ISSN:0975-9646

Bridging the Semantic Gap

Tanveer J. Siddiqui

Department of Electronics & Communication
University of Allahabad

Abstract— **Content-based image retrieval systems were introduced as an alternative to avoid the need of manual tagging in traditional keyword-based image retrieval systems. However, the representation of image using visual features only involves a loss of information which is referred to as semantic gap. A number of techniques have been proposed to deal with 'semantic gap'. This paper reviews existing approaches to handle the well-known 'semantic gap' problem in image retrieval systems with a particular focus to approaches based on text and image fusion.**

Keywords— **Image retrieval, Semantic Image Retrieval, CBIR, Text and Image fusion.**

I. INTRODUCTION

Image retrieval has always been an active area of research. Most popular web search engines are either based on keyword searching in the surrounding text or content-based. A keyword-based image retrieval (KBIR) system requires correct keyword description about the images, which is not always available in real situation and requires human annotator. Some image retrieval systems solve this problem using manual annotation of images with keywords. KBIR systems can accurately identify relevant images. They are also efficient in retrieving relevant images because it can be formulated as a document retrieval problem and therefore can be efficiently implemented using the inverted index technique. However, the performance of KBIR is highly dependent on the availability and quality of manual tags. Manual tagging requires too much time and are expensive to implement. As pointed out in [30], in order to minimize effort many users tend to describe the visual content of images by general, ambiguous, and sometimes inappropriate tags leading to noisy and incomplete tag. One way to handle this problem is to apply automatic image annotation techniques [28] [31] [6] to predict tags based on visual content of the image. Most image annotation algorithms are casted as a classification problem or as a machine-learning problem. Both the approaches require a large number of well-annotated images. An alternative to KBIR is to use content-based approach.

Content-Based Image Retrieval (CBIR) systems extract image features like color, texture, object shape, etc. and utilize them in the retrieval process. This approach however doesn't seem natural. People are not familiar to querying images based on low-level visual features. Instead, they focus on higher level entities and ignore small differences. They would like to query images using textual description. For example, they can phrase a query "Find images of tiger" or using a combination of query and textual keywords. In order to handle such queries a CBIR system requires semantic information which is lost when image is represented using visual features. This loss of information is referred to as semantic gap. Some systems exploit domain specific information from knowledge bases (KB) and ontologies to extract of meaning from images. Others use supervised or unsupervised learning techniques to associate low-level features with high-level concept. The state-of-the-art techniques for reducing semantic gap using high-level semantics as identified by Liu *et al.* [14] are as follows:

1. Using object ontology to define high-level concepts [18].
2. Automatic annotation techniques which uses supervised or unsupervised learning methods to associate low-level features with query concepts [4] [25] [10] [12] [15].
3. Introducing relevance feedback for continuous learning of users' intention [24][29] [36].
4. Generating semantic template to support high-level image retrieval [32].
5. Making use of both the textual information obtained from the Web and the visual content of images for Web image retrieval [22].

Many systems exploit one or more of the above techniques to implement high-level semantic-based image retrieval. In this chapter, we review some of these techniques with a particular focus on text and image fusion. Both the text and the image retrieval have long been an established research area. This isolated view of information processing fails to link related information from different modalities together. A number of image retrieval systems are already in place. However, efforts to combine them are only of recent interest. One of the main reasons to focus on these approaches is that it can be generalized for indexing and retrieval of multimodal document leading to an integrated view of information processing. The rest of the chapter is organized as follows. Section II reviews ontology-based approaches followed by automatic annotation techniques in section III. The next two sections discuss relevance feedback and semantic template based approaches. Section VI reviews existing approaches that attempt to integrate two important modalities: text and image. Finally, conclusions are made in section VII.

II. USING OBJECT ONTOLOGY TO DEFINE HIGH-LEVEL CONCEPTS [6-7].

As discussed in the preceding section, CBIR systems use image features to retrieve images. However, they suffer from the well-known semantic gap problem. Ontology has been used to bridge the semantic gap between the image feature and high-level concept [5] [18] [15][20] [26]. Ontology is a tool for structuring shared knowledge.

