International Journal of Research in Computer Science
eISSN 2249-8265 Volume 5 Issue 2 (2015) pp. 1-6
www.ijorcs.org, A Unit of White Globe Publications

MULTIMODAL IMAGE RETRIEVAL BASED ON QUERY WEIGHTING

Ebrahim Naderi[1], Elham Nikookar[2], Ali Tayebi Babookani[3]

[1] *ICT Department, National Iranian Gas Company, Ahvaz, IRAN*
Email: naderi@nigc-khgc.ir
[2]*Computer Department, Faculty of Engineering, Shahid Chamran University, Ahvaz, IRAN*
Email: e.nikookar@scu.ac.ir
[3] *ICT Department, National Iranian Gas Company, Ahvaz, IRAN*
Email: tayebi@nigc-khgc.ir

Abstract: *Image retrieval methods try to increase the performance of retrieval by adjusting weights of relevant and non-relevant images. Adjusting weights of queries instead of images is studied in this paper to see whether it increases or decreases performance of retrieval system. Evaluation measure of the method is calculated for each query instead of each sample. In proposed method, textual and visual features of images are used to retrieve ranked results. To train visual module, query weighting and updating weights in each iteration are used to concentrate on the queries which did not provide acceptable results and also cause performance drop of final retrieval. Compared with other previous researches, the results of applying the proposed method to IAPR TC-12 dataset indicate high efficiency of proposed method. Image retrieval methods like this can be used in a lot of applications such as gasometer digit recognition using gasometer digits image etc.*

Keywords: *content-based image retrieval, learning to rank, multimodal retrieval, query weighting, ranking model, resampling, text-based image retrieval.*

I. INTRODUCTION

Nowadays, a lot of researches are done in the field of information retrieval. One of the most important reasons is that World Wide Web extremely uses XML documents because of richness of them. XML documents include not only textual information but also other media informations such as images, videos and sounds. It is clear that in the meantime and despite all the works done on images, they are still playing their role as the most important source of information in human life. Generally, there are two frameworks for image retrieval; text-based and content-based. In text-based systems, images are annotated manually or using annotation algorithms and then, annotations are used to perform the retrieval process. Running text-based image retrieval alone has its own drawbacks specially the need to annotate images. Another disadvantage of this approach is due to the different views of people to a certain image that makes the image annotations in different forms. In content-based image retrieval, a number of visual features - including color, texture, etc. – are extracted from the images and then, images are compared to each other based on their visual features. As there is no access to image annotations in content-based image retrieval, extraction of semantic features is impossible and a semantic gap appears between what the user is looking for and what visual features express. Therefore, since the strength of each method is coverable by the other one, the two methods complement each other and can be combined together to improve retrieval results.

For example, in [1] an image-text multimodal neural language model is presented in which images are retrieved given complex sentence queries. In [2] a feature-independent context estimation method is applied which automatically annotates images and then searches the images. In [3], a combination method is proposed in which content-based image retrieval module creates the vector of image histogram features with 43376 bins and uses it to compare images and evaluate the performance of the system. A method is proposed in [4] in which each image is divided into regions and each region is replaced with a concept like water, rock and etc. and then these concepts are used to compare images. A method based on clustering is proposed in [5] in which images are clustered based on their feature vectors and the result is used to direct the retrieval to increase the probability of finding relative images.

In this paper, we propose an effective method which tries to improve retrieval results using multimodal image retrieval and ranking model based on query weighting. Ranking model is a function

