ISSN: 2321-8169
3788 - 3792

Fusion of Text and Image in Multimedia Information Retrieval System

Trupti S. Atre
Department of Computer Engineering
MET IOE-BKC, Savitribai Phule Pune University
Nasik, India
truptiatre.2089@gmail.com

Prof. K. V. Metre
Department of Computer Engineering
MET IOE-BKC, Savitribai Phule Pune University
Nasik, India
kvmetre@gmail.com

***Abstract*—**Multimedia Information Retrieval is very useful for any application in our daily work. Most of the applications consist of Multimedia data that are images, text, audio and video. Multimedia information retrieval system is used to search an image. There are same meanings for different data which is also known as semantic gap. This problem is solved by fusion of text based image retrieval and content based image retrieval. Weighted Mean, OWA and WOWA are aggregation operators used in this system for the fusion of text and image numeric values. The Scale invariant feature transforms and speeded up robust feature are two algorithms for feature extraction. To increase the speed of system, the speeded up robust feature algorithm is used. Bag of Words and Bag of Visual Word approaches are used in this system for retrieving images.

Keywords*-Content Based Image Retrieval, Fusion, Multimedia Information Retrieval, Text Based Image Retrieval.*

*****

I. INTRODUCTION

Today, in most of the applications multimedia data is used. Multimedia information retrieval is the retrieval of text, audio, video and image or combination of them. This system uses images and text of those images to retrieve them. Semantic gap is the problem in retrieval system which is the same meaning of different data[1][2]. Text Based Image Retrieval (TBIR) and Content Based Image Retrieval (CBIR) are two subsystems in Multimedia Information Retrieval System. Text Based Image Retrieval uses tags of an image to retrieve images by giving text query. Content Based Image Retrieval uses image query to search based on their color, texture and shape features. BM25 similarity is used for similarity matching. Bag of Visual Word (BoVW) approach is used in content based image retrieval to retrieve relevant images. Feature extraction is one of the important technique in content based image retrieval which uses Speeded Up Robust Feature (SURF) algorithm for retrieving images in fast and robust manner. For better result of image retrieval, fusion algorithm is used to combine both TBIR and CBIR.

Multimedia fusion uses each mode and the different sources as corresponding information to get a particular image from the text and image dataset. The semantic gap which is problem in multimedia information retrieval is solved by using the textual prefiltering and image re-ranking by using different techniques. Various late fusion algorithms are used as; Product, OWA operators, Enrich, MaxMerge and FilterN. Two algorithms of feature extraction are used. First is the SIFT algorithm (Scale Invariant Feature Transform) is used for extracting distinctive invariant features from images [3].

Second is, as Herbert Bay et. al. [4][5] first introduced the SURF algorithm as a new scale and rotation invariant interest point detector and descriptor. Also SURF produces a set of interest points for each image and a set of 64-dimensional descriptors for each interest points of an image. The combination rules are used as; combMAX(the maximum combination), combMNZ (the product of maximum and non-zero numbers) and combSUM(the sum combination)[6][14]. There are two approaches to fuse the information; the approach to fuse the information at the feature level is early fusion and other one is decision level or late fusion which merges multiple modalities in the semantic space of the multimedia information retrieval system. Combination of these approaches is known as the hybrid fusion of multimedia information. It uses the late fusion approach for combining both textual and visual information of image retrieval. It provides scalability, flexibility and simplicity of retrieval of information [1][7]. The TBIR (Text-Based Image Retrieval) system enhances the conceptual meaning of the query than the CBIR (Content-Based Image Retrieval) system. For TBIR textual pre-filtering approach is used[1]. In the CBIR system it emphasizes that the images visually similar from the low-level visual features but with different conceptual meaning of an image[1][8]. The CBIR process will be significantly reduced in terms of time and computation by using SURF technique. The fusion techniques in image retrieval are based on combining textual and visual results. The decisions obtained from text and visual-based systems by means of aggregation functions or classical combinations algorithm. It uses weighted factors to assign different levels of confidence to each mode (textual or visual) of an image[9]. The fusion approach used for multimedia retrieval shows improved results with improved performance for information retrieval [1].

