Cross-Modal Retrieval using Random Multimodal Deep Learning

Hemanth Somasekar, Kavya Naveen

Department of Computer Science and Engineering, RNS Institute of Technology, Chennasandra, Uttarahalli-Kengeri Road, Karnataka, India

hemanth.shantha@gmail.com, npkavya@gmail.com

Corresponding Author: Hemanth Somasekar
Email: hemanth.shantha@gmail.com

https://doi.org/10.26782/jmcms.2019.04.00016

Abstract

In multimedia community, cross modal similarity search based hashing received extensive attention because of the effectiveness and efficiency of query. This research work contributes large scale dataset for weakly managed cross-media recovery, named Twitter100k. Current datasets namely Wikipedia, NUS Wide and Flickr30k, have two main restrictions. First, these datasets are deficient in content diversity, i.e., only some pre-characterized classes are secured. Second, texts in these datasets are written informal dialect, that leads to irregularity with practical applications. To overcome these disadvantages, the proposed method used Twitter100k dataset because of two major points, first, it has 100,000 content-image pairs that are randomly crawled from Twitter and it has no importance in the image classifications. Second, text in Twitter100k is written in informal language by the clients. Since strongly supervised strategies use the class labels that might be missing in practice, this paper mainly concentrates on weakly managed learning for cross-media recovery, in which only text-image sets misused during training. This paper proposed a Random Multimodal Deep Learning (RMDL) based Recurrent Neural Network (RNN) for cross-media retrieval. The variety of input data such as video, text, images etc. are used for cross-media recovery which can be accept by proposed RMDL in weakly dataset. In RMDL, the various input data can be classified by using RNN architecture, to improve the accuracy and robustness of the proposed method, RMDL uses the specific RNN structure i.e. Long Short-Term Memory (LSTM). In the experimental analysis, the results demonstrated that the proposed RMDL-based strategy achieved 78% of Cumulative Match Characteristic (CMC) compared to other datasets.

Keywords: Cross modal similarity search, Twitter dataset, class labels, strong supervised methods, NUS Wide, Random Multimodal Deep Learning.
I. Introduction

Nowadays, the growth of multimedia data has significantly increased the demand for more sophisticated retrieval technologies and multimedia indexing [XXII] [XXIII] [XVI]. Particularly, the cross-modular recovery issue [II] [XX] coordinates one modular information with other modular information that has received much interest because of the heterogeneous sources of the mixed media information, i.e., writings, pictures and recordings. Nonetheless, different modular information can’t be straightforwardly coordinated because of the intrinsic diversity in various modalities. To handle this issue, a few hashing-based methodologies [XVIII] have gathered significant intrigue and has shown a better execution on cross-modular recovery issue. These hashing based strategies locate the linear projections to embed the heterogeneous information into a typical Hamming space, where the multi-modular features are recorded as binary codes. These strategies can deal with large scale information with low memory cost and computational expense, because the learning hash codes can be stored by using hamming distance algorithm which is calculated with bit-wise XOR tasks. These techniques are unsupervised, as the significant class label data has not been utilized yet [XXVII], [III]. The most challenging issues in cross-media recovery consists of eliminating the heterogeneity and mining the semantic connection between modalities [IV] [IX] [XXI]. Another test is to accelerate the way to find applicable information for an inquiry, which can be resolved by indexing systems [XVII] [XIX] [XXV]. This research work focuses on the primary test on weakly data and mostly discuss about picture content recovery. In this task, the aim is to seek an important picture (writings) from a large gallery that define the similar content with a picture query given by the user.

