Automatic Image Annotation Using Fuzzy Cross-Media Relevance Models

Mohamed Alkaoud, Ibrahim AshShohail, and Mohamed Maher Ben Ismail
King Saud University/Computer Science Department, Riyadh, KSA
Email: {maalkaoud, ibra.sho, maher.benismail}@gmail.com

***Abstract*—In this paper, the authors propose a novel automatic image annotation approach which relies on two main components: (i) identification of homogenous image regions, which share the same semantics using fuzzy clustering algorithm, and (ii) membership-based cross media relevance model to learn the association between keywords and image regions. The proposed fuzzy version of the Cross Media Relevance Model (CMRM) yields promising results. They use standard image collection to compare their approach to the original CMRM. The obtained results show that the proposed approach outperforms the original technique.**

***Index Terms*—image annotation, unsupervised learning, fuzzy logic, image retrieval**

I. Introduction

A lot of information can be conveyed in one picture, and any person, no matter what language he speaks, can understand it. Sometimes, it is even impossible for someone to express himself using words. Images can hold more information than text and have more expressive-power as the old Chinese proverb says: "A picture is worth a thousand words". People's interest and usage of images has exploded during this decade. The growth of the Internet and the popularity of electronic devices equipped with built-in cameras are the main factors behind this phenomenon.

Retrieving an image of particular interest from a dataset has become a challenging task with the increase of the volume of generated images in the age of the Internet. In fact, millions of images are captured and shared on the internet every day. Unlike text, images do not have a structured form. For instance, the word "tiger" is always written as [t, i, g, e, r]. On the other hand, a picture of a tiger can have thousands of different forms. This has led to an interest within the computer science community to try to explore different ways of understanding the semantics of images in order to improve image searching, indexing and retrieval processes. Yet the performance of text-based retrieval approach is limited by the semantic gap. For instance, given an image labeled with the keyword 'apple'. This label can refer to apple the fruit, or Apple the computer company which affects the retrieval accuracy of these systems. Manual annotation can be an alternative approach. However it is a labor-intensive and time consuming.

Lately, Content Based Image retrieval (CBIR) emerged as an alternative to text-based approach. CBIR consists of retrieving images based on the relevance of their visual content. This approach too suffers from semantic drawback; think about an image of a red apple, it can be described as a round red object. Now, think about how an image of red ball would be described, a round red object! This semantic gap problem remains a major limitation for CBIR approach [1].

Combining text-based and content-based image retrieval approaches would exploit the best of both worlds. However, as stated before, most images are not annotated and manual annotation is expensive, undesirable, and time consuming. It seems we have reached a dead end. Fortunately, computer scientists, love automation! Let us teach the computer to annotate these images automatically. That is what we mean by 'Automatic Image Annotation'. Automatic Image Annotation can be defined as the process of automatically assigning captions (annotations) to images [2].

The most popular image annotation techniques proposed in the literature are: the Co-occurrence Model [3], Translation Model [4], and the Cross-Media Relevance Model [5]. Mori et al [3] proposed the Co-occurrence approach. It tackles the automatic image annotation problem as the problem of assigning probabilities to each word and image region, then annotating the image with words with the highest probability. Duygulu et al [4] proposed the Translation Model. It solves the image annotation problem by assuming that the annotation process can be modeled as a translation problem. That is, translating from a vocabulary of blobs to a vocabulary of words. CMRM [5] was proposed as extension of the previous models, and succeeded to outperform them.

In this paper, the authors propose a novel automatic image annotation approach which relies on two main components: (i) identification of homogenous image regions, which share the same semantics using fuzzy clustering algorithm, and (ii) membership-based cross media relevance model to learn the association between keywords and image regions. Fig. 1 shows an overview of the system's different components. The offline part builds

Manuscript received January 11, 2014; revised May 12, 2014.

doi: 10.12720/joig.2.1.59-63

the model that is used by the online part. Initially, the offline part is provided with a set of manually annotated images *T*. Each image in *T* is segmented into regions using a segmentation algorithm. Then, the system extracts features from these regions. Given the vector of the regions correspondence to images' regions, the system uses a clustering algorithm to cluster all the vectors into homogenous categories/blobs. Then, it builds a model by estimating the probability that a given word w will appear in each blob *b* and the probability that each word *w* occurs in given blob *b*. The offline part role ends here. The online part uses the model generated by the offline part to assign labels to the unknown image. First, the system segments the image into regions, and then it extracts visual features from it, resulting in a group of vectors. Then, it finds the closest blob to each vector (remember that each vector corresponds to a region). Now, the result is a group of blobs $\{b_1, b_2, b_3, \dots b_m\}$. For each blob *b*, it finds the probability of each word occurring in *b* (using the model generated by the offline part). Finally, it returns the top N words with the highest probabilities.