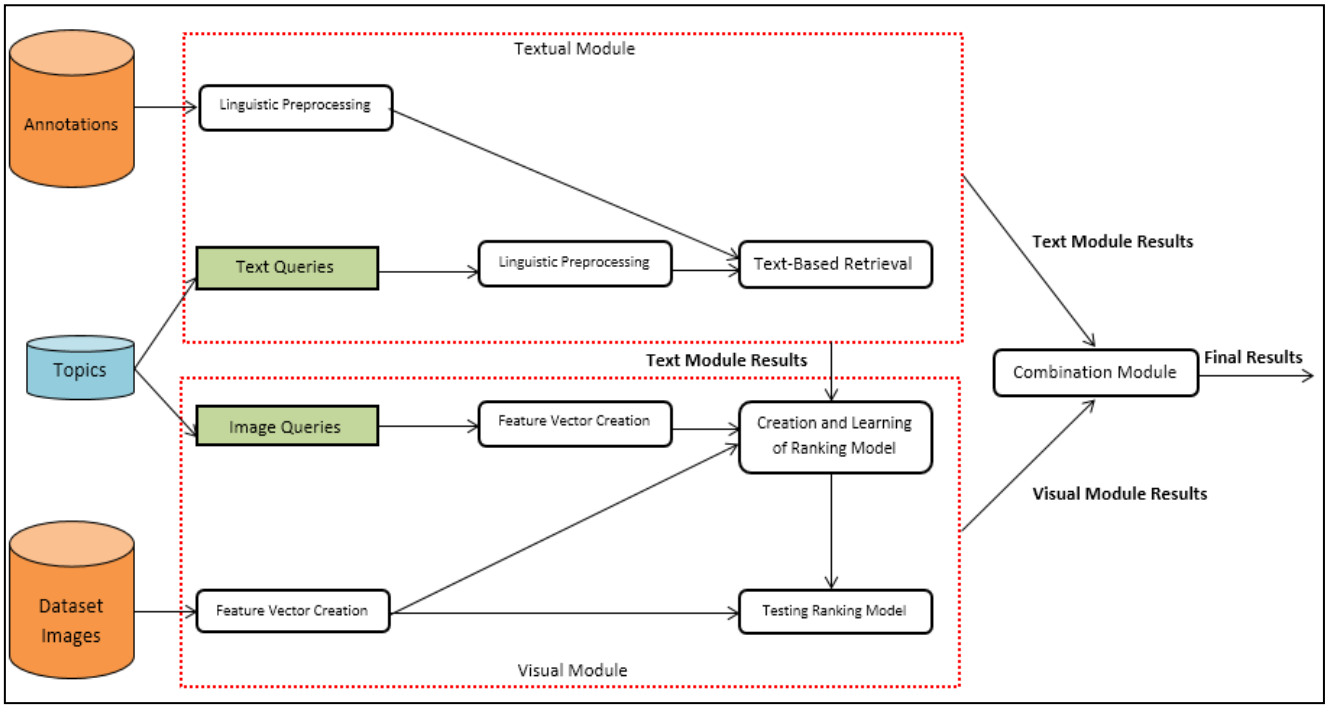

Figure 1: Proposed method framework

which receives query and dataset images as input and sends the ranked results to the output. Loss function is minimized exactly based on desired evaluation measure using boosting [6] method. The rest of paper is organized as section II in which the proposed method is explained, section III in which experimental results of applying proposed method to the dataset are depicted and section IV in which a conclusion is expressed on important aspects of the method and results.

II. METHOD

The proposed method, as shown in fig. 1, has three main modules: Textual module, Visual module and Combination module.

In textual module, linguistic preprocessing is applied on image annotations and preprocessed annotations are sent to text-based retrieval machine. Textual queries of topics are used to apply text-based retrieval and to evaluate performance of text module. Then, a copy of textual module result is sent to combination module and another copy is used to train ranking model in visual module.

In visual module, dataset images are processed to extract visual attributes and to create visual characteristics feature vector and this process is also applied on visual queries images. Training of ranking model is done in visual module using results of text module and dataset and query images. After evaluation of results of visual module, results are sent to combination module to combine with textual module results and create final results of system.

