Received: August 3, 2018

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Dictionary Learning Features Based Cross Media Retrieval Using Minkowski Distance

Alugoju Sreelatha [1]* **Chagunta Venkata Guru Rao**[2] **Porika Sammulal**[3]

[1]*Department of Computer Science,*
University Arts & Science College, Kakatiya University, Warangal (Urban), India
[2]*S.R. Engineering College, India*
[3]*Department of Computer Science Engineering,*
Jawaharlal Nehru Technological University, Hyderabad, India
* Corresponding author's Email: sreelathaalugoju@gmail.com

Abstract: Cross Media Retrieval (CMR) is one of the emerging research areas in the field of internet services and multimedia technology. The primary objective of this paper is to develop an effective dictionary learning methodology to reduce the semantic gap between the low-level and high-level features. In this experimental research, the image and text data in the Wikipedia dataset are separated, after separating the image and text data, feature extraction is performed individually. A suitable combination of feature extraction methods (normalized histogram colour feature and bag-of-words) are under-taken for image and text feature extraction. The normalized histogram colour feature has very quick generation, compared to other feature vectors. In addition, Bag-of-words measures the repetition of the words in the bag and retrieves the data similar to the query data. Also, it removes the conjunction words and the remaining words are named as keywords. These keywords are used as the query text and it compared to the proposed dictionary learning. Then, retrieves the text and image related to the keywords based on Minkowski distance measure. Finally, the experimental outcome shows that the proposed approach delivers better performance in terms of Mean Average Precision (MAP) value, retrieval efficiency, precision and recall. The proposed methodology improved the MAP value up to 0.24-0.20 compared to the existing methods.

Keywords: Bag-of-words, Cross media retrieval, Mean average precision, Minkowski distance, Normalized histogram colour feature.

1. Introduction

In recent decades, the information contents such as text, video and audio are easily generated by everyone, due to the development of internet techniques and digital capturing system [1]. So, the search for the required data in the large stored data, increases the difficulty and consumes more time [2]. At the same time, the number of digital images per day is also growing exponentially at an alarming rate, with the help of digital cameras or other devices [3]. To overcome these difficulties, most of the researchers focus on the retrieval of information such as, text retrieval, image retrieval, video retrieval and audio retrieval [4, 5]. For instance, content based image retrieval and text based image retrieval systems are used to retrieve the images from the database based on the query image or query text, which are processed by the similarity measurement [6]. These traditional image retrieval systems are based on the images with manual text annotation, and it is a keyword search, which consumes more man-power and resource materials. To avoid these difficulties, a new technology field is introduced named as CMR technology [7, 8].

CMR technology retrieves the various media data such as text, audio, image and video by using a single query. Lots of methods are available in the CMR system such as feature extraction, indexing, similarity measure, etc., for retrieving the relevant data based on user query [9-11]. In this experimental research, CMR system is performed on the reputed dataset: Wikipedia database. The proposed methodology contains six phases such as, data acquisition, pre-processing, feature extraction, similarity measure, dictionary learning and testing phase. In the first phase, the image and text data in the Wikipedia dataset are separated, after separating the image and text data, feature extraction is accomplished by employing normalized histogram colour feature and bag-of-words methods, which helps to retrieve the relevant keywords. These feature extraction methods are simple, which helps to reflect the depth and smoothness of image structure and also the feature locations are selected intelligently. Finally, these keywords are utilized as the query text and it matched with the proposed dictionary learning. Then, retrieves the text and image related to the keywords based on Minkowski distance measure.

Proposed dictionary learning: In training phase, two types of dictionaries are created such as, Text to Image (T2I) and Image to Text (I2T) dictionary. In T2I dictionary, text is considered as index and images are denoted as value. For an individual text, similar image features are learned as values. Correspondingly, in I2T dictionary, image is considered as index and text are represented as value. For an individual image, similar text features are learned as values. In testing phase: when an individual text or image data is tested, corresponding top k images or text are retrieved based on the similarity using minkowski distance measure. Experimental results on Wikipedia dataset shows good retrieval performance in terms of MAP, precision and recall compared to other existing approaches in CMR system.