Moreover, the data are mostly deficient in content diversity which is the major drawbacks of cross-media datasets. For instance, the Pascal VOC 2012 dataset [VII] has 20 distinct classes, namely dog, horse, plane and so on. Nonetheless, recovery technique includes different spaces under sensible Internet conditions. Recovery techniques trained on datasets of scanty areas may experience issues in dealing with inquiries from unknown areas. Second, each picture in an existing dataset is related with labels or content description. Yet, pictures and messages are approximately related in more sensible situations. Not every word in texts has a visual translation in pictures. Third, prevalent cross-media datasets comprising of sentences and pictures that might be imperfect in dataset scale, for example, the IAPR TC-12 dataset (20,000 examples) and the Wikipedia dataset [XIII] (2,866 examples). The lack of information in this dataset became more challenging to evaluate the strength of recovery techniques in substantial scale galleries. Considering the previously mentioned issues, this paper makes two noteworthy contributions. The first contribution is the gathering of new large scale cross-media dataset, named Twitter100k. It contains 100,000 picture content sets gathered from Twitter. It is recognized from two aspects of existing datasets: fluctuated domains and informal content dialect. In the result, this dataset gives a more sensible standard to cross-media examination. Another contribution is that the method should provide extensive investigations on the weakly supervised strategies by testing the performance of the proposed strategy on the new dataset along with the Wikipedia. Because of the characteristics of Twitter100k data, the method uses the text in images which are highly correlated with the paired tweets and implemented a RMDL-based
retrieval method for retrieving the text-image query. In this work, the RMDL approach uses the RNN deep learning method for making the proposed method more robust and accurate. This work uses the Adam optimization technique for stabilizing the classification task in training the models. In addition, RMDL has ability to process a variety of data types includes images, text and videos. The effectiveness of the technique checked by the investigation. The outcomes stated that RMDL method can accept a variety information to incorporate content, video, pictures, and emblematic as an input. This paper represents RMDL and showed test results for picture and content information for Twitter dataset. These test outcomes demonstrated that RMDL reliably delivered better execution than standard strategies over a wide scope of information types and classification issues.

The rest of this paper is organized as follows: Section II presents the survey of the existing methods in cross media retrieval. The Twitter100k dataset is presented in detail in Section III. This paper presents a RMDL-based retrieval method in Section IV. The results and experimental analysis are presented in Section V. Finally, Section VI describes the conclusion with future direction.

II. Literature Review

K. Ding, et al., [V] implemented a regression-based Rank-order Preserving Hashing (RoHP) method which has margin property and easy to upgrade. To solve the incited enhancement issue, the alternating descent strategy was embraced. The examinations on three benchmark datasets exhibited that RoPH essentially enhanced the positioning quality over the existing techniques. So, the position was exceptionally supported for boosting the execution of cross-modular hashing. This method was explicitly computed and consumed more memory because the training set was very large and had a number of triplets.

K. Ahmad, et al., [I] described the method QBIMF that retrieved the related images from the image dataset as an output to user. In this framework, each picture was stored in the database as a visual substance network and matching was performed by utilizing that lattice. In query based picture management framework (QBIMF), the essential representation of the picture was given by its shape, surface, and shading. The working rule behind QBIMF was totally different from indexing. This encourages QBIMF to get the nearest picture from the computerized picture datasets. This work was completely based on content-based image retrieval system which solved many trivial problems and also used in applications such as face, speech recognition, etc. The complexity of the method increased when the number of features increased.

H. Zhang et al., [XXIV] 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 with isomorphic feature sub-space and then extracted 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 this methodology is more significant in terms of recognition rate compared to existing approaches. This proposed approach is only applicable for datasets written in well-organized language. Otherwise, it leads to inconsistency with realistic applications.
B. Jiang, et al. [VIII] proposed a new methodology named as deep learning for recovering the important cross media (content and picture). The proposed method focused on two stages, such as feature extraction and distance discovery. The unnecessary element vectors were eliminated by this technique, once the feature data obtained. The exploratory research was performed on freely accessible datasets (i.e. Wiki content picture and NUSWIDE dataset) to approve its recovery exactness in terms of F-measure, accuracy and recall. In a few cases, this technique is not proficient in substantial database and furthermore not robust in substance interpretation of cross media documents.

