

## On the Design of a Content based Image Retrieval System

S. Farhan, M.A. Fahiem and H. Tauseef\*

Department of Computer Science, Lahore College for Women University, Lahore, Pakistan

### ARTICLE INFO

*Article history:*

Received: 19 July, 2018

Accepted: 15 June, 2019

Published: 21 June, 2019

*Keywords:*

Automated image tagging,

CBIR,

Computer vision,

Distance metric

### ABSTRACT

*With the abundance of multimedia content on the World Wide Web, research and learning of effective feature representation and similarity measures have become crucial. Image searching poses several challenges. Lately, many researchers have been exploring the field. Automatic annotation of images based on digital content processing proves to be an encouraging direction in the field. Content based image retrieval system development is an emerging field. Accuracy of the results of semantic search depends on the understanding of searcher's purpose, the meaning of conditions imposed in the search query and their mapping in the searchable data space. A visual content semantic search engine is proposed in this paper. The search engine employs digital image features for searching the image database. The presented algorithm produces promising results. The performance of our algorithm is tested on an extensive set of tags and queries resulting in accurate and efficient results.*

## 1. Introduction

In recent years, advancement in digital photography and consumer technology has led to popularity of personal photography and web based digital image collection. New challenges are faced for access, search and retrieval of images because of the ease in image sharing and retrieval. The main reason behind it is the huge growth in the volume of multimedia content over social media networks and instant messaging applications. 'Tags' has been used to annotate images for grouping, indexing and retrieval purposes but manually tagging an enormous ocean of images is a time consuming task. Moreover, improper tagging affects the search results [1]. Therefore, there is a need for an automated system that can tag images for the purpose of searching, organizing and browsing images accurately.

Searching through metadata usually presents irrelevant results. Tags can also be used for generating image description which will be helpful for semantic searching. Current content based search engines are a light in the dark but these systems still suffer from the gap between the visual image features and human lexical semantics. The extraction of low level features only for the significant semantic interpretation is not enough. Utilization of system generated descriptions enables content based search engines to produce better results.

Retrieval of relevant images from diverse image collections through World Wide Web is a challenging task. In image retrieval systems, there are two diversified research directions. Conventional information retrieval pursues information systems based on text and image indexing and retrieval is done employing numerous techniques and processes. The other significant direction of image retrieval systems is in the domain of computer vision by utilizing different techniques and algorithms for analysis and indexing

of images, relying on their visual content, for instance, color, grain, contour etc.

Content-Based Image Retrieval (CBIR) systems [2] are demarcated into various ranks of conceptualization. Image retrieval based on primitive features (low level features), lies at the lowest level [3-5]. Higher levels go into details of upper level of semantic characteristics, such as particular objects, their types and lastly events [6]. When differentiating the level wise features, the low level features are usually most popular because they are not complex in nature. Moving forward to the higher-level features, the complexity increases. Among many lower level features, color is most commonly used one. But as far as extraction of reliable results using CBIR technology is concerned, standalone color feature is not enough for image description.

Images, when serve as a communications medium; hold semantic information, e.g. 'a laughing child'. Aforesaid semantic features are a must for semantic level queries. Current CBIR techniques [7, 8] are unable to extract the semantic features proficiently, although primitive feature queries can be used to represent a few semantic level queries. A semantic-level query that searches for images of a green field can be represented by a primitive feature query. Such a query retrieves images that have the green color on bottom. Use of only primitive features have limited the capabilities of current CBIR systems as they fail to satisfy most semantic level query requirements. Although CBIR systems suffer from the mentioned distinctive shortcoming, its pragmatic outcomes still presents encouraging feedback, which motivates us for further research in this field with the hope of achieving desirable results.