This paper is composed as follows. Section II presents a broad survey of several recent papers on CMR strategies. In section III, Minkowski distance measure based dictionary learning is presented for effective CMR. In Section IV, comparative analysis is done for proposed and exiting methodologies. The conclusion is made in Section V.

2. Literature review

Several techniques are suggested by researchers in CMR system. In this scenario, a brief evaluation of some important contributions to the existing literatures is presented.

B. Jiang, J. Yang, Z. Lv, K. Tian, Q. Meng, and Y. Yan, [12] proposed a new approach named as deep learning for retrieving the relevant cross media (text and image). The proposed methodology concentrated on two phases such as feature extraction and distance detection. After obtaining the feature information, author sorts and eliminates the unnecessary feature vectors. The experimental research was performed on publicly available datasets (i.e. Wiki text image and NUSWIDE dataset) to validate its retrieval accuracy in terms of F-measure, precision and recall. In a few cases, the proposed method was not efficient in large scale database and also not robust in content interpretation of multimedia documents.

L. Huang, and Y. Peng, [13] addressed the problem of unified descriptive representation by proposing a new approach, fine grained correlation for CMR. The proposed approach initially constructed an entity level with fine-grained semantics between low-level features and high-level concepts. Then, by reducing the distance between media contents, a positive correlation entity level was achieved. The experimental outcome confirmed that the proposed methodology was more significant than existing approaches by means of retrieval rate. Whereas, the correlation between the entity levels were not fully exploited and also provide inefficient results for lower amount of data.

J. Yan, H. Zhang, J. Sun, Q. Wang, P. Guo, L. Meng, W. Wan, and X. Dong, [14] illustrated a new methodology named as joint graph regularization based CMR for addressing the problems of intra and inter modality similarities. In this literature, the proposed approach learns dissimilar projection matrices based on the intra and inter modality similarities for retrieving the different tasks. Extensive experiments were conducted and the efficiency of the proposed method was verified using Wikipedia, pascal sentence, INRIA-web search datasets. The complexity of the frame work is increased by considering both inter and intra modal similarities in a united frame work

H. Zhang, Y. Liu, and Z. Ma, [15] proposed a new CMR approach for multimedia data (image and audio) using the kernel based approach. At first, the samples of multimedia data were mapped into isomorphic feature sub-space and then extracts the features of multimedia data using linear regression approach. Using the extracted features, the semantic gap between the audio and images were calculated. The experimental outcome confirmed that the proposed methodology was more significant than existing approaches by means of recognition rate. The proposed approach was only applicable for datasets were written in well-organized language. Otherwise, it leads to inconsistency with realistic applications.

L. Xie, P. Pan, and Y. Lu, [16] combined the both latent and manifest combinative cross-modal semantic generation mode methodologies for constructing semantic correlation for CMR. The experimental research was performed on publicly available datasets (i.e. MIR Flickr show and Wikipedia featured articles) to validate its retrieval accuracy in terms of recall. Proposed approach does not give experimental results for Tri-Space and Ranking (TSR), TSR_{txt} and TSR_{img} in an acceptable time and also required high computational cost to implement.

To overcome the above mentioned drawbacks, Minkowski distance based dictionary learning methodology is proposed to enhance the performance of CMR system.

3. Proposed approach

Multi-media retrieval plays a crucial role in big-data usage. Existing researches focused only on single media retrieval. Whereas, the representation of media-type is inconsistent, so a new retrieval system: CMR system is developed. In order to further improve the retrieval efficiency in CMR system, a new dictionary learning methodology was proposed. The proposed CMR system consists of six steps such as, data acquisition, pre-processing, feature extraction, similarity measure, dictionary learning and testing phase. A general block diagram of CMR system is represented in the Fig. 1. The brief description about the proposed methodology is presented below.

3.1 Data acquisition and pre-processing

In the initial stage of CMR system, the media data (image and text) are taken from the standard benchmark datasets Wikipedia dataset, which is publicly available. Check the following link for Wikipedia dataset. "http://www.svcl.ucsd.edu/projects/crossmodal/"

It is the most widely used datasets for CMR system, which is based on Wikipedia "featured articles", a continually updated article collection. There are totally 29 categories in "featured articles", but only 10 most populated ones are actually considered. Each featured article is divided into numerous sections based on its headings, and this database is finally generated as a set of 2,866 image and text pairs. The sample data collection of Wikipedia dataset is represented in the Fig. 2. The acquired media data (image and text) in Wikipedia dataset are separated, after separating the image and text data, feature extraction is performed.