L. Malliga, and K. B. Raja, [X] proposed an improved CBMIR dependent on Modified FCM (MFCM) grouping method in order to achieve high efficiency and low computation cost. At first, the Haralick and Texture spectrum features were removed from the medicinal database pictures to recover the input query pictures. In this way, the extracted features from the trained database pictures were grouped by the FCM grouping strategy. The most relevant images were obtained by using distance values. Subsequently, CBMIR-MFCM method efficiently recovered the most important medical pictures through MFCM strategy with 83% of recovery rate and 65% of accuracy rate. The technique gave less accuracy at both retrieval and accuracy rate, which needs further enhancement in recovering the query pictures.

III Multi-Domain Dataset

In this segment, the proposed strategy presents a multi-scale dataset called Twitter100k, which contains 100,000 picture content sets gathered from Twitter.

Collection of Dataset

In data collection, there are four stages seed client gathering, client candidate generation, tweet accumulation and information pruning. The individual steps in information gathering are depicted as follows.

Seed user gatherings

To guarantee the diversity of the gathered information, the method gets seed clients by sending inquiries to Twitter with different topic words, for example, trip, feast, wellness, sports, and so forth. The randomly selected users are used as a seed for acquiring a number of user candidates.

User candidate generation

A web spider is created to crawl the records of the clients who are following the seed clients. The iteration of this step stops when the long list of other candidates is generated. The list becomes large when the users covers a number of domains.

Tweet collection

Another web spider gathers tweets with related pictures by visiting the home pages of the considerable number of clients in the list of candidates. The technique finds that around 1/3 of the tweets are combined with pictures.
Data Pruning

The image-tweet pair is pruned under any of the accompanying circumstances, for example, chaotic codes in tweets, tweets without words, tweets not written in English, reduplicate tweets with the same ID and error pictures.

The method finally obtains 100,000 image-text pairs in total. An image and text appearing in one piece of tweet are considered as a pair. Fig. 1 shows the sample images of Twitter dataset.

Characteristics of Tweets

The Twitter100k dataset is highlighted by few following aspects. In the first place, this dataset is gathered from web media, henceforth it covers an extensive variety of areas such as sport, engineering, nourishment, creature, news, plant, individual, publication, etc. Second, since informal dialect is typically utilized by Internet clients when posting tweets, messages in Twitter100k are recognized from different datasets by
the sentence structure and vocabularies. Third, the relationship between the picture and content is completely unrelated. Fourth, Twitter100k is a huge dataset, containing 100,000 picture content sets. An abundance of information can avoid over fitting during training. Also, it can be exploited to test the robustness of recovery strategies under huge information. Last, around 1/4 of the pictures in this dataset contains content which is highly related to the combined tweets. In tweets, a few words related to the content present in the pictures. In extraordinary cases, tweets can be identical to the content in pictures. To the best of knowledge, Twitter100k is the only cross media dataset with this trademark.

IV. Proposed Methodology

This work is utilizing RMDL models, like RNN system for content and picture characterization. The paper presents RMDL and afterward examine RNN methods of deep learning models for retrieving the query with the help of Twitter100K dataset. Next, the optimizer method is utilized in irregular models.

Feature Extraction

The component extraction is separated into two principle parts for RMDL (Text and picture). The feature space is structured for image datasets, text and sequential datasets are unstructured. The data from the Twitter consists of both structured and unstructured data, which are processed for retrieving query images given by the user. The input can be given to the RNN model for preprocessing and feature extraction to extract the components from the structured data in Twitter dataset.

Image and 3D Object Feature Extraction

Image features are $h \times w \times c$ where $h$ denotes the height of the image, $w$ represents the width of image, and $c$ is the color that has 3 dimensions (RGB). A 3D object in space contains $n$ cloud points in space and each cloud point has 6 features which are $(x, y, z, R, G, and B)$. The 3D object is unstructured due to number of cloud points from one object could be different with others. However, the method uses simple case of down/up sampling to generate the structured datasets.

