ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

Effective near -Duplication Detection and Cross Media Retrieval in Textual Information by Utilizing SCMH Method

[1]Umarani, [2]Sunitha Vanamala, [3]Shanagoda Sajitha

[1]Associate Professor, Department of CSE, School of Information Technology (JNTUH), Kukatpally, District Medchal, Telangana, India

[2] Assistant Professor, Department of Information Technology, Kakatiya Institute of Technology & Science, District Warangal, Telangana, India

[3]M. Tech Student, Department of CSE, School of Information Technology (JNTUH), Kukatpally, District Medchal, Telangana, India

ABSTRACT*:* Hashing techniques have proven to be useful for a different type of tasks and have attracted extensive attention in recent years. Hashing method is the one of the main method for searching same and different images based on hash code. For capturing similarities between textual, visual and cross media information; a hashing approaches have been proven. To address these challenges, in this paper we propose semantic level cross media hashing (SCMH) and deep belief network (DBN) is for a co-relation between different modalities.

I. INTRODUCTION

Recently, an experimental take a look at on pedestrian classification investigated the aggregate of several modern functions and classifiers. Some combinations performed better than others, but curiously, the benefit acquired with the aid of choosing the quality aggregate changed into much less pronounced than the advantage received via growing the training set (even though the latter was already quite big, related to many thou- sands of schooling samples). Methods to gather additional training samples of the non-target elegance are usually used, although it became observed that the performance has a tendency to saturate pretty speedy, after some bootstrapping iterations. The expansion of the schooling set with appreciate to the target magnificence yielded extra advantage, however this normally calls for time consuming (and therefore highly-priced) manual labeling. Uniquely connected generative discriminative approach to the pedestrian analysis and calculated at addressing the bottleneck caused by the scarcity of samples of the goal elegance; A generative version is learned from a pedestrian dataset captured in real city traffic and used to synthesize digital samples of the target class, for that reason enlarging the schooling set of a discriminative pattern classifier at little value. This set of virtual samples can be taken into consideration as a regularization time period to the real records to be equipped, which carries prior understanding approximately the target item elegance. The use of selective sampling, through method of probabilistic active information, to guide the training process towards the maximum informative samples;

Hashing method is one of the methods for searching a similar and different images based on hashing code. A mixed generative discriminative is the one of the searching type of image, means we in this we can search a both type of images by using the search keyword, it is helpful for display a combination of both content type and image type searches. Hashing-based methods, which create compact hash codes that preserve similarity, for single-model or cross-model retrieval on large-scale databases have attracted considerable attention. This method is based on the hash code, for searching a similarity of learning hashing functions in multi-model data for cross view similarity and novel hashing method which referred to called matrix factorization hashing (CMFH). It learns unified hash codes by collective matrix

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

factorization. Hashing based similarity of method can view a all based on category or cluster format. The word representations are learned by recurrent neural network language mode. A word embedding is a representation of words as continuous vector, for that for a particular word a particular hash code will generate.