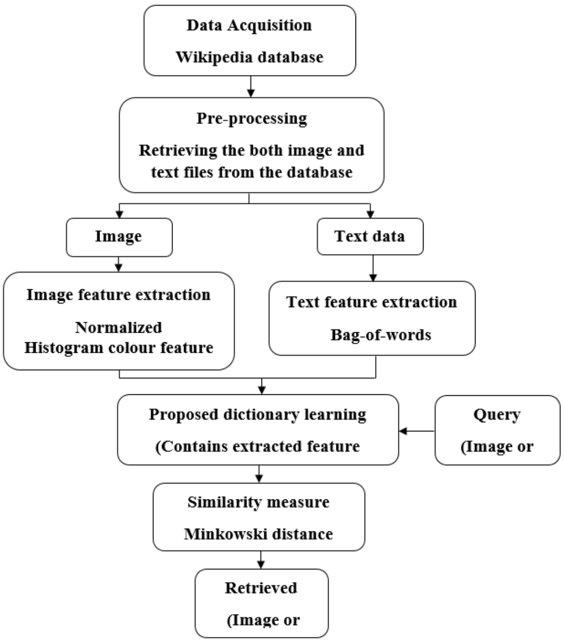
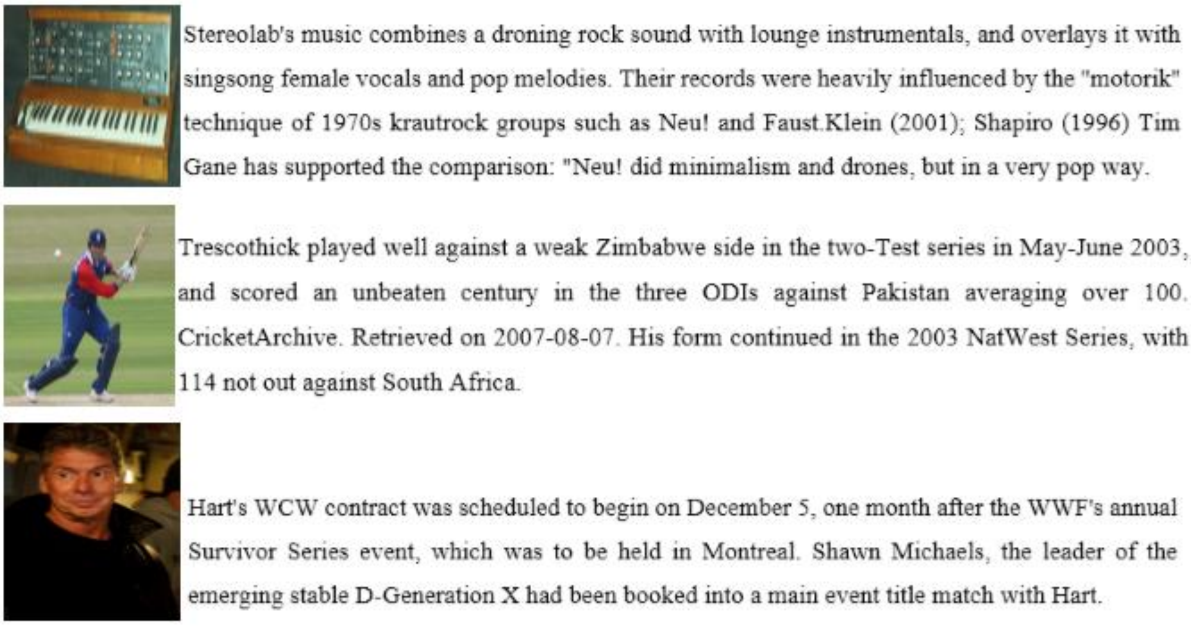


Figure.1 Block diagram of proposed methodology



Stereolab's music combines a droning rock sound with lounge instrumentals, and overlays it with singsong female vocals and pop melodies. Their records were heavily influenced by the "motorik" technique of 1970s krautrock groups such as Neu! and Faust.Klein (2001); Shapiro (1996) Tim Gane has supported the comparison: "Neu! did minimalism and drones, but in a very pop way.

