IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

A Fast Generative-Judicial Based Hashing Method Using Surf Descriptor

C. Maleappane Lawrence[1], M. Shanmugham[2], D. Sabarinathan[3], T. Banumathi[4]

Assistant Professor, Department of M.C.A., Christ College of Engineering and Technology, Pondicherry, India[1,2,3]

Final year Student, Department of M.C.A., Christ College of Engineering and Technology, Pondicherry, India[4]

Abstract: Hashing and Feature Detection methods have demonstrated to be useful for a variety of tasks and have appealed extensive tending in recent years. Various feature detection and hashing approaches have been purported to capture similarities between textual, visual, and cross-media information. However, most of the existing works use a bag-of-words methods to represent textual information and for feature detection use the SIFT algorithm. Since words with different forms may have similar meaning, semantic level text similarities cannot be well processed in these methods. To address these challenges, in this paper, we propose a novel method called semantic cross-media hashing (SCMH), which uses uninterrupted word representations to entrance the textual similarity at the semantic level and use a deep belief network (DBN) to fabricate the correlation between different sense modality and a technical report on feature detection and carrying out a SURF algorithm. To manifest the potency of the proposed method, we evaluate the proposed method on three commonly used cross-media data sets are used in this work. Experimental results show that the proposed method attains significantly better performance and speed than state-of-the-art approaches. Moreover, the efficiency of the proposed method is comparable to or better than that of some other feature detection and hashing methods.

Keywords: Hashing method, SURF, Word Embedding, Fisher Vector

I. INTRODUCTION

With the speedy elaboration of the World Wide Web, digital information has become much easier to access, modify, and duplicate. Hence, hashing based similarity calculation or estimate nearest neighbour searching methods have been proposed and received considerable attention in recent years.

Various applications such as information retrieval, near duplicate detection, and data mining are performed by hashing based methods. Due to speedy elaboration of mobile networks and social media sites, information input through multiple channels has also attracted increasing attention. Images and videos are associated with tags and captions. According to research published one Marketer, about 75 percent of the content posted by Facebook users contains photos. The relevant data from different modalities usually have semantic correlations.

Therefore, it is suitable to support the retrieval of information through different modalities. For example, images can be used to find semantically relevant textual information. On the other side, images textual descriptions are highly needed to be retrieved with textual query.Scale-Invariant Feature Transform, SIFT is a successful approach to feature detection introduced by Lowe [1]. The SURF-algorithm [2] is based on the same principles and steps, but it utilizes a different scheme and it should provide better results, faster. In order to detect feature points in a scale invariant manner SIFT uses a cascading filtering approach. Where the Difference of Gaussians (DoG) is calculated on progressively downscaled images.

In general the technique to achieve scale invariance is to examine the image at different scales, scale space, using Gaussian kernels. Both SIFT and SURF divides the scale space into levels and octaves. An octave corresponds to a doubling of, and the octave is divided into uniformly spaced levels.

Motivated by the success of continuous space word representations (also called word embedding's) in a variety of tasks, in this work we propose to incorporate word embedding's to meet these challenges. Words in a text are embedded in a continuous space, which can be viewed as Bag-of-Embedded-Words (BoEW). Since the number of words in a text is dynamic, we proposed a method to aggregate it into a fixed length Fisher Vector (FV), using a Fisher kernel framework.

However, the proposed methods only focus on textual information. Another challenge in this task is how to determine the correlation between multi-modal representations. Since we propose the use of a Fisher kernel framework to represent the textual information, we also use it to aggregate the SIFT descriptors of images. Through the Fisher kernel framework, both textual and visual information is mapped to points in the gradient space of a Riemannian manifold. However, the relationships that exist between FVs of different modalities are usually highly non-linear. Hence, to construct the correlation between textual and visual modalities, we introduce a DBN based method to model the mapping function, which is used to convert abstract

representations of different modalities from one to another.

II. RELATED WORK

An Along with the increasing requirement, extensive Hashing based methods have been proposed for cross-media retrieval. In this section, we briefly describe the related works, which can be categorized into the following four research areas: cross-media retrieval, text Reuse detection, and hashing methods.

A. Cross-Media Retrieval

Cross-media retrieval, in which the modality of input query and the returned results can be of different, has received considerable attentions. Wu et al. [3] introduced a Canonical Correlation Analysis based method to construct isomorphic subspace and multi-modal correlations between media objects and polar coordinates to judge the general distance of media objects. Due to lack of sufficient training samples, relevance feedback of user was used to accurately refine cross-media similarities. Yang et al. [4] proposed manifold-based method, in which they used Laplacian media object space to represent media object for each modality and a multimedia document semantic graph to learn the multimedia document semantic correlations.