\*Corresponding author: humaiftikhar@hotmail.com

Surveys on CBIR can be found elsewhere [6, 9, 10], that give a detailed insight of techniques existing in this domain including their strengths and limitations. Hierarchical, cure data and fusion algorithm clustering algorithms are effective and offer efficient image retrieval [9]. A new feature descriptor proposed by Gonde et al. [11] supports the use of frequency domain features for CBIR. It uses 3D local transform pattern and finds that the results are better as compared to spherical symmetric three dimensional local ternary pattern and recent methods. A prototype system Image Scape, issued for searching visual media over WWW. It accomplishes quick searching and browsing across high-dimensional spaces by utilizing vector quantization based image database compression and k-d trees [12]. Local features can be aptly used within the context of multiple image search and learning problems. Such a Pyramid Match Algorithm is defined [13] that demonstrates how to use local features for searching and learning. The pyramid match asserts on the matching between image local features to provide an effective and efficient method for the assessment of the similarity among images. Low level image features can be correlated to semantic properties and image content can be mapped to ontological data present on the WWW. To overpower the semantic gap between low and high level features, a system is proposed that characterizes the visual person ontology [14].

Use of machine learning techniques has gained popularity in recent years for the improvement of CBIR systems. A comprehensive study is conducted by Wan et al. [15] to investigate the use of deep learning methods in this regard and reported some encouraging directions. Audio, visual and text features are used to extract cues for semantic indexing. A similar knowledge-based approach is presented elsewhere [16] for compact disk content semantic indexing. A Knowledge-discovery approach is proposed by utilizing Content-based association rules [17]. Clustering based on hierarchical approach in large image databases is a renowned approach. It produces faster image retrieval with efficient and more relevant results. K-Means is a famous clustering method and is significant for its precision and efficacy in image retrieval [18].

Instead of using features from spatial and frequency domain separately, their combination is also investigated. Wavelet transform integration, Local Binary Patterns and moments for CBIR yielded promising results [19]. Combination of different types of features and their integration with the target to enhance such systems is also a hot research area these days. An overview of the techniques used in CBIR for fusion of different descriptors at different levels is given by Piras and Giacinto [20]. A hybrid features based CBIR [21], incorporating various spatial and frequency domain features using a number of distance metrics, reported that frequency domain features improved the results over spatial features. Search Images Effectively through Visual Elimination (SIEVE) is an improvement for the presentation of complete text-based web image search. It is an integration of the presented text-based image search engine with visual

features and is proposed by Liu et al. [22]. It is a substantial and considerable advancement over Google image search when taking into account retrieval precision and efficacy.

The current advancements and considerable challenges and issues that are faced by search associated to semantic gap in low level and high level features in images and videos are reported previously [23]. Where leading techniques, recent results and conclusion, and authentic applications and tools in semantic search and multimodal retrieval models are also reviewed. Acquisition of an enormous image dataset along with related descriptive features and development of image indexing and searching methods that are capable of balancing the considerable size of data are the two pivotal challenges of scalability relating to CBIR as discussed by Batko et al. [24]. It relies on the metric space model based on similarity and the philosophy of methodical peer-to-peer networks. The storage of unique images as thumbnails aid the search engines, especially developed for efficient and accurate representation of the output summary. Challenges and difficulties of CBIR systems that are encountered during recent years are addressed previously [25]. Issues that emerged during segmentation comprise of edge, boundary, region, color, texture and shape based feature extraction, object detection and identification. The main issues of semantic gap are dealt by the state of the art techniques.

## 2. Material and Methods

The images of famous personalities or movie celebrities are collected from multiple sources including internet. The images that contain only a celebrity from head to chest are selected for inclusion in the dataset. The system consists of two phases: (1) processing and storage of images and (2) searching and retrieval from database. The details of both phases are provided here and shown in Fig. 1.

### 2.1 Processing and Storage of Images

The main purpose of processing is automatic image annotation. A data dictionary is defined that contains a set of key terms for objects most commonly observed in celebrity images as shown in Table 1. These terms are used for tagging the images.