Trescothick played well against a weak Zimbabwe side in the two-Test series in May-June 2003, and scored an unbeaten century in the three ODIs against Pakistan averaging over 100. CricketArchive. Retrieved on 2007-08-07. His form continued in the 2003 NatWest Series, with 114 not out against South Africa.

Hart's WCW contract was scheduled to begin on December 5, one month after the WWF's annual Survivor Series event, which was to be held in Montreal. Shawn Michaels, the leader of the emerging stable D-Generation X had been booked into a main event title match with Hart.

Figure.2 Sample data of Wikipedia dataset

3.2 Feature extraction

In this research methodology, the feature extraction is performed on the pre-processed data (text and image). Feature extraction is defined as the action of mapping the data from data space to the feature space. In this scenario, the normalized histogram colour feature is utilized for extracting the features from the images and the Bag-of-words methodology is utilized for extracting the features from the text.

3.2.1. Normalized histogram colour feature

The Normalized histogram colour feature is a commonly used feature extraction method in image retrieval. It is very popular, because color histograms are computationally effective to compute. In this experimental research, all the images are scaled to 256×256 in the both height and width. So, there will be 256 different bins of colours, then take a summation of these colour bins and find the standard deviation and mean for the values. The general formula for finding the mean and standard deviation represented in the Eqs. (1) and (2).

$$E_x = \sum_{y-1}^{n} 1/n^{p_{xyj}} \tag{1}$$

$$\sigma_x = \sqrt{1/n \sum_{y=1}^{n} (p_{xy} - E_x)^2} \tag{2}$$

Where, p_{xy} is represented as the colour value of the $y - th$ image pixel from the $x - th$ channel, n is specified as the number of pixels in the image, E_x is denoted as the average value of the $x - th$ channel.

3.2.2. Bag-of-words

Bag-of-words is the simplified representation of sentence or words, mostly utilized in the information retrieval system. Bag-of-words considers text as the bag, which contains words irrespective of the grammar and order of words, also it measures the repetition of the words. These texts are used to retrieve the data similar to the query data. Bag-of-words removes the conjunction words in the text and it considers remaining words as the keywords. These keywords are used as the query text and it compared to the dictionary. Then, it retrieves the text and image related to the keywords based on Minkowski distance measure.

3.3 Similarity measure

After extracting the features from the normalized histogram colour feature and bag-of-words, extracted features are converted into row vector called template that consist of media data (image and text) features. These extracted features are stored in the dictionary based learning. Similarly, the corresponding media data (image and text) to be tested for its retrieval is also converted into a template. Then, the tested data template is matched with the dictionary to retrieve the relevant media data based on user query.

3.3.1. Minkowski distance

The Minkowski distance is defined as the distance between two points in a normed vector space. The general formula for Minkowski distance is denoted in the Eq. (3).

$$D(x, y) = (\sum_{i=0}^{n-1} |x_i - y_i|^p)^{1/p} \tag{3}$$

Special cases:

- When $P = 1$, the distance is known as Manhattan distance
- When $P = 2$, the distance is known as Euclidean distance.

3.3.1.1. Euclidean distance

The Euclidean distance is used to determine whether the two templates are matching or not and it calculates the distance between the features in the two templates. The Euclidean distance between the two templates are calculated by employing the formula given in the Eq. (4).

$$Euclidean\ distance = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4}$$

Where, $x\ and\ y$ are represented as Euclidean vectors, and n is represented as the position of the Euclidean point

3.3.1.2. Manhattan distance

Manhattan distance is also called as L_1 Distance or Manhattan length, If $u = (x1, y1)$ and $v = (x2, y2)$ are the two points, then the Manhattan distance between u and v is calculated by using the formula (5).

$$Manhattan\ distance\ (u, v) = |x1 - x2| + |y1 - y2| \tag{5}$$

In case, instead of two dimensions, if the points have n -dimensions such as $u = (x1, x2 \ldots.., xn)$

and $v = (y1, y2 \dots., yn)$. Then, the Eq. (5) is generalized by defining the Manhattan distance between u and v are analyzed using the Eq. (6).

$$Manhattan\ distance\ (u, v) = |x1 - y1| + |x2 - y2| + \cdots |xn - yn| = \sum |xi - yi|,\ for\ i = 1,2, .. n \quad (6)$$

The distance between the two points on a grid is based on the horizontal and vertical path. The Manhattan distance is the simple sum of the vertical and horizontal components. Whereas, the diagonal distance is calculated by employing Pythagorean Theorem.