In 8, a rich-media object retrieval method is proposed to represent data consisting of multiple modalities, such as 2-D images, 3-D objects and audio files. To tackle the large scale problem, a multimedia indexing scheme was also adopted.

Since the relationships across different modalities are typically highly non-linear and observations are usually noisy, Srivastava and Salakhutdinov [5] proposed a DeepBoltzmann Machine to learn joint representations of image and text inputs. The proposed model fuses multiple data modalities into a unified representation, which can be used for classification and retrieval.

The existing works described above focused on constructing the correlations between multiple modalities from different aspects. They usually use bag-of-words model to represent text. However, we in this work propose to use Fisher kernel framework to represent both textual and visual information and use a deep network to construct the correlations between the two manifolds.

B. Near-Duplicate Detection

The task of detecting near duplicate textual information has received considerable attentions in recent years. Previous works studied the problem from different aspects such as finger print extraction methods with or without linguistic knowledge, hash codes learning methods, different granularities, and so on. Border [9] proposed shingling method, which uses contiguous subsequences to represent documents. It does not rely on any linguistic knowledge. If sets of shingles extracted from different documents are appreciably overlap, these documents are considered exceedingly similar, which are usually measured by Jaccard similarity. In order to reduce the complexity of shingling, meta-sketches was proposed to handle the efficiency problem [10]. In order to improve the robustness of shingle-like signatures, Theobaldet al. [11] introduced a method, SpotSigs. It provides more semantic pre-selection of shingles for extracting characteristic signatures from Web documents. SpotSigs combines stop word antecedents with short chains of adjacent content terms.The aim of it is to filter natural-language text passages out of noisy Web page components. They also proposed several pruning conditions based on the upper bounds of Jaccard similarity. Different with these existing methods, in this paper, we propose to use aggregated word embeddings to capture the semantic level similarities to reduce the false matches.

C. Hashing-Based Methods

In recent years, hashing-based methods, which create compact hash codes that preserve similarity, for single-modal or cross-modal retrieval on large-scale databases have attracted considerable attention. For single-modal, Hintonand Salakhutdinov [12] proposed a two-layer network, which is called a Restricted Boltzmann machine (RBM), with a small central layer to convert high-dimensional input vectors into low-dimensional codes. Since we in this work learn the mapping functions between FVs of different modalities, all the hashing based methods for single modality can be incorporated into it.

D. Neural Networks for Representing

Image and Text The task of learning continuous space word representations have a long history. It has demonstrated outstanding results across a variety of tasks. Hinton and Salakhutdinov [13] introduced a deep generative model to learn word-count vector and binary code for documents. In [14], the word representations are learned by a recurrent neural network language model. The proposed architecture consists of an input layer and a hidden layer with recurrent connections. Probabilistic neural network language model (NNLM) simultaneously learns a distributed representation for each word and the probability function for word sequences. Although, in this work, we use word embeddings and SURF to represent texts and images respectively, the proposed method can also incorporate these representations.

III.EXISTING SYSTEM

Scale Invariant Feature Transform (SIFT) is an image descriptor for image-based matching and recognition. This descriptors as well as related image descriptors are used for a large number of purposes in computer vision related to point matching between different views of a 3-D scene and view-based object recognition. The SIFT descriptor is invariant to translations, rotations and scaling transformations in the image domain and robust to moderate perspective transformations and illumination variations. Experimentally, the SIFT descriptor has been

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

proven to be very useful in practice for image matching and object recognition under real-world conditions. In its original formulation, the SIFT descriptor comprised a method for detecting interest points from a grey-level image at which statistics of local gradient directions of image intensities were accumulated to give a summarizing description of the local image structures in a local neighbourhood around each interest point, with the intention that this descriptor should be used for matching corresponding interest points between different images. Later, the SIFT descriptor has also been applied at dense grids which have been shown to lead to better performance for tasks such as object categorization, texture classification, image alignment and biometrics.

SIFT key points of objects are first extracted from a set of reference images and stored in a database. An object is recognized in a new image by individually comparing each feature from the new image to this database and finding candidate matching features based on Euclidean distance of their feature vectors. From the full set of matches, subsets of key points that agree on the object and its location, scale, and orientation in the new image are identified to filter out good matches. The determination of consistent clusters is performed rapidly by using an efficient hash table implementation of the generalized Hough transform. Each cluster of 3 or more features that agree on an object and its pose is then subject to further detailed model verification and subsequently outliers are discarded. Finally the probability that a particular set of features indicates the presence of an object is computed, given the accuracy of fit and number of probable false matches. Object matches that pass all these tests can be identified as correct with high confidence.