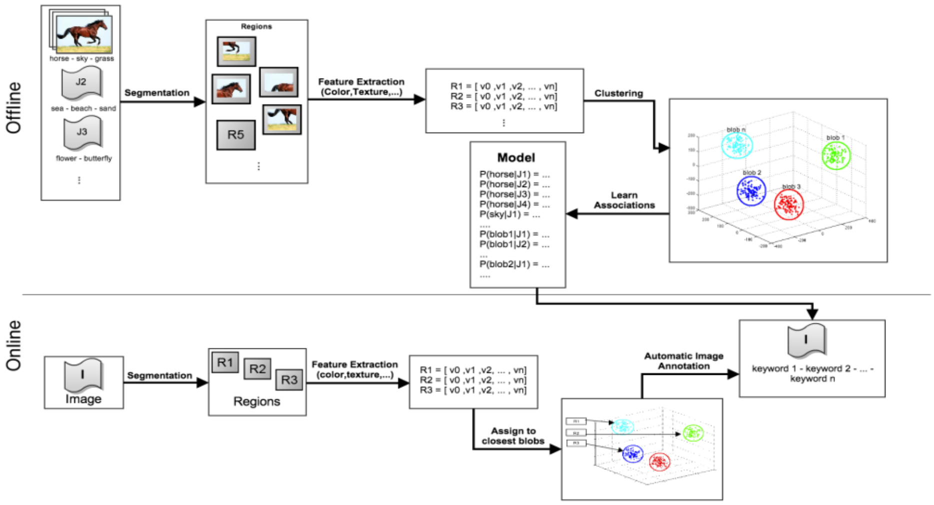


Figure 1. An overview of the different components in the system

II. Cross-Media Relevance Model

Jeon *et al.* [5] proposed the Cross-Media Relevance Model (CMRM) as an improvement to previous image annotation models [3], [4]. Visterms, which are the visual components of an image, are obtained by segmentation, and then grouped into blobs using clustering algorithm [6]. Assuming that images can be described using a small vocabulary of blobs. Each image is represented as a set of words $\{w_1, w_2, w_3, \dots w_m\}$ and blobs $\{b_1, b_2, b_3, \dots b_m\}$. The relevance model gives us the probability of generating a word *w* given the blobs in an image. The authors in [5] assume that the keywords and the blobs follow the distributions *P(w|I) and P(b|I`)*, respectively. *P(w|I)* allows the prediction of the best word *w* to annotate image *I*. The authors approximate *P(w|I)* using the conditional probability of observing *w* given that $b_1 \dots b_m$ are observed as a random sample from the same distribution:

$$P(\mathrm{w} \mid \mathrm{I}) \approx P(\mathrm{w} \mid \mathrm{b}_1 \dots \mathrm{b}_m) \tag{1}$$

$$P(\mathrm{w} \mid \mathrm{b}_1 \dots \mathrm{b}_m) = \sum_{J \in T} P(\mathrm{J})\, \mathrm{P}(\mathrm{w}, \mathrm{b}_1, \dots, b_m \mid J) \tag{2}$$

where *T* is a training set of manually annotated images and *J* is an image in the training set *T*. Assuming that the events *w* and $b_1 \dots b_m$ are conditionally independent, equation (2) becomes:

$$P(\mathrm{w} \mid \mathrm{b}_1 \dots \mathrm{b}_m) = \sum_{J \in T} P(\mathrm{J})\, \mathrm{P}(w \mid J) \prod_{i}^{m} P(\mathrm{b}_i \mid \mathrm{J}) \tag{3}$$

