CROSS-MEDIA HASHING WITH CENTROID APPROACHING

Ruoyu Liu, Yao Zhao, Shikui Wei, Zhenfeng Zhu*

Institute of Information Science, Beijing Jiaotong University, Beijing, 100044, China
Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, 100044, China
Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China
Three Gorges University, Yichang, Hubei 443002, China
Email:shkwei@bjtu.edu.cn

ABSTRACT

Cross-media retrieval has received increasing interest in recent years, which aims to addressing the semantic correlation issues within rich media. As two key aspects, cross-media representation and indexing have been studied for dealing with cross-media similarity measure and the scalability issue, respectively. In this paper, we propose a new cross-media hashing scheme, called Centroid Approaching Cross-Media Hashing (CAMH), to handle both cross-media representation and indexing simultaneously. Different from existing indexing methods, the proposed method introduces semantic category information into the learning procedure, leading to more exact hash codes of multiple media type instances. In addition, we present a comparative study of cross-media indexing methods under a unique evaluation framework. Extensive experiments on two commonly used datasets demonstrate the good performance in terms of search accuracy and time complexity.

Index Terms— Centroid, Cross-media, Hashing

1. INTRODUCTION

With the rapid development of social network and image sharing websites, users are contributing more and more information content. Since no unique rule is followed by users, the structure of content is informal and heterogeneous. That is, an information entity perhaps consists of multiple instances of different media types (heterogeneous instances). It is valuable to build the relationship among those heterogeneous instances and retrieve instances in a heterogeneous manner. To this end, a new research task, called cross-media retrieval [1, 2, 3, 4, 5], is attracting more and more attention. Its aim is to build semantic correlation among instances described by heterogeneous features, so as to directly perform similar search among them. A lot of schemes have been proposed to solve the task. In [1], canonical correlation analysis (CCA)is employed to build the relationship so that instances with the same semantic meaning but different media types have the maximum correlation. In [3], *He et al.* propose a novel Parallel Field Alignment to align heterogeneous instances, which achieves an outstanding performance. For explicitly employing the semantic information, *Xie et al.* propose the semantic generative model (SGM) [5]. To fully utilize correlation among heterogeneous instances, *Yang et al.* propose a structure called multimedia document (MMD) [4]. Recently, *Zhuang et al.* propose a multi-modal dictionary learning scheme called Supervised coupled-dictionary learning with group structures for Multi-Model retrieval (SliM^2) [2]. However, most cross-media retrieval schemes pay more attention on improving search accuracy, yet less effect is given to speed up the retrieval process. Due to the special properties of cross-media retrieval, the traditional indexing schemes are not suitable. Therefore, it is necessary to develop new indexing schemes for rapidly searching similar semantic heterogeneous instances.

This paper proposes a new cross-media hashing method, called Centroid Approaching Cross-Media Hashing (CAMH), which fully explores semantic information to improve the search accuracy of cross-media hash codes. Beside pairwise correlation information among heterogeneous instances, the proposed method also explicitly introduces the semantic category information to the training process of hash model. In this way, not only the hash codes of heterogeneous instances belonging to the same information entity, but also the hash codes of heterogeneous instances labeled as the same category are similar to each other. Besides, a comparative study of cross-media indexing methods is presented. To avoid the effects of datasets and preprocessing steps, all the state-of-the-art techniques are tested and compared under a unique framework, and both search accuracy and time complexity of indexing methods are evaluated.

2. RELATED WORK

Since the structure of content trends to be informal and heterogeneous, single-media based processing techniques like[6, 7, 8, 9, 10, 11] cannot handle such complex information en-

*Corresponding Author

tities. To build the relationship among heterogeneous instances, cross-media hashing [12, 13, 14, 15, 16] aims to learn several hash functions which project instances of varied media types to a shared binary space. Since similarity measure can be carried out by employing highly efficient hamming distance, the retrieval speed is very high. As a good attempt, *Bronsten et al.* propose a cross-modality similarity-sensitive hashing (CMSSH) [12] scheme, in which the value of each bit in hash codes is determined separately by a weak binary classifier. However, CMSSH does not take into account the intra-media similarity which is useful for distinguishing the instances of the same media type (homogeneous instances). To address this problem, *Kumar et al.* extend the spectral hashing from single modality to multiple modalities and proposes a Cross-view hashing (CVH) [13] scheme. The objective of CVH minimizes the hamming distances between both homogeneous and heterogeneous similar instances simultaneously. In fact, the authors prove that CVH is equivalent to CCA when the affinity matrix is not available.

