Liu *et al. EURASIP Journal on Image and Video Processing* (2017) 2017:17
DOI 10.1186/s13640-017-0167-4

EURASIP Journal on Image and Video Processing

REVIEW Open Access

Deep learning hashing for mobile visual search

Wu Liu[1*], Huadong Ma[1], Heng Qi[1], Dong Zhao[1] and Zhineng Chen[2]

Abstract

The proliferation of mobile devices is producing a new wave of mobile visual search applications that enable users to sense their surroundings with smart phones. As the particular challenges of mobile visual search, achieving high recognition bitrate becomes the consistent target of existed related works. In this paper, we explore to holistically exploit the deep learning-based hashing methods for more robust and instant mobile visual search. Firstly, we present a comprehensive survey of the existed deep learning based hashing methods, which showcases their remarkable power of automatic learning highly robust and compact binary code representation for visual search. Furthermore, in order to implement the deep learning hashing on computation and memory constrained mobile device, we investigate the deep learning optimization works to accelerate the computation and reduce the model size. Finally, we demonstrate a case study of deep learning hashing based mobile visual search system. The evaluations show that the proposed system can significantly improve 70% accuracy in MAP than traditional methods, and only needs less than one second computation time on the ordinary mobile phone. Finally, with the comprehensive study, we discuss the open issues and future research directions of deep learning hashing for mobile visual search.

Keywords: Mobile visual search, Deep learning hashing, Deep learning optimization, Mobile location recognition

1 Review

1.1 Introduction

The proliferation of increasingly capable mobile devices opens up exciting possibilities for massive mobile applications. Among them, mobile visual search, which can utilize mobile device to sense and understand what the users are watching at any time from any place, plays a key role in these applications. First of all, the always-on broadband connection makes users always online. In addition, the abundant sensors can accurately supply sufficient and effective information for mobile perception. More important, the increased computational ability of mobile device can instantly process the sensed information and fetch the related feedback. Therefore, we can conveniently sense where we are [1], what we are watching [2–4] or what happened with our surrounding [5, 6] with mobile visual search immediately.

However, compared with traditional visual search applications, mobile visual search faces the following unique challenges [7]. (1) *Large visual variance of query*—the visual query is naturally disturbed by varying visual qualities in the complex capture conditions, which needs robust visual signatures that can handle such significant variance in mobile visual search. (2) *Stringent memory and computation constraints*—as the cheaper CPU, GPU and memory of mobile devices, signatures with large memory costs or heavy computation cannot be utilized on mobile clients. (3) *Network bandwidth limitations*—as the unreliable and low bandwidth, signatures are expected to be as compact as possible to reduce network transmission latency. (4) *Instant search experience*—because mobile users care more about their experience, the visual search process is expected to be instant and progressive.

To solve the unique challenges of mobile visual search, the most existed searches focus on how to achieve high recognition bitrate, which considers the recognition performance with respect to the amount of data transmission between mobile devices and servers [8]. High recognition

*Correspondence: liuwu@bupt.edu.cn
[1]Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, 100876 Beijing, China
Full list of author information is available at the end of the article

bitrate leads to faster response time, lower network usage rate, and battery consumption, which are all important factors for real mobile visual search applications. According to the transferred query types, the existed works could be classified into four categories : transfer scaled-down images [9], transfer moderate features [10–12], transfer compressed features [13, 14], and transfer feature signature produced by hashing [15, 16]. Among them, as the high robust, lower transmission costs, less memory requirement, and cheaper computation, the hashing based feature compression method attracts the most attention. For example, He et al. [15] and Tseng et al. [17] propose to utilize the visual hashing bits to compact the raw visual descriptors, which contains two stages. Firstly, in the offline stage they learn the hash function from the large scale image database to maximally maintain the discriminative characteristic of raw features. Then in the online stage, the raw visual features are compressed into compact hash bits by the learned hash function to reduce the feature scale. However, as the existed hash-based methods all focus on how to compress the existed handcrafted features into binary codes, their performance is limited by the utilized features and the information loss in the compression process.

Recently, as the ability of automatic feature representation learning from large scale image dataset, deep learning methods are widely employed for content based image retrieval [18–21]. These studies achieved competitive results compared with the traditional methods. Consequently, how to take advantage of deep neural network to automatically learn the binary hash codes attracts massive researchers' attention. The primal works directly add a hidden hash layer into the well-trained deep neural network to map the learned feature representation into binary code. Although they achieve great improvement compared with the traditional hashing methods, the separation of feature and hash function learning cannot sufficiently exploit the power of deep neural network to learn effective binary code representation. Therefore, more researchers try to jointly learn the feature representation and hashing function to holistically exploit the power of feature learning in deep neural network, such as supervised deep learning hashing [22–25], unsupervised deep learning hashing [26], and triplet/pairwise similarities based deep learning hashing [27–29]. According to the evaluations, they all achieve better performance than the above separate methods.

Although the deep learning hashing methods show remarkable power in image and video retrieval, few of them have been applied to the mobile visual search. The main reasons are that the cheap hardware of mobile devices cannot meet the high computation and memory requirement of deep learning. On one side, the multiple convolutional layers in the network require massive computation resources, such as GPU cards with thousands CUDA cores. Moreover, the millions of parameters in the full connected layers need large space in storage and memory. As a result, more and more researchers try to accelerate the convolutional operation [30–33] and compress the parameters in full layers [34–38] to make the deep learning efficiently work on mobile devices. These optimizations make it is possible to run deep learning hashing methods for the mobile visual search [39, 40].

In this paper, we comprehensively investigate the possibility of exploiting the deep learning hashing for mobile visual search. First of all, we survey the existed four categories of mobile visual search methods to show that hashing based methods play significant roles to achieve the high recognition bitrat as their strong determinativeness, high compressed codes, easily transmitted and indexed. Then, we study the existed works of the deep learning hashing, and demonstrate their advantages for mobile visual search. We classify the existed works into three categories, analyze their network structure and training strategy. In addition, we also compare their performance on the CIFAR-10 dataset. Moreover, to solve the challenges of running deep learning hashing on the mobile device, we summarize and analyze the existed deep learning acceleration technologies in two categories: computation acceleration and model reduction. In particular, as shown in Fig. 1, we design a deep learning hashing based mobile location recognition system, which achieves 72% MAP improvements in accuracy, and needs less than one second computation on a common mobile device. The case study sufficiently demonstrates our hypotheses and conclusions. Finally, we discuss the emerging topics of deep learning hashing-based mobile visual search in the future.

The rest of the paper is organized as follows. Section 1.2 reviews mobile visual search work. Section 1.3 surveys the deep learning hashing methods. Section 1.4 introduces the newest deep learning optimization schemes. Section 2 demonstrates a case study and evaluations. Section 3 gives future research directions, followed by the conclusions in Section 4.