To estimate the maximum likelihood $P_{MLE}(\bullet|J)$, they count the occurrences of the term in the representation of *J* and normalize it by dividing by the total size of the representation. However, for terms that do not actually appear in image *J* we have P_{MLE} $(\bullet|J) = 0$, which means that associating the term with *J* entirely is impossible. This should be avoided as it means that the estimated probability distribution is unreliable. To overcome this, they consider some probabilities from words that do occur and distribute them among those which do not. This solution is formulated by interpolating the maximum likelihood estimates with the general relative frequency computed over the entire collection *T*. Thus, *P(b|J)* and *P(w|J)* for each training image *J* are estimated as follows:

$$P(\mathrm{w} \mid \mathrm{J}) = (1 - \alpha_j) \frac{\#(\mathrm{w}, \mathrm{J})}{|J|} + \alpha_j \frac{\#(\mathrm{w}, T)}{|T|} \tag{4}$$

$$P(\mathrm{b} \mid \mathrm{J}) = (1 - \beta_j) \frac{\#(\mathrm{b}, \mathrm{J})}{|J|} + \beta_j \frac{\#(\mathrm{b}, T)}{|T|} \tag{5}$$

where *#(w,J)* is the number of times the word *w* occurs in the image *J* (usually 0 or 1). *#(w,T)* is the total number of times the word *w* occurs in the captions of whole training set *T*. |*J*| equals the total numbers of blobs and captions in image *J*. |*T*| denotes the total size of the set. The parameters α_j and β_j are for interpolating between the two estimations.

III. Fuzzy C-Means Clustering

The major difference between traditional (crisp) and fuzzy clustering is that in traditional clustering every element can belong only to one cluster, whereas in fuzzy clustering each element belongs to all clusters with varying degrees-of-membership; hence the name fuzzy clustering [7]. Fuzzy c-means (FCM) [8] algorithm minimizes the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}{}^{m} \parallel \mathrm{x_i} - \mathrm{c_j} \parallel^2 \tag{6}$$

where C is the number of clusters, N is the number of observations to be categorized, x_i is the i^{th} observation, c_j is the j^{th} centroid and u_{ij} is the degree of membership of observation x_i in cluster c_j ($0 \le u_{ij} \le 1$). m is a real number that controls the degree of fuzziness ($m \ge 1.0$). As the value of m increases, the influence of the degree of memberships becomes larger. The membership degree u_{ij} and the centroids c_j are iteratively updated using:

$$u_{ij} = 1 / \sum_{k=1}^{C} \left(\frac{\parallel \mathrm{x_i} - \mathrm{c}_j \parallel}{\parallel \mathrm{x_i} - \mathrm{c_k} \parallel} \right)^{\frac{2}{m-1}} \tag{7}$$

$$c_j = \sum_{i=1}^{N} u_{ij}{}^{m} x_i \,/\, \sum_{i=1}^{N} u_{ij}{}^{m} \tag{8}$$

FCM starts by selecting C centroids randomly, then iterating over: the calculation of u_{ij} using (7) and recalculating the centroids using (8) until convergence.

IV. Fuzzy Cross-Media Relevance Model

We propose a new version of the Cross-Media Relevance Model called Fuzzy Cross-Media Relevance Model (FCMRM). The original CMRM [5] used K-means [9] clustering algorithm which generates crisp membership values. However, since we expect blobs to be overlapping, we propose to use FCM algorithm which reflects better the structure of the blobs in the feature space. Also, we include the membership functions generated by FCM in the estimation of *P(b|J)*. More specifically, we propose a membership-based version where we replace *#(b,J)* in (5) by *f(b,J)* which is computed as follows:

$$f\,(\mathrm{b,J}) = \sum_{r \in J} u_{rb} \tag{9}$$

With r any region in image J, b is the blob whose membership in being computing, u_{rb} is the degree-of-membership between region r and blob b. Thus, $P(b|J)$ becomes:

$$P(b \mid J) = (1-\beta_j)\frac{f\,(b,\mathrm{J})}{|\mathrm{J}|} + \beta_j \frac{\#(b,\mathrm{T})}{|\mathrm{T}|} \tag{10}$$

V. Experimental Results

The authors conducted experiments on a subset of the manually annotated Corel Data Set that Duygulu *et al.* used in their work [5]. The subset used contains 3690 training images and 409 testing images. Each image is labeled with 1 to 5 keywords. The total vocabulary size is 374 words.

The authors divided each image to a 6*4 grid, generating 24 segments for each one. After generating regions, they describe each region using one feature vector. Namely, they used the standard deviation, and skewness of the RGB values, and standard deviation, skewness and average of the CIE-Lab values as low-level features. Let p be a pixel in image I, the mean, standard deviation, and skewness are calculated as follows:

$$E = \frac{1}{N}\sum_{j=1}^{N} p_i \tag{11}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(p_i - E)^2} \tag{12}$$

$$s = \sqrt[3]{\frac{1}{N}\sum_{j=1}^{N}(\mathrm{p_i} - \mathrm{E})^3} \tag{13}$$

Then, they ran FCM algorithm with 300 centroids on the obtained features. They tested different values of α and β and concluded that setting α to 0.1 and β to 0.7 gives the best results. Notice that the metrics used for evaluating the performance of the system are precision, recall and f-score [10].