Besides the geometrical methods above, some probabilistic methods are employed to construct cross-media hash functions. In [14], *Zhen et al.* propose a multimodal latent binary embedding (MLBE) scheme, which regards hash codes as the binary latent factors of a generative model. Recently, a new cross-media hashing method, called linear cross-modal hashing (LCMH) [16] is proposed by *Zhu et al.* LCMH decreases the time complexity by representing each instance using the distances to K centroids.

In fact, only the pairwise correlation is taken into account by the abovementioned cross-media hashing methods. Nevertheless, the intra-category correlation is ignored, leading to weak distinguishing capability of hash codes. To address this issue, we fully introduce semantic category information into the learning process of hash functions so that both heterogeneous instances belonging to the same entity and instances labeled as the same category are close in the shared binary space.

3. CENTROID APPROACHING CROSS-MEDIA HASHING

3.1. Problem Description

Assume we have N information entities, each entity is carried by a pair of heterogeneous instances from different media types: $\{x_i^{(1)}, x_i^{(2)}\}_{i=1}^N$, where $x_i^{(1)} \in R^{D(1)}$ and $x_i^{(2)} \in R^{D(2)}$. For example, $x_i^{(1)}$ can be the SIFT feature extracted from an image, and $x_i^{(2)}$ can be the *Latent Dirichlet allocation* (LDA) feature extracted from a text document.

The goal of cross-media hashing is to project heterogeneous instances into a shared binary space, in which both the intra-media similarity and inter-media similarity can be measured directly. For the case of two media types (or modalities), the key is to learn two hash functions, which can be formulated as follows:

$$\begin{aligned} h^{(1)} &: R^{D(1)} \longmapsto \{0,1\}^L \\ h^{(2)} &: R^{D(2)} \longmapsto \{0,1\}^L \end{aligned} \tag{1}$$

where $\{0,1\}^L$ is a commonly shared binary space with L dimensions. In this space, heterogeneous instances (e.g, image and text) can be directly measured by hamming distance.

Generally, the existing methods of cross-media hashing are to maximize the correlation of $x_i^{(1)}$ and $x_i^{(2)}$ as shown in Fig. 1, i.e., fully exploiting pair information. Although these methods guarantee that $x_i^{(1)}$ and $x_i^{(2)}$ are close after hash mapping, the heterogeneous instances belonging to the different entities but having the same semantic meaning will be scattered. In this paper, we attempt to address this problem by introducing semantic category information into the learning process.

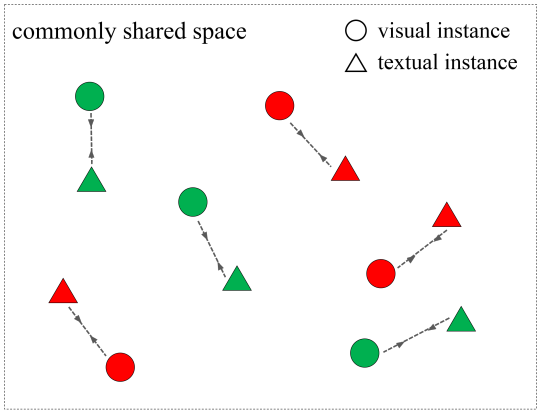

Fig. 1. Illustration of maximizing the correlation of the visual instance $x_i^{(1)}$ (circle) and the textual instance $x_i^{(2)}$ (triangle) belonging to the same information entity.

3.2. Formulation

Assume that each entity is labeled as one of M categories. The key idea of the proposed CAMH method is to minimize the distances between instances and centroids simultaneously, which introduces semantic category information into learning process. The CAMH scheme can be formulated as the following optimization problem:

$$\begin{aligned} \min_{h^{(1)},h^{(2)}} & \|\mathbf{B}^{(1)} - \mathbf{B}^{(2)}\|_F^2 + \lambda_1 \|\mathbf{B}_c^{(1)} - \mathbf{B}_c^{(2)}\|_F^2 \\ & + \lambda_2 \sum_{i=1}^{2} \|\mathbf{B}^{(i)} - \mathbf{B}_c^{*(i)}\|_F^2 \\ s.t., \ & \mathbf{B}^{(i)^T} e = 0, \ \ b(i) \in \{-1, 1\}, \\ & \mathbf{B}^{(i)^T} \mathbf{B}^{(i)} = \mathbf{I}_L, i = 1, 2 \end{aligned} \tag{2}$$

