Large scale visual search

Wu, S.

Author: Wu, S.
Title: Large scale visual search
Issue Date: 2016-12-22

Chapter 1

Introduction

Content-based image retrieval (CBIR) is one of the important and challenging problems in computer vision research. Frequently used search engines like Google or Yahoo represent a category of text-based information retrieval. However, the search accuracy of text-based information retrieval cannot satisfy the requirements of users when the text annotations are incomplete or incorrect. Moreover, due to the non-scalability of text-based information retrieval to large scale datasets, especially for the ever increasing multimedia data on the web, a high degree of manual effort is required to define the correct text annotations. Therefore, research on content-based information retrieval was proposed to address this issue.

CBIR aims to obtain more images related to the query by analyzing and exploring the content in images. The classic features (e.g., color, shape, texture, and etc.) are low-level features, which do not easily translate to the high level human concept vocabulary. In order to bridge the semantic gap, Bag-of-Words (BoW) [1] with local features was proposed to represent images and demonstrated high performance in image retrieval. Inspired by the BoW model, Fisher Vector (FV) [2], Vector of Locally Aggregated Descriptors (VLAD) [3], as well as their variants [4, 5, 6, 7] were proposed to seek a more informative image representation. Additionally, hash techniques were designed to map real valued image representations to binary codes such that large scale image search can be carried out in an efficient way.

Benefitting from the advantages of convolutional neural networks (CNNs) [8] trained on sufficiently large and diverse datasets such as ImageNet [9], image representations based on the activations within CNNs have shown significant high performance regarding the state-of-the-art of various computer vision applications. Recent work focuses on investigating effective ways to aggregate the activations within CNNs into compact and distinctive image representations for large scale image search.

Our main research objectives in this thesis are as follows:

- design a robust local descriptor to improve the discrimination of BoW based image representations.
- provide an effective way for large scale image search by exploring compact deep binary codes from deep layers within CNNs.
- propose a more powerful CNN architecture to improve the robustness of the image representation generated from the deep layers in CNNs.

1.1 Salient Point Methods

Generally, salient point methods consist of two parts: a local detector and a local descriptor. A robust local detector should have high repeatability. For compared images with the same object or scene, a high percentage of corresponding salient points should be detected on the scene part visible in both images. At the same time, they should be unaffected by various deformations, such as image blur, affine deformations, compression, and noise. Distinctive and robust local descriptors should be invariant and less sensitive to different deformations, such that they can be distinguished and effectively matched. Most of the existing salient point methods are invariant to transformations of scale, rotation, viewpoint and noise. All these properties are achieved by three distinct steps in salient point methods: scale space construction, orientation assignment and local descriptor generation.

Scale space construction: the construction of scale space aims to solve the challenges in computer vision where the vision information is captured at different scales. The commonly used scale space is the Gaussian smoothed image pyramid which can represent an image at multi-scales and multi-resolutions. The multi-scale representation is achieved by convolving the original image with different Gaussian smooth kernels, and the multi-resolution representation is achieved by down sampling or spatial pooling of the original image.

The experimental evaluation in the work by Lindeberg [10] demonstrated that the extreme points detected in a scale space constructed by scale-normalized Laplace-of-Gaussian (LoG) show high stability and repeatability. The representation of LoG scale space is constructed by smoothing the high resolution image with derivatives of Gaussian kernels of increasing size. Recent work focuses on the efficiency of scale space construction. An efficient framework of Difference-of-Gaussian (DoG) [11], which is an approximation of the Laplace-of-Gaussian (LoG) was proposed. A Difference-of-Gaussians pyramid is computed from the differences between the adjacent levels in the Gaussian pyramid. Then, the points at which the DoG values assume extrema of the differences with respect to both the spatial coordinates in the image domain and the scale level in the pyramid can be considered as candidate salient points. Compared to LoG, the DoG operation could significantly decrease the computational complexity. As the second order Gaussian derivatives (Hessian matrix) can also be used to approximate the LoG and the Hessian matrix can be further approximated by a box-filter, the box-filter based scale space [12] makes use of the box-filter and integral images to calculate the determinant of the Hessian matrix at a very low computational cost. The Gaussian smoothing operation in LoG, DoG as well as box-filter based scale space construction is linear and each pixel in the Gaussian image pyramid layer is convolved with the same Gaussian kernel. Hence, it can cause blurring as well as loss of the boundary details of the object in images. The introduction of nonlinear scale space can reduce noise but retain the object boundary structure such that we can obtain accurate positions of extreme points. The framework of nonlinear scale space first convolves the image with a Gaussian kernel of standard deviation. Then it builds the nonlinear scale space in an iterative way via the additive operator splitting (AOS) scheme [13]. Since the extreme points are detected in discrete scale space, the final scale and location information of these extreme points can be refined by a quadratic fitting operation.