In ref. [20], a semantic medical image retrieval framework has been proposed that supports query by

concept and semantic query by image in addition to usual query by keyword approach. Query by concept is achieved by mapping keywords typed by user to ontological concept. The disambiguation, if needed, is achieved with the help of user. In order to support semantic query by image, the framework uses a data analysis component which extracts concepts from medical ontology to annotate the input image. In ref. [5] image ontology and description logic is used for semantic-based reasoning and image retrieval. In Mezaris et al. [18] an ontology-based system is presented. Their system segments each image into a number of regions and describes each region of an image using color, its position in horizontal and vertical axis, its size and shape. These low-level indexing features are translated into intermediate level descriptors which describes each region qualitatively. These intermediate-level descriptors form the object-ontology and are used in the system to filter irrelevant regions. The final ranking is still dome using low-level features but the user has to manipulate only human-centered intermediate-level descriptors. Ref. [19] further extends this system by using a relevance feedback mechanism, based on support vector machines and using the low-level descriptor, to rank potentially relevant regions to produce results.

The effectiveness of ontology was successfully demonstrated in a study by Wang *et al.* [26] in which keyword-based image retrieval and ontology-based image retrieval was compared. They constructed ontology on a combination of text annotation and image feature and demonstrated experimentally that combining both text and image features in multi-modality ontology helps in improving image retrieval.

A framework on multimodality ontology approach to image retrieval is proposed in [2]. In order to provide shared semantic interpretations of images from sport news domain they combine three ontologies: the text based ontology, the visual description from image annotation and feature extraction; and the domain ontology, extracted from the DBpedia ontology. In a subsequent paper [1] the multi-modality ontology image retrieval system was compared with single modality ontology. The results show that multi-modality ontology IRS give high precision and recall compared to visual-based ontology and keyword-based ontology. Manzoor [16] proposed Ontology based Image Retrieval (OIR) system, which uses domain specific ontology for retrieving images. The system was trained and tested on mammal's domain.

III. AUTOMATIC ANNOTATION

Most of the searchable image database, such as Flickr™, Picassa™, use associated tags for retrieving images. However, a large amount of images do not have any tag and hence are never retrieved. Manual tagging of this 'invisible content' is not possible. Automatic image annotation (AIA) techniques can help making this content visible. The objective of AIA is to associate high-level semantic features (keywords) with images automatically. These keywords can be used to propose tags for a new image being uploaded by users or to retrieve images in response to a textual query submitted to the image database. In order to achieve AIA, semi-supervised and supervised learning approach has been widely employed. Barnard and Forsyth [3] proposed a statistical modal which organizes image database using semantic information provided by associated text and visual information provided by image features. Duygulu *et al.* [9] considered the problem of automatic image annotation as the task of translation from a vocabulary of blobs to a vocabulary of words. The blobs correspond to image regions. In [13], Jeon *et al.* proposed a cross-media relevance model (CMRM). They represent an image as a composition of certain number of blobs and learn the joint distribution of blobs. In contrast to the translation model, they do not assume one-to-one correspondence (alignment) between the blobs and the words in an image. Instead, their model takes advantage of the joint distribution of words and blobs and annotates each test image with a vector of probabilities for all the words in the vocabulary.

A semi-automatic image annotation algorithm is proposed [17] which uses three-layer architecture. The bottom layer contains visual feature vector representation of images in the database. The middle layer comprises of set of keywords that have been used in annotation. The visual information in bottom layer is mapped keywords with the help of a Bayesian network. The keywords map to specific slots in the domain-specific schema(s) contained in the top layer. The system learns both from the ontologies and schemas as well from the joint occurrence of visual features and keywords. The extracted knowledge (joint probabilities) is used to suggest additional tags to user and to prevent inconsistent or contradictory annotation.

