Multimedia Retrieval: Survey Of Methods And Approaches

Shweta Satish Kadam
Student
Dept. of Computer Engg.
DGOI, COE, Daund, Pune
ashwet.kdm@gmail.com

Usha Sunil Gound
Student
Dept. of Computer Engg.
DGOI, COE, Daund, Pune
goundusha9@gmail.com

Prof. Tanaji A.Dhaigude
Assistant Professor
Dept. of Computer Engg.
DGOI, COE, Daund, Pune
tanajidhaigude@gmail.com

Avinash Shivaji Gaikwad
Student
Dept. of E & TC Engg.
SCOE, Vadgaon (Bk),Pune
avinash.gaikwad12@gmail.com

Abstract**—** As we know there are numbers of applications present where multimedia retrieval is used and also numbers of sources are present. So accuracy is the major issue in retrieval process. There are number of techniques and datasets available to retrieve information. Some techniques uses only text-based image retrieval (TBIR), some uses content-based image retrieval (CBIR) while some are using combination of both. In this paper we are focusing on both TBIR and CBIR results and then fusing these two results. For fusing we are using late fusion. TBIR captures conceptual meaning while CBIR used to avoid false results. So final results are more accurate. In this paper our main goal is to take review of different methods and approaches used for Multimedia Retrieval.

Keywords- multimedia retrieval, text based image retrieval, content based image retrieval, late semantic fusion

*****

I. Introduction

Nowadays, multimedia data are present everywhere i.e. from large digital libraries to the web content, we are often use multimedia information both in the context of our professional or personal activities.

But there is a challenge that makes multimedia information retrieval to face problem like “semantic gap”. On the one hand, multimedia data such as images, videos, are stored in machines into a computational representation which consists of low-level features. On the other hand, humans fire queries by using high-level concepts such as keywords. So it becomes a difficult task to compare user query with items in large collection. So it is very challenging to automatically extract the semantic content of an image and to retrieve more accurate result from a huge database. Traditional systems uses text based retrieval. Here we are discussing number of techniques used to retrieve information like textual retrieval, visual retrieval etc. Textual information is used to capture meaning of query and visual information is used to retrieve more accurate result. Textual information means metadata of document and visual information means low level features like color, texture etc. also fusion techniques are used to combine both textual and visual result.

Types of Fusion Techniques:

Fusion techniques can be classified in three sub types called early fusion, late fusion and transmedia fusion [2].

Early Fusion [9]:

In early fusion approach feature representation of text and image are fused together using Joint features model [2]. Early fusion based on extracted features of information sources and combination of it. Advantage of early fusion approach is the correlation between multiple features and there is only one learning phase [1].

Late Fusion [1]:

In late fusion algorithm the similarity scores are drawn from features of sources. Textual similarity is calculated from textual feature and visual similarity is calculated from visual features. The fusion carried out at decision level calculated from features is called late fusion. And after that some aggregation functions are used to combine these two similarities [2]. Aggregation function include mean average, product etc. Advantages of late fusion are Simplicity, scalability and flexibility [1].

Transmedia Fusion:

The difference between late fusion and transmedia fusion lies in fusion function used. Instead of aggregation process diffusion process is used for fusion. This technique first uses one of the modalities and retrieve relevant documents and then to switch to the other modality and aggregate their results [5].

Datasets Used:

1. TRECVID[3]:

TRECVID dataset is mainly used for video-based fusion. Dataset includes information about broadcast news video, sound and vision video, BBC rushes video, and test dataset annotations for surveillance event detection.

2. Biometric Dataset:

There are number of datasets present which are used for biometric retrieval. These includes-

- BANCA [3] which includes face and speech modalities
- XM2VTS [3] which contains video and speech data
- BIOMET [3] that contains face, speech, fingerprint, hand and signature modalities
- MYCT [3] that contains fingerprint and signature.

3. ImageCLEF[1]:

ImageCLEF runs as a part of Cross Language Evaluation Forum (CLEF) and used as cross-language image retrieval [1]

(i) IAPR:

The IAPR TC-12 photographic collection [2] consists of 60 topics and 20,000 images which are taken from nature. This includes pictures of various actions, photographs of people, animals, cities, landscapes etc. Image has caption which is nothing but title of image, the location from which the photograph was taken, and a semantic description of the image.