Processing task consists of two stages i.e. training and testing. Each image in the training set is divided into three Region of Interests (ROIs) as shown in Fig. 2. The details of automated ROI extraction can be found elsewhere [26]. Different regions are extracted by detecting feature points on the silhouette of head to chest image of a celebrity. The technique uses a shape coding algorithm to represent the human body contour in images. The first or upper ROI is used for face region. The second or middle ROI is used for the neck region (below the face but above the chest). The third or lower ROI is used for torso or chest. Further each ROI is segmented into different objects (paired with key terms/tags) and a feature set based on texture, color, shape and position is calculated for each object using MaZda code [27]. A Neural

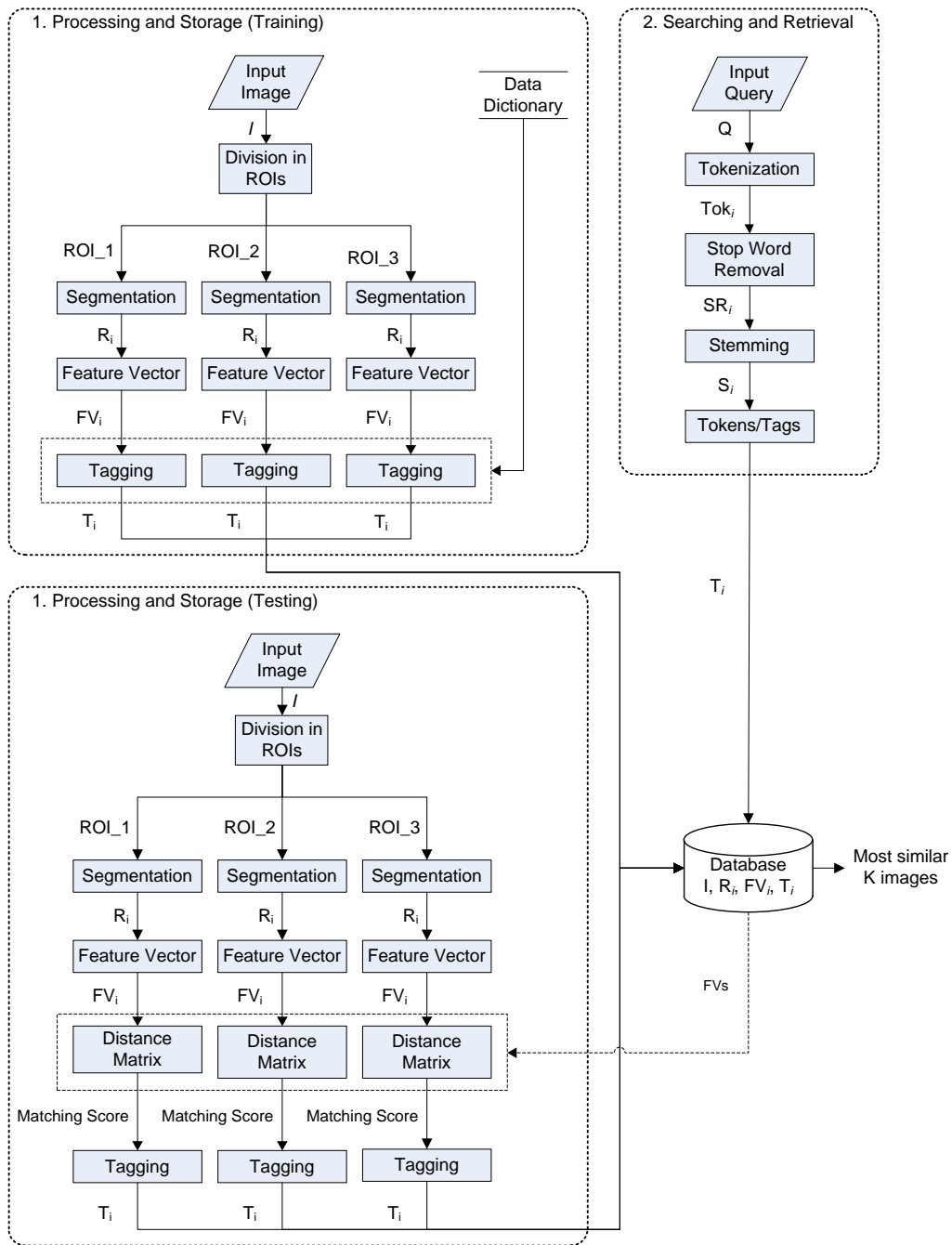


Fig. 1: Flow of work for the proposed approach.

