Multimedia Tools and Applications (2020) 79:14465–14489
https://doi.org/10.1007/s11042-018-6915-3

Unsupervised semantic-based convolutional features aggregation for image retrieval

Xinsheng Wang[1] · Shanmin Pang[1] · Jihua Zhu[1] · Jiaxing Wang[1] · Lin Wang[2]

Received: 30 July 2018 / Revised: 26 September 2018 / Accepted: 19 November 2018 /
Published online: 28 November 2018

Abstract

Deep features extracted from the convolutional layers of pre-trained CNNs have been widely used in the image retrieval task. These features, however, are in a large number and probably cannot be directly used for similarity evaluation due to lack of efficiency. Thus, it is of great importance to study how to aggregate deep features into a global yet distinctive image vector. This paper first introduces a simple but effective method to select informative features based on semantic content of feature maps. Then, we propose an effective channel weighting method (CW) for selected features by analyzing relations between the discriminative activation and distribution parameters of feature maps, including standard variance, non-zero responses and sum value. Furthermore, we provide a solution to pick semantic detectors that are independent on gallery images. Based on the aforementioned three strategies, we derive a global image vector generation method, and demonstrate its state-of-the-art performance on benchmark datasets.

Keywords Image retrieval · Deep convolutional features · Selection and aggregation · Unsupervised object localization · VGG16

✉ Jihua Zhu
zhujh@xjtu.edu.cn

Xinsheng Wang
wangxinsheng@stu.xjtu.edu.cn

Shanmin Pang
pangsm@xjtu.edu.cn

Jiaxing Wang
csuwjx@stu.xjtu.edu.cn

Lin Wang
wanglin@nwu.edu.cn

[1] School of Software Engineering, Xi'an Jiaotong University, Xi'an, People's Republic of China

[2] School of Information Science and Technology, Northwest University, Xi'an, People's Republic of China

1 Introduction

Artificial intelligence plays an import role in daily life and economic activities nowadays [22]. It refers to kinds of fields, such as speech recognition [37], image processing [38], video processing [9], anomaly detection [23] etc. Image retrieval, including the text-based image retrieval (TBIR) [42], content-based image retrieval (CBIR) [28] and cross-modal retrieval [13, 49, 50], is an important application of artificial intelligence. CBIR also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR), is aimed to efficiently search similar images from a large-scale image dataset for a given query image. One of key problems for the task of CBIR is to represent images effectively and efficiently. Image representations [14, 16] based on hand-crafted local descriptors (e.g. SIFT [21]) has been extensively investigated for over a decade in CBIR. However, with deep networks popularized by Krizhevsky et al. [20] in 2012, recent research focus has begun to shift to deep learning based methods, especially the convolutional neural networks (CNNs).

Image representations based on convolutional networks are increasingly permeating in various application domains including image classification [6, 24, 32, 44], object detection [4, 19, 35], semantic segmentation [25, 39, 54], image processing [26, 27, 52] etc. After training a CNN on a huge annotated dataset, e.g. ImageNet [36], the activations of the convolutional or fully connected layers capture semantic information of images, and therefore can be used for representing images. In the field of CNNs-based image retrieval, early works [3, 33] directly adopted global features obtained from fully connected layers to represent images. In order to improve the invariance of CNNs representations, Gong et al. [10] proposed a multi-scale orderless pooling (MOP-CNN) method, in which the representations extracted from the fully connected layers for local patches were at multiple scale levels, and the representations were performed orderless VLAD pooling before concatenating the features.

With the further research on image retrieval based on CNNs, recent works demonstrated that convolutional layers contain more visual information on edges, corners, patterns, and structures which are suitable for image retrieval [1, 2]. In other words, relevant information contained in convolutional layers that may be not suitable for classification is still preserved for instance retrieval. However, deep features extracted from convolutional layers are usually in a large number, and are hardly for similarity computation without aggregation due to large memory footprint and low efficiency. Thus, it is popular to aggregate derived deep convolutional features into a global descriptor. Off-the-shelf CNN features extracted from convolutional layers can be directly aggregated via spatial max pooling or sum pooling [10]. Despite efficiency, image vectors generated by max- and sum-pooling are not discriminative enough to result in state-of-the-art performance. Several recent works [5, 11, 47] have demonstrated that it is quite important to select features inside the region-of-interest (RoI) and to employ appropriate weighting schemes for the final aggregation.

Some recent works have focused on applying supervised fine-tuning to pre-trained CNN models [11, 12, 31]. When suitable training data is available, the image representations can be re-trained end-to-end. The fine-tuning process can significantly improve the performance of specific tasks. However, fine-tuning usually needs to spend large efforts on collecting, annotating and cleaning of suitable training dataset, which is not always feasible.

Inspired by aggregation methods with feature selection and weighting, we propose a feasible semantic-based image representation method in this paper. As shown in Fig. 1, the proposed global image vector generation method called RCSA contains three components: RoI selection, channel weighting and semantic-based aggregation. The RoI selection scheme,

which is denoted by RSC, is based on specific channels according to discriminative semantics and achieves excellent performance. The channel weighting (CW) scheme is obtained by analyzing the relations between the activation and various parameters (e.g. non-zero response) of feature maps. The final aggregation process dubbed CSBA is similar with SBA [17], while the difference is that, in the current work, we make it be independent of datasets. By incorporating the schemes of RSC, CW and CSBA, we finally derive our image representation method RCSA to aggregate deep convolutional features into a global image vector.

To be clear, we summarize the major contributions of this paper as follows:

1. Based on the different semantics of various feature maps, we select several specific channels of features to obtain an unsupervised RoI selection method RSC. This process is implemented prior to the aggregation of features. We will demonstrate that RSC performs very well and even better than the handle ground-truth box of query images on the Oxford5k dataset [29].
2. The relations between the discriminative activation and several parameters of feature maps, including standard variance, non-zero response and sum value, are analyzed in the work. Based on this research, we successfully propose an effective channel weighting method CW, and demonstrate its remarkable performance with both sum-pooling and CroW [18].
3. We improve and generalize the process of picking the semantic detectors in SBA [17]. Compared with SBA, our method CSBA chooses semantic detectors based on images sampled from Flickr, rather than the gallery images. It is worth noting that the new semantic detectors are obtained in once and can be applied for different datasets, which is more general than that in the original SBA in which the detectors should be calculated every time for each dataset. Besides, the new method even performs better on the instance retrieval task.
4. Finally, we present our image representation method through the combination of RSC, CW and CSBA. Extensive experiments demonstrate that RCSA achieves the state-of-the-art performance on benchmark image retrieval datasets.