3.4 Proposed dictionary learning

After extracting the features from normalized histogram colour feature and bag of words. The extracted feature values are transformed into row vectors, which is named as template. These extracted feature vectors/templates are stored in the dictionary based learning. Then, the tested image or text is matched with dictionary based learning in order to retrieve the relevant image or text. The proposed dictionary based learning contains two types of dictionaries such as, T2I and I2T dictionary. In T2I dictionary, the text data are considered as index and the image data are denoted as collection of values. Here, the similar image features are learned as values for an individual text. On the other hand, in T2I dictionary, the text data are denoted as index and the image data are represented as values. Then, the similar text features are learned as values for an individual image.

While testing an individual image or text, corresponding top k images or text is retrieved by using the similarity measure: minkowski distance. The experimental validation of the proposed methodology is determined in the upcoming section.

4. Experimental outcome

For experimental simulation, Net-Beans (version 6.2) was employed on PC with 3.2 GHz with i5 processor. In order to estimate the efficiency of proposed algorithm, the performance of the proposed method was compared with Adaptive Heterogeneous Similarity Measure (AHSM) [19], Accumulated Reconstruction Error Vector (AREV) [18], Semantic Matching (SM), Correlation Matching (CM), Semantic Correlation Matching (SCM) [17], deep learning [12], and Distance-Preserving Entity Projections (DPEP) [13] on the reputed dataset (Wikipedia dataset). Theoretical explanation about the existing methodologies are described below,

N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, and N. Vasconcelos, [17] proposed cross-modal matching algorithms like CM, SM and SCM with an effective feature extraction methodology: Scale Invariant Feature Transform (SIFT) for retrieving the cross model documents. Finally, the correlation between the components were learned with canonical correlation analysis. Correspondingly, K. Liu, S. Wei, Y. Zhao, Z. Zhu, Y. Wei, and C. Xu, [18] proposed a new semantic representation methodology named as AREV for retrieving the relevant media data based on user query. Instead of directly learning correlation relationship among media data, the developed methodology shared original feature space among media types.

X. Zhai, Y. Peng, and J. Xiao, [19] proposed a new cross-media similarity measure (AHSM), which considers the both inter and intra-media correlation. Inter media correlation concentrates on negative and positive correlations between the dissimilar media types and intra media correlation concentrates on semantic category data within each media. For mining the intra-media correlation, heterogeneous similarity measure was developed with nearest neighbours. This measure computes the probability for two media objects, which belongs to the similar semantic category. For mining the inter-media correlation, a correlation propagation methodology was developed to deal with negative and positive correlation between the media objects of dissimilar media types.

4.1 Performance measure

The relationship between the input and output variables of a system understand by employing the suitable performance metrics like precision and recall. The general formula for calculating the precision and recall of CMR system is given in the Eqs. (7) and (8).

$$Precision = \frac{TP}{TP+FP} \times 100 \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \times 100 \quad (8)$$

Where, TP is represented as true positive, FP is denoted as false negative, TN is represented as true negative and FN is stated as false negative.

MAP is the measure of statistical variability and a description of random errors. The general formula of MAP for determining CMR system is given in the Eq. (9).

$$MAP = \frac{\sum_{q=1}^{Q} Ave\ P(q)}{QQ} \quad (9)$$

Where, Q is represented as the number of queries.

4.2 Quantitative analysis

In this experimental analysis, Wikipedia dataset is assessed for comparing the performance evaluation of existing methods and the proposed scheme. In Table 1, the MAP value of the proposed and existing methodologies is compared for both the image and text query. The average MAP value of existing methods (CM [17], SM [17], SCM [17], AREV [18], and DPEP [13]) in image query delivers 0.194, 0.230, 0.235, 0.267, and 0.409. The proposed approach delivers 0.503 of average MAP in image query. Similarly, the average MAP value of existing methods in text query delivers 0.171, 0.208, 0.212, 0.224, and 0.290. The proposed approach delivers 0.303 of average MAP in text query. The graphical representation of MAP evaluation of existing and proposed method is represented in the Fig. 3.