Carneiro *et al.* [6] formulated the problem of image annotation as Supervised Multiclass Labeling (SML) problem where each of the semantic concepts of interest defines an image class. They used hierarchical Gaussian Mixture Model (GMM) to model class distribution using and two-level EM algorithm to estimate the distribution. Their model enhances the efficiency of the learning but is computationally expensive due to the requirement of one feature vector extraction at pixel level. More recently, Shi *et al.* [23] solve the problem of SML in greater generality using region-based supervised annotation technique. Unlike [6], they estimated regional-class distribution and calculate image-level posterior probabilities by combining region-level posterior probabilities. Following human perception behavior which uses context for object understanding, they attempt to improve annotation by modifying posterior probabilities using the possible labels for other regions in the image. Guillaumin *et al.* [11] proposed the TagProp model for automatic image annotation based on nearest neighbor methods that predicts tags by taking a weighted combination of the tag absence/presence among neighbors. The model allows the integration of metric learning. To boost the recall of rare words, they introduced introduce word-specific logistic discriminant models that increases the probability of rare tags and decreases it for frequent ones. In FastTag algorithm[7] supervised multi-label classification problem is casted as un-labeled multi-view learning and two linear classifiers is learnt to predict tag annotations: an image classifier to predict complete tag set from image features and another to estimate which tags are

likely to co-occur with those already in existing tag vector. Both the classifiers are forced to agree.

IV. Relevance feedback

Relevance feedback (RF) mechanism was first introduced in information retrieval field to improve retrieval performance. The basic idea is to perform an initial retrieval to get a ranked list of documents and then use direct or indirect feedback to refine the query with a hope to move closer to users' ideal query. The improved query representation is then used in subsequent retrieval. The same technique has been transformed and introduced in image retrieval systems in 90's and has gained much attention in last few years. Relevance feedback technique involves user in the retrieval loop. During the iteration of feedback, user is required to mark the retrieved images as to whether they are relevant or irrelevant to their request. These images are used to refine the original query. Another round of retrieval is performed using the improved query representation. The process may continue up to several rounds to achieve desired results. The feedback implicitly helps in moving query representation closer to user's preference which helps in narrowing the gap between high-level image semantics and low-level image features. Early works on relevance feedback in image retrieval includes [24][36]. In [24] a feedback is used to adapt the retrieval system continuously to the changing requests of the user whereas the ImageRover system[36] uses feedback in the form of the relevant images specified by the user to select appropriate Minkowski distance metrics on the fly. Yin *et al.* [29] proposed an image relevance reinforcement learning model for integrating multiple RF techniques. Their model selects optimal RF technique for a query automatically during feedback iteration.

V. Use Of Semantic Templates

Semantic template is a map between semantic concept of high level and visual feature of low level. In Ref. [8], Cheng et al. first introduced the idea of using semantic visual templates to bridge the gap between user's information needs and what the systems can deliver in CBIR system. They developed algorithms to interact with user to identify a set of possible low-level feature combinations which might represent their semantic query. The system then identifies regions of primitive feature space templates and generates an initial set of query icons. These templates represent a personalized view of concepts. In contrast to [8], Zhuang *et al.*[32] generate templates automatically in in the process of relevance feedback with the help of WordNet. They define the semantic template as a triplet:

ST={C,F,W}

Where *C* represent the concept of the user, *F* is the feature vector correspond to *C, W* is the weight of feature vector. The WordNet was used to extract ordered list of words semantically related to a keyword typed by user. For every term in the list, the system finds its corresponding semantic template, and uses the *F* and *W* to query similar image.

Liu *et al.[*15] proposed a decision tree-based learning algorithm named DT-ST (Decision Tree-Semantic Template) which uses semantic templates to discretize continuous-valued region features. They build a DT to associate the low-level features of image regions with 19 high-level concepts selected from natural scenery image database. Their semantic image retrieval system also allows users to retrieve images using both query by region of interest and query by keywords.