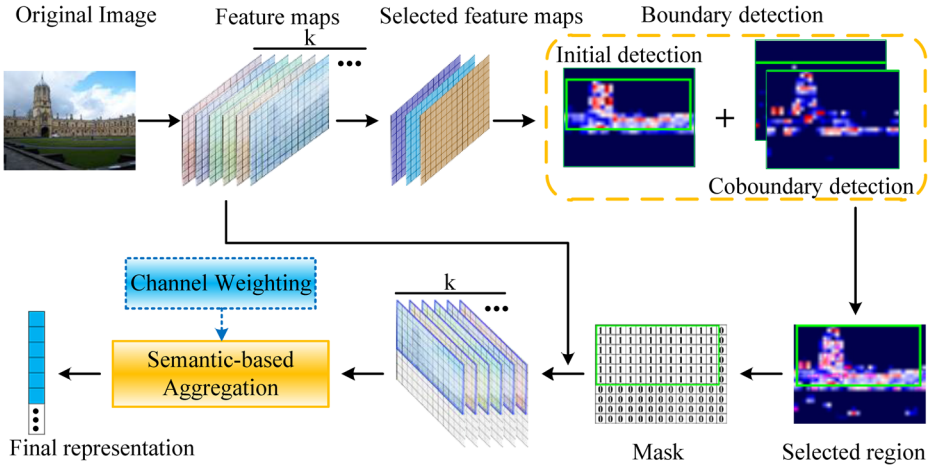

Fig. 1 The whole framework of our proposed method. The mask, which is used to select the RoI, is generated by several specific channels of features based on semantics. The semantic-based aggregation constituted by RSC, CW and CSBA is applied to generate the final representation

This work is an extension of our previous work [45], and this paper introduces more related methods and contains more details about the proposed methods. Besides, performance of each part of proposed methods in the retrieval task is conducted. Furthermore, the possibility to further improve the performance is discussed in the current work.

The remaining of this paper is organized as follows. Section 2 discusses related works. Section 3 presents the details of our main contributions, including the method to select the RoI, the strategy to get the channel weights, the process to get the CSBA features and the way to get the final semantic-based aggregation vector. Section 4 gives a wide range of experiments to comprehensively evaluate the proposed methods. Section 5 discusses the possibility to further improve the aggregate scheme. Section 6 concludes the current work.

2 Related works

To get global representation for image retrieval, RoI selection and weighted aggregation are often applied on the convolutional features. The related works of these two lines will be briefly reviewed in this section.

2.1 Selection of RoI

The discriminative information about the object semantics is useful for the selection of RoI [48, 56, 57]. As for the feature maps of CNNs, the semantic meaning has been analyzed by some works [51, 55]. Wei et al. [47] proposed a selective convolutional descriptor aggregation (SCDA) method based on the activated region of feature maps, and this method got a good performance on the fine-grained image retrieval. In this method, they added up the last convolutional layer activation tensor through the depth direction to get a 2-D tensor named "aggregation map", and the region on the aggregation map with value larger than the average value was selected as the interesting region.

Do et al. [8] attempted three different masking schemes for selecting RoI, including SIFT-mask, SUM-mask, and MAX-mask. Among those methods, the SUM-mask scheme was similar with that proposed by Wei et al. [47], and the MAX-mask, which was generated by the maximum local feature of each feature map, provided the best performance on image retrieval tasks.

Both "aggregation map" and "MAX-mask" seem to perform well on images containing single object. However, we find that these methods may not work when images not only contain the search objects but also have some other notable objects. To illustrate this phenomenon, Fig. 2 presents two images of Oxford5k dataset and their aggregation maps. For better visualization, the aggregation maps are overlaid to their original images by a multi-layered process. The first sampled image has rare noisy objects and most space of the image are filled by the objects of interest. It can be seen that the RoI selection method with the aggregation map works very well on the image. While the aggregation map to select the RoI fail to work on the second image, in which not only the search object but also other objects are contained. In this case, the most obvious activated part by the aggregation map is not the building but the red car and the flag. In order to avoid the drawback mentioned above on selecting RoI, in this work, we propose a simpler but more effective method based on the semantic meaning of feature maps.

2.2 Weighted aggregation

Babenko and Lempitsky [53] found that a simple global representation based on sum-pooling convolutional features and centering prior principle (SPoC) performed remarkably well without high-dimensional embedding. Razavian et al. [34] adopted maximum activations of the whole convolutional layers (MAC) as an image representation, in which the discriminative activations might be suppressed due to global max-pooling leading to a poor performance compared with sum-pooling [53]. Later, Tolias et al. [41] proposed a method to get regional maximum activation of convolutions (R-MAC), in which a strategy was used to aggregate the maximum activation over multiple spatial regions sampled on the convolutional layer using a fixed layout. Hoang et al. [43] embedded the selected local convolutional features to higher-dimensional space using various embedding methods before implementing aggregation with democratic aggregation method [15]. The results showed that the T-emb [15] embedding method and democratic aggregation achieved the most outstanding performance on the task of instance retrieval. Most recently, Chen et al. [5] further improved the performance of R-MAC. They proposed a method to generate regions through feature clustering based on feature similarity, which was different with the regions in original R-MAC that were square in shape and defined independent of the image content.

As for the recent researches of sum-pooling, Kalantidis et al. [18] proposed a non-parametric method to learn weights for both spatial locations and feature channels. In their work, the spatial weight derived from the spatial activation and the channel weight derived

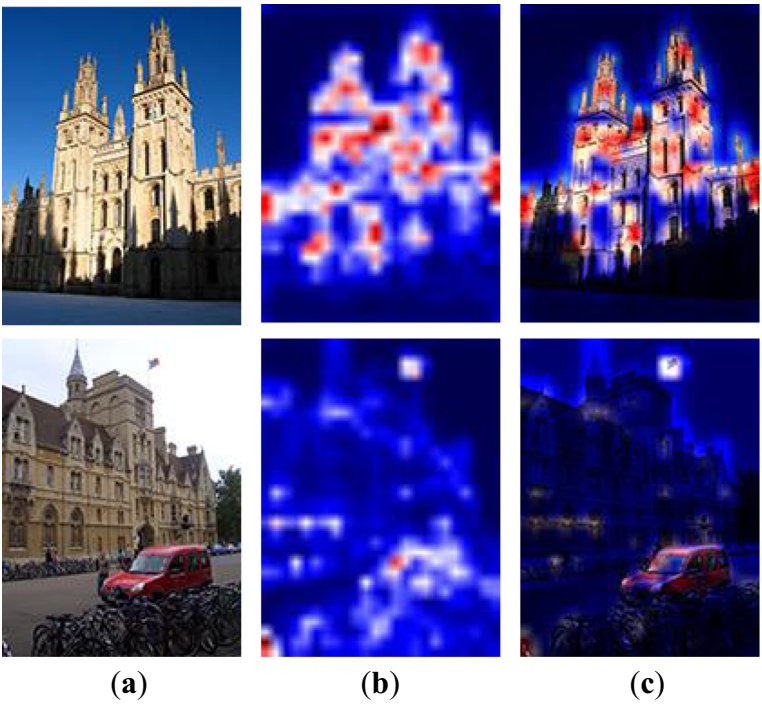

Fig. 2 Visualization of aggregation maps. (**a**) Images sampled from Oxford5k (**b**) Heat maps of aggregation maps, the warm (red) region is the activated region (**c**) Original images multiplied by the corresponding heat maps

from channel sparsity were used on the aggregation process accompanying the sum-pooling. This approach (CroW) obviously improved the performance of sum-pooling on convolutional features. Inspired by the features of SPoC and CroW, Wang et al. [46] improved the original CroW significantly. They extended SPoC by adaptively determining the center point of RoI for the spatial weight, and proposed element-value sensitive channel weighting strategy to obtain channel weights.

Most recently, based on the semantic content of feature maps, Xu et al. [17] proposed a method to create image representation via semantic-based aggregation (SBA). In their method, N discriminative channels were chosen as semantic detectors which were called as "probabilistic proposals", and then these detectors were used to weight the feature maps respectively obtaining N group regional features. After aggregation, N group features were concatenated to one final representation. This method gets rid of the limitation of representation dimensionality and makes it possible to obtain representations with higher dimension by simple aggregation and concatenation.