A. Textual Module

Lemur [7] is used as text retrieval engine through which images are retrieved according to textual part of queries and annotation of dataset images. Firstly, in language preprocessing, stop-words such as *about* and *with* were excluded. Next, stemming applied on textual data in which different forms of a word, sit next to that word. For example, if *swim* is in an annotation, *swimming* and *swimmer* will be added to that annotation. Then, results of text-based retrieval are sorted descending, based on the similarity to query. A copy of textual module result is sent to visual module and another copy is kept to be combined with results of visual module to produce final results.

B. Visual Module

In visual module, firstly, feature vector is formed by extracting visual features of images. It is necessary to note that in this section, feature extraction means creating feature vector for each image and this process is different from the feature extraction algorithms which reduce the size of feature vector, such as PCA and etc.

Features that have been used in the proposed method are low-level features that in the later stages, combine with each other and form high-level features. By definition, high-level features are obtained from combination of some low-level features or applying a

function on them. Low-level features which we used in this paper include:

(1) *Color histogram:* each dataset and query image, is divided to nine equal blocks. As a result, for each image, nine RGB histograms are formed. The reason of choosing RGB color space than other color spaces is that although it requires less processing, but produces similar results [8]. Each histogram has 512 bins that the number of bins is obtained by considering color cube so that each channel consists of eight colors; 8*8*8=512. Choosing eight for colors of each channel is because of doing a compromise between storage space usages and processing time to have acceptable results. Finally, general histogram of each image is formed with 4608 (=9*512) bins.

(2) *Local Binary Pattern*: LBP is a texture feature which has been used in some systems [9]. As shown in fig. 2, to calculate LPB of each pixel, its intensity value is considered as the reference. Then, intensity values of each of eight neighboring pixels are compared to reference, and pixels with higher (or equal) and lower intensity values are replaced by 1 and 0, respectively. Finally, an eight bit string for each pixel is formed as LPB of pixel. As each 8 bit string can have 256 different combinations (2^8=256), a texture histogram with 256 bins is formed for each image.

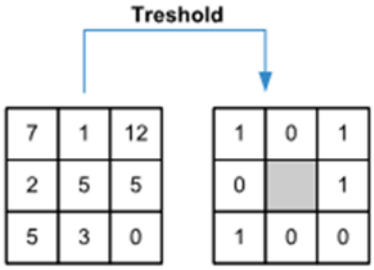

Figure 2: Calculating LBP for a pixel

Considering color and texture features, a 4864 (256+4608) bin histogram is formed for each image. Then, a normalized histogram is formed in which for each image, summation of all feature values equals to 1.

a) Distance measure

Distance measure is used to sort retrieval result of a query. Since histogram is a frequency distribution, a statistical measure is needed to calculate distance between two image histograms. Assuming histogram of query image and dataset image as H_q and H_d respectively, distance measure (Relative Bin Deviation) is calculated based on (1):

$$D_{RBD}(H_q,H_d)=\sum_{n=1}^{N}\frac{\sqrt{(H_q-H_d)^2}}{\frac{1}{2}(\sqrt{H_q}+\sqrt{H_d})} \quad (1)$$

where N is number of histogram bins.

b) Training and Testing Algorithm

The proposed method is an iterative method based on training and testing. In training section, ranking function is formed using half of queries and in testing section, final performance measure of system is calculated using remaining half of queries. Performance measure used in this paper is Precision at 20 (P@20). This measure specifies that how many images are relevant to the query in first twenty retrieved results. In each iteration of training process, the goal is maximization of performance measure.

c) Training section

At first, initial weighting of all training queries is done evenly. It means all training queries get equal weights. Then, P@20 is calculated for all training queries. To reduce calculation time, features are divided into 100 feature subsets. Therefore, 49 feature subsets are formed. At the end of training section, 10 feature subsets are selected from which a feature vector with 1000 or 964 features is created.