Table 1: Data Dictionary defined for the proposed approach.

ROIs	Region	Key Terms/Tags
ROI_1	Face	Hair, Glasses, Clip, Band, Scarf, Earrings, Moustache, Beard, Hat, Cap, Flower, Hair_color, Skin_tone
ROI_2	Neck	Scarf, Flower, Necklace
ROI_3	Chest	Tie, Shirt, Bow, Coat, Jacket, Pen, Scarf, Brooch, Necklace, Pocket_square, Flower, Ribbon, Dress_color, Buttons, Tie_pin, Tack_pin

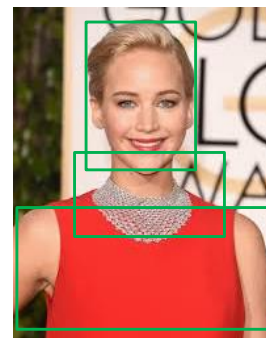


Fig. 2: Dividing image into 3 ROIs.

Network (NN) is trained to assign tags to each segment in each ROI based on their expected position and key features. Images, associated feature sets and assigned tags are stored in the database. The NN training is done using (error) Back Propagation (BP). The connecting weights between the layers are updated during this training process. The adaptive updating of the NN is continued till satisfactory results are achieved. This training of the NN is done in two passes. One is a forward pass and the other is a backward calculation with error determination.

During the testing stage, input image is divided into three ROIs and each ROI is segmented into objects as is done in training stage. Further feature set is calculated for each object. Next matching of the feature set is done with all instances in the database using Euclidean distance metrics. Depending on the matching score, tags are assigned to the input image. Image, associated feature vector and assigned tags are stored in the database.

### 2.2 Searching and Retrieval from the Database

Input to the system comes in the form of a text based query, which requires natural language processing to extract meaningful words. These words are then matched against annotations/tags stored with each image in the database. To extract such meaningful words from a user query, a number of preprocessing steps are performed including word extraction (tokenization) (Fig. 3), stopwords removal (Fig. 4) and stemming (Fig. 5). In first step, input query is tokenized and words are extracted from it. In second step, stopwords removal is done. Stopwords are usually the words with high frequency that are not providing any additional information regarding tags. The purpose for stopwords removal is to improve efficiency as stopwords are not useful for searching. In English, stopwords are like 'the', 'a', 'is', 'which', 'that', 'in' etc. Input query is processed for removal of three types of stop words; (1) determiners ('the', 'an', 'another' etc.), (2) coordinating conjunctions ('for', 'an', 'nor', 'but', 'or' etc.) and (3) prepositions ('in', 'under', 'towards' etc.).

```

Tokenize (Q)
Input:
    Q: User provided query
Output:
    Tok: List of tokens
Algorithm:
    Initialize Stream=Q, Curr_Pos= Start of Stream
    Initialize Delimeter = {' ' . ; : ! ? ( ) < > + - * /}
    Initialize whitespace={\n \endf space}, i=0
    While Curr_Pos != EOF
    {
        While Curr_Pos!= Delimeter OR Curr_Pos!=
        whitespace
        {
            Append value(Curr_Pos) to Tok[i]
            Curr_Pos ++
        }
        i++
    }
    Return Tok
    
```

Fig. 3: Tokenization of the query.

```

Stopword_removal (Tok)
Input:
    Tok: List of tokens
Output:
    S: List of selected tokens
Algorithm:
    Initialize Determiners = { the, an, another, ...}
    Initialize Coordinating_conjunctions = { for, an, nor, but, or,
    yet, so,...}
    Initialize Prepositions = { in, under, towards, before, ...}, i=0
    While !endof Tok
    {
        If Tok[i] != Determiners AND Tok[i]
        !=Coordinating_conjunctions
        AND Tok [i] != Prepositions
        Append Tok [i] to S
        i++
    }
    
```

Fig. 4: Stop word removal from tokens.

```

Stemming (S)
Input:
    S: List of selected tokens
Output:
    T: Term List
Algorithm:
    Stemming with Porter Stemming Algorithm []
    Append all words after stemming to T
    
```

Fig. 5: Stemming of tokens.