A. Disadvantages of Existing System

- Still quite slow.
- Generally doesn't work well with lighting changes and blur.
- Same object under differing illumination.

IV. PROPOSED SYSTEM

Speeded up Robust Features (SURF) is a local feature detector and descriptor that can be used for tasks such as object recognition or registration or classification or 3D reconstruction. It is partly inspired by the scale-invariant feature transform (SIFT) descriptor. The standard version of SURF is several times faster than SIFT and claimed by its authors to be more robust against different image transformations than SIFT. To detect interest points, SURF uses an integer approximation of the determinant of Hessian blob detector, which can be computed with 3 integer operations using a precomputed integral image. Its feature descriptor is based on the sum of the Haar wavelet response around the point of interest. These can also be computed with the aid of the integral image.

SURF descriptors can be used to locate and recognize objects, people or faces, to make 3D scenes, to track objects and to extract points of interest. SURF is a detector and a descriptor for points of interest in images where the image is transformed into coordinates, using the multi-resolution pyramid technique, to make a copy of the original image with Pyramidal Gaussian or Laplacian Pyramid shape to obtain an image with the same size but with reduced bandwidth. Thus a special blurring effect on the original image, called Scale-Space, is achieved. This technique ensures that the points of interest are scale invariant.

Fig 1. Images Textual Descriptions.

The processing flow of the proposed semantic cross-media hashing (SCMH) method is given a collection of text-image bi-modality data; we firstly represent image and text respectively. Through table lookup, all the words in a text are transformed to distributed vectors generated by the word embeddings learning methods. For representing images, we use SURF detector to extract image key points. SURF descriptor is used to calculate descriptors of the extracted key points. After these steps, a variable size set of points in the embeddings space represents the text, and a variable size set of points in SURF descriptor space represents each image. Then, the Fisher kernel framework is utilized to aggregate these points in different spaces into fixed length vectors, which can also be considered as points in the gradient space of the Riemannian manifold. Henceforth, texts and images are represented by vectors with fixed length.

Finally, the mapping functions between textual and visual Fisher vectors (FVs) are learned by a deep neural network. We use the learned mapping function to convert FVs of one modality to another. Hash code generation methods

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

are used to transfer FVs of different modalities to short length binary vectors.

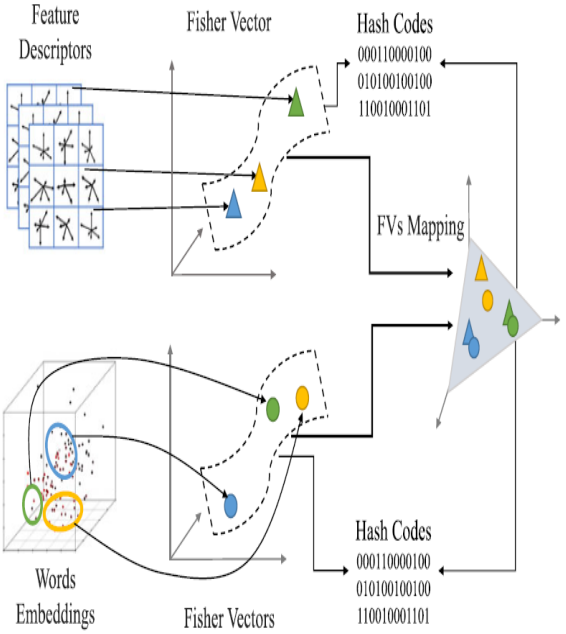

Fig 2. SCMH for cross-media retrieval.

The SIFT approach uses cascaded filters to detect scale-invariant characteristic points, where the difference of Gaussians (DoG) is calculated on rescaled images progressively. In SURF, square-shaped filters are used as an approximation of Gaussian smoothing. Filtering the image with a square is much faster if the integral image is used. SURF uses a blob detector based on the Hessian matrix to find points of interest. The determinant of the Hessian matrix is used as a measure of local change around the point and points are chosen where this determinant is maximal.