In the first iteration of training algorithm, performance measure is calculated for all "training query, feature set" pairs from which a matrix is created in which each row corresponds a query and each column corresponds a feature set. Then, in each iteration of training algorithm, weights of queries with worse performance measures increases more than better queries. It is like the process of AdaBoost algorithm:

$$W_{t+1}^{i}=W_{t}^{i}+\frac{1-PM_i}{1+\sum_{Q}^{i=1}PM_i} \quad (2)$$

where W_{t+1}^{i}, W_{t}^{i}, PM_i and Q are the weight of *i*th query in *(t+1)*th iteration, weight of *i*th query in *t*th iteration, performance measure of *i*th query and total number of queries respectively. Algorithm is continued until stop condition happens. Stop condition occurs when performance measure starts to decrease.

d) Testing Section

In testing section, remaining half of queries is sent to ranking function. Here, retrieval is done based on features that were selected in training section. At the end, final results of visual module are sent to combination module.

C. Combination Module

In combination module, results of visual and textual modules are combined to improve results of single modules and form final results of system. For each testing query of visual module, first 20 images are checked and then, added to results of textual module if they are relevant to query and not included in results of text-based retrieval. At the end of this section, final P@20 of whole system is calculated.

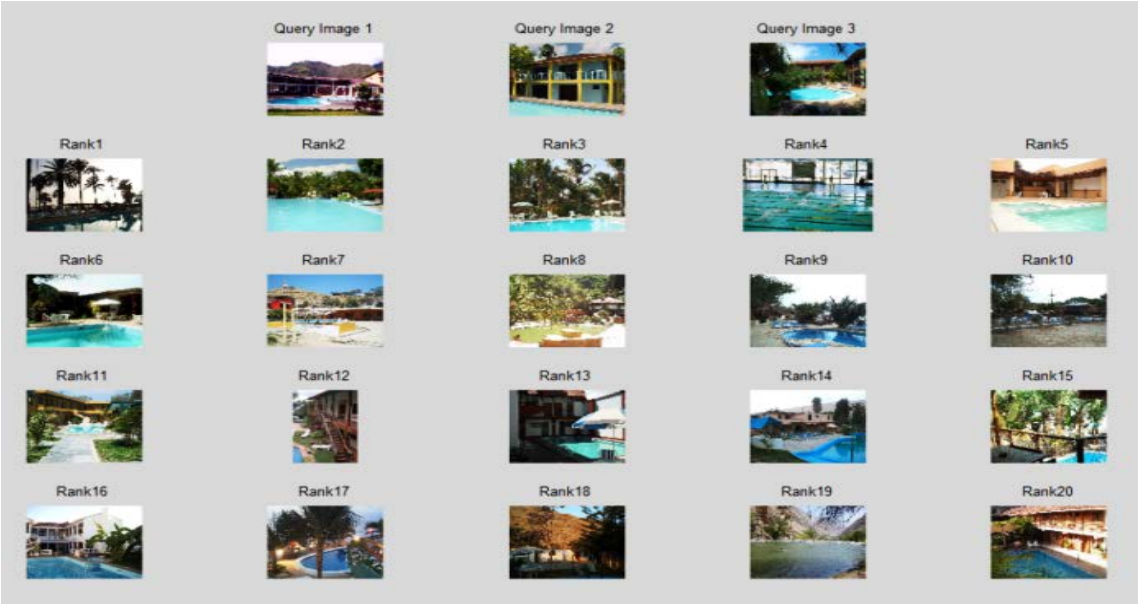

Figure 3: Result of applying method on one of queries. Text-query is "accommodation with swimming pool"

III. EXPERIMENTAL RESULTS

In this study, IAPR TC-12[10] dataset including 20000 images is used. There are 60 queries in the dataset that each of them includes 3 images and an annotation. Images and textual part of queries are used as image and text queries respectively. Totally, 180 image queries and 60 text queries are placed in IAPR TC-12 dataset. Recently, many researches ([11], [12]) are done on IAPR TC-12 dataset.

Results of textual module of proposed method, along with results of other text-based retrieval systems which applied on IAPR TC-12 is shown in table I.