In addition, the retrieval efficiency of the proposed and existing methodologies is compared for both the image and text query, which is stated in the Table 2. The retrieval efficiency of existing methods (CM [17], SM [17], SCM [17], AREV [18], and deep learning [12]) in image query delivers 20.82s, 22.42s, 22.71s, 30.99s and 20.12s. The proposed approach delivers 8.34s of retrieval efficiency in image query. Correspondingly, the retrieval efficiency of existing methods in text query delivers 21.09s, 23.5s, 22.85s, 31.31s and 20.88s. The proposed approach delivers 12.3s of retrieval efficiency in text query. The graphical representation of retrieval efficiency evaluation of existing and proposed method is represented in the Fig. 4.

Table 1. MAP evaluation of proposed and existing method

Methods	Image query			Text query		
	Run 1	Run 2	Average	Run 1	Run 2	Average
CM[17]	0.195	0.192	0.194	0.175	0.166	0.171
SM [17]	0.226	0.233	0.230	0.201	0.214	0.208
SCM [17]	0.238	0.231	0.235	0.214	0.209	0.212
AREV [18]	0.269	0.265	0.267	0.223	0.225	0.224
DPEP [13]	0.342	0.476	0.409	0.282	0.298	0.290
Proposed methodology	0.403	0.61	0.503	0.407	0.20	0.303

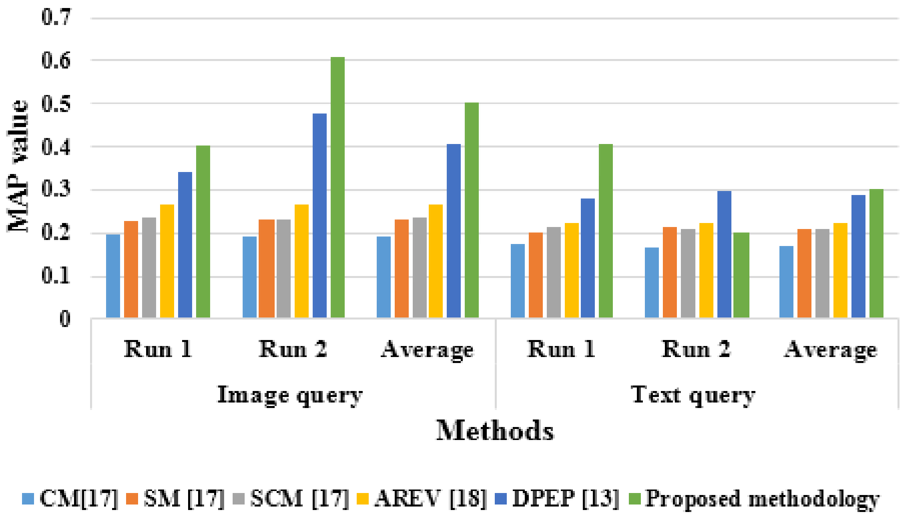


Figure.3 Graphical representation of MAP evaluation

Table 2. Evaluation on retrieval efficiency of Wikipedia dataset

Methods	Image query (seconds)	Text query (seconds)	Average (seconds)
CM [17]	20.82	21.09	20.96
SM [17]	22.42	23.5	22.96
SCM [17]	22.71	22.85	22.78
AREV [18]	30.99	31.31	31.15
Deep learning [12]	20.12	20.88	20.50
Proposed methodology	8.34	12.3	10.32