The SBA method achieves comparable performance with the state-of-the art methods even with the same dimensional representation after dimensionality reduction. However, in the original SBA, the channels used as detectors were obtained based on gallery images. Firstly, they extracted features of all images in the dataset, and then aggregated each of them to a 512-dimensional vector by sum-pooling. Variance of each channel of those vectors was calculated and channels with top N variance value were selected as the detectors. That means for each dataset, the original method has to analyze the dataset to get channels. Fortunately, in this work, we found that the standard variance of each feature map has a strong correlation with semantics. Based on which, we propose an effective method to select weighting channels independent on the datasets.

3 Methodology

Figure3 shows the detail framework of our method which would be presented in this section. As shown in Fig. 3, the branch 1 is the selection of RoI, which will be presented in the section 3.2; the branch 2 is the process of getting channel weights, and this method will be presented in the section 3.3; the selection of maps working as detectors in branch 3 and the subsequent process to combine these methods to get final representation will be discussed in section 3.4. For a start, we will introduce the notations used in the paper.

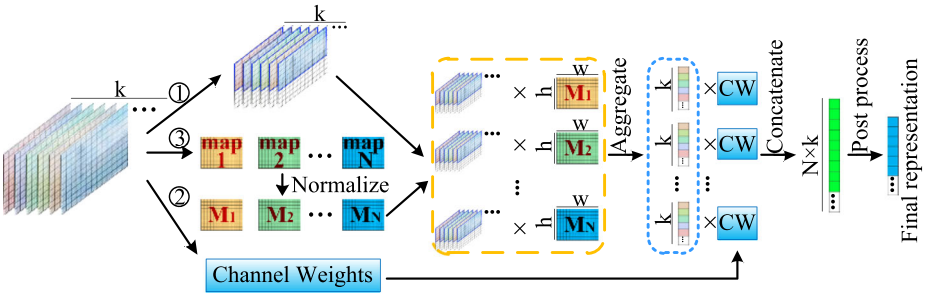

Fig. 3 The detailed framework of the proposed image representation method

3.1 Preliminary

Notations used in this paper would be introduced in the following. The term "feature map" is one channel of convolutional features; "features" indicates feature maps of all channels in a convolution layer; and the term "representation" indicates the final d-dimensional vector of aggregated features used for retrieval.

Features extracted from a convolutional layer is an order-3 tensor T with $h \times w \times k$ elements, which includes a set of 2-D feature maps $\boldsymbol{S} = \{S_n\}$ ($n = 0,1,\ldots,k\text{-}1$). S_n of size $h \times w$ is the n-th feature map of the corresponding channel (the n-th channel). We denote the deep convolutional features as $\boldsymbol{V} = \{v_0(i,j), v_1(i,j),\ldots,v_n(i,j),\ldots, v_{k\text{-}1}(i,j)\}$, where (i,j) is position on one feature map ($i \in \{1,\ldots,h\}$, $j \in \{1,\ldots,w\}$).

3.2 Selecting RoI

In the following, we propose our RoI selection method, and then present the selection results. Note that this work is only based on the pre-trained model VGG16 [40] and none fine-tuned model is adopted.

Method It has been reported that each channel of feature map is activated by special patterns according to fixed semantic content [17]. To illustrate this, several feature maps extracted from the geometries image are visualized in Fig. 4. It can be seen that the 146th channel tend to be activated by the columnar structure, and the 343rd and 394th channels are mostly activated by the cone and arc respectively, while the 447th channel is only activated by the sphere. Thus, it is possible to use those feature maps to locate the object. We do not mean to adopt all the channels related to the search object but utilize only a very small number of feature maps which can locate the object. Since a specific channel is activated by a specific pattern, the feature maps selected to locate the object would not share the same channel for different kinds of query objects. In the current work, we focus on the retrieval task on buildings. However, if it works on the building it would be possible to work on other kinds of object research, and one need to do is to select appropriate channels for the specific task.

For getting the reasonable channels to select the RoI, we visualized feature maps and chose three channels which are most useful for the selection of RoI. Figure 5 shows sampled images of Oxford5k dataset and heat maps of specific channels. Consistent with what reported by Xu et al. [51], the 360th feature map is most activated by the body of building. Thus the 360th feature map is possible to be used to detect the building region. It can be seen from the Fig. 5 that the activated region of the 360th feature map is not always continuous on the building regions but may be sparse such as that of third image in the left. Therefore, the 360th feature map is hard to be used as the detector to locate the object directly. A possible solution is to

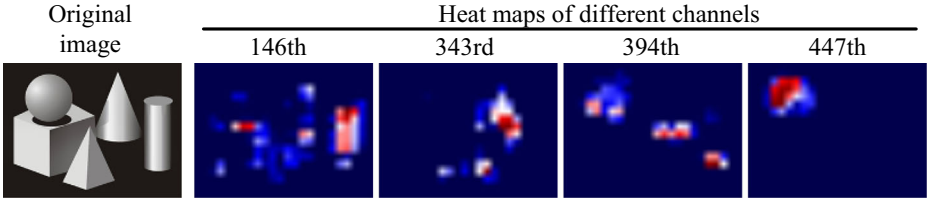

Fig. 4 Feature maps visualization of several geometries

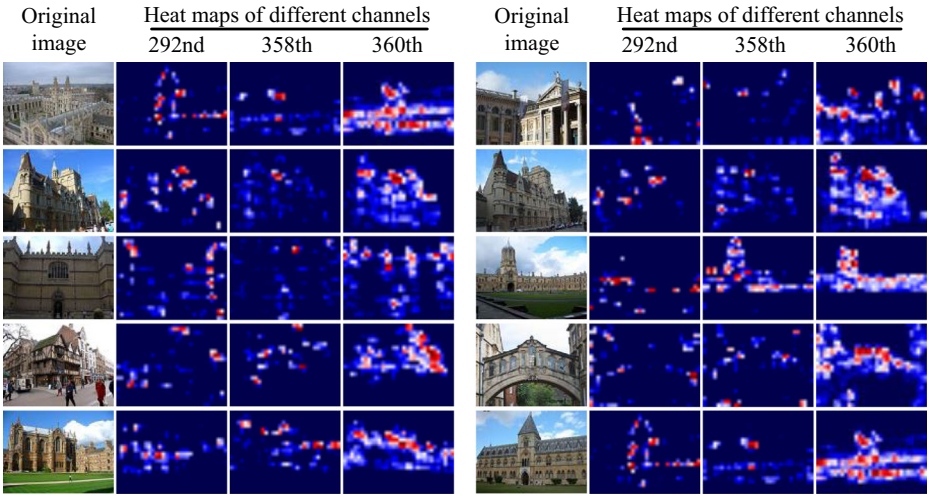

Fig. 5 Visualization of feature maps that can be used to select the RoI. The first column of each subfigure are the input images sampled from Oxford5k dataset, and the following images are heat maps of three specific channels of pool5 layers from VGG16

detect the boundary of the object and then select it with a continuous box. This process is shown in the first line of Fig. 1 with the label of "Initial detection". Another problem is that, as shown in the Fig. 5 and the initial box selection result in the Fig. 1, even though the 360th feature map is activated by the body of buildings, it fails to be activated by the rooftops which are always to be conical and sharp. Based on the fact that different feature maps are activated by different parts of the object, we choose two other feature maps which are activated by the conical shape and sharp shape respectively as shown in Fig. 5. It can be seen that the 292nd feature map is almost activated by the conical shape and the 358th feature map is always activated by the sharp shape. With the 292nd and 358th feature maps, we can successfully detect the upper boundary of the building as shown in the first line of Fig. 1 with the label of "Coboundary detection". Then, updating the selection box detected by the 360th feature map with the upper boundary detected by 292nd and 358th feature maps (the uppermost boundary is adopted) we can obtain the final selection box. The schematic of the selection process is shown in the first line of Fig. 1.