Text and Sequences Feature Extraction

Techniques like GloVe and Word2vec are used for text feature extraction is known as word embedding. In this paper, word vectorization methods utilized for separating the features. Then, the strategy also utilizes N-gram representation as features for neural deep learning. Documents enter into RNN models through separated features from the content. In RNN, the method utilized the content vector-space model that uses 200 measurements as presented in GloVe [XI]. A vector-space model is a scientific mapping of the word space, characterized as below Eq. (1).

$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{i,j}, \ldots, w_{l_j,j})$$ (1)

Where $l_j$ is the length of the document $j$, and $w_{i,j}$ is the GloVe word embedding vectorization of word $i$ in document $j$. 

190
Random Multimodal Deep Learning

RMDL is a novel method that can be used in any sort of dataset for characterization. The quantity of layers and nodes for these Deep learning multi models produced arbitrarily (e.g. 9 Random Models in RMDL constructs 3 RNNs, each of them are novel due to arbitrary creation)

\[ M(y_{11}, y_{12}, ..., y_{1n}) = \left[ \frac{1}{2} + \frac{\sum_{j=1}^{n} y_{ij} - \frac{3}{2}}{n} \right] \]  

Where \( n \) is the number of random models, and \( y_{ij} \) is the output prediction of model for data point \( i \) in model \( j \) (Eq. (2) is used for binary classification, \( k \in \{0 \ or \ 1\} \)). Output space uses majority vote for final \( \hat{y}_i \). Therefore, \( \hat{y}_i \) is given as Eq. (3).

\[ \hat{y}_i = [\hat{y}_{i1}, \hat{y}_{i2}, ..., \hat{y}_{in}]^T \] (3)

Where \( n \) is the number of random models, and \( \hat{y}_{ij} \) shows the prediction of label of document or data point of \( D_i \in \{x_i, y_i\} \) for model \( j \) and \( \hat{y}_{i,j} \) is defined in Eq. (4)

\[ \hat{y}_{i,j} = \arg \max_k [\text{softmax}(y_{i,j})] \] (4)

After all, RMDL models are trained, the final prediction calculated using majority vote of these models.

Deep Learning of Recurrent Neural Network in RMDL

RNN allocate more weights to the previous information points of grouping. Accordingly, this procedure is an incredible technique for content, string and successive information grouping and also utilized for picture characterization. In RNN, the neural system considers the data of past nodes in an extremely refined technique which allows better semantic examination of the structures of the dataset. The general formulation of this concept is given in Eq. 5 where \( x_t \) is the state at time \( t \) and \( u_t \) refers to the input at step \( t \).

\[ x_t = F(x_{t-1}, u_t, \theta) \]  

Specifically, the method needs to use weights to formulate the Eq. (5) with specified parameters in Eq. (6)

\[ x_t = W_{rec} \sigma(x_{t-1}) + W_{in} u_t + b \]  

Where \( W_{rec} \) refers to recurrent matrix weight, \( W_{in} \) refers to input weights, \( b \) is the bias and \( \sigma \) denotes an element-wise function. Several problems arise from RNN when the error of the gradient descent algorithm back propagated through the network: vanishing gradient and exploding gradient. To deal with these problems, LSTM is a special type of RNN that preserve long term dependency in a more effective way compared to the basic
RNN. This is particularly useful to overcome vanishing gradient problem [XII]. Although LSTM has a chain-like structure similar to RNN, LSTM uses multiple gates to carefully regulate the amount of information that will be allowed into each node state.

**Optimization**

In RNN, Adam optimizer is used which is a kind of stochastic gradient optimizer that uses only first two gradient moments (i.e. \( \nu \) and \( m \)) are described in Eq. 7 - 10. It can deal with non-stationary of target work as in RMSProp while solving the sparse angle issue which is a bottleneck of RMSProp.

\[
\theta \leftarrow \theta - \frac{\alpha}{\sqrt{1+\epsilon}} \hat{m}
\]

\[
g_{t,t} = \nabla_{\theta} J(\theta_t, x_t, y_t)
\]

\[
m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_{t,t}
\]

\[
m_t = \beta_2 \nu_{t-1} + (1 - \beta_2) \hat{g}_{t,t}^2
\]

Where \( m_t \) is the first moment and \( \nu_t \) indicates second moment and both are estimated as \( \hat{m}_t = \frac{m_t}{1-\beta_1} \) and \( \hat{\nu}_t = \frac{\nu_t}{1-\beta_2} \).