II. RELATED WORK

S. Kumar and R. Udupa proposed Cross-view Hashing which maps similar objects to similar codes across the views to enable similarity search. In this work, a hashing-based approach for solving the cross-view similarity search problem is used where each view of a multi-view data object as a compact binary codeword is represented. To support this similarity search, we need the code words of a data object to be similar if not identical. Later, code words of similar data objects should also be similar. Assuming that we can somehow map data objects to binary code-words, cross-view similarity search can be reduced to the much simpler problem of retrieving all data objects using hamming distance code word, the codeword for the query. Discriminative coupled dictionary hashing generates a coupled dictionary for each modality based on category labels. In this paper, they introduced a discriminative coupled dictionary hashing approach, coupled dictionary for each modality based on category labels which helped in fast cross-media retrieval. Multi view discriminative coupled dictionary hashing (MVDCDH) is extended from DCDH with multi-view representation to enhance the representing capability of the relatively “weak” modalities. Latent semantic sparse hashing uses Matrix Factorization J. Zhou, G. Ding, and Y. Guo, proposed the use of Factorization to represent text and sparse coding to capture the salient structures of images. LSSH requires the use of both visual and textual information to construct the data set. In this paper Collective matrix factorization hashing (CMFH) generates unified hash codes for different modalities of one instance through collective matrix factorization with latent factor model collective matrix factorization.
H. Zhang, J. Yuan, X. Gao and Z. Chen delivered move media retrieval Boosting through function analysis and relevance remarks. This feature analysis is visual - auditory analysis which provides the boosting in retrieval. And in paper it's been defined about harmonizing hierarchical manifolds for multimedia file semantics knowledge and go - media retrieval. In this paper cluster - based totally correlation evaluation (CBCA) to make the most the relation between extraordinary sorts of multimedia objects, and to degree semantic similarities; Based on a collection of files that are multi - media CBCA first carry out clustering on uni - media characteristic areas to produce several semantic clusters for each modality. After that, by using the co - incidence information of semantic clusters of extraordinary modalities, CBCA constructs a move - modal cluster graph (CMCG) to represent the similarities among clusters. Yuxin Peng , Xiaohua Zhai , Yunzhen Zhao, Xin Huang , In this paper, they cognizance on how to study pass - media features for distinct media sorts is a key venture. Actually, the facts from exceptional media sorts with the same semantic class are complementary to every different, and jointly modeling them is able to improve the accuracy of cross - media retrieval. In addition, despite the fact that the categorized information is accurate, they require a lot of human labor and for this reason are very scarce. To resolve such problems a semi - supervised cross - media function getting to know algorithm with unified patch graph regularization (S2UPG). Heterogeneous Feature Augmentation (HFA), in this paper, they make use of a brand new area adaptation to clear up Heterogeneous domain model (HDA) trouble i n cross - media retrieval the use of Heterogeneous Feature Augmentation (HFA). First, distinctive dimensions of capabilities are transformed right into a not unusual subspace with the aid of studying an intermediate variable, and augmented the transformed records with their original features and ones; 2nd, in retrieval stage, we compute the similarity and rank the question results by means of bag - based re - rating technique.

III. PROPOSED WORK

A. Overview of Proposed Framework

The processing flow of the proposed semantic cross media hashing (SCMH) method in that we give a collection of text and images. First we represent an image and text respectively, for representing text a text are transformed to distributed vectors by the word embedding learning methods. And for representing images we use a SIFT detector to extract image key points. After these steps a fisher vector will generate for a particular word and particular descriptor. For that vector a hash code will generate and text and images are represented by vectors with fixes length.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

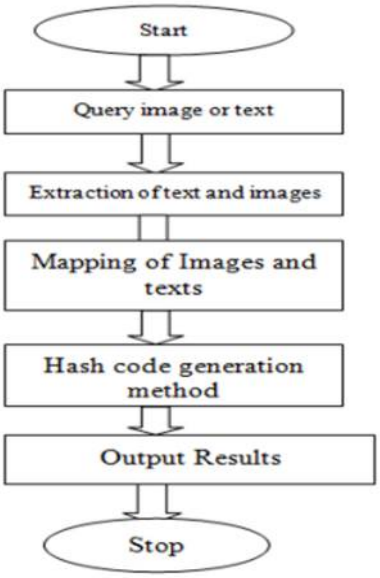


Fig1. Flow chart for SCMH Framework

Finally, the mapping functions between textual and visual fisher vector (FVs) are learned by a deep neural network. We used a learned mapping function to convert FVs of one modality to another.

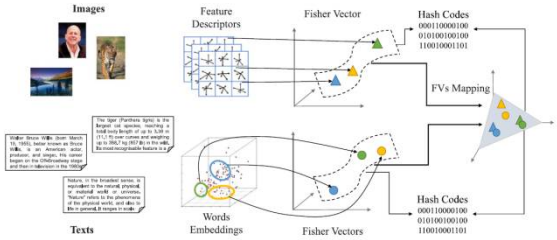


Fig2. Framework for SCMH

B. SIFT Descriptor

SIFT descriptor is used to calculate descriptors of the extracted key points. After these steps, a variable length set of points inside the embeddings space represents the text, and a variable length set of factors in SIFT descriptor space represents each picture. Then, the Fisher kernel framework is applied to mixture those factors in distinctive areas into constant period vectors, which can also be taken into consideration as factors within the gradient space of the Riemannian manifold. Henceforth, texts and images are represented by way of vectors with fixed duration.

C. Hash Code Generation in SCMH Framework

Hash code generation methods will be used to transfer different modalities. Following are the sequence of steps:

1. Hash code Generation
2. Matching

Various hashing methods are used to create compact hash codes for cross-media retrieval which preserves similarity. In this project, we will use semantic hashing which will create hash codes for the information; information may be visual or textual. Thus hash code generation will be used to transfer different modalities.