1.2 Traditional mobile visual search methods

Recently, as the basic function for numerous mobile applications, mobile visual search attracts massive researchers' attention [41]. As described in Section 1.1, the existed mobile visual search works mostly focused on how to achieve high recognition bitrate [8], which could be classified into four categories according to their transferred query types: transfer scaled-down or compressed images [9, 42, 43], transfer moderate features [10–12], transfer compressed features [13, 14, 44, 45], and transfer feature signature produced by hashing [15, 16, 46–48].

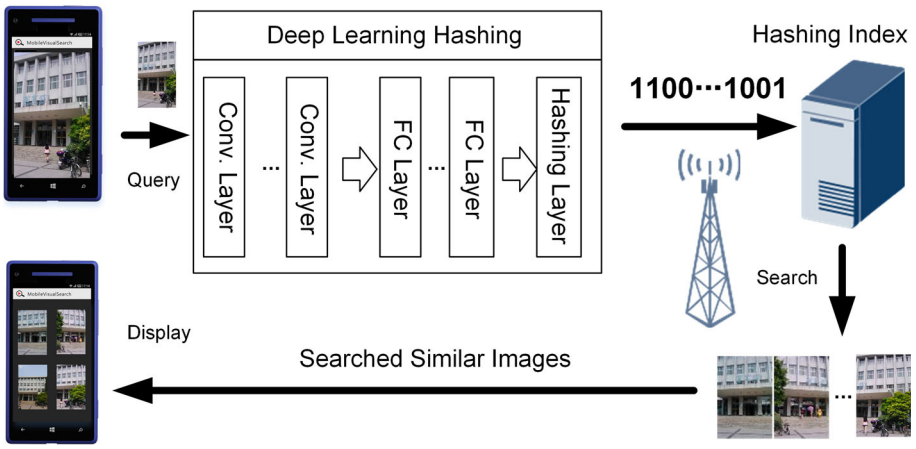

Fig. 1 The framework of deep learning hashing-based mobile visual search

1.2.1 Transfer scaled-down or compressed images

Early mobile visual search systems (e.g., Google Goggles) always try to send the compressed query image to the server, and apply the feature extraction and search on the server side [9]. For instance, Yue et al. propose a cloud-based image coding method to compress the images, and reconstruct them from a large-scale image database via the descriptions on the server side [42]. Moreover, Tan et al. propose a query image resize method that can preserve the robust local features in the image [43]. However, these methods which neglect the increasing computing capacity of mobile clients are limited by the low and unreliable bandwidth of network [49, 50]. Therefore, more works prefer to directly extract moderate features on the mobile device and transfer the features instead of images.

1.2.2 Transfer-moderate features

The moderate features-based methods always utilize the popular visual features applied in desktop visual search, such as speeded-up robust features (SURF) [51] and bag-of-words (BoW) [52]. In addition, the special extraction or transmission improvement is implemented on mobile device to speed up the visual search. For instance, Yang et al. propose to accelerate SURF extraction on the mobile device by content-aware tiling and gradient moment based orientation operator [10]. The content-aware tiling divides the image into tiles. Then, the feature detection is performed on each tile individually to reduce memory traffic. The heterogeneous tile size can be automatically selected by the gradient moment based orientation operator. Besides, Chandrasekhar et al. [12] and Xia et al. [11] propose that they can significantly reduce the data size transmitted over the network to decrease the retrieval latency with the progressive transmission of local features.

1.2.3 Transfer-compressed features

In order to further decrease the transferred feature scale, many recent methods try to compress the extracted moderate features on the mobile device. For instance, the Compressed Histogram of Gradients [13] encodes the raw features with an entropy-based coding method on the mobile client, and approximately decodes the compressed features on the server. Similarly, Ji et al. [14] take advantage of rich contextual cues to compress the raw BoW on the client with a multiple-channel coding scheme. Moreover, Chen et al. [44] develop a compact and discriminative global signature to characterize each image. The global signature employs an optimized local feature count derived from a statistical analysis of the retrieval performance. Finally, Bianco et al. investigate the use of different detectors and color descriptors in the compact descriptors for visual search framework, and demonstrate the advantages of using color descriptors on six benchmark datasets [45].

1.2.4 Transfer feature signature produced by hashing

Different from feature compressing, He et al. [15] and Tseng et al. [17] have suggested to utilize the visual hashing bits to present raw visual descriptors. This kind of method includes the (1) offline hashing function learning and (2) online binary code extraction stages. Compared to transfer the compressed features, without decoding, the hash bits can be directly searched and indexed on the server. Therefore, this kind of methods has lower transmission costs, cheaper memory and computation than others. Moreover, Zhou et al. [48] propose a codebook-free algorithm for large-scale mobile visual search, which firstly employs a novel scalable cascaded hashing scheme to ensure the recall rate of local feature matching, and enhances the matching precision by an efficient verification with the binary signatures of these local features.

In addition, Zhu et al. [46] propose a topic hypergraph hashing for mobile image retrieval, which learns hashing codes with high order semantic correlation preserving, and simultaneously leverages the associated textual modality to enrich semantics of hashing codes. Besides, to mitigate the information loss from binary codes, based on the hashed binary codes transmitted to the server, Kuo et al. [47] propose a de-hashing process that reconstructs the BoW by leveraging the computing power of remote servers. In addition, Liu et al. [7] also tries to further decrease the transmission size of binary code with progressive transmission strategy.

From the above works, we can find that more and more researchers try to transfer feature signatures produced by hashing in mobile visual search, because of its good balance among computation and memory requirements, training efficiency, quantization complexity, and search performance. However, the exited hash based methods for mobile visual search all try to compress the exited classical handcrafted features into binary code. None of them try to automatically learn the effective binary code feature with deep neural network from the large scale image dataset. There are two main reasons: (1) lack of effective deep learning hashing method; and (2) high computational complexity of deep neural network. Next, we will survey the existed methods that try to solve the two main problems.

1.3 Deep learning hashing

1.3.1 Background

Hashing, a widely-studied solution to approximate nearest neighbor search, aims to transform the data item to a low-dimensional representation, or equivalently a short code consisting of a sequence of bits, called hash code [53]. Hashing methods have been intensively studied and widely used in many different fields, including computer graphics, computational geometry, telecommunication, computer vision, especially for mobile visual search. The existed hashing methods can be divided into two categories: data-independent method and data-dependent method (i.e., learning to hash method). The goal of learning to hash is to learn data-dependent and task-specific hash functions that yield compact binary codes to achieve good search accuracy, where sophisticated machine learning tools and algorithms have been adapted to the procedure of hash function design [54, 55]. The existed learning to hash methods can be divided into eight categories: (1) unsupervised hashing, (2) supervised hashing, (3) ranking-based hashing, (4) multi-modal hashing, (5) online hashing, (6) quantization for hashing, (7) distributed hashing, and (8) deep learning hashing. We classified these methods into eight categories to emphasize some important categories, such as deep learning hashing and quantization for hashing. Actually the classification boundaries are not strict with each other. Traditional image search systems based on learning to hash mainly involve two steps: Firstly, the system extracts a vector of hand-crafted descriptors such as HoG, SIFT, SURF, etc. Next, the hashing function learning is posed as either a pointwise or a pairwise optimization problem to preserve the pointwise or pairwise label information in the learned Hamming space [27, 56–58]. However, as the above two steps are mostly studied as two independent problems, the learned feature representation may not be tailored to the objective of hashing function learning.