II. LITURATURE SURVEY

Multimedia information retrieval system is the best developed technology for better result of multimedia. It works with multimedia (textual and visual) information. Here, two techniques can used to determine the robust regions in the image that are Scale Invariant Feature Transform (SIFT) and Speeded Up Robust Features (SURF) [3]. These techniques are used to find the salient regions in the image prior to the embedding process and to reveal the possible differences in their performance. So, SURF is faster and robust than SIFT technique. These both approaches detect interest points or features and also propose a method for creating an invariant descriptor. These can be used to identify the found interest points and match them even under a variety of disturbing conditions like scale changes, rotation, changes in illumination or viewpoints or an image noise[3][4][5][10]. Support vector machine (SVM) is used for data classification. In this system SVMs used for different tasks including categorization of text and feature, face detection and modality fusion, concept classification etc. The SVM is used as an optimal binary linear classifier in which a set of input data vectors are partitioned as belonging to either one of the two classes [11]. In the MIRS, the system can be queried by: Query by example, Metadata based quires, queries based on data patterns or features and Annotation-based queries (event based queries) for searching an image.

A. Filtering and Normalization

In the multimedia information retrieval system for text retrieval, query based semantic filtering as a first level of information fusion is used. And it is categorized into semantic relationships such as relevance scores and multimedia similarities [1].

B. Visual Re-ranking

Visual Re-ranking is also known as Image Re-ranking. In Visual or image re-ranking technique; Images which are visually similar have similar relevance scores. Two approaches are used to re-arrange the top retrieved images by the text similarities. First is, text based similarities in order to find the most relevant objects from a semantic viewpoint and then second employ the visual similarities between objects of the database in order to refine the textual similarities based ranking by the text query[7][13].

In the previous multimedia information retrieval system they used 5 algorithms of late fusion[1]. IDRA(Indexing and retrieving automatically) tool for preprocessing in TBIR and SIFT technique for feature extraction in CBIR. IDRA tool performed operations on text that are Special characters, stopwords and stemming. The framework is presented by Lowe[3] for object recognition. There are 4 steps in SIFT technique are as:

1. Scale space extrema detection
2. Keypoint localization
3. Orientation assignment
4. Keypoint description

Large number of features are extracted by using SIFT technique. To find the correct match this technique is used. There are two techniques for detecting the robust region of an image. First is the SIFT and another one is SURF(Speeded Up Robust Features). For detecting correct robust regions the SURF technique is better than SIFT[3]. There are three important steps in SURF technique that are Detection, Description and Matching[5].

III. Proposed System

There are three subsystems in the proposed multimedia information retrieval system as shown in Fig. 1. The three subsystems are: shows the architecture with Text Based Image Retrieval (TBIR), Content Based Image Retrieval (CBIR) and the Fusion of both the subsystems[1][7]. Pre-filtering is done by textual module i.e. TBIR. In this subsystem, each section gets a ranked list based on a similarity score or probability score also, Score of the text (Pt) and score of an image (Pi) is calculated. Score of the both CBIR and TBIR are merged by using the fusion technique that is OWA[16].

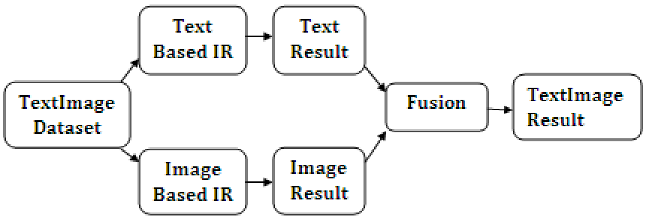


Fig. 1: System Architecture

A. TBIR

As shown in Fig.2. TBIR subsystem works. The TBIR subsystem uses the Bag of Words approach. The TBIR subsystem uses the BM25 tool that is ranking function. BM25 is a Bag of Words retrieval function. This function is uses in this system for retrieving matching documents according to relevance of the given query. It is also known as preprocessing of textual information[1][15].

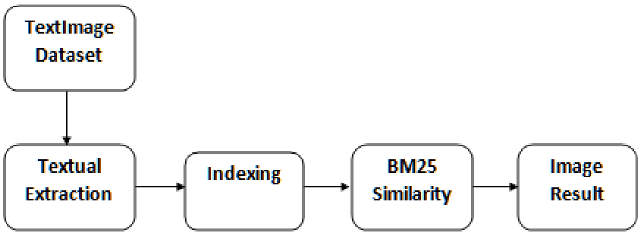


Fig. 2: TBIR Subsystem

These TBIR and CBIR subsystems are used to generate a ranked list with a certain probability and merged in the fusion module which gives final result that is fused score. Join and merging algorithms are used for the TBIR subsystem to fuse different textual result lists from monolingual preprocessing, and other fusing techniques are used for the CBIR subsystem.