By comparing the descriptors obtained from different images, matching pairs can be found. For each descriptor, find a match and then verify matches.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

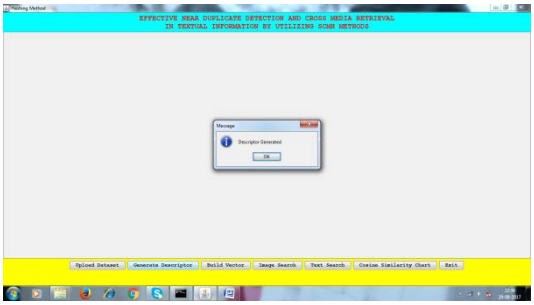
(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

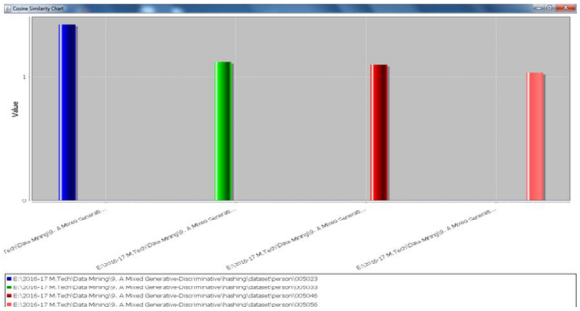
Vol. 6, Issue 8, August 2017

IV. EXPERIMENTAL RESULTS

We implemented a mixed generative discriminative based hashing method. This is mainly used for finding the co relation between the image and text to avoid the duplication.
In this experiment, we need to upload the dataset and to that dataset we need to generate the descriptors by using SIFT descriptor. After getting the descriptors for the dataset, we can build the vectors.



If we search image then output will be the text and if we search text then we get images as output.



Finally, we can see the cosine similarity chart for given input.

V. CONCLUSION

In this work we have designed hashing method to perform SCMH, cross media retrieval task. We have proposed to use a set of word embeddings to represent textual and visual information with fixed length vectors. For mapping the fisher vectors of different modalities, a deep belief network is proposed to perform the task. We evaluate the proposed method SCMH on three used data sets. SCMH achieves better results than state-of-the-art methods with different lengths of hash code.

REFRENCES

[1] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, Heng Tao Shen, “Inter-media hashing for large-scale retrieval from heterogeneous data sources”, 22-27 June 2013
[2] Deming Zhai, Hong Chang, Yi Zhen, Xianming Liu, Xilin Chen, “Parametric local multimodal hashing for cross-view similarity search”, 03-09 August 2013.
[3] T. Mikolorv, M. Karafiat, L. Burget, L. Cernocky, S. Khudanpur, “Recurrent neural network based language model”, 20 July 2010.
[4] Jile Zhou, Guiguang Ding, Yucheu Guo, “Latent semantic sparse hashing for cross-modal similarity search”, 06-11 July 2014.
[5] Zhou Yu, Fei Wu, Yi Yang, Qi Tian, Jiebo Luo, Yueting Zhuang, “Discriminative coupled dictionary hashing for fast cross-media retrieval”, 6-11 July 2014.
[6] Shaishav Kumar, Raghavendra Udupa, “Learning hash functions for cross-view similarity search”, 12 July 2011.
[7] Guiguang Ding, Yuche Guo, Jile Zhou, “Collective Matrix Factorization Hashing for Multimodal Data”, June 2014.
[8] David G. Lowe, “Object recognition from local scaleinvariant features”, 20-25 September 1999.
[9] Graham W. Taylor, Rob Fergus, Yann LeCun, Christoph Bregler, “Convolutional learning of spatio-temporal features”, 05-11 September 2010.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

[10] Yuncho Gong, Svetlana Lazebnik, Albert Gordo, Florent Perronnin, "Iterative quantization: "A Procrustean approach to learning binary codes for large-scale retrieval", 20-25 June 2011.
[11] Shaishav Kumar, Raghavendra Udupa, "Learning hash functions for cross-view similarity search", 12 July 2011.
[12] Zhen, Yi Yeung, Dit Yan, "A probabilistic model for multimodal hash function learning", 12-16 August 2012.