Because it demonstrated as an effective image content understand and tackle scheme, deep learning technology [59, 60], also known as deep neural networks, attracts growing interests in the fields of image and video search. Deep learning is a biologically-inspired variant of multilayer perception, which supplies an end-to-end framework for feature extracting and classifier training on large scale dataset [61–63]. Furthermore, features extracted by deep learning model show extraordinary performance over overwhelming majorities of existing hand-crafted features [64–66]. Therefore, many researches try to propose an end-to-end deep learning hashing framework to automatically learn effective binary code representations for images [67]. As shown in Table 1, the existed deep learning hashing methods can be divided into three categories: (1) supervised deep learning hash with single network; (2) unsupervised deep learning hash with single network; and (3) pairwise/triple similarity based deep learning hash with parallel network. The image search results on the CIFAR-10 dataset of different deep learning hashing methods are shown in Table 2 for comparation. The accuracy in terms of MAP with different hash bits length, which are all collected from their papers.

1.3.2 Learning feature and hashing function separately

The early works on deep learning hashing continue the traditional hashing strategy. First of all, these works train the deep neural network on large scale image datasets to learn the effective features for image search. Then a hidden hash layer is added to learn the hashing function which maps the learned features into binary code. For example, Lin et al. [68] propose an effective deep learning framework to generate binary hash codes for images with labels, which consists of two main components. The first step is the supervised pre-training of a convolutional neural network on the ImageNet to learn rich mid-level image representations. Then, they add a latent layer to the network and have neurons in this layer learn hashes-like representations while fine-tuning it on the target domain dataset. In addition, Liong et al. propose to learn the hash layer under three constraints: (1) the loss between the original feature descriptor and the learned binary vector is minimized, (2) the binary codes distribute

Table 1 Characteristics of recent deep learning hashing methods

ID	Method	Category	Network	Layer	Feature+hashing
1	Deep learning of binary hash [68]	Supervised	Single	8	Separate
2	Deep hashing for compact binary codes [69]	Unsupervised/supervised	Single	3	Separate
3	Unsupervised deep neural networks hashing [23]	Unsupervised	Single	16	Separate
4	Supervised deep hashing [22]	Supervised	Single	5	Together
5	Semantics-preserving deep hashing [23]	Supervised	Single	8	Together
6	Deep semantic ranking hashing [24]	Supervised	Single	8	Together
7	Binary deep neural network hashing [26]	Unsupervised/supervised	Single	5	Together
8	Bit-scalable deep hashing [27]	Triplet	Parallel	10	Together
9	One-stage deep hashing [28]	Triplet	Parallel	10	Together
10	Deep pairwise-supervised hashing [29]	Pairwise	Parallel	8	Together
11	Deep hashing network [72]	Pairwise	Parallel	8	Together
12	Deep semantic-preserving and ranking-based hashing [73]	Triplet	Parallel	19	Together

evenly on each bit, and (3) different bits are as independent as possible. However, their neural network only has three layers [69]. Moreover, besides the three constraints, the unsupervised deep neural networks hashing also add the rotation invariant into the learning of the binary descriptors, which further improve the performance of unsurprised deep learning hashing [70]. However, as the above works treat feature learning and hashing function learning as two separate stages, they also have the similar problem as the traditional learning hashing method. That is, the quality of produced hash codes heavily depends on the quality of handcrafted features or learned deep learning features, which weakens the feature learning ability of deep neural network and generates less effective hash codes.

1.3.3 Learning feature and hashing function simultaneously

Aim to jointly learn the feature and hashing function simultaneously, Xia et al. [22] propose a supervised deep learning hashing to integrate image feature and hashing value learning into a joint learning model. The model firstly consists a stage of learning approximate hash codes with given supervised information and then trains a deep CNN that outputs continuous hash values. Such hash values can be generated by activation functions like sigmoid, hyperbolic tangent or softmax, and then quantized into binary hash codes through appropriate thresholds. As the power of CNNs, the joint model can simultaneously learn image features and hash values from raw image pixels. However, the above work still requires separately learning approximate hash codes firstly to guide the next subsequent learning of image representation and finer hash values. Differently, [23] and [24] both propose end-to-end deep learning frameworks to learn the hashing function with semantic information of images. The supervised semantics-preserving deep hashing in [23] constructs the hash functions as a latent layer between image representations and classification outputs in CNN. Then binary codes can be learned by the minimization of an objective function defined over classification error, with additional constraints on the learning objective to make each hash bit carry as much information as possible. Therefore, the learned binary code encourages semantically similar images to have small Hamming distance. Differently, the works in [24] utilize the multi-label images to learn deep semantic ranking based hashing. Moreover, Li et al. directly propose a binary deep neural network, which designs one layer to directly output the binary code instead of involving the *sgn* or step function in [22–24]. Besides, the authors also alternate and relax the optimization object to solve the NP-hard problem of optimizing the binary code with similarity preserving, independence, and balance properties together.

As a kind of particular supervised hashing, similarity-preserving hashing is also a widely utilized method for large-scale image search tasks. In training, the input of similarity-preserving hashing is in the form of triplets or pairwise similar/dissimilar images. The binary codes are learned to keep the original similarities of the input tripltes/pairs. For example, Lai et al. [28] propose a "one-stage" supervised deep hashing architecture that has three parts: (1) shared stacked convolution layers to capture the image representations, (2) divide-and-encode modules to divide intermediate image features and map them into multiple hash codes, and (3) a triplet ranking loss function which is designed to keep triple relationship on images. Similarly using triple images, Zhang et al. propose a supervised learning framework to generate compact and bit-scalable hashing codes directly from raw images [70]. Besides the similarity-preserving, each bit of the hashing codes is unequally weighted so that the hashing framework can manipulate the code lengths by truncating

Table 2 Image search results on the CIFAR-10 dataset for different deep learning hashing methods