(ii) BELGA:

The Belga News Collection [2] contains 498,920 images from Belga News Agency. Belga News Agency is an image search engine for news photographs. The caption of an image can contain the date and the place where the image was captured.

(iii) WIKI:

The Wikipedia collection [1][2] consists of 70 topics and 237,434 images and user-supplied annotations in English, German and/or French. In addition, the collection contains the original Wikipedia pages in wikitext format from where the images were extracted.

(iv) MED:

The medical image collection [2] consists of 16 topics and 77,477 medical images like CT, MR, X-Ray, PET microscopic images but also graphical plots and photos. In the ad-hoc retrieval task [20], the participants were given a set of 16 textual queries with 2-3 sample images for each query. The queries were classified into textual, mixed and semantic queries, based on the methods that are expected to yield the best results. In our experiments we did not consider this explicit query classification, but handled all queries in the same way.

II. ARCHITECTURAL DESCRIPTION:

Three sub-systems are used in most of the paper. These are TBIR (Text-Based Image Retrieval), CBIR (Content-Based Image Retrieval), and Fusion subsystem [1] [2] [11].

(i) Text based image retrieval(TBIR) sub-system[1]:

TBIR takes input from metadata and articles used in Wikipedia collection and text from topics. From that it calculates relevance score (St). Four steps are used for retrieval: Textual information Extraction, Textual preprocessing, Indexing and search.

a) Textual Information Extraction[1]: The metadata and the articles are used as sources for this step. The metadata XML tags are extracted including <name>, general <comment> ,<description>, and <caption>.

b) Textual Preprocessing[1]: This component processes the selected text in three steps: 1) characters which has no statistical meaning, like punctuation marks or accents, are eliminated 2) elimination of stopwords and 3) stemming

c) Indexation[1][9]: After textual preprocessing information is indexed using Lucene .

d) Search[1]: After preprocessing textual results list with the retrieved images ranked by their similarity score (St).

(ii) Content based image retrieval(CBIR) sub-system[1]:

CBIR takes input from images used in Wikipedia collection and topics and also it uses textual pre-filtered list to reduce dataset. From that it calculates relevance score (Si). Two steps are used for retrieval: Feature extraction and similarity module.

a) Feature Extraction [9][1]: The visual low-level features for all the images in the database for the example images for each topic are extracted using the SIFT.

b) Similarity module [1]: The similarity module uses own logistic regression relevance feedback algorithm [14] to calculate the Similarity (Si) of each of the images of the collection to the query.

(iii) **Fusion sub-system[1]:**

Numbers of fusion techniques are used to fuse two different lists of TBIR and CBIR.

These techniques are MaxMerge [1] [4] [9] [11], Enrich [1] [5] [9] [11], OWA operator [1] [4] [9], FilterN [1] [5], Text-Filter [5], Join [9] and Product [1].

III. REVIEW OF MULTIMEDIA RETRIEVAL:

In this section we have reviewed the papers given in the references section.

1. In [6] author proposed metasearch model based on an optimal democratic voting procedure, the Borda Count and based on Bayesian inference and also investigated a model which obtains upper bounds on the performance of metasearch algorithms.

2. In [4] author presented experiments in ImageCLEF 2010 Campaign. Author assumes that textual module better captures the conceptual meaning of a topic. So that, the TBIR module works firstly and acts as a filter for CBIR, and the CBIR system starts working by reordering the textual result list. The CBIR system presents three different algorithms: the automatic, the query expansion and a logistic regression relevance feedback.

3. In [2] author proposed different techniques i.e. author semantically combines text and image retrieval results to get better fused result in the context of multimedia information retrieval. Using these techniques some observations are drawn that image and textual queries are expressed at different acceptable levels and that an only image query is often unclear. Overall, the semantic combination techniques overcome a conceptual barrier rather than a technical one: In these methods there is combination of late fusion and image reranking and also proposed techniques against late and cross-media fusion using 4 different ImageCLEF datasets.

4. In [11] author introduced a new task i.e. ImageCLEF 2009 Campaign used to retrieve photo. Author proposed an ad-hoc management of the topics delivered, and also generates different XML files for large number of caption of photos delivered. For this two different merging algorithms to merge textual and visual results were developed. Author's best run is at position 16th, in the 19th for MAP score of performance metrics, at position 11th, for a total of 84 submitted experiments of diversity metrics.