Before comparing the two annotation methods, they noticed the following problem; the centroids obtained for FCMRM (using FCM) and CMRM (using K-means) were different. If this discrepancy is kept, the comparison of the automatic image annotation will not be objective. In other words, if one of the methods yields better annotation performance, should this be attributed to the clustering performance or to the annotation approach? Therefore, they decided to use the same centroids for both CMRM and FCMRM in order to ensure that the clustering does not affect the comparison.

To compare the performance of both approaches, annotate a set of ~400 test images which were not used during the training phase using the proposed approach and the method in [5].

In order to compute recall and precision, each test image is annotated using the top five keywords using FCMRM and CMRM. The obtained performance measure values show that FCMRM outperforms the traditional CMRM. The mean precision of FCMRM is 18.8%, while the CMRM is 16.38%. The mean recall of the FCMRM is 26.96% against 24.08% for the CMRM. F-score measure is calculated by combining both the recall and the precision. In terms of f-score, FCMRM achieves 22.15% against 19.50% for CMRM. Table I shows some sample images automatically labeled using FCMRM. In the horse image, the top five generated keywords (tree, horses, foals, mare, and garden) are good annotations. This means that the clustering algorithm discovered the corresponding category properly and that FCMRM learned the blob-word association accurately. In Fig. 2, word accuracies are calculated for both FCMRM and CMRM. For almost all words FCMRM gives better accuracy e.g. the proposed method overcomes CMRM by over 200% for the accuracy of the word: 'plane'. Also, notice that for many words, CMRM results in an accuracy of 0. The large difference between FCMRM and CMRM in terms of word accuracies can be attributed to the fact that CMRM often generates one or more of the most frequent words in the data set (e.g. 'sky', 'water', 'tree') for a given image; while this approach will help in improving the precision and recall, it will impact the accuracy in a negative way since the accuracy is computed as the ratio of the number of all correct annotations to the number of all annotations.

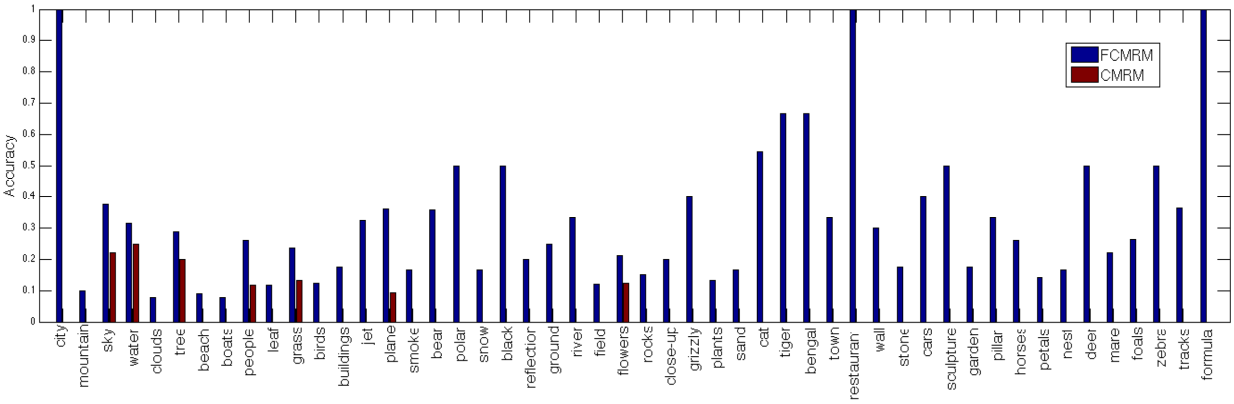


Figure 2. Word accuracies for both FCMRM and CMRM

While the proposed model improves over the previous models, it is still not perfect. Table II shows few samples where the model fails to assign suitable captions. For the second image in Table II the captions: (people, waves, Oahu, water, tree) were generated. It is clear that these are not really good annotations. This can attributed either to the fact that the clustering is not perfect and that the clustering was not efficient for the blobs of the image, or that there were not enough training samples for FCMRM to learn the association. The Table III shows a comparison between FCMRM and the real (manual) annotations of some images.

VI. Conclusion

The authors proposed a new annotation approach that integrates fuzzy logic to cross-media relevance models [5]. The obtained results show that the proposed approach outperforms the state-of-the-art method. This can be attributed to the fact that their approach exploits the fuzzy membership functions generated by the clustering step, and learns efficiently the association between image regions and labeling keywords.

