue 4, September 2014

4. Duygulu, P., Barnard, K., de Freitas, N., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In Seventh European Conf. on Computer Vision, pp. 97--112 (2002)
5. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. In: Proceedings of the 26th Intl. ACM SIGIR Conf., pp. 119--126 (2003)
6. "Segmented and Annotated IAPR TC-12 dataset", http://imageclef.org/SIAPRdata
7. Hugo Jair Escalante, Carlos A. Hernández, Jesus A. Gonzalez, A. López-López, Manuel Montes, Eduardo F. Morales, L. Enrique Sucar, Luis Villaseñor and Michael Grubinger, "The segmented and annotated IAPR TC-12 benchmark ",Computer Vision and Image Understanding, Volume 114, Issue 4, April 2010, Pages 419-428
8. IAPR TC-12 Benchmark", http://imageclef.org/photodata
9. Lavrenko V., Manmatha R., Jeon. J.: A model for learning the semantics of pictures. In Proceedings of Advances in Neural Information Processing Systems (NIPS) (2004)
10. Feng S. L., et al. Multiple bernoulli relevance models for image and video annotation. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1242–1245 (2004).
11. Jin, R., Chai, J.Y., Si, L.: Effective automatic image annotation via a coherent language model and active learning. In Proceedings of ACM International Conference on Multimedia (ACM MULTIMEDIA), pp. 892--899 (2004)
12. Burdescu, D., Brezovan, M., Ganea, E., Stanescu, L.: A New Method for Segmentation of Images Represented in a HSV Color Space. In Lecture Notes in Computer Science, 5807, pp. 606-617 (2009)
13. J. R. Smith, S. F. Chang.: "Tools and Techniques for Color Image Retrieval", Symposium on Electronic Imaging. In: Science and Technology - Storage & Retrieval for Image and Video Databases IV, volume 2670, San Jose, CA, February 1996. IS&T/SPIE. (1996)
14. P.F. Felzenszwalb, W.D. Huttenlocher: "Efficient Graph-Based Image Segmentation", Intl. Journal of Computer Vision, 59(2), pp. 167–181 (2004)