Orientation assignment: assigning each detected salient point an orientation. The local descriptor of the salient point in images can be represented relative to the calculated orientation and therefore be invariant to image rotation. The orientation assigned by a good measurement can make the generated local descriptor more robust to even large image rotation changes. The general orientation measurements can be categorized as: histogram of gradients [14], Haar-wavelet response [12], intensity centroid [15], average gradient of sampling pairs in a pre-designed pattern [16, 17, 18]. The histogram of gradient based orientation is calculated according to the gradient of each pixel within a patch around the salient point $\theta(x, y) = \arctan(\bigtriangledown I(x, y))$. Then, all $\theta(x, y)$ values are counted to generate a histogram and the maximum bin in the histogram is considered to be the orientation of the salient point. The Haar-wavelet response based orientation is computed according to the responses of Haar-wavelets in both horizontal and vertical directions in a circle region around the salient point. The dominant orientation of the local region is estimated by calculating the sum of all responses within a sliding orientation window covering an angle of 60 degrees, and the largest response is considered as the orientation of the salient point. The intensity centroid approach assumes that the intensity of a salient point is an offset from its center, and it can be used to compute the moments of a patch and further to find its centroid. The orientation is defined as the direction of the vector from the salient point position to the centroid position in the local patch. Some other methods assign orientations to the salient points by averaging the sum of local gradients of the defined pairs in the pre-defined structure.

Local descriptor generation: the local descriptor generation is performed on image data that has been transformed relative to the assigned scale, location, and orientation such that they can be distinguished and matched under these transformations. The existing local descriptor can be categorized as a hand-crafted based scheme and a machine learning based scheme.

Hand-crafted local descriptors: the hand-crafted schemes mainly explore the intensity patterns around the detected salient points. The most representative descriptors are the distribution-based local descriptors, which include successful representations such as: histogram of gradients, histogram of gradient orientations, and Haar-wavelet responses distributions, which represent distinctive visual information according to the distributions of pixel intensities in the local patch. Binary local descriptors were proposed with an emphasis on minimizing computational and storage costs. The binary string representations make use of simple pair-wise pixel intensity comparisons. Different binary string local descriptors advocate different pre-defined structures to select the pixel pairs and the generated binary codes can be very efficiently matched with low computational cost using the Hamming metric (bitwise XOR followed by a bit count).

Machine learning based local descriptors: machine learning has been applied to improve both the efficiency and accuracy of local binary descriptor generation. Hashing is one of the most effective techniques which aims to construct a set of hash functions to map the original input space to compact and similarity-preserving local binary codes. Therefore, similar input spaces could be projected to similar binary codes in the Hamming space. Existing hashing approaches can be divided into two categories: data-independent and data-dependent methods. Data-independent methods randomly generate a projection matrix to map image features into binary representations without training data [19]. The representative methods are locality-sensitive hashing (LSH) [20] as well as its variants [21, 22]. Data-dependent hashing which is also referred to as a hashing learning approach focuses on learning hash functions from a specific training dataset.

Hash learning approaches involve two main steps. First, the training data is represented as hand-crafted real valued local features. Second, optimize an objective function to learn the hash function and use the learned hash function to convert the real valued input space representation into a binary representation. Generally, the process of hash functions learning is either done in an unsupervised, semi-supervised or supervised manner. Unsupervised hash function learning makes use of the unlabeled training data, and learns the compact binary descriptors whose Hamming distance is correlated with data similarity in the original input space

[23, 24, 25, 26, 27]. Supervised and semi-supervised hashing approaches take advantage of semantic label information of training data to preserve the ground truth similarity during the construction of binary hash codes. Supervised hashing fully exploits the labeled training data to seek a linear transformation. A loss function is usually defined that penalizes the reconstruction errors between the distances of original data and the distances of corresponding binary data to learn the hash functions [28, 29, 30, 31, 32]. For the case of semi-supervised hashing learning, both unlabeled data and labeled data are utilized to learn the hash functions. For example, the semi-supervised hashing frameworks proposed by Wang et al. [33, 34] minimize the empirical error on the labeled data while maximizing the variance over labeled and unlabeled data for binary representations. Recently, the deep supervised hashing methods were proposed to learn binary hashing codes. Deep supervised hashing trains a deep hierarchical and nonlinear transformation model and projects the original local descriptors into local binary codes [35, 36, 37, 38].