ID	Method	12/16-bits	32-bits	48-bits	64-bits
1	Deep learning of binary hash [68]	89.30%	89.72%	89.73%	-
2	Deep hashing for compact binary codes [69]	46.75%	51.01%	-	52.50%
3	Unsupervised deep neural networks hashing [23]	19.43%	24.86%	-	27.73%
4	Supervised deep hashing [22]	46.5%	52.1%	53.2%	-
5	Semantics-preserving deep hashing [23]	-	-	89.97%	-
6	Binary deep neural network hashing [26]	67.32%	69.62%	-	-
7	Bit-scalable deep hashing [27]	55.2%	55.8%	58.1%	-
8	One-stage deep hashing [28]	61.46%	62.87%	63.05%	63.26%
9	Deep pairwise-supervised hashing [29]	71.3%	74.4%	75.7%	-
10	Deep hashing network [72]	55.5%	60.3%	62.1%	-
11	Deep semantic-preserving and ranking-based hashing [73]	≈ 78%	≈ 78%	≈ 77%	-

The accuracy in terms of MAP with different hash bits length, which are collected from their papers

the insignificant bits. Moreover, Li et al. propose a deep pairwise-supervised hashing to firstly perform simultaneous feature learning and hash-code learning for applications with pairwise labels [71]. Compared with [28, 70], the main difference of the deep pairwise-supervised hashing is that the triplet ranking loss is replaced by the pairwise ranking loss, which is similar to Siamese Neural Network. In addition, Zhu et al. extend the original pairwise rank loss to the pairwise cross-entropy loss and a pairwise quantization loss together [72]. Besides, Yao et al. present a novel deep semantic-preserving and ranking-based hashing architecture, which jointly learns projections from image representations to hash codes and classification [73].

1.3.4 Discussion

In conclusion, although many state-of-the-art deep learning hashing methods have been proposed to demonstrate the power of binary code features learning in deep neural networks, few of them have been implemented in the mobile visual search. The main reasons can be concluded as follows:

- As introduce in Section 1.1, due to the complex capture conditions, the query image is naturally noisy with varying visual qualities, such as flashing, occupy, rotation, blur, affine transformation, etc. Therefore, except the three hash code constraints in the learning process, how to handle these specific noisy in the mobile visual search is another big challenge for the deep learning hashing-based mobile visual search.
- Undoubtedly, the deep neural networks have very high computation requirement for the hardware, such as GPU card and large memory. In particular, the comprehensive and sufficient binary code feature learning needs very deep neural network. Although the training process can be solved on the cloud, the feature extraction in the search process still gives big challenges for the memory and computation limited mobile devices. Therefore, how to accelerate the deep learning hashing computation and compress the model on the mobile device are very pressing problems. Fortunately, some researches have tried to transfer the deep learning technology into the mobile device, which will be introduced in the next session.

1.4 Deep learning optimization on mobile devices

Although deep learning technology achieves great success in a wide range of visual applications, its high computation and large memory requirement also create big problems for many applications such as mobile visual search. The existed effective deep neural networks mainly deponed on their deeper network architectures, which can only be implemented on the server with GPU card and high memory. Therefore, many existed works try to further improve their efficiency, which can be divided into three categories.

1.4.1 Speed up the convolutional layers

Speeding up the computation in the convolutional layer is a common method to accelerate the deep neural network [30–33, 74]. For example, Lebedev et al. propose a two-step framework to speed up convolution layers based on tensor decomposition and discriminative fine-tuning [32]. The tensor decomposition uses non-linear squares to compute a low-rank CP-decomposition to decompose the full kernel tensor. Then the original convolutional layers are replaced by four convolutional layers with small kernels. After that, the new network will be fine-tuned on the training dataset again. The evaluations show that the new network achieves a 8.5× CPU speedup of the whole network with only very little accurate drop. Moreover, Zhang et al. try to accelerate the

very deep convolutional networks with the nonlinear asymmetric reconstruction, which achieves a 4× speedup with merely a 0.3 percent increase of top-5 center-view error [33].

1.4.2 Compress the parameters to reduce the model size

As the millions of parameters in the deep learning model, how to reduce the parameter number and compress the model is another research topic to accelerate the deep neural network [34–38]. For instance, Chen et al. use a hash function to randomly group network connections into hash buckets to make the connections in same hash bucket share the same weight [34]. Moreover, Han et al. use pruning, trained quantization and huffman coding to compress the deep model [36]. Differently, Srinivas et al. directly remove the similar and redundant neurons [37]. The deep pried convnet is an end-to-end trainable network which tries to replace the fully connected layers of the network with an Adaptive Fastfood transform [38].

1.4.3 Accelerate the network on mobile devices

The above works are still implemented on the servers. For mobile devices, Wu et al. propose a Quantized CNN to simultaneously speed-up the computation and reduce the storage and memory of CNN models [39]. They employ the approximate inner product computation to estimate the response of both convolutional and fully-connected layers. Then the estimation error is also considered in the training process. According to their evaluations, the Quantized CNN achieves 4× acceleration and 15× compression for the common CNN network, with only less than 1% drop in the top-5 classification accuracy. On Huawei Mate 7 smartphone (i.e., 1.8GHz Kirin 925 CPU), the practical running time, storage, and memory consumption of optimized AlexNet is 0.95s, 12.60MB, and 74.65MB respectively. In addition, Rallapalli et al. [40] try to accelerate very deep neural networks (i.e., YOLO [75] with 27 layers) on mobile devices. Their strategy is a range of memory optimizations which includes: (1) reducing useless variables in the network; (2) using managed memory in the GPUs; (3) slitting the FC layer into subparts, which are loaded and executed sequentially; and (4) offloading the FC layer to the CPU while pipelining CPU and GPU computation. Then implementation on NVIDIA Jetson TK1 board shows that the optimized YOLO needs 0.262*s* with 2.2% accuracy loss.

From the existed works, we can find that speeding up the convolutional layer and reducing the weights in full layers are two common methods to accelerate the deep neural network on the mobile devices. Although the networks continue to become deeper, we believe that with the development of hardware on the mobile devices and improvement of speed-up technologies, the deep learning hashing can be commonly used on the mobile devices.

2 A case study: mobile visual search for location recognition

Location recognition (i.e., logical localization) is one of the most important applications for mobile visual search. Different from the physical localization which gives the location of users or devices, the location recognition is to localize the objects captured by the mobile devices. For example, when the traveller takes a photo of one building in the city, the location recognition can instantly recognize and tag the photos with the name and location of building [76]. Although GPS embedded with mobile device can easily give the user/device location, how to localize the building from crowd buildings is a great challenge. In the past, mobile visual search was a main method to solve this problem. Given the images of the captured objects, the system will search the similar images which have location labels in the dataset to determine its location. Different from the existed mobile visual search scheme described in section 2, we try to use the deep learning hashing to solve the challenges of mobile visual search.

To evaluate the proposed method, we invited seven volunteers to collect 8,062 images which contains 162 object locations in and around our campus. As shown in Fig. 2, the objects contains buildings, trees, statues, restaurant, supermarket, library, dormitories, and so on. Then we randomly select 6442 images as source images and 1620 images as query images. More details of the dataset can be found in [77, 78].