With the selected channels we can acquire the exact coordinates of the predicted bounding box. As mentioned before, each feature map is a $h \times w$ 2-D tensor. We denote the $v_n(i,j)$ as the value at the position (i,j) on the n-th feature map S_n; denote $X_{n,j}$ and $Y_{n,i}$ as the sum of v_n along the column and row direction respectively, and $\overline{X}_n, \overline{Y}_n$ denote the average value of $X_{n,j}$ and $Y_{n,i}$:

$$\overline{X}_n = \frac{1}{w}\sum_{j=1}^{w} X_{n,j} = \frac{1}{w}\sum_{j=1}^{w}\sum_{i=1}^{h} v_n(i,j) \tag{1}$$

$$\overline{Y}_n = \frac{1}{h}\sum_{i=1}^{h} Y_{n,i} = \frac{1}{h}\sum_{i=1}^{h}\sum_{j=1}^{w} v_n(i,j) \tag{2}$$

The 360th feature map is used in the initial detection process of box boundaries. For the detection of left boundary, let j of $X_{360,j}$ increase from 1 till the value of $X_{n,j}$ satisfies

$X_{360,j} \geq \alpha \overline{X}_{n,j}$, where α is a coefficient. Then the coordinate of left boundary is obtained which is the value of j at this position. With the same process we can get the other three boundaries. The coefficient α is set as 0.6 in initial detection process for all boundaries. The next is to update the upper boundary with the 292nd and 358th channels. Since the activated region by the peak of conical or sharp is always tend to be a point, the average value of whole feature map is used in this section, which is defined as:

$$\Omega_n = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} v_n(i,j) \tag{3}$$

Let $\phi_{n,i}$ to be the max value of $v_n(i, j)$ in the *i-th* row, which is:

$$\phi_{n,i} = \max\{v_n(i,1), v_n(i,2), \cdots, v_n(i,j), \cdots, v_n(i,w)\} \tag{4}$$

Similar with the detection method in the initial detection process, make i of $\phi_{n,i}$ increase from 1 till the value of $\phi_{n,i}$ satisfies $\phi_{n,i} \geq \beta \Omega_n$, and the coordinate of upper boundary is the value of i at this position. Both of the 292nd and 358th channels are used for the above process, and the uppermost boundary is adopted as the new upper boundary to update the bounding box obtained in the initial process. The coefficient used in this process β is 0.05.

Qualitative evaluation The evaluation of the proposed method to select the RoI is presented in this section. Since that the query images of Oxford building supply the ground-truth bounding boxes, it is desirable to compare the given ground-truth bounding boxes with that predicted by our proposed method.

We qualitatively evaluate the proposed method to select the RoI on all query images of Oxford building dataset. Figure 6 presents the comparing results, where red boxes are ground-truth boxes and the green are predicted by our method. According to these figures, one can see that the predicted bounding boxes can cover the object building very well. Compared with the ground-truth boxes, the predicted boxes even get better selection in some cases. For instance, in the third image from last of the first row, the predicted box is smaller than the given one, and covers the object building with fewer noisy background; in the sixth image of the second row,

Fig. 6 Object localization bounding box of all query images of Oxford building. The ground-truth bounding box is marked as the red rectangle, while the predicted one is marked in the green rectangle

the predicted box also covers the object building with fewer irrelevant background information, and these results preliminarily demonstrate the reliability of our method.

3.3 Systematic investigation on channel weights

As mentioned above, various feature maps of different channels are activated by different parts of the object. We can observe all of those feature maps and choose several feature maps to be the detectors of RoI, however, it is impossible for us to analyze all feature maps to learn the weight of each channel individually. Therefore, a unified standard of measurement for the channel weight is needed. Kalantidis et al. [18] implemented the channel weight in their work with the sparsity of each channel. Wang et al. [46] improved this channel weight by replacing the sparsity of each channel with sum value. However, comprehensive researches on the channel weight are still lacking. In this section we tend to give a systematic investigation on this aspect and to get a more reasonable method to generate channel weights.

Besides the sparsity and sum value of channels used in the previous works [18, 46], the standard variance is also taken into account in the current work. In order to get intuitive relation between the feature map and those parameters, we sampled several images from Oxford5k dataset and draw heat maps of all feature maps. Heat maps with top 10 value of corresponding parameters are shown in Fig. 7, and those feature maps are arranged in descending order. For the direct comparison with the rest two parameters, the sparsity is replaced by the non-zero response, which is defined as:

$$\Psi n = \sum_{i=1}^{h} \sum_{j=1}^{w} \{1 | v_n(i,j) > 0\} \tag{5}$$

As shown in these figures, there are many feature maps activated by the irrelevant background (boxed by green boxes) in these channels with top 10 non-zero response. This is because the non-zero response or sparsity only considers the activated area and ignores the intensity of activation, leading to the result that the background may take a large weight in these feature maps with large value of non-zero response. For instance, in the third image of Fig. 7 (a), the feature map with the largest non-zero response is almost activated by the background, since the background (sky and ground) takes a larger proportion than the object building in the original image. That means the non-zero response or the sparsity might not be an optimal parameter for the setting of channel weights. Compared with non-zero response, the standard variance and sum value show better relevance with semantics of channels. As shown in Fig. 7 (b), only two feature maps are mainly activated by the ground which is not related to the object building. As for the sum value, as Fig. 7 (c) shows, the number of feature maps activated by the noisy objects is more than that in Fig. 7 (b) while it shows much better than that presented by the non-zero responses.

According to the Fig. 7, the standard variance and sum value seem to be more appropriate in generating the channel weight comparing with the sparsity. In the method of CroW [18], the channel weight is set by a logarithmic function of sparsity, where the weight of channel shows a positive correlation with the channel sparsity. The explanation given by the authors of CroW about this positive relation is that channels with frequent features occurrences are already strongly activated while infrequently occurring features could provide important signals. However, we tend to think that those feature maps only activated by a small part of the object, such as the 292nd and 358th