B. CBIR

The CBIR subsystem uses its own low-level features or the CEDD (Color and edge directivity descriptor) features [11]. This system also uses its logistic regression relevance feedback algorithm for query re-formulation [12]. The proposed system uses Lucene indexing[17]. Feature extraction is very important step in image retrieval system. SURF technique is used for feature extraction. In existing system it uses SIFT(Scale Invariant Feature Transform) technique. SURF technique is robust than SIFT[3][4][10].

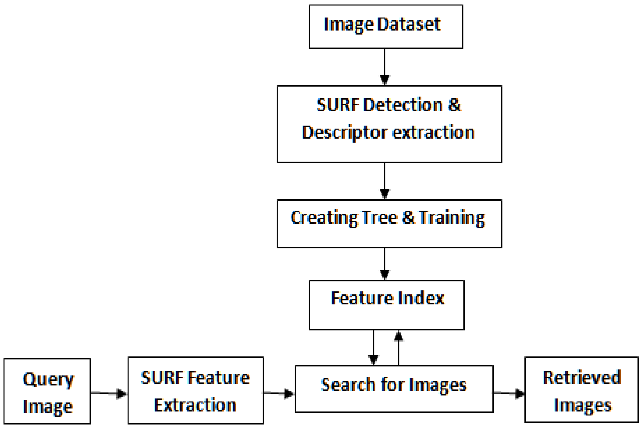


Fig. 3. CBIR Subsystem

As shown in Fig. 3. , SURF technique is used in proposed system. Firstly it reads each image from the dataset and interest point descriptor is applied to each image. SURF descriptor extraction is applied on each detector point. And it stores the descriptors in the form of tree to create clusters of similar vectors. After clustering, the system takes query image and SURF descriptor of it to search its closest match[4][5].

The 3 steps of SURF technique are as;

1. Detection
2. Description
3. Matching

In first step Hessian matrix is used for finding the point of interest.

Hessian-based interest point localization :

$$H = \begin{pmatrix} L_{xx} & L_{xy} \\ L_{xy} & L_{yy} \end{pmatrix} \qquad (1)$$

$L_{xx}(x, y, \sigma)$ is the Laplacian of Gaussian of the image.It is the convolution of the Gaussian second order derivative with the image[4][5]. we use D_{xx} to approximate L_{xx}

$$\det\left(H_{approx}\right) = D_{xx}D_{yy} - (\omega D_{xy})^2 \qquad (2)$$

For measure speed up integral images are used which is also known as intermediate representation:

$$I_{\Sigma(x,y)} = \sum_{i=0}^{i\le x}\sum_{j=0}^{j\le y} I(i,j) \qquad (3)$$

In second step it fixes the orientation based information then constructing square region and haar wavelet is used for interest point orientation.

In the third step it compares two images as; points and descriptor of first image and points of second image are compared for matching[4][5].

Bag of Visual Word(BoVW) approach is used in the system for indexing the filtered interest points. By this approach classification is also improves. It was firstly introduced for object recognition in clustering SIFT features[5][18].

IV. Implementation Details

This system uses late fusion approach that is based on Combining the TBIR and CBIR subsystems. Here Decisions that will be in the form of numerical similarities (probabilities or scores). The score or probabilities (Pt from textual-based retrieval and Pi from the visual-based retrieval) merged or fused by means of aggregation functions. The late fusion algorithms are better than those of early fusion technique. A technique is called as image re-ranking, which retrieves a set of ranked objects from textual subsystem that is followed by a reorder step of these objects according to the visual score (Pi). The CBIR subsystem which computes the visual scores (Pi) working only on the selected objects of the TBIR subsystem[1]. The late fusion algorithm is as follows:

A. *OWA Operators:*

The ordered averaged weighted operator (OWA) provides a finite number of inputs to perform a single output. None of the weight is associated with any particular input and the relative magnitude of the input by the OWA operator[16]. It decides which weight corresponds to each input provided by an operator. The inputs are in the form of textual and image scores (Pt and Pi), that can provide us the best information. The OR (max) and AND (min) operators can be used to find orness to characterize the degree to which the aggregation is like operation:

$$orness(wt) = \frac{1}{n-1}\sum_{i=1}^{n}(n-i)\, wt_i \qquad (4)$$

OWA operators with many of the weights close to their highest values will be *or-like* operators that is $orness(Wt) \le 0.5$, while those operators with most of the weights close to their lowest values will be *and-like* operators that is $orness(Wt) \ge 0.5$[1][16].