V. Experimental Result

This paper mostly examines the weakly managed recovery strategies, in which only the co-occurrence data of the picture and content were misused. This experimental section presents a brief description of Twitter100k dataset.

**Dataset Description**

This work utilized the recovery techniques on the Twitter100k datasets. The Twitter100k datasets consists of 50,000 images-text pairs for training and remaining images-text pairs for testing purposes. The Twitter100k dataset gives a more realistic benchmark for cross-media recovery. Since it gathered from Twitter, cross-media recovery on this dataset is applicable for some application scenarios. For instance, Twitter only gives predefined emoticons for clients to choose during posting a tweet. In fact, the cross-media recovery becomes more advantageous and fascinating to extend the scope of emoticons and prescribe reasonable pictures for clients as per the substance of the tweets.

**Evaluation Criteria**

Since, the pre-defined category labels are not available, text and image in a pair are considered as a ground-truth match. This is, in the given a query text (image), only one ground-truth image (text) exists in the gallery. As a consequence, this paper takes the following evaluation metrics namely Cumulative Match Characteristic (CMC), percentage of precision and recall (i.e. PER) and Mean Rank.
Cumulative Match Characteristic

CMC often utilize as a measurement in the field of face acknowledgment and individual re-distinguishing proof [XXVI] [VI]. It calculates how well a distinguishing proof framework ranks the personalities in the selected database with respect to ‘unknown’ test picture. For cross-media recovery, CMC describes the need for finding the correct match in the best $n$ coordinates and can be presented by a combination of average recovery precision regarding rank.

Mean Rank

Mean rank is the average of the ranks of the correct matches for a series of queries $Q$. The equation for mean rank is defined in the following Eq. (11).

$$\text{Mean Rank} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \text{rank}_i$$

Where $\text{rank}_i$ refers to the rank position of the correct match for the $i^{th}$ query.

Precision and Recall

The probability of success for an image retrieval system can be measured with the help of precision, whereas the percentage of retrieved images from the system is measured by using recall. The two evaluation indicators can be explained by using mathematical expression in Eq. (12) and (13),

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Where TN is True Negative, TP is True Positive, FN is False Negative and TN is True Negative.

Evaluation of the proposed method

This section describes the performance of proposed method by comparing with the existing techniques that are verified from the experimental results. Canonical correlation analysis (CCA) [XIII] found a premise of authoritative segments, i.e., directions, along which the information is maximally corresponded. Partial Least Square (PLS) [XV] utilized the minimum square strategy to associate the subspaces of CCA in order to avoid data scattering during the process of distinctive modular relationships. The BiLinear model (BLM) [XV] isolate style and content by utilizing Singular Value Decomposition (SVD). Generalized multi-view extension of Marginal Fisher Analysis (GMMFA) [XV] was a multi-view augmentation of Marginal Fisher Analysis. It endeavored to isolate the distinctive class and compress same-class tests in the element space. Table 1 and 2 shows the execution of Mean Rank of the proposed technique for picture and content inquiry with Twitter100k dataset. The graphical representations are given in Figure 2 and 3.
Table 1. Mean Rank of the correct matches for image query

| Image Query | Methods       | Different value of $\alpha$ on Twitter dataset |
|-------------|---------------|----------------------------------------------|
|             |               | 0.   | 0.2 | 0.3 | 0.4 | 0.5 |
|             | CCA [XIII]    | 1.   | 1.42| 1.4 | 1.38| 1.36|
|             | RoHP [V]      | 1.   | 1.39| 1.40| 1.38| 1.35|
|             | PLS [XV]      | 1.   | 1.339| 1.335| 1.33| 1.32|
|             | BLM [XV]      | 1.   | 1.34 | 1.33 | 1.29 | 1.25 |
|             | GMMFA [XV]    | 1.   | 1.38 | 1.385| 1.37 | 1.35 |
|             | Proposed      | 1.   | 1.27 | 1.26 | 1.25 | 1.24 |