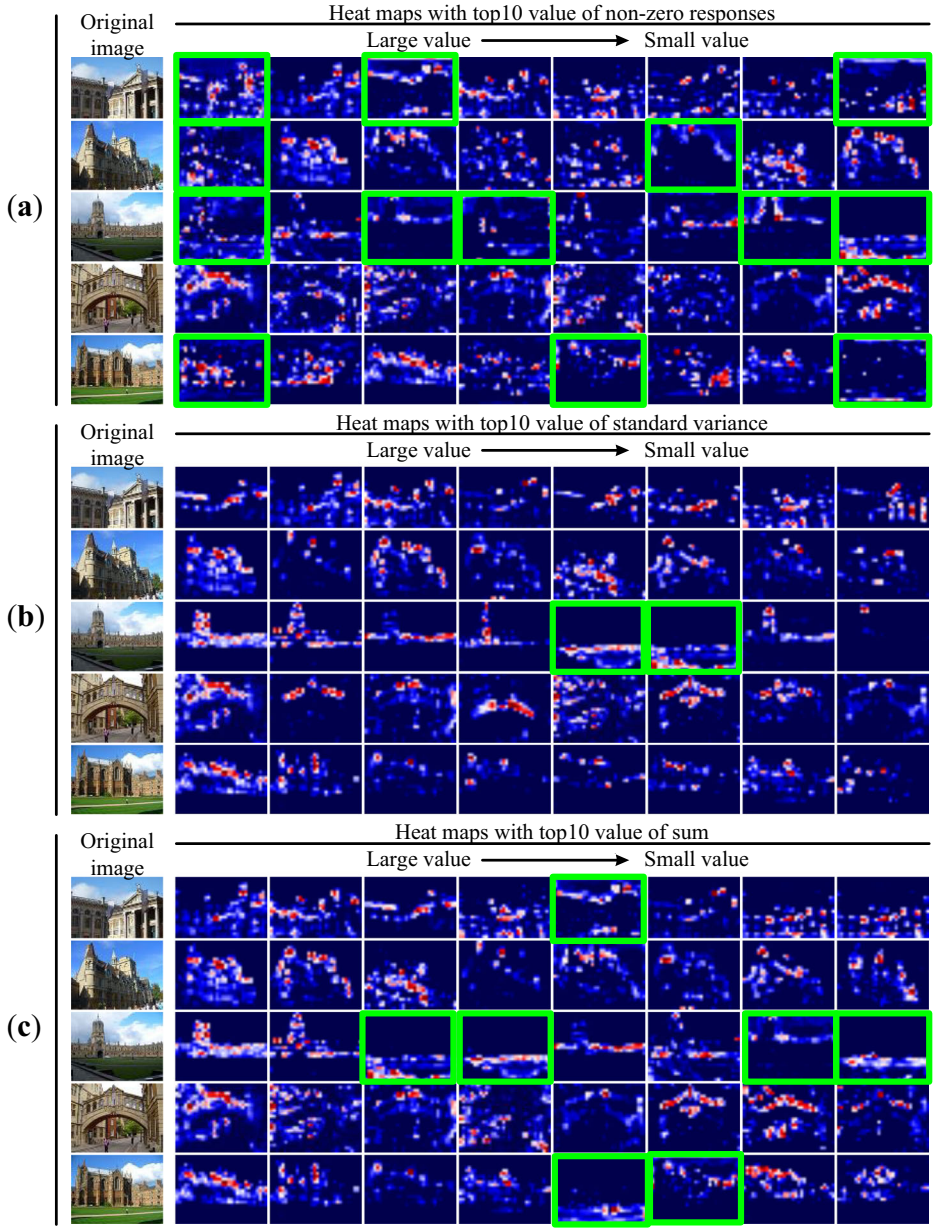

Fig. 7 Images sampled from Oxford5k and corresponding heat maps arranged by various parameters. (**a**) Arranged in descending order according to non-zero responses, (**b**) Arranged in descending order according to standard variance (**c**) Arranged in descending order according to sum value

channels which are activated by the conical shape and sharp shape, are still very import since they may represent the key features of object but are more sparsity and have a smaller value of standard variance and sum value. Whatever, expressions which present negative correlation between the channel weight and standard variance or sum value are needed. Besides the logarithmic function

which is similar with that used in CroW, linear function, exponential function and Gaussian function will be tested in the current work. Expressions of these functions are listed in Eqs. 6–9.

Linear function:

$$B_n = -\sigma Q_n^* + 1 \tag{6}$$

Exponential function:

$$B_n = \exp\left(-Q_n^*\right) \tag{7}$$

Logarithmic function:

$$B_n = \log\left(\frac{\varepsilon + \sum_i Q_i^*}{\varepsilon + Q_n^*}\right) \tag{8}$$

Gaussian function:

$$B_n = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{Q_n^{*2}}{2\sigma^2}\right\} \tag{9}$$

where B_n is the *n-th* channel weight; σ is a coefficient which optimized value for linear function and Gaussian function is 0.6 and 0.8 respectively; ε is a small constant added for numerical stability like that in CroW; Q^*_n is a normalized parameter defined as:

$$Q_n^* = \frac{Q_n - Q_{\min}}{Q_{\max} - Q_{\min}} \tag{10}$$

where Q_n is the concerned variable of the *n-th* channel and it can represent the standard variance or sum value of the channel; $Q_{\min}$ and $Q_{\max}$ represent the minimum and maximum value of interested variable (standard variance or sum value) among all channels of one image features extracted from the pool5 layer.

3.4 Semantic-based aggregation

As mentioned in the section of related works, in the original aggregation method SBA, the channels used as semantic detectors are obtained based on gallery images. In the following, an improved method (CSBA) which can get rid of the dependence on the dataset is presented. When this aggregation method proposed in this section is combined with the previous proposed RSC and CW, the whole semantic-based aggregation method can be obtained.

Algorithm 1. Selection of discriminative channels

Input: 60 building images picked from Flickr
Output: N channel number
f_c=0
For feature maps of 60 images
  For channel C of feature maps
    If standard variance of channel C appears among top 50 value of feature maps
      f_c+=λ
$\boldsymbol{f}$=$\{f_0, f_1, \ldots, f_c, \ldots, f_{511}\}$
N channels $\boldsymbol{C}$ are selected with top N value of f_c

It has been shown that the standard variance value of each 2-*D* feature map has obvious relation with the semantic content, which makes it possible to use the channels whose feature maps are in a great value of standard variance to act as the semantic detectors in SBA. In the current work, we sampled 60 building images[1] from Flickr to pick the detector channels. With these images, we count the frequency of each channel when their standard variance value of the feature map appears among the top 50 in all feature maps of each image. Then N channels with top N frequency are selected to replace those channels acting as semantic detectors in SBA. The detail process of selection is shown in Algorithm 1. Since the feature map with larger standard variance has more discriminative content, as shown in Fig. 7, a coefficient λ is adopted based on their standard variance value of each feature map:

$$\lambda = \log\left(\sum_{n=1}^{50} n \Big/ n\right) \tag{11}$$

where n means the order of a feature map in the descending-ordered feature maps sorted by standard variance of one image. The weighted frequency is shown in Fig. 8, and the number of channels (N) is 25 as recommended in the SBA. These channels are used to replace that acting as semantic detectors in the original SBA for different datasets.

After getting the detector channels, 25 feature maps can be acquired from features $\boldsymbol{S}$ of each image according to the selected channel number. With the selected feature maps, series weighted and sum-pooled representations can be obtained:

$$\psi_n(I) = \sum_{i=1}^{h}\sum_{j=1}^{w} w_n(i,j)S(i,j) \tag{12}$$

The coefficients w_n are the normalized weights based on the activation values $v_n(i, j)$ of one selected feature map:

$$w_n(i,j) = \sqrt{\frac{v_n(i,j)}{\left(\sum_{i=1}^{h}\sum_{j=1}^{w} v_n(i,j)^2\right)^{1/2}}} \tag{13}$$

Concatenating these representation, a global $25 \times k$-dimensional representation can be obtained:

$$\psi(I) = [\psi_1(I), \psi_2(I), ..., \psi_{25}(I)] \tag{14}$$

The final CSBA representation $\psi_{\mathrm{CSBA}}(I)$ is generated when the global representation $\psi(I)$ is processed by $l2$-normalization and PCA whitening. This representation can derive the RSCA representation, which combines the RSC, CW and CSBA, when the features $\boldsymbol{S}$ in Eq. 12 is multiplied by the mask of RSC and each representation $\psi_n(I)$ is weighted by channel weights (CW).