1.2 Visual Word based Image Search

Content-based image search is still a challenging problem in computer vision. This is mainly due to the existing variations in image appearance, such as the changes of scale, orientation, viewpoint and illumination. In addition, with the increasing amounts of image data on the web, a robust image representation with the approximate nearest neighbor (ANN) search has been widely used for large scale image retrieval. This method is mainly benefiting from the robustness of local descriptors to various geometric transformations and the applicability of different similarity measures.

The Bag-of-Words (BoW) model which is inspired from simple document retrieval systems and based on the analogy of visual words, has been widely applied in content-based image retrieval. In the BoW model, salient regions are first detected from each image in the training dataset and a high dimensional descriptor is calculated for each region. These descriptors are then clustered to form a vocabulary of visual words. Therefore, an image is finally represented as a histogram over a set of learned visual words after quantizing each of the local descriptors to the nearest visual word. Early systems [1] used a flat K-means clustering to generate the visual vocabulary, but it was difficult to scale to large vocabularies generation and large scale datasets. The later works [39, 40] show that flat K-means can be scaled to similarly large vocabulary sizes by the use of approximate nearest neighbor methods.

The Fisher Vector (FV) [2] image representation seeks to capture the probability distribution of features. The generative model Gaussian mixture model (GMM) is utilized in FV to estimate a parametric probability distribution over the feature space from a large representative set of local descriptors. The local descriptors extracted from the image dataset are assumed to be sampled independently from this probability distribution. Each local descriptor is represented by the gradient of the probability distribution at that feature with respect to its parameters. Gradients corresponding to all the features with respect to a particular parameter are summed. The final FV representation is the concatenation of the accumulated gradients. They achieve a fixed length vector from a varying set of features that can be used in various discriminative learning activities. Compared with the K-means cluster algorithm, GMM delivers not only the mean information of visual words, but also the shape of their distribution.

Jegou et al. proposed Vector of Locally Aggregated Descriptors (VLAD) [3] which can be viewed as a simplified non-probabilistic version of Fisher Vector. Similar to the BoW model, a vocabulary with C visual words is first learned via K-means cluster. Each local descriptor is associated to its nearest visual word in the vocabulary. The idea of the VLAD representation is to accumulate the residuals belonging to each of the visual words. This characterizes the distribution of the vectors with respect to the center. A number of variants of VLAD have also been designed to enhance the image representation by considering vocabulary adaptation and intra-normalization [4], residual normalization and local coordinate systems [5], geometry information [6] and multiple vocabularies [7].

The relationship among the models of BoW, FV and VLAD can be described as: BOW encodes the 0-order statistics of the distribution of local descriptors, the

Fisher vector extends the BOW by encoding high-order statistics (first-order and, optionally, second-order), and VLAD is a non-probabilistic equivalent of Fisher Vector. During the past decade, the visual words based image representation has been successfully applied in various computer vision applications.

1.3 Convolutional Neural Networks

Generally, the convolutional neural network (CNN) is a hierarchical architecture [8] which consists of several stacked convolutional layers (optionally followed by a normalization layer and a spatial pooling layer), fully connected layers and a loss layer on top. The convolutional layers generate feature maps by linear convolutional filters followed by nonlinear activation functions (Rectifier, Sigmoid, TanH, etc.). The fully connected layer has full connections to all activations in the feature maps and the resulted one dimensional vector can be fed into the loss layer for loss function optimization.

There are two main stages for training the convolutional neural network: a forward stage and a backward stage. First, the forward stage is to represent the input image with the current parameters (weights and bias) in each layer. Then the output from the last layer is used to compute the loss function with the ground truth labels. Second, based on the loss cost, the backward stage computes the gradients of each parameter with chain rules. All the parameters are updated based on the gradients, and are prepared for the next forward computation. After sufficient iterations of the forward and backward stages, the network could be optimized. The convolutional neural network has been applied in diverse computer vision applications and demonstrated their significant advantages and high performance.