Fig. 2. Mean Rank for Image Query
This section presents the performance of the proposed RMDL-based retrieval method on Twitter100k. Above all, the impact of the weight parameter $\alpha$ on the retrieval performance is discussed. The mean rank of the correct matches with different values of $\alpha$ on Twitter100k are provided. These results revealed that the mean rank of the correct matches declines with $\alpha$ when $\alpha < 1$ for all the baselines. It shows that text on image plays a dominant role in the retrieval. But the performance deteriorates when $\alpha$ is set to 1, because the information about the color and the shape of the image is lost. After a certain point, the mean rank will start to increase slowly.

The following table 3 and 4 represent the CMC curve on Twitter dataset with existing methods. The graphical representation is shown below in Fig. 4 and 5.
Table 3. CMC curve on Twitter100k Dataset for text query

| Methods  | Rank of CMC curve on Twitter dataset |
|----------|--------------------------------------|
|          | 0.5  | 1.0  | 1.5  | 2    | 2.5  |
| CCA [XIII]| 10   | 30   | 45   | 60   | 68   |
| PLS [XV]  | 10   | 30   | 50   | 62   | 73   |
| BLM [XV]  | 10   | 28   | 43   | 55   | 65   |
| GMMFA [XV]| 10   | 30   | 45   | 58   | 65   |
| Proposed  | 10   | 35   | 52   | 64   | 78   |

Fig. 4. CMC curve for Text Query

Table 4. CMC curve on Twitter100k Dataset for image query

| Methods  | Rank of CMC curve on Twitter dataset |
|----------|--------------------------------------|
|          | 0.5  | 1.0  | 1.5  | 2    | 2.5  |
| CCA [XIII]| 12   | 28   | 47   | 52   | 58   |
| PLS [XV]  | 18   | 27   | 48   | 54   | 63   |
| BLM [XV]  | 10   | 28   | 45   | 57   | 66   |
| GMMFA [XV]| 11   | 30   | 43   | 58   | 70   |
| Proposed  | 21   | 32   | 52   | 63   | 74   |
From Fig. 4 and 5, it can be concluded that the performance of proposed method provides better performance compared to existing methods. In other words, average retrieval accuracy improved by incorporating RMDL-based cross medial retrieval method. But, the existing method BLM in text to image retrieval achieved only 43.10% retrieval accuracy. This enhancement described the facts that the tweets and the texts on the images are highly correlated, the proposed methods can utilize the text information of the images besides the shape and color information. Table 5 represents the evaluation of PER in both image-text query for proposed method is explained.

Table 5. Evaluation of proposed method in percentage of precision and recall

| Methodology   | PER          |            |            |
|--------------|--------------|------------|------------|
|              | Image-query-Text | Text-query-Image |            |
| MTRM [VIII]  | 0.3735       | 0.4041     |            |
| DLM [VIII]   | 0.4121       | 0.4097     |            |
| Proposed RMDL | 0.4268       | 0.4125     |            |

When compared with the existing methods such as Mutual Topic Reinforce Model (MTRM) and Deep Learning Method (DLM) [VIII], the proposed method achieved better percentage in precision and recall for both text-image query. The DLM method achieved 0.4121 PER for image query, whereas the proposed RMDL method achieved 0.4268 PER. For text query, the MTRM method achieved less PER whereas the DLM achieved 0.4097 PER and RMDL achieved nearly 0.42 PER. Fig. 6 [XIV] shows that the sample retrieved images and text queries by using proposed RMDL method.
Fig. 6.(i) Text to image query (ii) Image to text query

From Fig. 6(i), it is clearly shows that the proposed RMDL method was successful in leaning background, colors and class information. In addition, Fig. 6(ii) shows the tagging results retrieved by the proposed method on some test images. The proposed RMDL-based method may be further developed by utilizing superior similarity metrics and advanced text retrieval methods.