In the implementation, as the highest accuracy and open-source codes, we choose Deep Learning of Binary Hash (DLBH) [68] and Supervised Semantics-Preserving Deep Hashing (SSDH) [23] as our deep learning hashing methods. We implement DLBH to learn deep learning hashing for mobile location recognition in two main steps. Firstly, the AlexNet model provided by [68] is fine-tuned on our source dataset with the label of 162 different objects. In this step, we can learn sufficient mid-level image representations for location recognition. Next, the latent hash layer is added to the AlexNet, and learn hashes-like representations to minimize the loss between the original feature descriptors and the learned binary vectors. Different from DLBH, except the classification loss, SSDH also adds the constraints that: (1) the binary codes distribute evenly on each bit, and (2) different bits are as independent as possible. Moreover, the feature and hashes-like representations are learned together.

On the server, we build the HDIdx [1] proposed in [79] as our search index. We use HDIdx as its index building process corporates an adaptive bits partition algorithm into the original multi-index hashing framework [80], which can separate the highly correlated bits into different code segments and greatly improve the search speed. In the testing process, to accelerate the hashing methods on

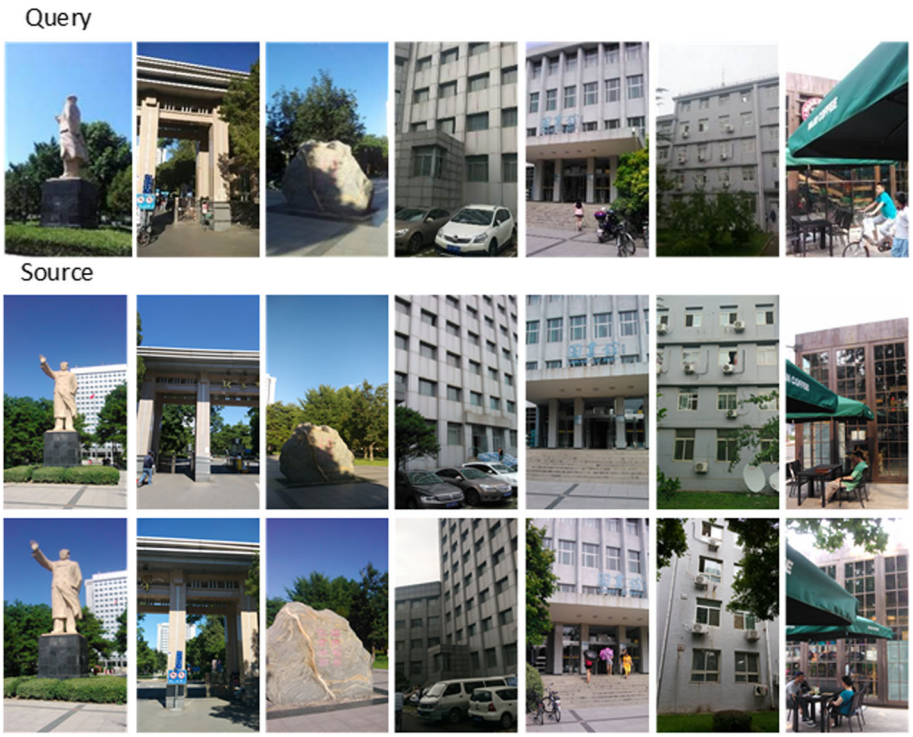

Fig. 2 The examples of the query and source images in the mobile location recognition dataset [77, 78]

the mobile device, we utilize the Quantized CNN [2] to optimize the DLBH and SSDH on the Huawei Mate 7 smartphone with 3G RAM and 1.8GHz Kirin 925 CPU. The framework of the system can be found in Fig. 1.

In the evaluation, we use MAP as the evaluation criterion, which is computed as the Eq. 1

$$MAP = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk}), \quad (1)$$

where Q is the query set, m_j is the number of positive images in each locations, $Precision(R_{jk})$ is the average precision at the position of returned *kth* positive images. To evaluate the effectiveness of deep learning hashing, we compared our methods with the visual hash bits (VHB) [81] and space-saliency fingerprint selection based hash codes (SSFS) [77], which are both state-of-the-art methods for mobile location recognition. The results can be found in Table 3. From the results, we can find that the accuracies of deep learning hashing method greatly outperform the traditional hashing methods, which demonstrate the power of deep learning technology in binary code learning. Furthermore, as the SSDH learns the feature representations and binary code simultaneously and add more constrains for binary code learning, it achieves higher performance than DLBH.

3 Open issues and future directions

In this paper, we just give a preliminarily practice for deep learning hashing based mobile visual search, several major issues remain open to be solved in the future.

3.1 Improve the accuracy of deep learning hashing based mobile visual search

As introduced in Section 1.1, mobile visual search is seriously disturbed by the noise of captured images or videos, such as flashing, occupy, rotation, blur, affine transformation, and so on. How to design robust features for more accurate search is still a great challenge. However, the existed deep learning hashing methods for desktop images search mainly focus on how to mine the discriminative features for images having a similar labels, which neglect these invariance properties. Therefore, in the future, the

Table 3 The MAP of different hashing methods for mobile visual search based location recognition

Method	MAP			
	16-bits	32-bits	64-bits	128-bits
VHB [81]	-	-	19.36%	-
SSFS [77]	-	-	20.22%	-
DLBH+QCNN	59.80%	78.68%	87.15%	90.67%
SSDH+QCNN	78.26%	91.82%	92.43%	93.21%

deep learning hashing method designed for mobile visual search must handle these specific noise in the learning process to further improve the accuracy, such as add the transformation invariance in the loss function and so on. In addition, large scale mobile visual search dataset is also needed to learn effective features.

3.2 Explore the ability of unsupervised deep learning hashing

As shown in Table 2, the performance of the unsupervised deep learning hashing methods is significantly worse than the supervised ones. However, in most of the case, it is hard to label all the images/videos for large scale visual search, which cannot use supervised hashing. Therefore, how to design the unsupervised deep learning hashing models to further improve the accuracy of unsupervised hashing is another important research topic in the future.

3.3 Further accelerate the computation and reduce the model size

Although many existed works have tried to optimize the deep learning technology on the mobile device, it is far from satisfactory. Until now, we can only run deep neural network with limited layers as the strong constraints of computation and memory on mobile devices. Moreover, the computation time is still more than ten times longer than on servers. In the future, more powerful hardware, such as GPU card with more CUDA cores and large graphic memories, are needed to be developed on the mobile devices. Moreover, the specific speed and memory optimization methods for deep learning hashing on mobile devices are also deserved more attentions.