4 Experiments and results

The performance of each part of the proposed semantic-based aggregation method, including the RoI selection method RSC, channel weighting method (CW) and improved aggregation

[1] https://github.com/ShawnWXS/filckr_building

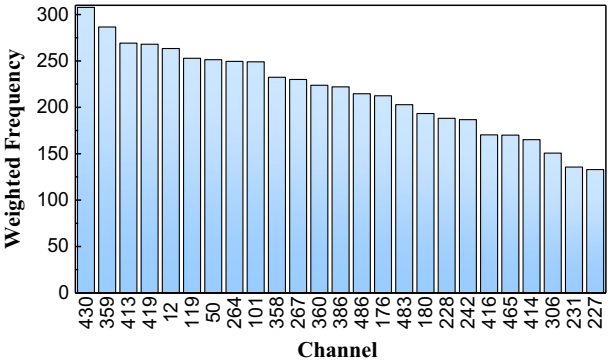

Fig. 8 Channels with top 25 weighted frequency are used to replace those channels acting as semantic detectors in SBA

methods CSBA, on image retrieval task is tested respectively. The retrieval results of the final representation generated by the whole semantic-based aggregation process are compared with previous state-of-the-art results.

4.1 Experiment setting

The proposed methods are evaluated on four benchmark datasets, including Oxford5k [29], Paris6k [30], Oxford105k and Paris106k. Oxford5k dataset contains 5063 building photos with 55 queries including 11 landmarks, and Paris6k contains 6392 building photos with 55 queries including 11 landmarks. Oxford105k and Paris106 are extensions of Oxford5k and Paris6k respectively by adding other 100,000 distractor images collected from Flickr.

The off-the-shelf pre-trained VGG16 [40] is used in this paper. Deep convolutional features maps are extracted from the pool5 layer and the number of channels is $k = 512$. For fair comparison with the related retrieval methods, we learn the PCA and whitening parameters on Oxford when testing on Paris and vice versa. The mean Average Precision (mAP), which is defined as the average percentage of same class images in all retrieved images after evaluating all queries, is used to evaluate the retrieval performance. Additionally, all images are in the original size and not any resize process is adopted in this paper.

4.2 Implementation details

When the RSC is adopted in the retrieval task, the feature maps $\boldsymbol{S}$ obtained from the pool5 layer would be treated by a mask map before the aggregation. The mask map M with the same size as a feature map is generated by the RSC:

$$M_{i,j} = \begin{cases} 1 & \text{if } x_0 \le j \le x_1 \text{ and } y_0 \le i \le y_1 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

where x_0 and x_1 are the positions of left and right boundaries of predicted bounding box; y_0 and y_1 are the locations of upper and lower boundaries of the box. Then M is used to select the deep convolutional features within the predicted box. The descriptor $v(i, j)$ should be kept when $M_{i,j} = 1$, while $M_{i,j} = 0$ means the position (i, j) is not inside the box:

$$F = \left\{ v(i,j) | M_{i,j} = 1 \right\} \tag{16}$$

where F is the selected descriptor set which will be aggregated into the final representation for image retrieval.

The aggregation processes of CSBA is similar with the original works. The difference between the CSBA and the original SBA is just the selection method of channels which are used as semantic detectors. Note that in the current CSBA, the selection of detectors never depends on the datasets. As for the implementation of the proposed channel weighting (CW) method, it is applied to the features after aggregation like most of similar works [18, 46]. In the whole semantic-based aggregation process (RCSA), the channel weights are implemented on the aggregated vector treated by each detector of CSBA before concatenation.

When the process is carried out on an Intel® Core™ i7–4790 quadcore CPU running at 3.6GHz and 8GB of RAM, the aggregation process takes around 81 ms for a query image with resolution of 768×1024 pixels, which can be capable to meet the real-time online research.

4.3 Performance of RSC

The retrieval performance of RSC is presented in Table 1. The simple sum-pooling and CroW are adopted in the aggregation process. Under the item of method of selecting region, "None" means features are extracted from the original images, "Ground-truth box" means features are extracted from the queries within the ground-truth box and "RSC" means features extracted from original images are treated by the RSC selection process. As shown in this table, since the ground-truth box filters parts of background noises, the use of it can significantly improve the retrieval performance for both cases with simple sum-pooling aggregation and CroW aggregation. Compared with the handled ground-truth box, when the proposed RSC is only adopted on the query images, this unsupervised RoI selection method even gets better retrieval performance with the aggregation method of sum-pooling. When the CroW is implemented, although the performance of RSC is slightly poor than ground-truth box on the dataset of Oxford5k, it still gets better performance on dataset of Paris6k. If the RSC is implemented on both of query images and index images, encouraging results are presented on both cases with aggregation methods of simple sum-pooling and CroW.

Note that the performance with RSC and simple sum-pooling method even better than the original CroW method (in which the representation of query image is obtained within the ground-truth box). Combining the RSC and CroW would greatly enhance the performance as

Table 1 Performance of RSC on the image retrieval task

Method of selecting region		Aggregation	Dim	Dataset	
Query Images	Index Images			Oxford5K	Paris6K
None	None	Sum	512	68.0	77.7
Ground-truth box	None	Sum	512	69.6	78.6
RSC	None	Sum	512	69.7	78.6
RSC	RSC	Sum	512	72.4	80.9
None	None	CroW	512	68.3	79.1
Ground-truth box	None	CroW	512	70.8	79.7
RSC	None	CroW	512	70.6	81.4
RSC	RSC	CroW	512	**74.1**	**83.8**

shown in the last line of the table. These results further indicate the effectiveness of the proposed method to select RoI and filter the information of irrelevant background.

4.4 Impact of the parameters of channel weights

The parameters of channel weights (CW) contains two parts, namely, the variable Q (standard variance or sum value) and the weighting function B_n, i.e., linear function (linear), exponential function (exp), logarithmic function (log) and Gaussian function (Gauss). To test the performance of channel weighting method with different parameters, the aggregation method Crow is adopted in this section, and the channels weights of Crow are replaced by the proposed weights. Table 2 shows the performance of different variables (standard variance and sum value) and weighting functions on retrieval task. The best result of same dimensionality is in bold. It can be seen that the performance under almost every weighting function with standard variance and sum value is better than that of the original CroW, which demonstrates the effectiveness of standard variance and sum value on the calculation of channel weights. Particularly, for the dataset of Paris6k, when the channel weight is acquired by sum value and exponential function, the mAP increases by at least 2% compared with CroW with 512, 256, and 128 dimensional representations. Overall, the channel weight based on sum value outperforms that based on standard variance, and the exponential function performs better than other functions. Therefore, the sum value and exponential function is adopted for the channel weights (CW) generating method.