VI. Text And Image Fusion For Image Retrieval

With the increased availability of multimodal information on the web, multimedia fusion has gained much attention of researchers in recent times. It refers to "*integration of multiple media, their associated features, or the intermediate decisions in order to perform an analysis task* [38]*."* In this chapter, we restrict our discussion to fusion of text and image for the task of image retrieval. Both the text and the image retrieval have long been an established research area. A number of text and image retrieval systems are already in place. However, efforts to combine them are only of recent interest. As discussed earlier (Section 1), representation of an image using visual features results in the loss of semantic information which can be compensated by textual features. Visual and textual features complement each other and hence this integration provides improved understanding of images. In literature, fusion of textual and visual features is performed generally at two levels: feature level or early fusion and decision level or late fusion. In early fusion, first textual and visual features are extracted. Then, the extracted feature vectors are combined. In late fusion approach, retrieval is performed using textual and visual features independently and then the result is fused into a joint result. Few reported work apply fusion at both the level.

One of the early efforts in combining visual and textual features in image retrieval is the work by Sclaroff *et al.* [22], in which textual and visual statistics from HTML documents are combined into a unified index vector. Text statistics are captured in vector form using latent semantic indexing (LSI) and image statistics are computed using color histogram and texture orientation distribution. The use of LSI supports semantic matching of keywords. The visual statistics is computed over six regions leading to 12 visual statistics vector per image. The LSI and different image feature vectors are then combined into a global similarity measure using a linear combination of normalized Minkowsky distances. The relative weightings of the individual features are determined using relevance feedback.

Latent semantic analysis has been used by Pham *et al.* [21] for early fusion of image features and keywords.

In ImageCLEF 2009 Medical Retrieval Track, Simpson *et al.* [35] experimented with a number of approaches utilizing textual and visual features and their combination. In each test run involving fusion of text and image, first an initial search was performed using textual or visual search and then either re-ranking was done using visual search or another phase of retrieval was done using modified query vector. In many cases they experienced a drop in performance when text-based approaches were combined in

a multimodal scheme. However, they observed overall best result for the case when the text- and content-based approaches were combined in relevance feedback retrieval scenarios.

An early fusion image retrieval approach based on single pLSA (probabilistic latent semantic analysis) model is presented in [37]. Each image is represented using a set of visual-textual words generated by fusing the visual descriptors and textual descriptors using pLSA model. The experiments conducted on ImageCLEF2009 Medical Image Retrieval dataset shows better retrieval performance in fusion approach than retrieval using textual features or visual features alone in a similar setting.

In Ref. [34], Caicedo *et al.* (2010) proposed a strategy based on Latent Semantic kernels to fuse visual and textual features in a medical mage retrieval system. They used Latent Semantic Kernels to define latent concepts that merge visual patterns and textual terms. Their visual-text fused approach improves retrieval performance as compared to using visual information only when tested on medical image collection from ImageCLEFmed08 challenge.

A hybrid fusion approach was followed in Ref. [33] in a multimodal image retrieval system which supports text only query, single image query and multimodal query (text and multiple images). Early fusion was used to concatenate multiple image descriptors. Assisted query semantic-expansion based on medical thesaurus was used to expand text query. Text and image data was indexed and searched separately, resulting in several ranks per query. The ranked lists obtained by text-based retrieval and image-based retrieval were fused using a novel technique, called Inverse Square Rank.

VII. Conclusion

This paper reviews existing techniques to address the semantic gap problem. With the increasing amount of multimedia content fusion-based approaches are gaining increased attention. Approaches involving fusion of text and images for image retrieval discussed in this paper can be generalized for other modalities as well.

Acknowledgement

This work was supported by Department of Science & Technology (grant no. SR/FTP/ETA-136/2011 dt 03/08/2012).

References

[1] Y. I. Aspura, M. Khalid, S. Azman. And M. Noah, "Improving the performance of multi-modality ontology image retrieval system using DBpedia" Global Journal on Technology, Vol 3 (2013): 3rd World Conference on Information Technology (WCIT-2012) p. 1515-1523.

[2] Y. I. Aspura, M. Khalid, S. A. Noah and S. N. S. Abdullah , "Towards a Multimodality Ontology Image Retrieval", In H. Badioze Zaman *et al.* (Eds.*), Second International Visual Informatics Conference, IVIC 2011, Part II, LNCS 7067*, 2011, Springer-Verlag, p. 382–393,.