TABLE I. Top Five Keywords Generated by FCMRM

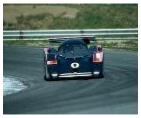
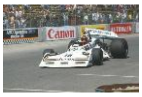
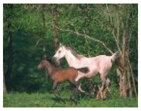
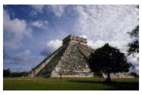
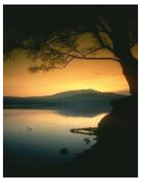
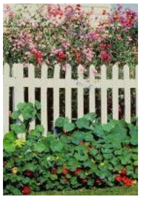
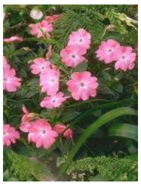
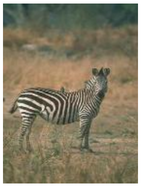
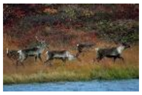
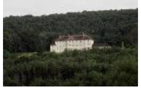
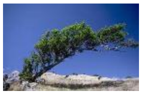
Image	Top five annotations	Image	Top five annotations
	cars, tracks, turn, prototype, water		cars, tracks, water, wall, formula
	tree, horses, foals, mare, garden		water, buildings, skyline, sky, tree
	people, sky, water, sunset, beach		tree, flowers, grass, bush, leaf
	flowers, petals, grass, water, tree		tree, plane, grass, zebra, water

TABLE II. Inaccurate Annotations Obtained Using FCMRM

Image			
Top five annotations	tree, field, hills, Kauai, water	people, waves, Oahu, water, tree	birds, flight, plane, jet, sky

Future works to improve the results may consist in using more sophisticated FCM based clustering algorithms. For instance, assigning different relevance weights to the different low-level features could be a good alternative when clustering the image region collection [11]. Regarding the number of clusters, their optimal number could be found automatically using a competitive agglomeration approach [12].

TABLE III. FCMRM VS. MANUAL ANNOTATIONS

Image	FCMRM	Manual annotation
	bear, snow, polar ice	bear, polar, snow, tundra
	sky, plane, jet, f-16	jet, plane, sky, smoke
	water, bear, black, reflection	bear, black, reflection, water
	sky, tree, sand, island	beach, palm, people, tree
	grass, cat, tiger, bengal	bengal, cat, grass, tiger

ACKNOWLEDGMENTS

The authors would like thank K. Barnard and P. Duygulu [4] for making their dataset available for other researchers to use.

REFERENCES

[1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, 1st Ed. Cambridge: Cambridge University Press, 2008, pp. 155-156.

[2] R. Datta, D. Joshi, J. Li and J. Z. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” *ACM Computing Surveys*, vol. 40, pp. 1-60, 2008.

[3] Y. Mori, H. Takahashi, and R. Oka, “Image-to-word transformation based on dividing and vector quantizing images with words,” in Proc. *International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[4] P. Duygulu, K. Barnard, N. Freitas, and D. Forsyth, “Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary,” *ECCV*, 2002.

[5] J. Jeon, V. Lavrenko, and R. Manmatha, “Automatic image annotation and retrieval using cross-media relevance models,” in *Proc. 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, New York, 2003.

[6] L. Ballesteros and D. Petkova, “A clustered retrieval approach for categorizing and annotating images,” *Lecture Notes in Computer Science*. vol. 4022, pp. 662-672, 2006.

[7] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd Ed. Waltham, MA: Morgan Kaufmann, 2012.

[8] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, 1981.

[9] S. Lloyd, “Least squares quantization in PCM,” *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129-137, March 1982.

[10] D. L. Olson and D. Delen, *Advanced Data Mining Techniques*, 1st Ed. Springer, 2008.

[11] H. Frigui and O. Nasraoui, “Simultaneous clustering and attribute discrimination,” in *Proc. 9th IEEE Int. Conf. on Fuzzy Systems*, 2000.

[12] H. Frigui and R. Krishnapuram, “Clustering by competitive agglomeration,” *Pattern Recognition*, vol. 30, no. 7, pp. 1109-1119, 1997

Doctor Mohamed Maher Ben Ismail is assistant professor at the computer science department of the College of Computer and Information Sciences at King Saud University. He received his PhD. degree in Computer Science from the University of Louisville in 2011. His research interests include Pattern Recognition, Machine Learning, Data Mining and Image Processing.