We will first present an overview of the different types of layers and then briefly review the CNN based computer vision applications.

Convolutional layers: the convolutional layers in the CNN architecture utilizes k filters (or kernels) to convolve the input image to generate k feature maps. There are three main advantages of the convolution operation [41]: first, the parameter sharing mechanism is used in convolutional layers such that the number of parameters could be significantly reduced. Second, local connectivity learns correlations among neighboring pixels. Third, it is invariant to the location of the object. Due to these benefits introduced by the convolution operation, some well-known research papers also use it as a replacement for the fully connected layers to accelerate the learning process [42, 43].

Pooling layers: a pooling layer is an optional layer following a convolutional layer which sub-samples its input. Average pooling and max pooling are the most commonly used pooling operations. The reason to use a pooling operation in the convolutional neural network is that: first, it can reduce the dimensions of the output and the number of parameters of the network, while keeping the most salient information. Second, a pooling operation also provides basic invariance to translating (shifting) and rotation. For max pooling and average pooling, Boureau et al. [44] provided a detailed theoretical analysis of their performances. Scherer et al. [45] further conducted a comparison between the two pooling operations and found that max-pooling can lead to faster convergence, selection of superior invariant features and improve generalization.

Fully-connected layers: after several convolutional and max pooling layers, the high-level reasoning in the convolutional neural network is done via the fully connected layers. A fully connected layer takes all neurons in the previous layer (be it fully connected, pooling, or convolutional) and connects it to every single neuron it has. Fully connected layers are not spatially located, as the input feature maps are converted to a one dimensional feature vector. The one dimensional feature vector could either feed forward the vector into a loss layer or take it as a feature representation for follow-up processing [46]. The drawback of the fully connected layer is that it contains many parameters, which results in large computational and storage costs. Therefore, GoogleNet [42] designed a deep and wide network while keeping the computational budget constant, by switching from fully connected to sparsely connected architectures. The Network in Network (NIN) [47] architecture replaces the fully connected layer by a global average pooling layer.

Recently, deep learning, especially for the CNNs, produced state-of-the-art performance on various computer vision applications, such as image classification, image search, object detection, semantic image segmentation, human pose estimation, etc.

Image classification: Krizhevsky et al. [48] set a milestone for large-scale image classification when training a large CNN on the ImageNet dataset [9], thus proving that CNN could perform well on natural image classification. OverFeat [49] proposed a multi-scale and sliding window framework, which could find the optimal scale of the image and fulfill different tasks simultaneously, e.g., image classification, object detection and localization. Because the existing CNNs require fixed-size image data as input, the SPP-Net [50] model eliminated this restriction via a spatial pyramid pooling strategy in the CNNs and improved the classification accuracy of a variety of CNN architectures despite their different designs. The later proposed VGGNet [51] and GoogleNet [42] significantly improved the performance of image classification by increasing the width and depth of the network architectures.

Object detection: a general scheme for high performance object detection systems is to generate a large number of candidate object region proposals and classify them using their high performance CNN features. The most representative approach is the regions with CNN features (RCNNs) [46]. It utilizes selective search [52] to generate object region proposals, and extracts the CNN features for each candidate region. The features are then fed into a SVM classifier to decide whether the related candidate windows contain the object or not. RCNNs improved the benchmark of object detection by a large margin, and became the base model for many other promising algorithms [53, 54, 55]. Also, the original RCNNs were computationally expensive, the recent works [50, 56] employed the strategy of sharing convolutions across the region proposals to reduce the computation cost. The latest frameworks of Fast RCNNs [56] and Faster RCNNs [57] achieve near real-time rates using very deep networks.

Image retrieval: The success of AlexNet [48] suggests that CNNs can be used as high level and universal feature extractors. The features emerging in the fully connected layers of the CNN can serve as a high level image representation for image classification. Motivated by this, many recent studies make use of the activations of the top fully connected layers in CNNs for image retrieval [58, 59, 60, 61, 62, 63]. Recent researches also suggested to explore the features from the deep convolutional layers in CNNs. These features have very useful properties: first, they can be efficiently extracted from an image of any size and aspect ratio. Second, features from the convolutional layers have a natural interpretation as descriptors of local image regions corresponding to receptive fields of the particular features. Finally, simple pooling operations can aggregate feature maps from deep convolutional layers into low dimensional and highly distinctive features [58, 60, 63, 64, 65]. The CNNs based image representations have demonstrated their competitive and even better results compared with the traditional visual words methods, such as BoW, VLAD and Fisher Vector.