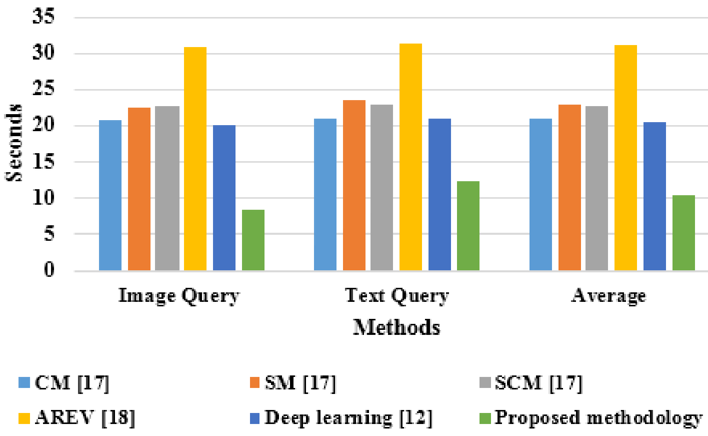


Figure.4 Graphical representation of retrieval efficiency evaluation

Table 2 clearly shows that the proposed approach improves the retrieval efficiency in the CMR system upto 10seconds compared to the existing methods in the Wikipedia dataset. Here, the proposed approach determines the non-linear characteristics of the data and preserves quantitative relationships between the low level and high level features. The evaluation metrics confirms that the proposed scheme performs significantly in CMR system compared to previous methods by means of retrieval efficiency and MAP.

The Figs. 5 and 6 shows the graphical representation of image query and text query for the value of precision vs recall. The precision value is measured with the gradual increase in the recall value. In image query, the higher precision value is obtained as 0.562, when the recall value is set as 0.1. Likewise, in text query, the higher precision value is obtained as 0.623, when the recall value is set as 0.3. Compared to the existing study, the proposed approach shows better result by means of precision vs recall.

4.3 Discussion about proposed methodology

The proposed dictionary learning consists of two types of dictionaries such as, T2I and I2T dictionary. In this dictionary learning, each individual text or image data is tested by using minkowski distance measure. This action effectively increases the matched pairs and dramatically reduces the un-matched pairs of media data. The effectiveness of the proposed methodology is shown in the Tables 1 and 2. The proposed system achieves better retrieval rate compared to the existing approaches by means of MAP value, retrieval efficiency, precision and recall. The proposed approach has additional advantages like improving the retrieval accuracy by reducing the quantization error and delivers an effective outcome in both CMR and single media retrieval systems.

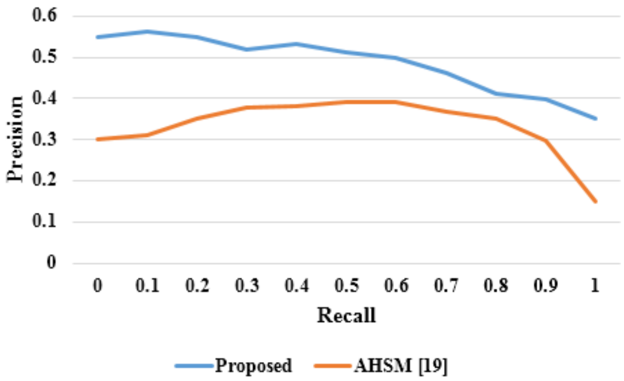


Figure.5 Graphical representation of image query

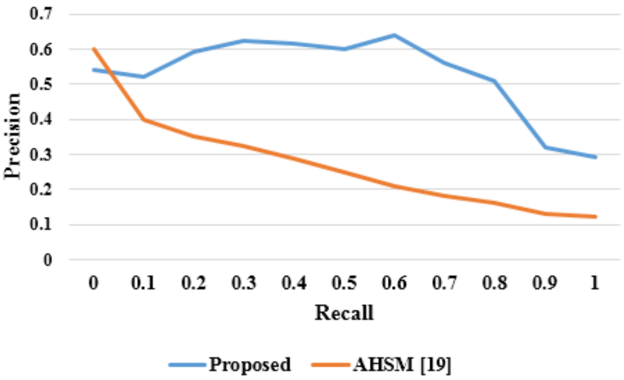


Figure.6 Graphical representation of text query

5. Conclusion

CMR system is utilized for classifying the documents to obtain different media result based on user search. In this experimental research, normalized histogram colour feature and bag-of-words methods are under-taken for image and text feature extraction. Bag-of-words methodology measures the repetition of the words, also it removes the unwanted conjunction words and the remaining words are considered as keywords. These keywords are considered as the query text for retrieving the related image or text based on proposed dictionary learning using minkowski distance measure. This experimental research is verified on a publicly available database: Wikipedia dataset shows the superiority of the proposed approach. Compared to other obtainable approaches in CMR system, the proposed methodology showed 0.24-0.20 of enhancement in MAP value than the existing methods. In future work, a new combination of features will be developed with an adaptive distance measure for further improving the retrieval efficiency.