where $\|.\|_F$ means a Frobenius norm, λ_1, λ_2 are two tuning parameters, e is a $N \times 1$ vector whose entries are all 1 and $\mathbf{I}_L$ is a $L \times L$ identity matrix. $\mathbf{B}^{(1)}, \mathbf{B}^{(2)} \in R^{N \times L}$, in which rows represents the hash codes of heterogeneous instances $x^{(1)}$ and $x^{(2)}$ respectively. $\mathbf{B}_c^{(1)}, \mathbf{B}_c^{(2)} \in R^{M \times L}$, are two centroid hash code matrices of M categories. $\mathbf{B}_c^{*(1)}, \mathbf{B}_c^{*(2)} \in R^{N \times L}$, where each row is the hash code of a centroid corresponding to an instance. The constraint $\mathbf{B}^{(i)^T} e = 0$ requires each bit has equal change to be -1 or 1, and the constraint $\mathbf{B}^{(i)^T} \mathbf{B}^{(i)} = \mathbf{I}_L$ requires the bits to obtained independently.

For the previous works, only the first term in Eq.2 is optimized, which preserves only the inter-media similarity of the instances belonging to the information entity as mentioned in section 3.1. In order to introduce the semantic category information, the second and the third items are added into the objective function. In this way, the heterogeneous instances with same semantic category will approach to each other. Fig. 2 illustrates the optimizing process.

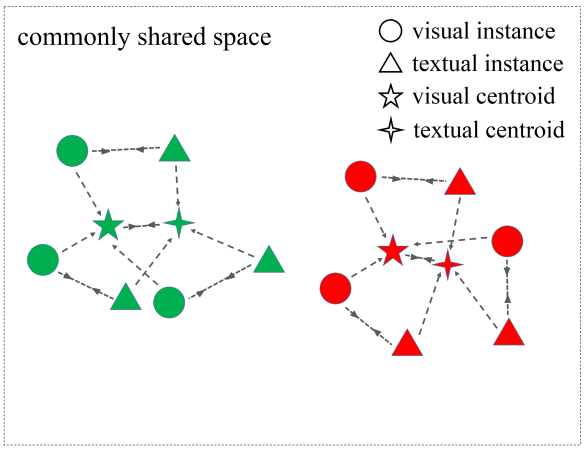

Fig. 2. Illustration of minimizing the distance between the visual instance $x_i^{(1)}$ (circle) and textual instance $x_i^{(2)}$ (triangle) belonging to the same information entity, the distance between visual centroid (pentagram) and textual centroid (cross-star), and the distance between instance and centroid.

3.3. Optimization

The optimization problem in Eq.2 is equal to the issue of balanced graph partition, which is NP hard. Therefore, we relax it into a real-valued case, and then the objective function of CAMH is changed to learn two linear functions:

$$\begin{aligned} f^{(1)}(z_i^{(1)}) &= \mathbf{W}^{(1)^T} z_i^{(1)} \\ f^{(2)}(z_i^{(2)}) &= \mathbf{W}^{(2)^T} z_i^{(2)} \end{aligned} \tag{3}$$

where $\mathbf{W}^{(1)}, \mathbf{W}^{(2)} \in R^{K \times L}$ are two linear projection matrices. $z_i^{(1)}, z_i^{(2)} \in R^K$ are the feature representation of $x_i^{(1)}, x_i^{(2)}$, which is individually obtained by concatenating their distances to K cluster centroids as in [16].