5. In [7] author gave an overview of different features used in content-based image retrieval and compares them quantitatively on four different tasks: stock photo retrieval, personal photo collection retrieval, building retrieval, and medical image retrieval. Five different available image databases are used for this experiments and the performance of image retrieval is investigated in detail. Due to this comparison of all features is possible and in future possibility of comparison of newly proposed features to these features.

6. In [5] author introduced new merging techniques to fuse text-based retrieval and content-based retrieval results, and improved the text-based results while using one of the three merging algorithms although visual results are lower than textual ones. In this MIRACLE-FI textual retrieval is used using TF-IDF weight and CBIR uses different low level features based on color and texture information. The main conclusion of this paper is that the Mahalanobis distance works better than the Euclidean one, and the best aggregation method is the AND operator.

7. In [3] author introduced survey paper which provides no of fusion techniques for multimedia researchers used to combine different results which are used for multimedia retrieval and its analysis purpose. Paper also gives observations based on the reviewed literature. These observations will be useful for different readers to understand which fusion technique to be used and at which level.

8. In [9] author focused on applying different strategies of merging multimodal information i.e. textual and visual information by using both early and late fusion approaches. In this system, the TBIR module works firstly and acts as a filter, and then CBIR system works only on filtered TBIR results i.e. on reduced database to get better result. The two ranked lists are fused using its own probability in a final ranked list. The best run of the TBIR system is in position 14 with a MAP of 0.3044, and TBIR system uses IDRA tool and Lucene for indexing, fusing monolingual experiments carried out with IDRA preprocessing of text and Lucene search engine, with some extra information from Wikipedia articles. For CBIR system uses logistic regression relevance feedback algorithm and CEDD low-level features for similarity modularity.

9. In [10] author proposed that indexing and classification of multimedia data an efficient information fusion of the different modalities is essential for the system's overall performance. Since information fusion, its influence factors and performance improvement boundaries have been lively discussed in different research communities. Author most importantly point out that exploiting the features and modality's dependencies will yield to maximal performance.

10. In [12] author introduced a method for extracting distinctive invariant features from images which are then used to perform reliable matching between different views of an object. The features are invariant to image scale and rotation, and are shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination. The features are highly distinctive, in the sense that a single feature can be correctly matched with high probability against a large database of features from many images. This paper also describes an approach to using these features for object recognition. The recognition proceeds by matching individual features to a database of features from known objects using a fast nearest-neighbor algorithm, followed by a Hough transform to identify clusters belonging to a single object, and finally performing verification through least-squares solution for consistent pose parameters.

11. In [5] author proposed that results obtained by using text-based retrieval are much better than content-based result. Author introduced three different merging techniques to combine textual and visual results and proves that visual results are lower than text based result.

12. In [10] author presents a method to extract distinctive features of images. These features are used to match different views of an object. Author also describes an approach of using these features for identification of an object.

13. In [14] author deals with the problem to retrieve image from huge database of images. During retrieval process retrieved images must be same as user's mind and also considering user's positive or negative feedback preference for images. Author presented a novel algorithm which considers the probability of an image belonging to the set of those sought by the user, and models the *logit* of this probability as the output of a generalized linear model whose inputs are the low-level image features. The image database is ranked by the output and given to the user, who selects a few positive and negative samples. This process is repeated in an iterative manner until user is satisfied.

14. In [17] author surveyed about an overview of the resources and topics of the Wikipedia Retrieval task at ImageCLEF 2010 and also summarizes the retrieval approaches given by the participating groups, and provides an analysis of the main evaluation results.

15. In [18] author proposed a relevance feedback based interactive retrieval approach which considers some characteristics of CBIR. During the retrieval process the user's high level query and perception are captured by dynamically updated weights which are based on the user's feedback.

IV. CONCLUSION:

This paper gives a detailed description and analysis of multimedia retrieval and also using some textual pre-filtering techniques. Due to textual pre-filtering techniques size of multimedia database is reduced so improving the final fused retrieval results. Large numbers of papers prefer late semantic fusion i.e. decision level fusion than early fusion. Due to its simplicity, flexibility and scalability late fusion is advantageous. Numbers of datasets are used in different papers for experimentation. Numbers of fusion algorithm are used out which Product algorithm gives best result.