Semantic image segmentation: semantic image segmentation can be referred to as a problem of pixel-level classification or labeling. Recently, the CNNs and probabilistic graphical models were utilized to address this task and yielded promising progress [66, 67, 68, 69, 70]. The CNN based semantic image segmentation methods usually convert an existing CNN architecture constructed for classification to a fully convolutional network (FCN). This is mainly because the FCN architecture accepts a whole image as an input and performs fast and accurate inference. Long et al. [68] replaced the last fully connected layers of a CNN by convolutional layers, and obtained a coarse label map from the network by classifying every local region in the image, then performed a simple deconvolution, which is implemented as a bilinear interpolation, for pixel-level labeling. DeepLab [69] proposed a similar FCN based model which obtained denser score maps within the FCN framework to predict pixel-wise labels and refined the label map using the fully connected conditional random field (CRF). Instead of using CRF as a post-processing step, Lin et al. [70] jointly trained the FCN and CRFs by efficient piece wise training.

Human pose estimation: estimating the human pose by locating human body joints or facial landmarks is a challenging task, because of the changes in pose, illumination, occlusion and etc. As CNNs have shown outstanding performance on visual classification and object localization, human pose estimation can be formulated as a CNN-based regression problem towards human body joints. The representative projects [71, 72] proposed to use a cascade of CNN-based regressors to reason the positions of body joints or facial landmarks. The cascade of CNNs can be viewed as a kind of holistic process which takes the full image as the input and output the ultimate position of body joints or facial landmarks in the image without using any explicit graphical model or part detectors. The part-based processing methods propose to detect the human body parts individually, followed with a graphical model to incorporate the spatial information. Rather than training the network using the whole image as input, Chen et al. [73] utilized the local part patches to train a CNN, in order to learn conditional probabilities of the part presence and spatial relationships. By incorporating with graphical models, the algorithm gained promising performance. Tompson et al. [74] designed multi-resolution ConvNet architectures to perform heat-map likelihood regression for each body part, followed with an implicit graphical model to further promote joint consistency. The model was further extended and improved [75], which argues that the pooling layers in the CNN would limit spatial localization accuracy and try to recover the precision loss caused by the pooling process. Additionally, Fan et al. [76] proposed a dual-source convolutional neutral network (DS-CNN) to integrate the holistic and partial view in a two branche CNN architecture. It takes part-patches and body-patches as inputs to combine both local and contextual information for more accurate pose estimation.

1.4 Thesis Overview

This thesis is based on first-authored conference papers that have been published or journal papers are currently under review. The research has been carried out during the four-year period of the PhD research. The focus of this thesis has been on developing and analyzing techniques to improve the state-of-the-art of large scale image search.

Chapter 2: A Comprehensive Evaluation of Salient Point Methods

A survey is presented that evaluates the performance of a wide set of salient point detectors and descriptors. First, the survey compares diverse salient point algorithms with regard to the repeatability of salient point detectors, recall and precision of salient point descriptors. Then, it integrates the salient point methods with the framework of fully affine space and evaluates the performance under major viewpoint transformations. The presented comparative experimental results can benefit researchers in choosing an appropriate detector and descriptor for different computer vision applications. An early version of this work was presented at:

- 21st ACM international conference on Multimedia (MM 2013), in Barcelona, Spain.

Chapter 3: RIFF: Retina-inspired Invariant Fast Feature Descriptor

We first propose the Retina-inspired Invariant Fast Feature, RIFF, which was designed for invariance to scale, rotation, and affine image deformations. The RIFF descriptor is based on pair-wise comparisons over a sampling pattern loosely based on the human retina and introduces a method for improving accuracy by maximizing the discriminatory power of the point set. A performance evaluation with regard to Bag-of-Words based image retrieval on several well-known benchmark datasets demonstrates that the RIFF local descriptor has competitive performance to the state-of-the-art local descriptors. Additionally, a popular approach from the literature is to use visual words (or Bag-of-Words) constructed from real valued descriptors (SIFT and SURF). To accommodate large scale data sets, we used the approximate nearest neighbor (ANN) based cluster approach to both real valued local descriptors and binary string local descriptors (BRIEF, ORB, BRISK and FREAK) and the results on the test datasets reveal that some of the recent binary string approaches outperformed notable descriptors such as SIFT and SURF. This approach has been presented at the following conferences:

- 22nd ACM international conference on Multimedia (MM 2014), in Orlando, FL, USA.
- 4th ACM International Conference on Multimedia Retrieval (ICMR 2014), in Glasgow, Scotland.