Then we can rewrite Eq.2 to:

$$\begin{aligned} \min_{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}} & \|\mathbf{Z}^{(1)}\mathbf{W}^{(1)} - \mathbf{Z}^{(2)}\mathbf{W}^{(2)}\|_F^2 \\ & + \lambda_1 \|\mathbf{Z}_c^{(1)}\mathbf{W}^{(1)} - \mathbf{Z}_c^{(2)}\mathbf{W}^{(2)}\|_F^2 \\ & + \lambda_2 \sum_{i=1}^{2} \|\mathbf{Z}^{(i)}\mathbf{W}^{(i)} - \mathbf{Z}_c^{*(i)}\mathbf{W}^{(i)}\|_F^2 \\ s.t., \quad & \mathbf{W}^{(1)^T}\mathbf{W}^{(1)} = \mathbf{I}, \quad \mathbf{W}^{(2)^T}\mathbf{W}^{(2)} = \mathbf{I} \end{aligned} \tag{4}$$

where $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}$ are two feature matrices, and each row is a sample of $z_i^{(1)}, z_i^{(2)}$. $\mathbf{Z}_c^{(1)}, \mathbf{Z}_c^{(2)}$ are two feature matrices of centroids. $\mathbf{Z}_c^{*(1)}, \mathbf{Z}_c^{*(2)}$ are also two centroid feature matrices, where each row is corresponded to an instance.

Eq.4 can be reduced to the following generalized eigenvalue problem:

$$\max_{\mathbf{W}} \; tr(\mathbf{W}^T \mathbf{Z} \mathbf{W}) \quad s.t., \quad \mathbf{W}^T \mathbf{W} = \mathbf{I} \tag{5}$$

The matrix $\mathbf{Z}$ is a block matrix constructed by four matrices $\mathbf{Z}_{ij}, \; i, j = 1, 2$:

$$\begin{aligned} \mathbf{Z}_{11} = & -[\mathbf{Z}^{(1)^T}\mathbf{Z}^{(1)} + \lambda_1 \mathbf{Z}_c^{(1)^T}\mathbf{Z}_c^{(1)} \\ & + \lambda_2 (\mathbf{Z}^{(1)} - \mathbf{Z}_c^{*(1)})^T (\mathbf{Z}^{(1)} - \mathbf{Z}_c^{*(1)})] \\ \mathbf{Z}_{22} = & -[\mathbf{Z}^{(2)^T}\mathbf{Z}^{(2)} + \lambda_1 \mathbf{Z}_c^{(2)^T}\mathbf{Z}_c^{(2)} \\ & + \lambda_2 (\mathbf{Z}^{(2)} - \mathbf{Z}_c^{*(2)})^T (\mathbf{Z}^{(2)} - \mathbf{Z}_c^{*(2)})] \\ \mathbf{Z}_{12} = & \mathbf{Z}^{(1)^T}\mathbf{Z}^{(2)} + \lambda_1 \mathbf{Z}_c^{(1)^T}\mathbf{Z}_c^{(2)} \\ \mathbf{Z}_{21} = & \mathbf{Z}^{(2)^T}\mathbf{Z}^{(1)} + \lambda_1 \mathbf{Z}_c^{(2)^T}\mathbf{Z}_c^{(1)} \end{aligned} \tag{6}$$

Then $\mathbf{W}^{(1)}, \mathbf{W}^{(2)}$ are calculated as following:

$$\mathbf{W}^{(1)} = \mathbf{W}(1:K, :), \mathbf{W}^{(2)} = \mathbf{W}(K+1:end, :) \tag{7}$$

where $\mathbf{W}$ is constructed by calculating the L largest eigenvalues' corresponding eigenvectors of Eq.5 as columns.

3.4. Binarization

After two functions of Eq.3 have been learned, we can easily project the entities represented by heterogeneous features into a shared real-valued space. The next step is to binarize the commonly shared and real-valued feature vectors into hash codes. To this end, we employ the same strategy reported in LCMH. Firstly, we relax $\mathbf{B}^{(i)}$ into its real-valued form $\mathbf{Y}^{(i)}$, which is calculated as follows:

$$\mathbf{Y}^{(i)} = \mathbf{Z}^{(i)}\mathbf{W}^{(i)} \tag{8}$$

where $i = 1, 2$. Then we calculate the binarization threshold using the $median$ function:

$$u^{(i)} = median(\mathbf{Y}^{(i)}) \tag{9}$$

where $u^{(i)} \in R^L$.

Finally, we binarize $\mathbf{Y}^{(i)}$ as follows:

$$\begin{cases} b_{jk}^{(i)} = 1 & if \ \ y_{jk}^{(i)} \geq u_k^{(i)} \\ b_{jk}^{(i)} = -1 & if \ \ y_{jk}^{(i)} < u_k^{(i)} \end{cases} \tag{10}$$

where $\mathbf{Y}^{(i)} = [y_1^{(i)}, ..., y_N^{(i)}]^T$, $i = 1, 2$, $j = 1, ..., N$ and $k = 1, ..., L$. j is the index of instances and k is the index of the elements of $y^{(i)}$ and $b^{(i)}$, where $y^{(i)}$ is the real-valued relaxation of $b^{(i)}$. Note that the calculation of median vectors is based on the training samples.