V. REFERENCES

[1] Xaro Benavent, Ana Garcia-Serrano, Ruben Granados, Joan Benavent, and Esther de Ves "Multimedia Information Retrieval Based on Late Semantic Fusion Approaches: Experiments on a Wikipedia Image Collection" in Computer Science Department, Universidad de Valencia, Valencia 46022, Spain, 2013.

[2] S. Clinchant, G. Csurka, and J. Ah-Pine, "Semantic combination of textual and visual information in

multimedia retrieval," in Proc. 1st ACM Int. Conf. Multimedia Retrieval, New York, NY, USA, 2011.

[3] P. K. Atrey, M. A. Hossain, A. El Saddik, and. S. Kankanballi, "Multimedia Fusion for Multimedia Analysis: A Survey," Multimedia Syst., vol. 16, pp. 345–379, 2010.

[4] J. Benavent, X. Benavent, E. de Ves, R. Granados, and A. García-Serrano, "Experiences at ImageCLEF 2010 using CBIR and TBIR mixing information approaches," in Proc. CLEF 2010, Padua, Italy, 978-88- 904810-2-4, Notebook papers.

[5] A. García-Serrano, X. Benavent, R.Granados, and J. M.Goñi-Menoyo, "Some results using different approaches to merge visual and text based features in CLEF'08 photo collection," in Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. Berlin, Germany: Springer-Verlag, 2009, pp. 568–571.

[6] J. A. Aslam and, M. Montague, "Models for metasearch," in Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, New Orleans, LA, USA, 2001, pp. 276–284.

[7] T. Deselaers, D. Keysers, and H. Ney, "Features for image retrieval: An experimental comparison," Inf. Retrieval, vol. 11, pp. 77–107, Apr. 2008.

[8] J. Kludas, E. Bruno, and S. Marchand-Maillet, "Information fusion in multimedia information retrieval," in AMR Int. Workshop Retrieval, User Semantics, 2007.

[9] R. Granados, J. Benavent, X. Benavent, E. de Ves, and A. Garcia-Serrano, "Multimodal Information Approaches for the Wikipedia Collection at ImageCLEF 2011," in Proc. CLEF 2011 Labs Workshop, Notebook Papers, Amsterdam, The Netherlands, 2011.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International J. Comput. Vision, vol. 60, no. 2, pp. 91–110, 2004.

[11] A. García-Serrano, X. Benavent, R. Granados, E. de Ves, and J. Miguel Goñi, "Multimedia Retrieval by Means of Merge of Results from Textual and Content Based Retrieval Subsystems," in Multilingual Information Access Evaluation II. Multimedia Experiments: 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers. Berlin, Germany: Springer-Verlag, 2010, pp. 142–149.

[12] E. A. Fox and J. A. Shaw, "Combination of multiple searches," in Proc. 2nd Text Retrieval Conf., 1993, pp. 243–252.

[13] M. Grubinger, "Analysis and Evaluation of Visual Information Systems Performance," Ph.D. thesis, School Comput. Sci. Math., Faculty Health, Engi., Sci., Victoria Univ., Melbourne, Australia, 2007.

[14] T. Leon, P. Zuccarello, G.Ayala, E. de Ves, and J. Domingo, "Applying logistic regression to relevance feedback in image retrieval systems," Pattern Recog., vol. 40, pp. 2621–2632, Jan. 2007.

[15] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," ACM Trans. Multimedia Comp., Commun., Appl., vol. 2, no. 1, pp. 1–19, Feb. 2006.

[16] "ImageCLEF: Experimental Evaluation in Visual Information Retrieval," in The Information Retrieval Series, H. Müller, P. Clough, T. Deselaers, and B. Caputo, Eds. New York, NY, USA: Springer-Verlag, 2010, vol. 32.

[17] A. Popescu, T. Tsikrika, and J. Kludas, "Overview of the wikipedia retrieval task at ImageCLEF 2010," in Proc. CLEF 2010 Labs Workshop, Notebook Papers, Padua, Italy, 2010.

[18] Y. Rui, S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," IEEE Trans. Circuits Syst. Video Technol., vol. 8, no. 5, Sep. 1998.