4.5 Performance of CSBA

Comparison of proposed method (CSBA) and the original SBA is presented in Fig. 9, and the performance of CSBA combined with the proposed RoI selection method (RSC) and channel weighting method (CW) is also shown in this figure. One can see that the proposed method based on 60 building images sampled from Flickr not only achieves the performance of SBA

Table 2 Performance of different variables (standard variance and sum value) and channel weighting functions on retrieval task

Q	B_k	Datasets					
		Oxford5k			Paris6k		
		Dim			Dim		
		512	256	128	512	256	128
Standard variance	linear	71.1	68.6	63.2	82.0	78.8	76.8
	exp	71.2	68.6	63.5	81.9	78.7	76.7
	log	71.2	68.2	63.8	80.6	76.9	75.2
	Guss	71.2	68.7	63.2	**82.0**	78.8	76.7
Sum value	linear	71.7	68.6	64.3	81.8	79.1	77.1
	exp	**71.9**	68.8	**64.4**	81.9	**79.3**	**77.3**
	log	71.8	**68.9**	64.2	81.5	78.4	76.7
	Guass	71.4	68.1	64.2	81.5	78.9	76.7
Baseline (CroW)		70.8	68.4	64.1	79.7	76.5	74.6

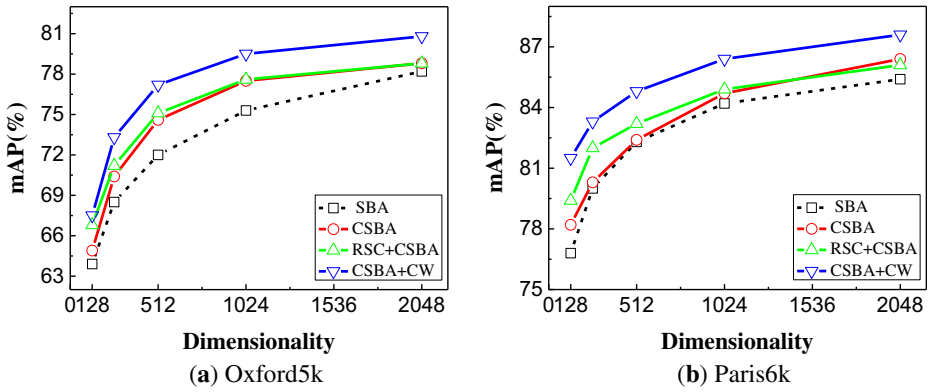

Fig. 9 Performance of proposed CSBA on image retrieval task. CSBA combined with proposed RSC and CW is also tested. The setting of CSBA almost same with original SBA except for the selection of detector channels

which based on gallery images, but even performs better on both Oxford5k and Paris6k under all dimensionalities. Particularly, in the dataset of Oxford5k, compared with the original method the proposed method exceeds 2.6% on mAP using 512 dimensional features. Most importantly, the proposed CSBA makes the original SBA get rid of the dependence on the datasets and makes it simpler and more effective.

The implementation of the proposed channel weighting method (CW) significantly improves the performance of CSBA, and the smaller dimensionality the larger improvement will be. With 128-D presentation, the utilization of CW on CSBA make the mAP increase by 2.6% and 3.3% comparing with CSBA on datasets of Oxford5k and Paris6k respectively, and the least gain with all dimensionality is 2.0% and 1.2% respectively. As for the performance of RSC on CSBA, when the dimensionality of presentation is smaller, the improvement of mAP

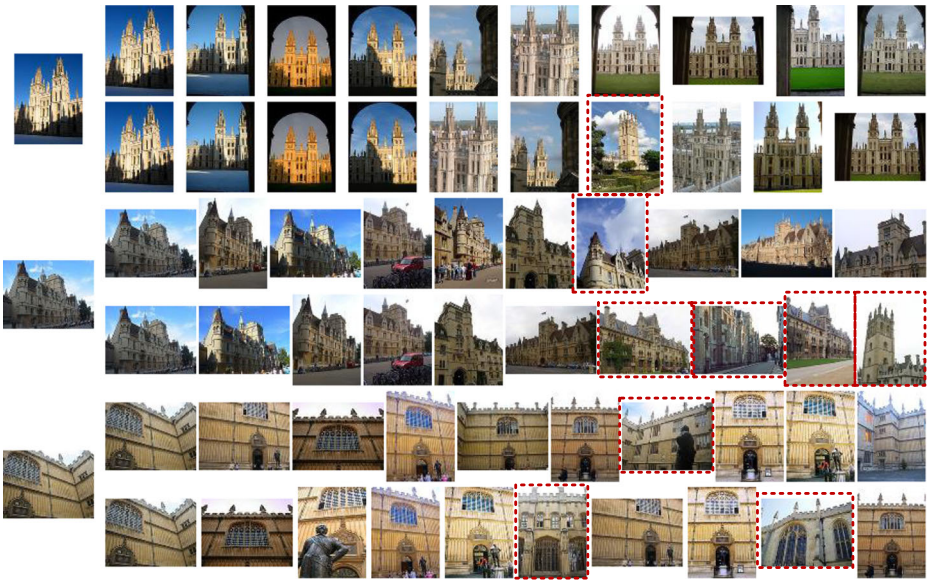

Fig. 10 Three example queries (on the far left) from Oxford5k and the corresponding top10 results of RCSA (top) and SBA (bottom). The false results are marked with red dashed borders

achieved by the implementation of RSC is significant, while the RSC even plays a negative role when the dimensionality is larger than 1024. It is perhaps caused by that more semantic information can be contained in the representations with higher dimensionality, and in this case those background could be helpful for the image retrieval. Overall, the CSBA can be enhanced by both of RSC and CW.

We obtain the whole semantic-based aggregation process RCSA by combining RSC, CW and CSBA. Several retrieved results of RCSA and SBA are presented in Fig. 10, which visually show the better performance of the proposed method. The detail performance of the proposed RCSA and the comparison results with other methods will be presented in the next section.

4.6 Comparison with the state-of-the-art

The comparison of the proposed RCSA with outstanding methods is presented in Table 3. Among these methods, the improved R-MAC proposed by Chen et al. [5] achieves the state-of-

Table 3 Accuracy comparison with the state-of-the-art unsupervised methods

Method	Dim	Dataset			
		Oxford5k	Paris6k	Oxford105k	Paris106k
CroW [18]	128	64.1	74.6	59.0	67.0
SBA [17]	128	63.9	77.0	59.1	69.0
Wang et al. [46]	128	65.8	77.9	61.4	70.5
Previous state-of-the-art	*128*	*65.8*	*77.9*	*61.4*	*`70.5*
RCSA	128	**70.3**	**81.8**	**66.1**	**74.8**
CroW [18]	256	68.4	76.5	63.7	69.1
SBA [17]	256	68.5	80.0	63.8	72.2
R-MAC [41]	256	56.1	72.9	47.0	60.1
Wang et al. [46]	256	70.7	80.5	66.5	74.0
Previous state-of-the-art	*256*	*70.7*	*80.5*	*66.5*	*74.0*
RCSA	256	**75.0**	**84.2**	**71.2**	**77.8**
CroW [18]	512	70.8	79.7	65.3	72.2
R-MAC [41]	512	66.9	83.0	61.6	75.7
SBA [17]	512	72.0	82.3	66.2	75.8
Hoang et al. [43]	512	65.7	81.6	60.5	72.4
Chen et al. [5]	512	73.8	83.9	69.7	76.4
Wang et al. [46]	512	72.8	83.0	68.1	76.3
Previous state-of-the-art	*512*	*73.8*	*83.9*	*69.7*	*76.4*
RCSA	512	**77.5**	**85.2**	**74.8**	**79.1**
Hoang et al. [43]	1024	72.2	83.2	67.9	76.1
SBA [17]	1024	75.3	84.2	69.3	78.2
Previous state-of-the-art	*1024*	*75.3*	*84.2*	*69.3*	*78.2*
RCSA	1024	**79.8**	**86.3**	**77.2**	**80.4**
SBA [17]	2048	78.2	85.4	71.1	79.7
RCSA	2048	**80.9**	**87.5**	**78.4**	**81.8**
CroW+QE [18]	512	74.9	84.8	70.6	79.4
R-MAC + AML + QE [41]	512	77.3	86.5	73.2	79.8
SBA + QE [17]	512	74.8	86.0	72.5	80.7
Previous state-of-the-art	*512*	*77.3*	*86.5*	*73.2*	*80.7*
RCSA+QE	512	**81.9**	**87.6**	**79.6**	**82.0**
SBA + QE [17]	1024	77.9	87.8	76.7	82.8
RCSA+QE	1024	**83.9**	**88.8**	**81.8**	**83.7**
SBA + QE [17]	2048	80.7	88.7	79.3	83.9
RCSA+QE	2048	**84.9**	**90.2**	**82.8**	**84.9**