3.5. Extension

The proposed CAMH scheme can be easily extended to more than two media types, while only the case of two media types is discussed above. Details are given in supplemental material.

4. EXPERIMENTAL ANALYSIS

In this section, we present a comparative study of schemes for cross-media hashing. Two popular benchmark datasets, i.e., **Wiki** and **NUS-WIDE**, are employed, and each dataset is divided into two sets, i.e., query set and database set. Due to the experiment environment limit, we only use part of database set for training, i.e., training set.

4.1. Experimental Setup

In our evaluation framework, **Wiki** and **NUS-WIDE** are employed for testing, and mean Average Precision(mAP)is used for performance measure. Time cost in off-line and on-line phases is used for measuring the speed of various methods.

Wiki is generated from a group of 2,866 Wikipedia documents. Each document in is an image-text pair and is uniquely tagged with one of 10 labels. The images are represented by SIFT histograms and the text articles are represented by LDA model. We download the version used in MLBE from [17], which selects 2,289 data points as the database set and treats the rest 577 data points as the query set. In our experiments, we randomly select 300 image-text pairs from the database set as the training set. Note that we treat the labels as the categories.

NUS-WIDE [18] is a dataset downloaded from Flickr, which includes 269,648 images with associated tags. In addition to tags, each image is also assigned to one or several classes. After removing the images without any class, we randomly select 17,600 image-tags pairs from the first 20 classes which have the most data points as a new database due to the internal memory capacity limit, which we call it **NUS-min** database. The images and tags are represented by bag-of-word [19] histogram. Each image-tags pair is labeled with at least one of the 20 classes. In our experiments, **NUS-min** is partitioned into the database set with 10,560 data points and the query set with the remaining 7,040 data points. We randomly select 600 image-tags pairs from the database set as the training set.

We use **mean Average Precision** (mAP) as the performance measure as in MLBE.

We use the **time cost in off-line and on-line phases** to measure the speed of different methods. The time cost in off-line phase contains the time used for training and computing the hash codes of database set. The time cost in on-line phase contains the time spent on computing the hash codes of query set and calculating the hamming distance for cross-media retrieval.

4.2. Comparison Methods

Four state-of-the-art methods are fully tested and compared with the proposed CAMH scheme, which includes CMSSH, CVH, MLBE, and LCMH.

Similar to previous works, we evaluate all the cross-media indexing algorithms on two cross-media retrieval tasks. One is to use a text query to search relevant images in the visual media type (shorted for "Text query *vs.* Image data"), and the other is to use an image query to search relevant texts in the textual media type (shorted for "Image query *vs.* Text data").

4.3. Parameters' Setting

There are three key factors for the LCMH and CAMH schemes when calculating $z_i^{(1)}, z_i^{(2)}$, i.e., the number of clusters K, the number of reserved distances S and the tuning parameter σ of Gaussian function. The K value is set according to the size of training set. In our experiment, We find that $K = 40$ is proper for **Wiki**, and $K = 80$ for **NUS-min**. The value of S is set according to the conclusion in LCMH, here we set $S = 5$. As discussed in LCMH, the tuning parameter σ should make each element of $z_i^{(1)}, z_i^{(2)}$ fits a Gaussian distribution. Because of the different feature representations in two databases, we set $\sigma = 1$ for **Wiki** and $\sigma = 100$ for **NUS-min**. For the two turning parameters in CAMH, we find that $\lambda_1 = 3, \lambda_2 = 2$ is proper.

For CMSSH, CVH and MLBE, since the program codes are provided by authors, we use the default parameters in these algorithms.

The length of hash codes L in our experiments we set 8, 16, and 32, which is the same as in LCMH. We set these values because the hash codes can be easily stored in bytes and measured by fast bitwise operations.

4.4. Accuracy Evaluation

The experimental results on **Wiki** and **NUS-min** are shown in Table 1 and Table 2, repsectively. Clearly, CAMH outperforms the existing methods in most of cases. This means that

introducing semantic category information into cross-media hashing indeed improves the discriminative capability of hash codes.