VI. Conclusion

The classification assignment is an important issue to address in machine learning because of the growing number and size of datasets that require sophisticated arrangement. This paper presents a new methodology called as RMDL for the classification that consolidates RNN deep learning approaches to deal with random characterization models. Additionally, this paper presents a large scale cross-media dataset called Twitter100k, which gives a more sensible benchmark towards weakly supervised content picture recovery. By considering the characteristic of the new dataset, this paper proposed to enhance the retrieval performance based on RMDL-based RNN technique. The experimental evaluation acquired from the Twitter dataset showed higher accuracy than the existing strategies. Its conclude form the outcomes that the deep learning strategies can give upgrades to characterization and provides flexibility to group datasets with the help of majority vote. The proposed RMDL method achieved 0.4268 and 0.4125 PER values for both image-text-queries. The proposed method achieved 1.24 and 1.22 Mean Rank in both image-text-queries for $\alpha = 0.5$. In future work, we intend to structure a better evaluation protocol for this dataset and consider the correlation between the pictures and messages from high level semantic perspectives, such as opinion and sentiment views.
References

I. Ahmad, Khaleel, Monika Sahu, Madhup Shrivastava, Murtaza Abbas Rizvi, and Vishal Jain., “An efficient image retrieval tool: query based image management system,” International Journal of Information Technology, pp. 1-9, 2018.

II. Ballan Lamberto, Tiberio Uricchio, Lorenzo Seidenari, and Alberto Del Bimbo, “A cross-media model for automatic image annotation”, In Proceedings of International Conference on Multimedia Retrieval, pp. 73, 2014.

III. Ding Guiguang, Yuchen Guo, and Jile Zhou, “Collective matrix factorization hashing for multimodal data,” Proceedings of the IEEE conference on computer vision and pattern recognition, 2014.

IV. Deng Cheng, Xu Tang, Junchi Yan, Wei Liu, and Xinbo Gao, “Discriminative dictionary learning with common label alignment for cross-modal retrieval,” IEEE Transactions on Multimedia, vol. 18, 2, pp. 208-218, 2016.

V. Ding, Kun, Bin Fan, Chunlei Huo, Shiming Xiang, and Chunhong Pan, “Cross-modal hashing via rank-order preserving,” IEEE Transactions on Multimedia, vol. 19, no. 3, pp. 571-585, 2017.

VI. Hauptmann, A. G., Yi Yang, and L. Zheng, “Person Re-identification: Past, Present and Future,” 2016.

VII. Hwang Sung Ju, and Kristen Grauman, “Reading between the lines: Object localization using implicit cues from image tags,” IEEE transactions on pattern analysis and machine intelligence vol. 34, no.6, pp. 1145-1158, 2012.

VIII. Jiang Bin, Jiachen Yang, Zhihan Lv, Kun Tian, Qinggang Meng, and Yan Yan, “Internet cross-media retrieval based on deep learning”, Journal of Visual Communication and Image Representation, vol.48, pp. 356-366, 2017.

IX. Kang Cuicui, Shiming Xiang, Shengcai Liao, Changsheng Xu, and Chunhong Pan, “Learning consistent feature representation for cross-modal multimedia retrieval,” IEEE Transactions on Multimedia, vol. 17, no. 3, pp. 370-381, 2015.

X. L. Maliga, and K. Bommanna Raja, “A Novel Content Based Medical Image Retrieval Technique with Aid of Modified Fuzzy C-Means Clustering (CBMIR-MFCM),” Journal of Medical Imaging and Health Informatics vol. 6, no. 3, pp. 700-709, 2016

XI. Pennington Jeffrey, Richard Socher, and Christopher Manning, “Glove: Global vectors for word representation,” Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014.