3.4 Design the deep learning hashing for particular mobile visual search applications

The existed deep learning hashing methods all focus on extracting the binary codes from the images. In particular, there are diverse sensors on the mobile devices to support multi-modality fusion based visual search [82, 83]. Specifically, for location recognition, we can leverage the information from GPS, digital compass, accelerometer, and gyroscope to learn the multi-modality based deep learning hashing. Moreover, for mobile video search, we can holistically exploit the complementary nature of audio and video signals in the deep learning hashing. Therefore, more effective deep learning hashing for particular mobile visual search applications with multi-modality fusion will be more attractive in the future.

4 Conclusions

In this paper, we comprehensively survey the exited deep learning hashing technologies to demonstrate the necessity and sufficiency of deep learning hashing based mobile visual search. To achieve it, we analyze three different kinds of deep learning hashing methods in detail, and compare their performance on the CIFAR-10 dataset. Moreover, to efficiently implement the networks, we also discuss the deep learning optimization on mobile devices. Most important, according to our knowledge, we give one of the first attempts to design a deep learning hashing-based mobile visual search system for location recognition to evaluate our conclusions. Finally, after sufficient investigation, we give the emerging topics of deep learning hashing based mobile visual search in the future.

Endnotes

[1] "HDIdx", https://github.com/hdidx/hdidx.

[2] "Quantized CNN", https://github.com/yingxiaosan/quantized-cnn.

Funding

This work is partially supported by the national key research and development plan (No. 2016YFC0801005), the National Natural Science Foundation of China (No. 61602049, 61332005), the Funds for Creative Research Groups of China (No. 61421061), the Beijing Training Project for the Leading Talents in S&T (No. ljrc 201502), and the CCF-Tencent Open Research Fund (No. AGR20160113).

Authors' contributions

WL designed the algorithm, carried out the experiments, and drafted the manuscript. HDM gave suggestions on the structure of manuscript and participated in modifying the manuscript. QH participated in the implementation of the mobile location recognition system and carried out the experiments. DZ built the dataset and gave suggestions on the experimental analysis. ZNC participated in the survey of deep learning hashing and gave suggestions on the experimental analysis. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, 100876 Beijing, China. [2] Institute of Automation, Chinese Academy of Science, 100190 Beijing, China.

Received: 4 September 2016 Accepted: 5 February 2017
Published online: 21 February 2017

References

1. H Liu, T Mei, J Luo, H Li, S Li, in *Proceedings of the 20th ACM Multimedia Conference, MM '12, Nara, Japan, October 29 - November 02, 2012*. Finding perfect rendezvous on the go: accurate mobile visual localization and its applications to routing (ACM, New York, 2012), pp. 9–18
2. B Girod, V Chandrasekhar, DM Chen, N-M Cheung, R Grzeszczuk, YA Reznik, G Takacs, SS Tsai, R Vedantham, Mobile visual search. IEEE Signal Proc. Mag. **28**(4), 61–76 (2011)
3. H Li, Y Wang, T Mei, J Wang, S Li, Interactive multimodal visual search on mobile device. IEEE Trans. Multimed. **15**(3), 594–607 (2013)
4. H Chi, C Chen, W Cheng, M Chen, Ubishop: Commercial item recommendation using visual part-based object representation. Multimedia Tools Appl. **75**(23), 16093–16115 (2016)
5. E Miluzzo, ND Lane, K Fodor, R Peterson, H Lu, M Musolesi, SB Eisenman, X Zheng, AT Campbell, in *Proceedings of the 6th International Conference on Embedded Networked Sensor Systems, SenSys 2008, Raleigh, NC, USA, November 5–7, 2008*. Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application (ACM, New York, 2008), pp. 337–350