Table 1. Accuracy Evaluation on **Wiki**.

Task	Method	Code Length		
		$L = 8$	$L = 16$	$L = 32$
Image Query *v.s.* Text Database	CMSSH	0.2007	0.1692	0.1172
	CVH	0.2013	0.1571	0.1514
	MLBE	0.2198	**0.2191**	**0.1823**
	LCMH	0.2062	0.1666	0.1668
	CAMH	**0.2304**	0.2032	0.1791
Text Query *v.s.* Image Database	CMSSH	0.1570	0.1436	0.1645
	CVH	0.2639	0.2641	0.1997
	MLBE	0.2294	0.1503	0.0974
	LCMH	0.2210	0.2207	0.2503
	CAMH	**0.3071**	**0.3667**	**0.4143**

Table 2. Accuracy Evaluation on **NUS-min**.

Task	Method	Code Length		
		$L = 8$	$L = 16$	$L = 32$
Image Query *v.s.* Text Database	CMSSH	0.1774	0.1983	0.1523
	CVH	0.1950	0.1980	0.1972
	MLBE	0.1321	0.1577	0.2161
	LCMH	0.1946	0.1953	0.1974
	CAMH	**0.2443**	**0.2498**	**0.2577**
Text Query *v.s.* Image Database	CMSSH	0.2046	0.1925	0.1565
	CVH	0.1955	0.1961	0.2032
	MLBE	0.1828	0.2014	0.2104
	LCMH	0.1973	0.2044	0.2099
	CAMH	**0.2443**	**0.2372**	**0.2348**

From the point of view of model optimization, we can separate the five methods (one is ours) into two groups, the ones based on *iteratively optimization* and the ones based on *eigenvalue decomposition*. CMSSH and MLBE belong to the former group, and the other three methods belong to the latter group. However, the weakness is that their speed is extremely slower than the other methods.

For CVH, LCMH and CAMH, all of them transfer the hash functions learning problem into a two-step process: firstly map heterogeneous instances of different media types into a shared space, and then binarize each bit. They are all related with CCA, and the optimization is finally formulated as an eigenvalue decomposition problem. The performance of these methods mainly depends on how much information contained in the model. CAMH works best, because intra-category correlation is considered together with the pairwise correlation preserved in other four methods.

We present an example of CAMH's retrieval result of two tasks in supplemental material. The results show that CAMH works well, since correct results are ranked at the top of the result list.

4.5. Speed Evaluation

To quantitatively evaluate the time complexity of the state-of-the-art indexing methods, we make a statistic of their time cost in both off-line and on-line phases, respectively. Fig. 3 shows the results obtained on **Wiki** database and **NUS-min** database. Notice that the time on **Wiki** is expanded 100 times (denoted as "100×") for easy illustration.

Clearly, the time complex of the two methods in iteratively optimization group is much higher than that of the other three methods in eigenvalue decomposition group. MLBE has the longest time cost in off-line phase, because it needs to iteratively optimize the model until it is converged. CMSSH has the longest time cost in the on-line phase, because it uses the weighted hamming distance as similarity measure. Therefore, CMSSH and MLBE are not suitable methods in real-world applications when the dataset is large.

The methods in eigenvalue decomposition group have the similar speed in both off-line and on-line phases, which are much faster than the methods in iteratively optimization group. We can see that CAMH does not increase the time complexity.

5. CONCLUSIONS

This paper proposes a new cross-media hashing method, namely Centroid Approaching Cross-media Hashing (CAMH). The main idea is to introduce semantic category information into learning process, and preserve the correlation between heterogeneous instances and centroids as well. In this way, both intra-category and pairwise correlations are considered when learning the hash functions. Experiments on two commonly used datasets show the proposed cross-media hashing method outperforms state-of-the-arts in terms of accuracy and speed.

6. ACKNOWLEDGEMENTS

This work was supported in part by National Basic Research Program of China (No.2012CB316400), NationalNatural Science Foundation of China (No.61202241, No.61210006), Program for Changjiang Scholars and Innovative Research Team in University (No.IRT201206), Fundamental Research Funds for the Central Universities(No.2015JBM028), and Joint Fund of Ministry of Education of China and China Mobile (No.MCM20130421).