TABLE I. *COMPARE TEXTUAL MODULE WITH OTHER METHODS*

Method	P@20
Textual Module of proposed Method	0.3450
Villena-Roman et. al [13]	0.2990
Inoue et. al [14]	0.2409
Sarin et. al [8]	0.2700
Demerdash et. al [15]	0.3449

A copy of text module results is sent to visual module. Results of visual module of proposed method is compared to dimensionality reduction algorithms in table II and compared in Fig. 4.

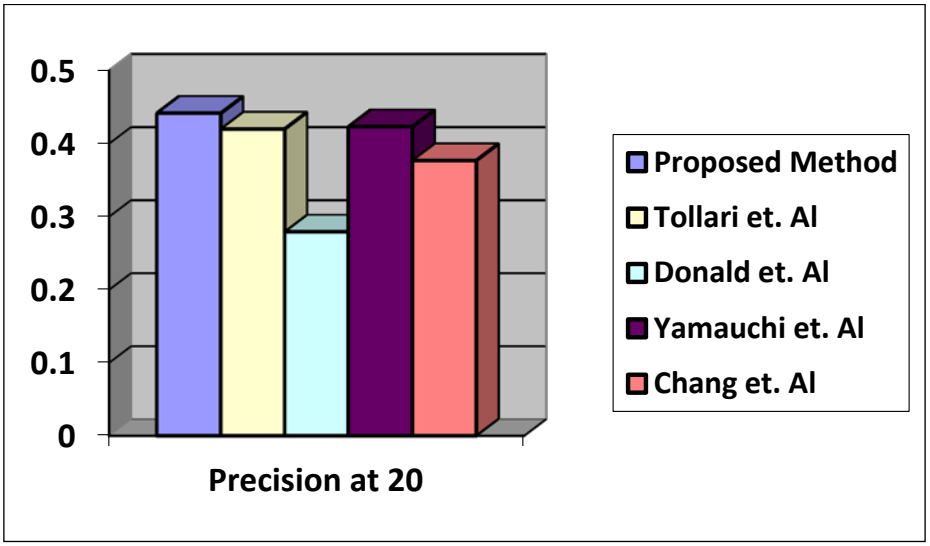

Figure 4: Comparison of results of applying method on one of queries. Text-query is "accommodation with swimming pool"

TABLE II. *COMPARE VISUAL MODULE WITH DR METHODS*

Method	P@20
Visual module of proposed method	0.2250
PCA (1000 Features)	0.1750
Forward Selection	0.1783
No dimensionality reduction	0.1566

Final result of proposed method, along with other methods which worked on IAPR TC-12 is shown in table III.

TABLE III. COMPARE FINAL RESULTS

Method	P@20
Proposed Method	0.4416
Demerdash et. al [15]	0.3744
Tollari et. al [16]	0.4200
Donald et. al [17]	0.2792
Yamauchi et. al [18]	0.4231
Chang et. al [19]	0.3769

Based on result comparison, results of proposed method shows better results than the other applied methods. Final result of proposed method for a query is illustrated in fig. 3.

Another point that should be considered is the size of feature vector in different methods. If two methods produce similar results, the one with less feature vector size is better, because less number of features leads to less calculation time and better time complexity. Therefore, the proposed method could be compared to other methods in terms of number of features. For proposed method uses only 1000 or 964 features.

IV. CONCLUSION

In this study we presented a multimodal model for combining results of text-based and content-based image retrieval in which inputs of textual and visual modules are queries instead of image samples whether textual annotation or image sample. Looking to queries as samples and trying to adjust weight of queries to achieve better retrieval result cause the system to produce better results than other state of the art methods. In the proposed method, ranking function served as a high level feature in comparing images and calculating similarity of dataset and query images. Generally, our algorithm tried to reduce the semantic gap and increase effectiveness of retrieval which, according to results it was almost successful.

V. REFERENCES

[1]. R. Kiros, R. Salakhutdinov, R. Zemel, Multimodal neural language models, Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014.