6. T Tsai, W Cheng, C You, M Hu, AW Tsui, H Chi, Learning and recognition of on-premise signs from weakly labeled street view images. IEEE Trans. Image Process. **23**(3), 1047–1059 (2014)
7. W Liu, T Mei, Y Zhang, Instant mobile video search with layered audio-video indexing and progressive transmission. IEEE Trans. Multimed. **16**(8), 2242–2255 (2014)
8. Y-C Su, T-H Chiu, Y-Y Chen, C-Y Yeh, WH Hsu, in *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*. Enabling low bitrate mobile visual recognition: a performance versus bandwidth evaluation (ACM, New York, 2013), pp. 73–82
9. Google Goggles. http://www.google.com/mobile/goggles/. Accessed 22 Dec 2016
10. X Yang, K-TT Cheng, in *Proceedings of the 20th ACM Multimedia Conference, MM '12, Nara, Japan, October 29 - November 02, 2012*. Accelerating SURF detector on mobile devices (ACM, New York, 2012), pp. 569–578
11. J Xia, K Gao, D Zhang, Z Mao, in *Proceedings of the 20th ACM Multimedia Conference, MM '12, Nara, Japan, October 29 - November 02, 2012*. Geometric context-preserving progressive transmission in mobile visual search (ACM, New York, 2012), pp. 953–956
12. VR Chandrasekhar, SS Tsai, G Takacs, DM Chen, N-M Cheung, Y Reznik, R Vedantham, R Grzeszczuk, B Girod, in *Proceedings of the 2010 ACM Multimedia Workshop on Mobile Cloud Media Computing*. Low latency image retrieval with progressive transmission of chog descriptors (ACM, New York, 2010), pp. 41–46
13. V Chandrasekhar, G Takacs, DM Chen, SS Tsai, Y Reznik, R Grzeszczuk, B Girod, Compressed histogram of gradients: a low-bitrate descriptor. Int. J. Comput. Vis. **96**(3), 384–399 (2012)
14. R Ji, L-Y Duan, J Chen, H Yao, Y Rui, S-F Chang, W Gao, in *Proceedings of the 19th ACM Multimedia Conference, MM '12, Scottsdale, AZ, USA, November 28 - December 1, 2011*. Towards low bit rate mobile visual search with multiple-channel coding (ACM, New York, 2011), pp. 573–582
15. J He, J Feng, X Liu, T Cheng, T-H Lin, H Chung, S-F Chang, in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16–21, 2012*. Mobile product search with bag of hash bits and boundary reranking (IEEE Computer Society, Washington, 2012), pp. 3005–3012
16. G-L Wu, Y-H Kuo, T-H Chiu, WH Hsu, L Xie, Scalable mobile video retrieval with sparse projection learning and pseudo label mining. IEEE Trans. Multimed. **20**(3), 47–57 (2013). http://doi.ieeecomputersociety.org/10.1109/MMUL.2013.13
17. K-Y Tseng, Y-L Lin, Y-H Chen, WH Hsu, in *Proceedings of the 20th ACM Multimedia Conference, MM '12, Nara, Japan, October 29 - November 02, 2012*. Sketch-based image retrieval on mobile devices using compact hash bits (ACM, New York, 2012), pp. 913–916
18. J Wan, D Wang, SCH Hoi, P Wu, J Zhu, Y Zhang, J Li, in *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*. Deep learning for content-based image retrieval: a comprehensive study (ACM, New York, 2014), pp. 157–166
19. A Babenko, A Slesarev, A Chigorin, V Lempitsky, in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I*. Neural codes for image retrieval (Springer International Publishing, Cham, 2014), pp. 584–599
20. Y Gong, L Wang, R Guo, S Lazebnik, Multi-scale orderless pooling of deep convolutional activation features. Comput. Sci. **8695**, 392–407 (2014)
21. AS Razavian, H Azizpour, J Sullivan, S Carlsson, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23–28, 2014*. Cnn features off-the-shelf: an astounding baseline for recognition (IEEE Computer Society, Washington, 2014), pp. 512–519
22. R Xia, Y Pan, H Lai, C Liu, S Yan, in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27–31, 2014, Québec City, Québec, Canada*. Supervised hashing for image retrieval via image representation learning (AAAI, Palo Alto, 2014), pp. 2156–2162
23. H Yang, K Lin, C Chen, Supervised learning of semantics-preserving hashing via deep neural networks for large-scale image search. CoRR. **abs/1507.00101** (2015)
24. F Zhao, Y Huang, L Wang, T Tan, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015*. Deep semantic ranking based hashing for multi-label image retrieval (IEEE Computer Society, Washington, 2015), pp. 1556–1564
25. A Krizhevsky, GE Hinton, in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12–17, 2016, Phoenix, Arizona, USA*. Using very deep autoencoders for content-based image retrieval (AAAI, Palo Alto, 2011)
26. T Do, A Doan, N Cheung, in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V*. Learning to hash with binary deep neural network (Springer International Publishing, Cham, 2016), pp. 219–234
27. R Zhang, L Lin, R Zhang, W Zuo, Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification. IEEE Trans. Image Process. **24**(12), 4766–4779 (2015)
28. H Lai, Y Pan, Y Liu, S Yan, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015*. Simultaneous feature learning and hash coding with deep neural networks (IEEE Computer Society, Washington, 2015), pp. 3270–3278
29. W Li, S Wang, W Kang, in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9–15 July 2016*. Feature learning based deep supervised hashing with pairwise labels (AAAI, Palo Alto, 2016), pp. 1711–1717
30. EL Denton, W Zaremba, J Bruna, Y LeCun, R Fergus, in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9–15 July 2016*. Exploiting linear structure within convolutional networks for efficient evaluation (AAAI, Palo Alto, 2014), pp. 1269–1277
31. M Jaderberg, A Vedaldi, A Zisserman, in *British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1–5, 2014*. Speeding up convolutional neural networks with low rank expansions, (2014)
32. V Lebedev, Y Ganin, M Rakhuba, IV Oseledets, VS Lempitsky, Speeding-up convolutional neural networks using fine-tuned cp-decomposition. CoRR. **abs/1412.6553** (2014)
33. X Zhang, J Zou, K He, J Sun, Accelerating very deep convolutional networks for classification and detection. IEEE Trans. Pattern Anal. Mach. Intell. **38**(10), 1943–1955 (2016)
34. W Chen, JT Wilson, S Tyree, KQ Weinberger, Y Chen, in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015*. Compressing neural networks with the hashing trick, (2015), pp. 2285–2294
35. Y Gong, L Liu, M Yang, LD Bourdev, Compressing deep convolutional networks using vector quantization. CoRR. **abs/1412.6115** (2014)
36. S Han, H Mao, WJ Dally, Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. CoRR. **abs/1510.00149** (2015)
37. S Srinivas, RV Babu, in *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7–10, 2015*. Data-free parameter pruning for deep neural networks, (2015), pp. 31.1–31.12
38. Z Yang, M Moczulski, M Denil, N de Freitas, AJ Smola, L Song, Z Wang, in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7–13, 2015*. Deep fried convnets (IEEE Computer Society, Washington, 2015), pp. 1476–1483
39. J Wu, C Leng, Y Wang, Q Hu, J Cheng, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016*. Quantized convolutional neural networks for mobile devices (IEEE Computer Society, Washington, 2016), pp. 4820–4828
40. S Rallapalli, H Qiu, A Bency, S Karthikeyan, R Govindan, B Manjunath, R Urgaonkar, Are very deep neural networks feasible on mobile devices? IEEE Trans. Circ. Syst. Video Technol. (2016). http://hgpu.org/?p=15652. Accessed 22 Dec 2016
41. J Sang, T Mei, Y-Q Xu, C Zhao, C Xu, S Li, Interaction design for mobile visual search. IEEE Trans. Multimed. **15**(7), 1665–1676 (2013)
42. H Yue, X Sun, J Yang, F Wu, Cloud-based image coding for mobile devices-toward thousands to one compression. IEEE Trans. Multimed. **15**(4), 845–857 (2013)
43. W Tan, B Yan, K Li, Q Tian, Image retargeting for preserving robust local feature: application to mobile visual search. IEEE Trans. Multimed. **18**(1), 128–137 (2016)
44. DM Chen, B Girod, A hybrid mobile visual search system with compact global signatures. IEEE Trans. Multimed. **17**(7), 1–1 (2015)
45. S Bianco, D Mazzini, DP Pau, R Schettini, Local detectors and compact descriptors for visual search: a quantitative comparison. Digit. Signal Process. **44**(1), 1–13 (2015)
46. L Zhu, J Shen, L Xie, in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, MM '15, Brisbane, Australia, October 26–30, 2015*.