Parameter Estimation by MLE[1]: Weight calculation by matrix as, WMx[k=g0] with addition of positive plus negative images for each group of characteristic:

Weight calculation is as matrix by,

$$W_{Mx[k/g0]} = (WP,WN) \qquad (5)$$

Where,

$$WP = (x^{p}_{ini},\ldots\ldots..,x^{p}_{fi},y_{p}) \qquad (6)$$
$$WN = (x^{n}_{ini},\ldots\ldots..,x^{n}_{fi},y_{n}) \qquad (7)$$

Parameter Estimation for data W is $\overline{\mu} = (\alpha, \beta_1, .., \beta_{k/g_0})$

by MLE estimator. Predict the probability for each image as; $\pi_r(I_i)$.

B. *Weighted Mean :*

It calculates the mean of weight as follows :

$$WM(a_1,\ldots,a_N) = \Sigma_{i=1}^{N}\, p_i\, a_i \qquad (8)$$

Where, p is the score of i th information or image[19][20].

C. WOWA :

It is aggregation of weighted mean and OWA operation. It is calculated as shown below:

$$\mathrm{WOWA}(a_1,\ldots,a_N) = \Sigma_{i=1}^{N}\ \omega_i\ a_{\sigma(i)} \qquad (9)$$

where σ corresponds to a permutation of the a_i so that they are ordered from the largest one to the lowest one, and where &omega is defined by:

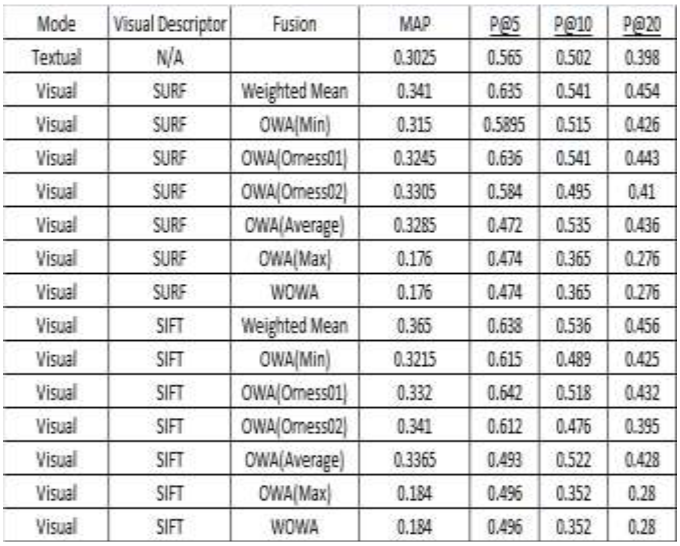
$$\omega_i = w^*(\Sigma_{j\leq i}\ p_{\sigma(j)}) - w^*(\Sigma_{j<i}\ p_{\sigma(j)})$$

where, as in OWA operator, w_i correspond to the weight of the ith data after ordering them (w is a weighting vector as in the weighted mean), and where w^* is a function that interpolates the points (i/n, $\Sigma_{j\leq i}p_j$) and the point (0,0), and that should be a straight line if the points can be interpolated in this way. The WOWA is calculates the fusion of two numerical values[19][20].

V. Experimental Setup and Result Analysis

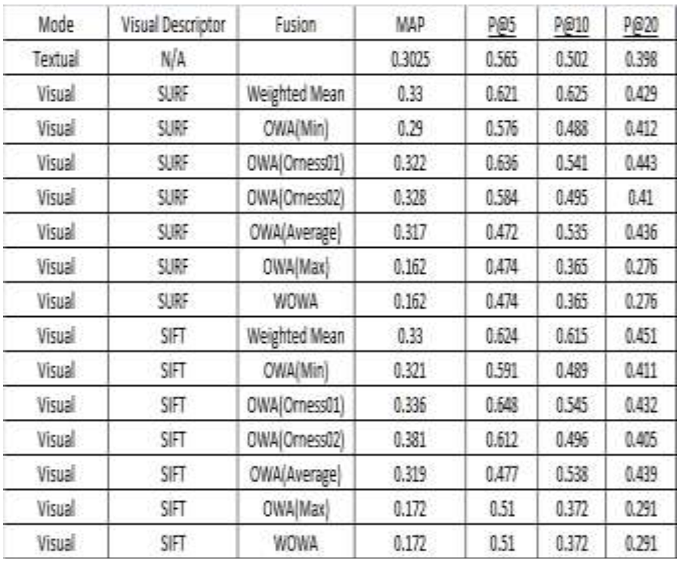
Precision is calculated and by using values of it Mean Average Precision(MAP) are also calculated. The proposed system gives result of fusion(Weighted Mean, OWA and WOWA) in textual and visual modes. Run ID specifies the number of runs that is query image uses for search. Visual descriptor SURF is used so it gives its result as precisions. Precisions calculated such as P@5, P@10 and P@20 that means number of documents or images(5,10,20) are seen that are calculated as shown in TABLE I and TABLE II. Graphical representation of fusion algorithm performance (MAP) using SIFT and SURF with and without textual prefiltering as shown in Fig. 4.