References

[1] Q. Ma, A. Nadamoto, and K. Tanaka, "Complementary information retrieval for cross-media news content", *Information Systems*, Vol.31, No.7, pp.659-678, 2006.

[2] Y. Hu, L. Zheng, Y. Yang, and Y. Huang, "Twitter100k: A Real-world Dataset for Weakly Supervised Cross-Media Retrieval", *arXiv preprint arXiv:* 1703.06618, 2017.

[3] Y. Yang, Y.T. Zhuang, F. Wu, and Y.H. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval", *IEEE Transactions on Multimedia*, Vol.10, No.3, pp.437-446, 2008.

[4] Y.T. Zhuang, Y. Yang, and F. Wu, "Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval", *IEEE Transactions on Multimedia*, Vol.10, No.2, pp.221-229, 2008.

[5] R. Ren and J. Collomosse, "Visual sentences for pose retrieval over low-resolution cross-media dance collections", *IEEE Transactions on Multimedia*, Vol.14, No.6, pp.1652-1661, 2012.

[6] X. Mao, B. Lin, D. Cai, X. He, and J. Pei, "Parallel field alignment for cross media retrieval", In: *Proc. of the 21st International Conf. On Multimedia*, ACM, pp. 897-906, 2013.

[7] Y. Xia, Y. Wu, and J. Feng, "Cross-Media Retrieval using Probabilistic Model of Automatic Image Annotation", *International Journal of Signal Processing, Image Processing and Pattern Recognition*, Vol.8, No.4, pp.145-154, 2015.

[8] D. Ma, X. Zhai, and Y. Peng, "Cross-media retrieval by cluster-based correlation analysis", In: *Proc. of 20th International Conf. on Image Processing*, pp. 3986-3990, 2013.

[9] Y. Yang, F. Wu, D. Xu, Y. Zhuang, and L.T. Chia, "Cross-media retrieval using query dependent search methods", *Pattern Recognition*, Vol.43, No.8, pp.2927-2936, 2010.

[10] A. Habibian, T. Mensink, and C.G. Snoek, "Discovering semantic vocabularies for cross-media retrieval," In: *Proc. of the 5th International Conf. on Multimedia Retrieval*, pp.131-138, 2015.

[11] J.C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G.R. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.36, No.3, pp.521-535, 2014.

[12] B. Jiang, J. Yang, Z. Lv, K. Tian, Q. Meng, and Y. Yan, "Internet cross-media retrieval based on deep learning", *Journal of Visual Communication and Image Representation*, Vol.48, pp. 356-366, 2017.

[13] L. Huang and Y. Peng, "Cross-media retrieval by exploiting fine-grained correlation at entity level", *Neurocomputing*, Vol. 236, pp. 123-133, 2017.

[14] J. Yan, H. Zhang, J. Sun, Q. Wang, P. Guo, L. Meng, W. Wan, and X. Dong, "Joint graph regularization based modality-dependent cross-media retrieval", *Multimedia Tools and Applications*, pp.1-19, 2017.

[15] H. Zhang, Y. Liu, and Z. Ma, "Fusing inherent and external knowledge with nonlinear learning for cross-media retrieval", *Neurocomputing*, Vol.119, pp.10-16, 2013.

[16] L. Xie, P. Pan, and Y. Lu, "Analyzing semantic correlation for cross-modal retrieval", *Multimedia Systems*, Vol. 21, No. 6, pp.525-539, 2015.

[17] N. Rasiwasia, J.C. Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," In: *Proc. of the 18th International Conf. on Multimedia*, pp.251-260, 2010.

[18] K. Liu, S. Wei, Y. Zhao, Z. Zhu, Y. Wei, and C. Xu, "Accumulated reconstruction error vector (AREV): a semantic representation for cross-media retrieval", *Multimedia Tools and Applications*, Vol.74, No.2, pp.561-576, 2015.

[19] X. Zhai, Y. Peng, and J. Xiao, "Cross-media retrieval by intra-media and inter-media correlation mining", *Multimedia systems*, Vol.19, No.5, pp.395-406, 2013.