The interest points can be found in different scales, partly because the search for correspondences often requires comparison images where they are seen at different scales. In other feature detection algorithms, the scale space is usually realized as an image pyramid. Images are repeatedly smoothed with a Gaussian filter, and then they are subsampled to get the next higher level of the pyramid. he goal of a descriptor is to provide a unique and robust description of an image feature, e.g., by describing the intensity distribution of the pixels within the neighbourhood of the point of interest. Most descriptors are thus computed in a local manner; hence a description is obtained for every point of interest identified previously. The dimensionality of the descriptor has direct impact on both its computational complexity and point-matching robustness/accuracy.

A short descriptor may be more robust against appearance variations, but may not offer sufficient discrimination and thus give too many false positives.

The first step consists of fixing a reproducible orientation based on information from a circular region around the interest point. Then we construct a square region aligned to the selected orientation, and extract the SURF descriptor from it.

Furthermore, there is also an upright version of SURF (called U-SURF) that is not invariant to image rotation and therefore faster to compute and better suited for application where the camera remains more or less horizontal.

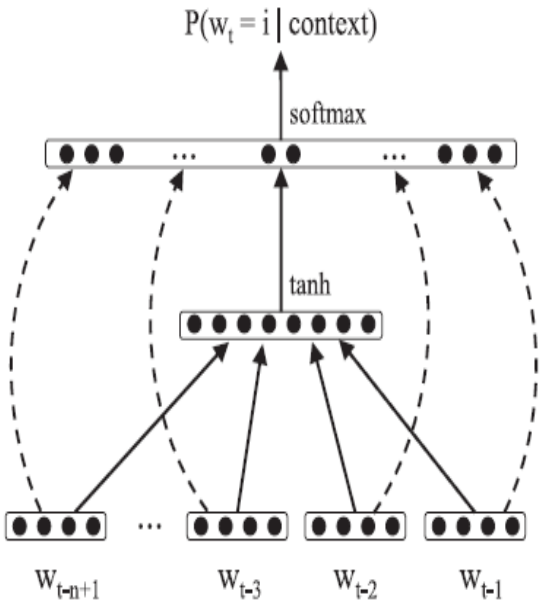

Fig 3. NNLM architecture predicates the probability of words based on the existing words .

A. Word Embedding's Learning

Representation of words as continuous vectors recently has been shown to benefit performance for a variety of NLP and IR tasks similar words tend to be close to each other with the vector representation. The learned word representations could capture meaningful syntactic and semantic regularities. Hence, in this work, we propose to use word embedding's to capture the semantic level similarities between short text segments.

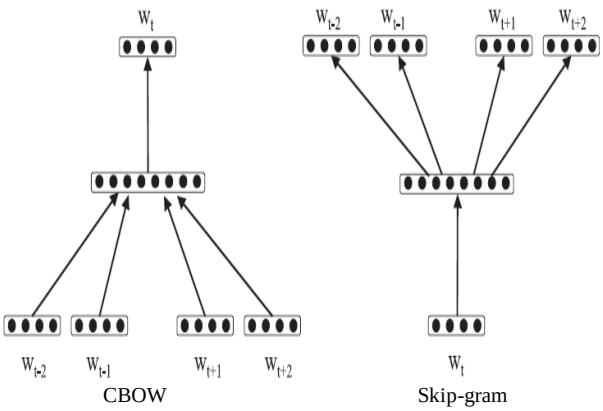

Fig 4. CBOW predicts the current word based on the context and Skip-gram predicts surrounding words given the current word.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

B. Fisher Kernel Framework

Fisher kernel framework was proposed to directly obtain the kernel function from a generative probability model. A parametric class of probability models for some positive integer l. If the dependence on u is sufficiently smooth, the collection of models with parameters from Q can be viewed as a manifold MQ. Though applying a scalar product at each point 2MQ, it can be turned into a Riemannian manifold.

C. Mapping Function Learning

To transfer the FVs of one modality to another, we propose to use a deep belief network with one hidden layer to achieve the task. The building block of the network used in this work is the Gaussian restricted Boltzmann machine. Because we have converted both textual and visual information into the gradient space of a Riemannian manifold, we into his work use a single hidden layer model to do it. The restricted Boltzmann machine is a kind of an undirected graphical model with observed units and hidden units. The undirected graph of an RBM has a bipartite structure. It can be understood as a Markov random field with latent factors which explain the input observed data using binary hidden variables. Let v be the L dimensional observed data, which can take real values or binary values. The dimension of stochastic binary units h is K. Each visible unit is connected to each hidden unit.