Chapter 4: Deep Binary Codes for Large Scale Image Retrieval

We present a novel and effective method to create compact binary codes (deep binary codes) based on deep convolutional features for image retrieval. Deep binary codes are generated by comparing the response from each feature map and the average response across all the feature maps on the deep convolutional layer. Additionally, a spatial cross-summing strategy is proposed to directly generate bit-scalable binary codes. As the deep binary codes on different deep layers can be obtained by passing the image through the CNN and each of them makes a different contribution to the search accuracy, we then present a dynamic, on-the-fly late fusion approach where the top N high quality search scores from deep binary codes are automatically determined online and fused to further enhance the retrieval precision. Two strengths of the proposed methods are that the generation of deep binary codes is based on a generic model which does not require additional training for new domain areas, and the dynamic late fusion scheme is query adaptive. Extensive experimental results on well known benchmarks show that the deep binary codes are competitive to state-of-the-art approaches in terms of the performance on large scale image retrieval. Moreover, the search accuracy is shown to be enhanced substantially by the dynamic late fusion scheme. The paper has been submitted to:

- Journal of Neurocomputing

Chapter 5: Comparison of Information Loss Architectures in CNNs

We propose a novel deep convolutional neural network called "Weighted Integration Architecture Network" (WIAN) which can effectively recover the information loss due to the pooling operation in the CNNs. The proposed WIAN reuses the information from the previous layers in the network and assigns a weight matrix to each layer and then integrates them via an element-wise sum operation to further enhance the performance of image classification. Several types of weight scheme such as adaptive weight learning framework as well as responses or entropy based weight learning schemes have been evaluated in this chapter. Extensive experiments on four standard benchmark datasets demonstrate the effectiveness as well as an improved performance of the proposed WIAN. The basis for this chapter is formed by the publication in the conference proceeding:

- 17th Pacific-Rim Conference on Multimedia (PCM, 2016) in Xi'an China.

These are the publications which were related to the contents of this thesis:

- **Wu S.**, and Lew M.S., "Evaluation of salient point methods." Proceedings of the 21st ACM International Conference on Multimedia. ACM, 2013.
- **Wu S.**, and Lew M.S., "Salient features for visual word based image copy detection." Proceedings of International Conference on Multimedia Retrieval. ACM, 2014.
- **Wu S.**, and Lew M.S., "RIFF: Retina-inspired Invariant Fast Feature Descriptor." Proceedings of the 22nd ACM International Conference on Multimedia. ACM, 2014.
- **Wu S.**, and Lew M.S., "Comparison of Information Loss Architectures in CNNs." Proceedings Pacific RIM Conference on Multimedia, 2016.
- **Wu S.**, and Lew M.S., "Image Correspondences Matching Using Multiple Features Fusion." Proceedings of European Conference on Computer Vision Workshop, 2016.
- **Wu S.**, Oerlemans A, Bakker E.M., and Lew M.S., "Deep Binary Codes for Large Scale Image Retrieval." Journal of Neurocomputing (submitted).
- **Wu S.**, Oerlemans A, Bakker E.M., and Lew M.S., "A Comprehensive Evaluation of Salient Point Methods." Journal of Computer Vision and Image Understanding (submitted).
- Guo Y., Liu Y., Oerlemans A., Lao S., **Wu S.**, and Lew M.S. "Deep learning for visual understanding: A review." Journal of Neurocomputing, vol 187, 2016.
- Zhang Q., Zaaijer S., **Wu S.**, and Lew M.S. "3D Image Browsing: The Planets". Proceedings of International Conference on Multimedia Retrieval. ACM, 2014.

- Guo Y., Bai L, Lao S., **Wu S.**, and Lew M.S. " A Comparison between Artificial Neural Network and Cascade-Correlation Neural Network in Concept Classification." Proceedings Pacific RIM Conference on Multimedia, 2014.
- Liu Y., Guo Y., **Wu S.**, and Lew M.S. (2015), "DeepIndex for Accurate and Efficient Image Retrieval." Proceedings of International Conference on Multimedia Retrieval. ACM 2015.