After stop word removal, last step is stemming [28], which is the process for reduction of inflected words to their root form. For example, 'fishing', 'fished', 'fisher' all have same stem, i.e., 'fish'. Stemming improves effectiveness of finding matching similar words. After preprocessing of input query, the resultant words are matched with stored tags for each image. Images with highest similarity index are displayed as output (Fig. 6).

```

Search (T, D)
Input:
    T: Term List
    D: Set of images, ROIs with tags in database
Output:
    K: List of matching images
Algorithm:
    Initialize the Matching images K= { }, i=0
    For all d ∈ D Do
    {
        For all t ∈ T Do
        {
            if t == d.tags [i]
            calculate d.similarityindex
            Append d to K
        }
    }
    Arrange K descending order (d.similarityindex)
    Return first five from K
    
```

Fig. 6: Semantic searching.





Image					
Original Annotation	blond hair, glasses, earrings, scarf, black, white dress, light skin tone.	black hair, glasses, hat, beard, shirt, tie, blue dress, medium skin tone.	brown hair, earrings, necklace, black dress, medium skin tone.	black hair, glasses, tie, tie pin, pocket square, flower jacket, shirt, black and white dress, light skin tone.	Black hair, earrings, tie, shirt, jacket, black and white dress, dark skin tone.
Automatic Annotation by Proposed System	blond hair, glasses, scarf, light skin tone.	black hair, glasses, hat, tie, shirt, blue dress, medium skin tone.	brown hair, necklace, black dress, medium skin tone.	black hair, glasses, tie, tie pin, pocket square, flower, black and white dress, light skin tone.	Black hair, tie, shirt, jacket, black and white dress, dark skin tone.

Fig. 7: Annotation of different images.

### 3. Results and Discussion

We developed our own dataset consisting of 3000 images of celebrities collected from the internet (by crawling google for images). The images contain only a single celebrity from head to chest. Dividing each image into three ROIs facilitates easy searching of the images. Segmentation is done for each ROI. Each image is assigned tags using NN based on feature vectors. The dataset is divided into two parts: 2,500 images in the training set and 500 images in the testing set.

The annotation performance is evaluated by retrieving images from the database using the stored tags. The relevance of the retrieved images can be easily judged by looking at the retrieved images (Fig. 7).

For a query  $Q$ , the appropriate and most relevant images are those that possess all tags extracted from  $Q$ . We have used recall, precision and F-measure as evaluation metrics. Recall measures the ratio of correctly retrieved images over the total number of relevant images (Eq. 1). Precision is the ratio of correctly retrieved images over the total number of images retrieved (Eq. 2). F-measure combines recall and precision to calculate the efficiency of a system (Eq. 3).

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$F = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (3)$$

Here

TP defines the number of correctly classified positive examples (True positives),

FN defines the number of incorrectly classified positive examples (False Negatives),

FP defines the number of incorrectly classified negative examples (False Positives)

Recall, precision and F-measure obtained for a few sample tags used as queries are shown in Table 2.

Table 2: Results achieved using different tags in proposed system.

Tags	Recall	Precision	F-measure
Tie	0.85	0.63	0.72
Hairclip	0.81	0.72	0.76
Man, Jacket	0.90	0.58	0.71
Woman, Necklace	0.87	0.67	0.76
Man, Tie, Tiepin	0.78	0.57	0.66
Woman, scarf, glasses	0.80	0.65	0.72

Searching is done on the basis of the ROI's already defined. The proposed system automatically annotates the images and stores them and their semantic properties in the database. Searching is done only in the stored images and the semantic properties defined in the database are considered. User may search whatever is in his/her mind by giving a query related to the semantic properties of the image for semantic searching. Semantic properties based search is done purely on the low and high level features. The query entered for searching images by the user can be like "man wearing black tie", "female having black hair". Searched images are ranked according to the scores calculated. Top five images are displayed to the user. The results of the search as retrieved images for sample queries are illustrated in Figs. 8-10.