[3] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures". *In International Conference on Computer Vision, Vol. 2*, pp. 408-415, 2001.

[4] M. D. Blei and M. I. Jordan, "Modeling annotated data", *In the proceedings of 26th Annual international ACM SIGIR conference*, 2003.

[5] H. Bo, S. Dasmahapatra, P. Lewis, and N. Shadbolt, "Ontology-based medical image annotation with description logics", *In proceedings of the 15th IEEE International conference on Tools with Artificial Intelligence*, IEEE, 2002, p. 77-82.

[6] G. A. Carneiro, B. Chan, P. J. Moreno and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval". *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 29(3), pp. 394 – 410, 2007.

[7] M. Chen, A. Zheng, and K. Q. Weinberger, "Fast Image Tagging", *Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia*, USA, 2013.

[8] S.-F. Cheng, W. Chen, H. Sundaram, Semantic visual templates: linking visual features to semantics, *International Conference on Image Processing (ICIP), Workshop on Content Based Video Search and Retrieval*, vol. 3, October 1998, p. 531–534.

[9] P. Duygulu, K. Barnard, N. de Freitas and D. Forsyth, "Object recognition as machine translation: learning a lexicon for a fixed image vocabulary", *In Proc. ECCV*, 2002, p. 97-112.

[10] H. Feng and T.-S. Chua, "A bootstrapping approach to annotating large image collection", *Workshop on Multimedia Information Retrieval in 5th ACM Multimedia*, November 2003, ACM, p. 55–62.

[11] M. Guillaumin T. Mensink, J. Verbeek and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation", I*n Computer Vision, 2009 IEEE 12th International Conference on*, 2009, IEEE, p. 309–316.

[12] C. Hudelot, N. Maillot and M. Thonnat, "Symbol Grounding for Semantic Image Interpretation: From Image Data to Semantics", *In Proceedings of Tenth IEEE International Conference on Computer Vision*, 2005.

[13] J. Jeon, V. Lavrenko and Manmatha, "Automatic Image annotation and retrieval using cross media relevance models. *In the Proceedings of SIGIR'03, Toronto, Canada*, 2003.

[14] Y. Liu, D. Zhang, G. Lu and W-Y. Ma, "A survey of content-based image retrieval with high-level semantics", *Pattern Recognition 40*, pp. 262–282, Elsevier, 2007.

[15] Y. Liu, D. Zhang and G. Lu, "Region-based image retrieval with high-level semantics using decision tree learning", *Journal of pattern recognition*, 41(8), Elsevier, 2008.

[16] U. Manzoor, "Ontology based image retrieval", *In the Proceedings of International Conference for Internet Technology and Secured Transactions, 10-12 Dec. 2012*, IEEE pp. 288 – 293.

[17] O. Marques and N. Barman, "Semi-automatic Semantic Annotation of Images Using Machine Learning Techniques", *In Dieter Fensel, Katia Sycara, John Mylopoulos (Eds.*), *The Semantic Web - ISWC 2003*, Lecture Notes in Computer Science Volume 2870, Springer, 2003, pp. 550-565.

[18] V. Mezaris, I. Kompatsiaris and M. G. Strintzis, "An ontology approach to object-based image retrieval", *Proceedings of the International conference on Image Processing 2003, Vol. 2*, IEEE, 2003, pp. 511–514.

[19] V. Mezaris, I. Kompatsiaris and M. G. Strintzis, "Region-based Image Retrieval using an Object ontology and Relevance Feedback", *EURASIP Journal on Applied Signal Processing*, Vol. 2004(6), 2004, pp. 886-901.

[20] M. Möller and S. Michael, "A Generic Framework for Semantic Medical Image Retrieval", *In Proceedings of 7th Korea-Germany Joint Workshop on Advanced Medical Image Pro In Stefan Brass, C. G.(editor)*, 2007.