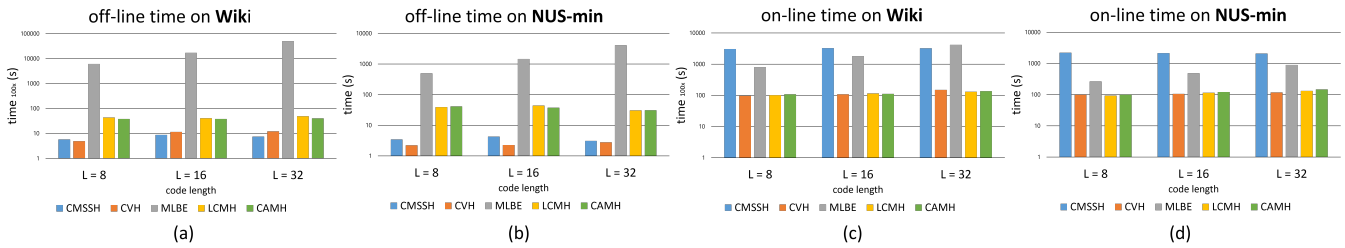

Fig. 3. Speed Evaluation on **Wiki** and **NUS-min**.

7. REFERENCES

[1] 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 *ACMMM*. ACM, 2010, pp. 251–260.

[2] Y.Zhuang, Y.Wang, F.Wu, Y.Zhang, and W.Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval.," in *AAAI*, 2013.

[3] X.Mao, B.Lin, D.Cai, X.He, and J.Pei, "Parallel field alignment for cross media retrieval," in *ACMMM*. ACM, 2013, pp. 897–906.

[4] 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.

[5] L.Xie, P.Pan, and Y.Lu, "A semantic model for cross-modal and multi-modal retrieval," in *ACM ICMR*. ACM, 2013, pp. 175–182.

[6] S.Wang, Q.Huang, S.Jiang, and T.Qi, "S3mkl: Scalable semi-supervised multiple kernel learning for real-world image applications," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1259–1274, 2012.

[7] S.Wang, Q.Huang, S.Jiang, and T.Qi, "Nearest-neighbor method using multiple neighborhood similarities for social media data mining," *Neurocomputing*, vol. 95, pp. 105–116, 2012.

[8] Shikui Wei, Dong Xu, Xuelong Li, and Yao Zhao, "Joint optimization toward effective and efficient image search," *IEEE Transactions on,Cybernetics*, vol. 43, no. 6, pp. 2216–2227, 2013.

[9] Shikui Wei, Yao Zhao, Zhenfeng Zhu, and Nan Liu, "Multimodal fusion for video search reranking," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 8, pp. 1191–1199, 2010.

[10] J.Li, D.Xu, and W.Gao, "Removing label ambiguity in learning-based visual saliency estimation," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1513–1525, 2012.

[11] J.Li, H.Tian, T.Huang, and W.Gao, "Multi-task rank learning for visual saliency estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 5, pp. 623–636, 2011.

[12] M.M.Bronsten, A.M.Bronstein, F.Michel, and N.Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *CVPR*. IEEE, 2010, pp. 3594–3601.

[13] S.Kumar and R.Udupa, "Learning hash functions for cross-view similarity search," in *IJCAI*, 2011, vol. 22, p. 1360.

[14] Y.Zhen and D.Y.Yeung, "A probabilistic model for multimodal hash function learning," in *ACM SIGKDD*. ACM, 2012, pp. 940–948.

[15] Y.Zhen and D.Y.Yeung, "Co-regularized hashing for multimodal data," in *NIPS*, 2012, pp. 1376–1384.

[16] X.Zhu, Z.Huang, H.T.Shen, and X.Zhao, "Linear cross-modal hashing for efficient multimedia search," in *ACMMM*. ACM, 2013, pp. 143–152.

[17] Y.Zhen, "Wiki dataset constructed by Yi Zhen," Website, `https://dl.dropboxusercontent.com/u/41679313/MLBE_data.zip`.

[18] T.S.Chua, J.Tang, R.Hong, H.Li, Z.Luo, and Y.Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *ACM ICIVR*. ACM, 2009, p. 48.

[19] J.Sivic and A.Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*. IEEE, 2003, pp. 1470–1477.

[20] D.Y.Yeung, "Dit-Yan Yeung publication list," Website, `http://www.cse.ust.hk/~dyyeung/paper/publist.html`.