TABLE I. Precision of image search with surf interest points And fusion with textual prefiltering



Mode	Visual Descriptor	Fusion	MAP	P@5	P@10	P@20
Textual	N/A		0.3025	0.565	0.502	0.398
Visual	SURF	Weighted Mean	0.341	0.635	0.541	0.454
Visual	SURF	OWA(Min)	0.315	0.5895	0.515	0.426
Visual	SURF	OWA(Orness01)	0.3245	0.636	0.541	0.443
Visual	SURF	OWA(Orness02)	0.3305	0.584	0.495	0.41
Visual	SURF	OWA(Average)	0.3285	0.472	0.535	0.436
Visual	SURF	OWA(Max)	0.176	0.474	0.365	0.276
Visual	SURF	WOWA	0.176	0.474	0.365	0.276
Visual	SIFT	Weighted Mean	0.365	0.638	0.536	0.456
Visual	SIFT	OWA(Min)	0.3215	0.615	0.489	0.425
Visual	SIFT	OWA(Orness01)	0.332	0.642	0.518	0.432
Visual	SIFT	OWA(Orness02)	0.341	0.612	0.476	0.395
Visual	SIFT	OWA(Average)	0.3365	0.493	0.522	0.428
Visual	SIFT	OWA(Max)	0.184	0.496	0.352	0.28
Visual	SIFT	WOWA	0.184	0.496	0.352	0.28

TABLE II Precision of image search with surf interest points And fusion without textual prefiltering



Mode	Visual Descriptor	Fusion	MAP	P@5	P@10	P@20
Textual	N/A		0.3025	0.565	0.502	0.398
Visual	SURF	Weighted Mean	0.33	0.621	0.625	0.429
Visual	SURF	OWA(Min)	0.29	0.576	0.488	0.412
Visual	SURF	OWA(Orness01)	0.322	0.636	0.541	0.443
Visual	SURF	OWA(Orness02)	0.328	0.584	0.495	0.41
Visual	SURF	OWA(Average)	0.317	0.472	0.535	0.436
Visual	SURF	OWA(Max)	0.162	0.474	0.365	0.276
Visual	SURF	WOWA	0.162	0.474	0.365	0.276
Visual	SIFT	Weighted Mean	0.33	0.624	0.615	0.451
Visual	SIFT	OWA(Min)	0.321	0.591	0.489	0.411
Visual	SIFT	OWA(Orness01)	0.336	0.648	0.545	0.432
Visual	SIFT	OWA(Orness02)	0.381	0.612	0.496	0.405
Visual	SIFT	OWA(Average)	0.319	0.477	0.538	0.439
Visual	SIFT	OWA(Max)	0.172	0.51	0.372	0.291
Visual	SIFT	WOWA	0.172	0.51	0.372	0.291

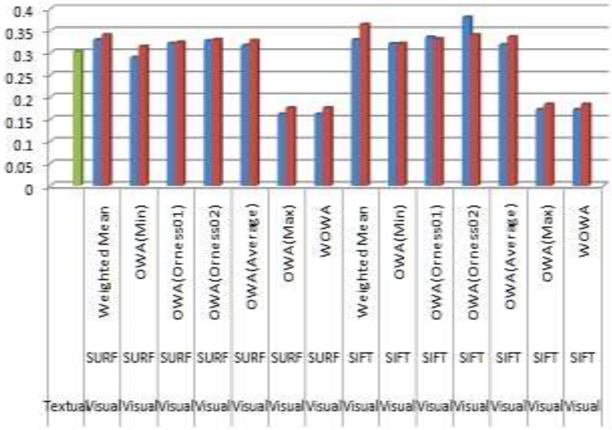


Fig. 4. Fusion algorithm performance (MAP) using SIFT and SURF with(red right bars)and without(blue left bars) textual prefiltering

VI. Conclusion

Feature extraction is important step in image retrieval system. A Speeded Up Robust First (SURF) technique is proposed for feature extraction from high large image dataset. The OWA algorithm provides single output to finite number of inputs and also decides which weight corresponds to each input provided by an operator. The Weighted Mean, OWA and WOWA algorithms are used for fusion. Bag of Visual Words(BoVW) approach is retrieval algorithm uses indexing and classifies images and also used for object recognition. Interest points and descriptors are combined by SURF technique. By textual pre-filtering and image re-ranking; the system is giving accurate and efficient result of image retrieval. This system can be used for any application for searching images like medical database, intrusion detection, educational environment and business.