Fig. 8: Retrieval in response to the text query "Tie".



Fig. 9: Retrieval in response to the text query "Man with Jacket".

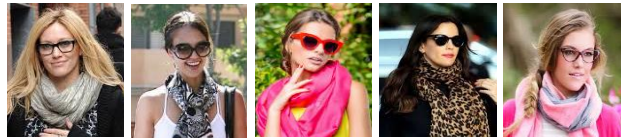


Fig. 10: Retrieval in response to the text query "Woman with Scarf and Glasses".

#### 4. Conclusions

A system is proposed to annotate and search images based on semantics. The results achieved are quite promising. The proposed system has been tested and validated for performance. Searching large number of images and tags associated is a challenge for performance of image annotation and retrieval systems. Improved feature extraction will further enhance the performance. Manual tagging can be merged to facilitate and improve the searching process.

#### References

[1] I.K. Sethi, I.L. Coman and D. Stan, "Mining association rules between low-level image features and high-level concepts", *Data Mining and Knowledge Discovery: Theory, Tools and Technology III*, vol. 4384, pp. 279-290, 2001.

[2] V.N. Gudivada and V.V. Raghavan, "Content based image retrieval systems", *Computer*, vol. 28, no. 9, pp. 18-22, 1995.

[3] C.H. Lin, R.T. Chen and Y.K. Chan, "A smart content-based image retrieval system based on color and texture feature", *Image and Vision Computing*, vol. 27, no. 6, pp. 658-665, 2009.

[4] P.S. Hiremath and J. Pujari, "Content based image retrieval using color, texture and shape features", *Proc. of the 15<sup>th</sup> Int. Conf. on Advanced Computing and Communications*, IEEE, Guwahati, India, December 18-21, 2007, California, pp. 780-784, 2007.

[5] H. Yu, M. Li, H.J. Zhang and J. Feng, "Color texture moments for content-based image retrieval", *Proc. of the Int. Conf. on Image Processing*, Rochester, NY, USA, September 22-25, 2002, IEEE, pp. 929-932, 2002.

[6] Y. Liu, D. Zhang, G. Lu and W.Y. Ma, "A survey of content-based image retrieval with high-level semantics", *Pattern Recognition*, vol. 40, no. 1, pp. 262-282, 2007.

[7] D. Papadias, M. Mantzourgiannis, P. Kalnis, N. Mamoulis and I. Ahmad, "Content-based retrieval using heuristic search", *Proc. of the 22<sup>nd</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR Berkeley, CA, USA, August 15-19, 1999, New York: ACM, pp. 168-175, 1999.

[8] J.S. De Bonet and P.A. Viola, "Structure driven image database retrieval", *Proc. of Conference on Advances in Neural Information Processing Systems*, Denver, CO, USA, December 2-4, 1997, M.I. Jordan, M.J. Kearns, S.A. Solla, Eds. USA: MIT Press Cambridge, pp. 866-872, 1997.

[9] S. Surya and G. Sasikala, "Survey on content based image retrieval", *IJCSE*, vol. 2, no. 5, pp. 691-696, 2011.

[10] R.C. Veltkamp and M. Tanase, "Content-based image retrieval systems: A survey", *Utrecht University, Netherlands*, pp. 1-62, 2002.

[11] A.B. Gonde, S. Murala, S.K. Vipparthi, R. Maheshwari and R. Balasubramanian, "3D Local Transform Patterns: A New Feature Descriptor for Image Retrieval", *Proc. of Int. Conf. on Computer Vision and Image Processing*, Roorkee, UP, India, September 9-12, 2017, B.B Chaudhuri, M.S. Kankanhalli, B. Raman, Eds. Singapore: Springer, pp. 495-507, 2017.

[12] M.S. Lew, "Next-generation web searches for visual content", *Computer*, vol. 33, no. 11, pp. 46-53, 2000.