XII. Pascanu Razvan, Tomas Mikolov, and Yoshua Bengio, “On the difficulty of training recurrent neural networks,” International Conference on Machine Learning. 2013.

XIII. Rasiwasia Nikhil, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos, “A new approach to cross-modal multimedia retrieval,” In Proceedings of the 18th ACM international conference on Multimedia, pp. 251-260, ACM.

XIV. Rehman Sadaqat Ur, Shanshan Tu, Yongfeng Huang, and Obaid Ur Rehman, “A Benchmark Dataset and Learning High-Level Semantic Embeddings of Multimedia for Cross-Media Retrieval,” IEEE Access, vol. 6, pp. 67176-67188, 2018.
XV. SharmaAbhishek, Abhishek Kumar, Hal Daume, and David W. Jacobs, “Generalized multiview analysis: A discriminative latent space,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2160-2167, 2012.

XVI. Shen Fumin, Chunhua Shen, Qinfeng Shi, Anton Van Den Hengel, and Zhenmin Tang, “Inductive hashing on manifolds,” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1562-1569, 2013.

XVII. Song Jingkuan, Yi Yang, Zi Huang, Heng Tao Shen, and Jiebo Luo, “Effective multiple feature hashing for large-scale near-duplicate video retrieval,” IEEE Transactions on Multimedia, vol. 15, no. 8, pp. 1997-2008, 2013.

XVIII. Song Jingkuan, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen, “Intermediate hashing for large-scale retrieval from heterogeneous data sources,” In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, pp. 785-796, 2015.

XIX. Wu Fei, Zhou Yu, Yi Yang, Siliang Tang, Yin Zhang, and Yueting Zhuang, “Sparse Multi-Modal Hashing,” IEEE Trans. Multimedia, vol. 16, no. 2, pp. 427-439, 2014.

XX. Xu Xing, Yang Yang, Atsushi Shimada, Rin-ichiro Taniguchi, and Li He, “Semi-supervised coupled dictionary learning for cross-modal retrieval in internet images and texts,” In Proceedings of the 23rd ACM international conference on Multimedia, pp. 847-850, 2015.

XXI. Yang Yi, Feiping Nie, Dong Xu, Jiebo Luo, Yueting Zhuang, and Yunhe Pan, “A multimedia retrieval framework based on semi-supervised ranking and relevance feedback,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 4, pp. 723-742, 2012.

XXII. Yang Yang, Zheng-Jun Zha, Yue Gao, Xiaofeng Zhu, and Tat-Seng Chua, “Exploiting web images for semantic video indexing via robust sample-specific loss,” IEEE Transactions on Multimedia, vol. 16, no. 6, pp. 1677-1689, 2014.

XXIII. Yang Yang, Hanwang Zhang, Mingxing Zhang, Fumin Shen, and Xuelong Li, “Visual coding in a semantic hierarchy,” In Proceedings of the 23rd ACM international conference on Multimedia pp. 59-68, 2015.

XXIV. Zhang Hong, Yun Liu, and Zhigang Ma “Fusing inherent and external knowledge with nonlinear learning for cross-media retrieval”, Neurocomputing, vol.119, pp.10-16, 2013.

XXV. Zha Zheng-Jun, Meng Wang, Yan-Tao Zheng, Yi Yang, Richang Hong, and Tat-Seng Chua, “Interactive video indexing with statistical active learning,” IEEE Transactions on Multimedia, vol. 14, no. 1, pp. 17-27, 2014.

XXVI. Zheng Liang, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian, “Mars: A video benchmark for large-scale person re-identification.” In European Conference on Computer Vision, pp. 868-884, Springer, 2016.

XXVII. Zhou Jile, Guiguang Ding, and Yuchen Guo, “Latent semantic sparse hashing for cross-modal similarity search,” In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, 2014.