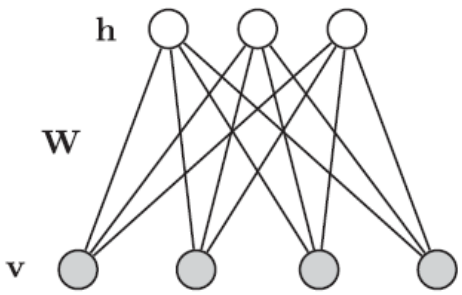

Fig 5. A graphical model representation of restricted Boltzmann Machine.

D. Hash Code Generation

Through the previous steps, a variable length of text segments or key points can be transferred to a fixed length vector. However, Fisher vectors are usually high dimensional and dense. It limits the usages of FVs for large-scale applications, where computational requirement should be studied. In this work, we propose to use hashing methods to address the efficiency problem. Fisher vectors of text segments or key points are the x in mapping function. A variety of existing methods have been proposed to achieve this task under this framework using different forms of f and different optimization objectives. Most of the learning to hash methods for dense vectors can be used in this framework.

V. CONCLUSION

In this work, we propose a novel hashing method, SCMH, to perform the near-duplicate detection and cross media retrieval task. We propose to use a set of word embedding's to represent textual information. Fisher kernel framework is incorporated to represent both 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 commonly used data sets. SCMH achieves better results than state-of-the-art methods with different the lengths of hash codes. Implementing the SURF algorithm has proven to be a challenge.

It has been interesting and time consuming to implement the algorithm from the ground up. If I were to employ the SURF algorithm to a real world problem in the future, this experience will be valuable when adapting a open source implementation to my needs. Having more eyes on the code can help optimize details and assure correct implementations. This would free resources to investigate different variations of parameters and strategies. It is able to detect and describe with consistent results and demonstrates the core principles of the SURF algorithms detection and description scheme. Experimental results demonstrate the effectiveness of the proposed method on the cross-media retrieval task.

REFERENCES

[1] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2014, pp. 2083–2090.
[2] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 12, pp. 2916–2929, Dec. 2013.
[3] F. Wu, H. Zhang, and Y. Zhuang, "Learning semantic correlations for cross-media retrieval," in Proc. IEEE Int. Conf. Image Process, 2006, pp. 1465–1468.
[4] Y. Yang, Y.-T. Zhuang, F. Wu, and Y.-H. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," IEEE Trans. Multimedia, vol. 10, no. 3, pp. 437–446, Apr. 2008.
[5] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 2222–2230.
[6] Q. Zhang, J. Kang, J. Qian, and X. Huang, "Continuous word embeddings for detecting local text reuses at the semantic level," in Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2014, pp. 797–806.
[7] D. G. Lowe, "Object recognition from local scale-invariant features," in Proc. Int. Conf. Comput. Vis., 1999, p. 1150.
[8] P. Daras, S. Manolopoulou, and A. Axenopoulos, "Search and retrieval of rich media objects supporting multiple multimodal queries," IEEE Trans. Multimedia, vol. 14, no. 3, pp. 734–746, Jun. 2012.
[9] A. Z. Broder, "On the resemblance and containment of documents," in Proc. SEQUENCES, 1997, p. 21.
[10] A. Z. Broder, "Identifying and filtering near-duplicate documents," in Proc. Combinatorial Pattern Matching, 2000, pp. 1–10.
[11] M. Theobald, J. Siddharth, and A. Paepcke, "Spotsigs: Robust and efficient near duplicate detection in large web collections," in Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 563–570.
[12] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, pp. 504–507, 2006.
[13] G. Hinton and R. Salakhutdinov, "Discovering binary codes for documents by learning deep generative models," Topics Cognitive Sci., vol. 3, pp. 74–91, 2010.

[14] T. Mikolov, M. Karafi at, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model," in Proc. INTERSPEECH, 2010, pp. 1045–1048.
[15] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," J. Mach. Learn. Res., vol. 3, pp. 1137–1155, 2003.

BIOGRAPHIES

C. Maleappane Lawrence is an Assistant Professor of Master of Computer Application in Christ College of Engineering and Technology affiliated to Pondicherry University, India.

M. Shanmugham is an Assistant Professor of Master of Computer Application in Christ College of Engineering and Technology affiliated to Pondicherry University, India.

D. Sabarinathan is an Assistant Professor of Master of Computer Application in Christ College of Engineering and Technology affiliated to Pondicherry University, India.

T. Banumathi is a final year Student of Master of Computer Application in Christ College of Engineering and Technology affiliated to Pondicherry University, India.

DOI 10.17148/IJARCCE.2016.55225