[2]. A. Tariq, H. Foroosh, Feature-Independent Context Estimation for Automatic Image Annotation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[3]. O. E. Demerdash, L. Kosseim and S. Bergler, CLaC at ImageCLEFphoto 2008, ImageCLEF 2008 Reports, 2008.

[4]. Julia Vogel, Semantic Modeling of Natural Scenes for Content-Based Image Retrieval, International Journal of Computer Vision, Springer, Vol 72, pp. 133-157, 2007.

[5]. D. Frossyniotis, A. Likas, A. Stafylopatis, A clustering method based on boosting, Pattern Reconition Letters, Vol. 25, pp. 641-654, Elsevier, 2004.

[6]. Y. Freund, R. Schapire, A short introduction to boosting, Journal of Japanese Society for Artificial Intelligence, vol. 14(5), pp. 771-780, September, 1999.

[7]. http://www.lemurproject.org/

[8]. Sarin, W. Kameyama, Targeting Diversity in Photographic Retrieval Task with Commonsense Knowledge, Clef- Campaign, Kameyama, 2008.

[9]. Andrew P. Berman and Linda G. Shapiro. A flexible image database system for content-based retrieval. Computer Vision and Image Understanding, Vol. 75, 1999.

[10]. M. Grubinger, P. Clough, H. Müller, T. Deselaers, The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems, Proceedings of International Workshop OntoImage'2006 Language Resources for Content-Based Image Retrieval, pp. 13-23, 2006.

[11]. A. Gupta, Y. Verma, C. Jawahar, Choosing Linguistics over Vision to Describe Images, AAAI, 2012.

[12]. Y. Verma, C, Jawahar, Im2Text and Text2Im: Associating Images and Texts for Cross-Modal Retrieval, Proceedings of the British Machine Vision Conference. BMVA Press. 2014.

[13]. J. Villena-Román, S. Lana-Serrano, J. González-Cristóbal, MIRACLE-GSI at ImageCLEFphoto 2008: Experiments on Semantic and Statistical Topic Expansion, Working Note for the ImageCLEFphoto task, 2008.

[14]. M. Inoue,P. Grover, Effects of Visual Concept-Based Post-retrieval Clustering in ImageCLEFphoto 2008, Working Note for the ImageCLEFphoto task, Denmark, 2008.

[15]. O. Demerdash, L. Kosseim,S. Bergler, CLaC at ImageCLEFPhoto 2008, Concordia University, CLEF 2008, 2008.

[16]. S. Tollari, M. Detyniecki, M. Ferecatu, H. Glotin, P. Mulhem, M. Amini, A. Fakeri-Tabrizi, P. Gallinari, H Sahbi, Z Zhao, Consortium AVEIR at ImageCLEFphoto 2008: on the Fusion of Runs, Working Note for the ImageCLEFphoto 2008 task, 2008.

[17]. K. Donald, G. Jones, Dublin City University at CLEF 2006: Experiments for the ImageCLEF Photo Collection Standard Ad Hoc Task, Lecture Notes in Computer Science, Springer, Vol. 4730/2007, pp. 633-637, 2007.

[18]. K. Yamauchi, T. Nomura, K. Usui, Y. Kamoi, T. Takagi, Meiji University at ImageCLEF2008 Photo Retrieval Task: Evaluation of Image Retrieval Methods Integrating Different Media, Lecture Notes in Computer Science, Springer, Vol. 5706/2009, pp. 551-559, 2009.

[19]. Y. Chang, H. Chen, Approaches of Using a Word-Image Ontology and an Annotated Image Corpus as Intermedia for Cross-Language Image Retrieval, Lecture Notes in Computer Science, Springer, Vol. 4730/2007, pp. 625-632, 2007.

How to cite

Ebrahim Naderi, Elham Nikookar, Ali Tayebi Babookani, "Multimodal Image Retrieval Based on Query Weighting". *International Journal of Research in Computer Science*, 5 (2): pp. 1-6, December 2015.