[21] T. Pham, N. E. Maillot, J. Lim and J. Chevallet, "Latent Semantic Fusion Model for Image Retrieval and Annotation"", *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.* ACM, 2007.

[22] S. Sclaroff, La Cascia, M. and S. Sethi, "Unifying textual and visual cues for content-based image retrieval on the world wide web", *Computer Vision and Image Understanding*, 75(1/2):p.86–98, July/August 1999.

[23] F. Shi, J. Wang and Z. Wang, "Region-based supervised annotation for semantic image retrieval", *International Journal of Electronics and Communications*, 65(11), p. 929– 936, 2011.

[24] P. T. Kurita and T. Kato, "Learning of personal visual impression for image database systems", *In Second Intl. Conf. on Document Analysis and Recognition*, 1993, pp. 547–552. , 2003.

[25] Vailaya, A., Figueiredo, M. A. T. , Jain, A. K, & Zhang, H. .J. (2001). Image classification for content-based indexing, *IEEE Transaction on Image Processing*, 10 (1), 117–130.

[26] H. Wang, S. Liu, and L-T, Chia, "Does Ontology Help in Image Retrieval? — A Comparison between Keyword, Text Ontology and Multi-Modality Ontology Approaches", *In Proceeding of the 14th annual ACM international conference on Multimedia* 2006, p. 109-112, ACM New York, NY, USA.

[27] L. Wu, R. Jin and A. K. Jain, "Tag Completion for Image Retrieval", *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 35(3), pp. 716–727, 2013.

[28] Wu, P., Hoi, S. C.-H., Zhao, P. & He, Y. (2011). Mining social images with distance metric learning for automated image tagging. *In Proceedings of the 4th ACM International Conference on Web Search and Data Mining* (pp. 197–206), ACM New York, NY, USA.

[29] P. -Y., Yin, B. Bhanu, K.-C. Chang, and A. Dong, "Integrating relevance feedback techniques for image retrieval using reinforcement learning", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10), pp. 1536 – 1551, 2005.

[30] N. Zhou, W. K. Cheung, G. Qiu, and X. Xue, "A hybrid probabilistic model for unified collaborative and content-based image tagging". *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 33:1281–1294, 2011.

[31] J. Zhuang, and S. C. Hoi, "A two-view learning approach for image tag ranking. *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, ACM, 2011, pp. 625–634.

[32] Y. Zhuang, X. Liu and Y. Pan, "Apply semantic template to support content-based image retrieval", *Proceedings of the SPIE, Storage and Retrieval for Media Databases, vol. 3972*, December 1999, (pp. 442–449).

[33] A. Mourão, F. Martins and J. Magalhães, "Multimodal medical information retrieval with unsupervised rank fusion", *Computerized Medical Imaging and Graphics* 00, pp. 1–13, 2014.

[34] J. C. Caicedo, J. G. Moreno, E. A. Niño and F. A. González, "Combining visual features and text data for medical image retrieval using latent semantic kernels, M*IR'10, March 29–31*, 2010.

[35] M.. S. Simpson, M. M., Rahman and D. Demner-Fushman, S. Antani and G. R. Thoma, "Text- and Content-based Approaches to Image Retrieval for the Image CLEF 2009 Medical Retrieval Track"**.** *CLEF (Working Notes)*. http://clef.isti.cnr.it/2009/working_notes/simpson-paperCLEF2009.pdf

[36] S. Sclaroff, L. Taycher, and M. L. Cascia, "ImageRover: a content-based image browser for theWorldWide Web," Technical Report 97-005, Boston University CS Dept., 1997.

[37] Y. Cao, H. Müller, C. E. Kahn, E. Munson, "Multi-modal Medical Image Retrieval", In Procc. SPIE Medical Imaging, 2011.

[38] A. Alpkocak, O. Ozturkmenoglu T. Berber, A. H. Vahid and R. G. Hamed, "DEMIR at ImageCLEFMed 2011: Evaluation of Fusion Techniques for Multimodal Content-based Medical Image Retrieval", 2011.