Acknowledgment

The author wish to thank their guide, parents, god and MET's Institute of Engineering Bhujbal Knowledge City Nasik, for supporting and motivating for this work because without their blessing this was not possible.

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, *IEEE Transactions On Multimedia*, vol. 15, No. 8, 2013.

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

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

[4] Nagham Hamid, Abid Yahya, R. Badlishah Ahmad, and Osamah M. Al-Qershi, “ A Comparison between Using SIFT and SURF for Characteristic Region Based Image Steganography” IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 3, May 2012.

[5] Herbert Bay, Tinne Tuytelaars, Luc Van Gool, ”SURF: Speeded Up
Robust Features, Computer Vision and Image Understanding (CVIU),
Vol. 110,No. 3,pp. 346-359,EECV,2008.

[6] Chandrika L, ”Implementation Image Retrieval and classification with
SURF Technique, International Journal of Innovative Science, Engineering and Technology, Volume 1, Issue 4, June 2014.

[7] J. A. AslamandM. Montague, ”Models for metasearch, in Proc. 24thAnnu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval, New Orleans, LA, USA, 2001, pp. 276-284.

[8] P. K. Atrey,M. A. Hossain, A. El Saddik, andM. S. Kankanballi, ”Multimedia Fusion for Multimedia Analysis: A Survey, Multimedia Syst., vol. 16, pp. 345 379, 2010.

[9] S. Clinchant, G. Csurka, and J. Ah-Pine, ”Semantic combination of textual and visual information in multimedia retrieval, Proc. 1st ACM Int. Conf. Multimedia Retrieval, New York, NY, USA, 2011.

[10] Gerald Kowalski, ”Information Retrieval Systems Theory and Implementation, The Kluwer international series on information retrieval ; 1,ISBN 0-7923-9926-9 (alk. paper).

[11] S. A. Chatzichristofis, K. Zagoris, Y. S. Boutalis, and N. Papamarkos, ”Accurate image retrieval based on compact composite descriptors and relevance feedback information, Int. J. Pattern Recog. Artif. Intell., vol. 24, no. 2, pp. 207 244, 2010.

[12] 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. 119, 2006.

[13] R. Granados, J. Benavent, X. Benavent, E. de Ves, Ana Garca-Serrano, ”Multimodal information approaches for the Wikipedia collection at ImageCLEF 2011, in Proc. CLEF 2011 Labs Workshop, Notebook Papers, Amsterdam, The Netherlands,2011.

[14] M. Montague and J.A. Aslam, ”Condorcet fusion for improved retrieval, in Proc 11th Int.,Conf. Inf, Knowledge Manage, McLean, VA, USA, 2002, pp. 538-548.

[15] Stephen Robertson, Hugo Zaragoza, ”The Probabilistic Relevance Framework: BM25 and Beyond, Foundations and Trends in Information Retrieval Vol. 3, No. 4 (2009) 333389, 2009.

[16] Qingjian Zhou,Jia Jiao, ”The Ordered Weighted Averaging Algorithm to Multiple Attribute Decision Making within Triangular Fuzzy Numbers, Applied Mathematical Sciences, Volume 8, no. 56, 2763 - 2766, Vol. 8, 2014.

[17] Apache Lucene. http://lucene.apache.org

[18] Ke Gao, Shouxun Lin, Yongdong Zhang, Sheng Tang, Huamin Ren,”Attention Model Based SIFT Keypoints Filtration for Image Retrieval, Seventh IEEE/ACIS International Conference on Computer and Information Science,May 2008.

[19] V. Torra, The Weighted OWA operator, Int. J. of Intel. Systems, 12 (1997) 153-166. [20] V. Torra, The WOWA operator and the interpolation function W*: Chen and Otto's interpolation method revisited, Fuzzy Sets and Systems, 113:3 (2000) 389-396.