Topic hypergraph hashing for mobile image retrieval (ACM, New York, 2015), pp. 843–846
47. YH Kuo, WH Hsu, De-hashing: server-side context-aware feature reconstruction for mobile visual search. IEEE Trans. Circuits Syst. Video Techn. **27**(1), 139–148 (2017)
48. W Zhou, M Yang, H Li, X Wang, Y Lin, Q Tian, Towards codebook-free: scalable cascaded hashing for mobile image search. IEEE Trans. Multimed. **16**(3), 601–611 (2014)
49. W Liu, T Mei, Y Zhang, J Li, S Li, in *ACM Multimedia*. Listen, look, and gotcha: instant video search with mobile phones by layered audio-video indexing (ACM, New York, 2013), pp. 887–896
50. H Shuai, D Yang, W Cheng, M Chen, Mobiup: An upsampling-based system architecture for high-quality video streaming on mobile devices. IEEE Trans. Multimed. **13**(5), 1077–1091 (2011)
51. H Bay, A Ess, T Tuytelaars, LJV Gool, Speeded-up robust features (SURF). Comp. Vision Image Underst. **110**(3), 346–359 (2008)
52. J Sivic, A Zisserman, in *9th IEEE International Conference on Computer Vision (ICCV 2003), 14–17 October 2003, Nice, France*. Video google: a text retrieval approach to object matching in videos (IEEE Computer Society, Washington, 2003), pp. 1470–1477
53. J Wang, T Zhang, J Song, N Sebe, HT Shen, A survey on learning to hash. CoRR. **abs/1606.00185** (2016)
54. J Wang, W Liu, S Kumar, SF Chang, Learning to hash for indexing big data-a survey. Proc. IEEE. **104**(1), 34–57 (2015)
55. L Weng, I Jhuo, M Shi, M Sun, W Cheng, L Amsaleg, in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, June 23–26, 2015*. Supervised multi-scale locality sensitive hashing (ACM, New York, 2015), pp. 259–266
56. R Salakhutdinov, G Hinton, Semantic hashing. Int. J. Approx. Reason. **50**(7), 969–978 (2009)
57. A Torralba, R Fergus, Y Weiss, in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24–26 June 2008, Anchorage, Alaska, USA*. Small codes and large image databases for recognition (IEEE Computer Society, Washington, 2008), pp. 1–8
58. Y Pan, T Yao, H Li, C Ngo, T Mei, in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9–13, 2015*. Semi-supervised hashing with semantic confidence for large scale visual search (ACM, New York, 2015), pp. 53–62
59. A Krizhevsky, I Sutskever, GE Hinton, in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3–6, 2012, Lake Tahoe, Nevada, United States*. Imagenet classification with deep convolutional neural networks, (2012), pp. 1106–1114
60. C Gan, N Wang, Y Yang, D Yeung, AG Hauptmann, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015*. Devnet: A deep event network for multimedia event detection and evidence recounting (IEEE Computer Society, Washington, 2015), pp. 2568–2577
61. J Deng, W Dong, R Socher, L Li, K Li, F Li, in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. Imagenet: A large-scale hierarchical image database (IEEE Computer Society, Washington, 2009), pp. 248–255
62. J Donahue, Y Jia, O Vinyals, J Hoffman, N Zhang, E Tzeng, T Darrell, in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21–26 June 2014*. Decaf: a deep convolutional activation feature for generic visual recognition, (2014), pp. 647–655
63. X Zhang, H Zhang, Y Zhang, Y Yang, M Wang, H Luan, J Li, T Chua, Deep fusion of multiple semantic cues for complex event recognition. IEEE Trans. Image Process. **25**(3), 1033–1046 (2016)
64. I Shen, W Cheng, Gestalt rule feature points. IEEE Trans. Multimed. **17**(4), 526–537 (2015)
65. D Riahi, G Bilodeau, in *Canadian Conference on Computer and Robot Vision, CRV 2014, Montreal, QC, Canada, May 6–9, 2014*. Multiple feature fusion in the dempster-shafer framework for multi-object tracking (IEEE Computer Society, Washington, 2014), pp. 313–320
66. H Yao, S Zhang, Y Zhang, J Li, Q Tian, Coarse-to-fine description for fine-grained visual categorization. IEEE Trans. Image Process. **25**(10), 4858–4872 (2016)
67. W Liu, Hashing by Deep Learning (2016). http://www.ee.columbia.edu//~wliu/WeiLiu_DLHash.pdf. Accessed 22 Dec 2016
68. K Lin, H Yang, J Hsiao, C Chen, in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops, Boston, MA, USA, June 7–12, 2015*. Deep learning of binary hash codes for fast image retrieval (IEEE Computer Society, Washington, 2015), pp. 27–35
69. VE Liong, J Lu, G Wang, P Moulin, J Zhou, in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015*. Deep hashing for compact binary codes learning (IEEE Computer Society, Washington, 2015), pp. 2475–2483
70. K Lin, J Lu, C-S Chen, J Zhou, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016*. Learning compact binary descriptors with unsupervised deep neural networks (IEEE Computer Society, Washington, 2016), pp. 1183–1192
71. S Chopra, R Hadsell, Y LeCun, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20–26 June 2005, San Diego, CA, USA*. Learning a similarity metric discriminatively, with application to face verification (IEEE Computer Society, Washington, 2005), pp. 539–546
72. H Zhu, M Long, J Wang, Y Cao, in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12–17, 2016, Phoenix, Arizona, USA*. Deep hashing network for efficient similarity retrieval (AAAI, Palo Alto, 2016), pp. 2415–2421
73. T Yao, F Long, T Mei, Y Rui, in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9–15 July 2016*. Deep semantic-preserving and ranking-based hashing for image retrieval (AAAI, Palo Alto, 2016), pp. 3931–3937
74. L Gao, J Song, F Zou, D Zhang, J Shao, in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, MM '15, Brisbane, Australia, October 26 - 30, 2015*. Scalable multimedia retrieval by deep learning hashing with relative similarity learning (ACM, New York, 2015), pp. 903–906
75. J Redmon, S Divvala, R Girshick, A Farhadi, in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016*. You only look once: unified, real-time object detection (IEEE Computer Society, Washington, 2016), pp. 779–788
76. H Liu, H Li, T Mei, J Luo, Accurate sensing of scene geo-context via mobile visual localization. Multimed. Syst. **21**(3), 255–265 (2015)
77. H Wang, D Zhao, H Ma, H Xu, in *Advances in Multimedia Information Processing - PCM 2016 - 17th Pacific-Rim Conference on Multimedia, Xi'an, China, September 15–16, 2016, Proceedings, Part II*. SSFS: A space-saliency fingerprint selection framework for crowdsourcing based mobile location recognition (Springer International Publishing, Cham, 2016), pp. 650–659
78. H Wang, D Zhao, H Ma, H Xu, X Hou, in *21st IEEE International Conference on Parallel and Distributed Systems, ICPADS 2015, Melbourne, Australia, December 14–17, 2015*. Crowdsourcing based mobile location recognition with richer fingerprints from smartphone sensors (IEEE Computer Society, Washington, 2015), pp. 156–163
79. J Wan, S Tang, Y Zhang, J Li, P Wu, SCH Hoi, Hdidx: High-dimensional indexing for efficient approximate nearest neighbor search. CoRR. **abs/1510.01991** (2015)
80. M Norouzi, A Punjani, DJ Fleet, Fast Exact Search in Hamming Space With Multi-Index Hashing. IEEE Trans. Pattern Anal. Mach. Intell. **36**(6), 1107–1119 (2014)
81. SF Chang, H Chung, TH Lin, T Cheng, X Liu, J Feng, J He, in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16–21, 2012*. Mobile product search with bag of hash bits and boundary reranking (IEEE Computer Society, Washington, 2012), pp. 3005–3012
82. L Zhang, Y Zhang, R Hong, Q Tian, Full-space local topology extraction for cross-modal retrieval. IEEE Trans. Image Process. **24**(7), 2212–2224 (2015)
83. L Chu, Y Zhang, G Li, S Wang, W Zhang, Q Huang, Effective multimodality fusion framework for cross-media topic detection. IEEE Trans. Circ. Syst. Video Technol. **26**(3), 556–569 (2016)