[13] K. Grauman, "Efficiently searching for similar images", *Communications of the ACM*, vol. 53, no. 64, pp. 84-94, 2010.

[14] E. Kim, X. Huang and J. Heflin, "Finding VIPs-A visual image persons search using a content property reasoner and web ontology", *Proc. of IEEE Int. Conf. on Multimedia and Expo, ICME 2011, Barcelona, Spain, July 11-15, 2011*. New York: IEEE, pp. 1-7, 2011.

[15] J. Wan, D. Wang, S.C.H. Hoi, P. Wu, J. Zhu, Y. Zhang and J. Li, "Deep learning for content-based image retrieval: A comprehensive study", *Proc. of the 22<sup>nd</sup> ACM Int. Conf. on Multimedia*, Orlando, FL, USA, November 3-7, 2014. New York: ACM, pp. 157-166, 2014.

[16] W. Adams, G. Iyengar, C.Y. Lin, M.R. Naphade, C. Neti, H.J. Nock and J. R. Smith, "Semantic indexing of multimedia content using visual, audio and text cues", *EURASIP J. Adv. Sig. Pr.*, vol. 2003, no. 2, pp. 170-185, 2003.

[17] A.S. Barb and C.R. Shyu, "Visual-semantic modeling in content-based geospatial information retrieval using associative mining techniques", *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 1, pp. 38-42, 2010.

[18] V. Murthy, E. Vamsidhar, J.S. Kumar and P.S. Rao, "Content based image retrieval using Hierarchical and K-means clustering techniques", *IJEST*, vol. 2, no. 3, pp. 209-212, 2010.

[19] P. Srivastava and A. Khare, "Integration of wavelet transform, local binary patterns and moments for content-based image retrieval", *J. Vis. Commun. Image R.*, vol. 42, pp. 78-103, 2017.

[20] L. Piras and G. Giacinto, "Information fusion in content based image retrieval: A comprehensive overview", *Inform. Fusion*, vol. 37, pp. 50-60, 2017.

[21] Y. Mistry, D. Ingole and M. Ingole, "Content based image retrieval using hybrid features and various distance metric", *Journal of Electrical Systems and Information Technology*, vol. 5, no. 3, pp. 1-15, 2017.

[22] Y. Liu, D. Zhang and G. Lu, "SIEVE- Search images effectively through visual elimination", *Proc. of Int. Workshop on Multimedia Content Analysis and Mining*, Weihai, China, June 30- July 1, 2007, N. Sebe, Y. Liu, Y. Zhuang, Th. S. Huang, Eds. Berlin: Springer, pp. 381-390, 2007.

[23] S.-F. Chang, W.-Y. Ma and A. Smeulders, "Recent advances and challenges of semantic image/video search", *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Honolulu, HI, USA, April 15-20, 2007, New York, pp. IV-1205-IV-1208, 2007.

[24] M. Batko, F. Falchi, C. Lucchese, D. Novak, R. Perego, F. Rabitti, J. Sedmidubsky and P. Zezula, "Building a web-scale image similarity search system", *Multimedia Tools and Applications*, vol. 47, no. 3, pp. 599-629, 2010.

[25] D.G. Thakore and A. Trivedi, "Content based image retrieval techniques- Issues, analysis and the state of the art", *Proc. of the Int. Symp. on Computer Engineering & Technology*, Mandi Gobindgarh, Punjab, India (March 19-20, 2010), pp. 1-5, 2010.

[26] Y.L. Lin and M.J.J. Wang, "Automated body feature extraction from 2D images", *Expert Systems with Applications*, vol. 38, no. 3, pp. 2585-2591, March 2011.

[27] P.M. Szczypiński, M. Strzelecki, A. Materka and A. Klepaczko, "MaZda- A software package for image texture analysis", *Comp. Meth. Prog. Bio.*, vol. 94, no. 1, pp. 66-76, April 2009.

[28] M.F. Porter, "An algorithm for suffix stripping", *Program*, vol. 14, no. 3, pp. 130-137, 1980.