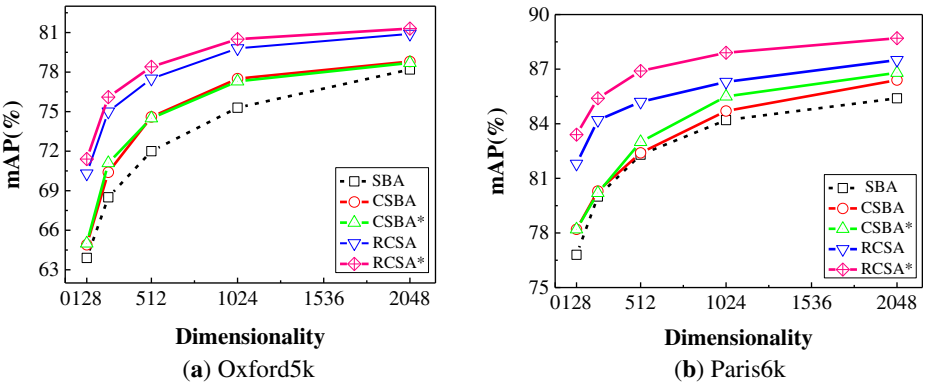

Fig. 11 Comparison of CSBA and CSBA*. The detector channels in CSBA* are selected dependent on the 55 query images of Oxford dataset and weighting formula Eq. 17. RCSA* is combination of RSC, CW and CSBA*

the-art performance with the most common representation dimension (512-D). With regard to other dimensions, the best performance is emphasized in the table. As shown in this table, the proposed methods RCSA outperforms the previous state-of-the-art on all four standard retrieval datasets and all dimensionalities from 128-*D* to 2048-*D*. On specifics, the gain is at least 3.9%, 3.7%, 1.3%, 2.1% and 2.1% in mAP from 128-*D* to 2048-*D* respectively for all datasets comparing with the state-of-the art, and the largest gain reaches 7.9% in mAP for Oxford105k with 1024-*D* representation. Note that the 256-*D* representation of this RCSA method achieves significantly better mAP than that of the previous state-of-the-art with 512-*D* representation.

The proposed methods combined with query expansion are also compared with other methods in the last part of Table 3. In the experiments, the average query expansion (QE) [7] computed by the top 10 query results is used. It can be seen from this table that with QE the RCSA still gets the best performance. Specifically, the gain with mAP reaches 6.4% for Oxford5k with 512-D representation when RCSA+QE is adopted. With higher dimensional representations, RSCA+QE increases the mAP at least 0.9% for all four datasets comparing with that of SBA + QE.

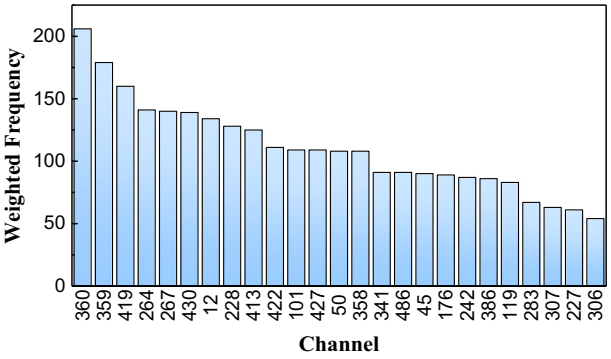

Fig. 12 Channels with top 25 weighted frequency acting as detectors in CSBA*

5 Discussion

In the process of selecting detector channels in CSBA, 60 building images sampled from Flickr and simple weighting equation Eq. 11 are adopted to make it be free of the dependence on the dataset and avoid the complex parameters in the weighting formula. However, we find that the channels selected with this method is not optimal. Figure 11 presents the performance of another process to improve the original SBA, which is donated by CSBA*. In the CSBA*, the images used to select detector channels are 55 query images of the Oxford dataset, and the weighting formula of Eq. 11 is replaced by the Eq. 17. The selected channels and their corresponding weighted frequency is shown in Fig. 12.

As shown in Fig. 11, RCSA* obviously improves the performance of RCSA, especially in the dataset of Paris6k. Table 4 shows the detail performance of RCSA and RCSA*. It can be seen that compared with RCSA, RCSA* gets better performance on all datasets and with all dimensionalities. The largest gain reaches 2.3% when combined with QE with 512-*D* representation on Paris106k. This means that it is possible to get appropriate detector channels in some way to further improve the performance.

$$\lambda = \begin{cases} 4 & n \leq 5 \\ 3 & 5 < n \leq 15 \\ 2 & 15 < n \leq 30 \\ 1 & n > 30 \end{cases} \tag{17}$$

Table 4 Accuracy comparison of RCSA and RCSA*

Method	Dim	Dataset			
		Oxford5k	Paris6k	Oxford105k	Paris106k
RCSA	128	70.3	81.8	66.1	74.8
RCSA*	128	**71.4**	**83.4**	**67.2**	**76.4**
RCSA	256	75.0	84.2	71.2	77.8
RCSA*	256	**76.1**	**85.4**	**72.5**	**78.4**
RCSA	512	77.5	85.2	74.8	79.1
RCSA*	512	**78.4**	**86.9**	**75.3**	**80.7**
RCSA	1024	79.8	86.3	77.2	80.4
RCSA*	1024	**80.5**	**87.9**	**77.5**	**81.7**
RCSA	2048	80.9	87.5	78.4	81.8
RCSA*	2048	**81.3**	**88.7**	**78.9**	**83.1**
RCSA+QE	512	81.9	87.6	79.6	82.0
RCSA* + QE	512	**82.2**	**89.6**	**79.9**	**84.3**
RCSA+QE	1024	83.9	88.8	81.8	83.7
RCSA* + QE	1024	**84.1**	**90.7**	**82.1**	**85.5**
RCSA+QE	2048	84.9	90.2	82.8	84.9
RCSA* + QE	2048	**85.4**	**91.1**	**83.6**	**86.3**

6 Conclusions

To get effective global representation in CBIR task based on deep convolutional features, this manuscript proposed and combined three strategies which formed a whole semantic-based aggregation (RCSA) method. This RCSA global representation outperformed the previous related works and achieved the state-of-the-art performance.

The first strategy is to select the RoI. Based on the fact that each channel of features is activated by special patterns according to semantic content, we proposed a simple but effective method to select RoI with several specific channels. This RoI selection method which is denoted by RSC showed excellent performance on instance retrieval tasks. Although the selected channels to predict bounding boxes in this paper specializes in the retrieval of buildings, the effectiveness of this method shows the possibility for other objects retrieval by choosing specific channels. The second is the channel weighting method. Comprehensive researches on the relation between semantic content and several parameters of each feature map, such as sum value, standard variance and non-zero responses, were conducted in this paper, according to which results, a channel weighting method (CW) for aggregated features was proposed. The implementation of CW on several aggregation methods significantly improved the retrieval performance, which indicates the availability of the proposed CW. The last is that we successfully improved an aggregation method (SBA). Compared with the original SBA, the improved method (CSBA) gets rid of the dependence on the datasets, which makes it simpler and more effective.

Our future research will pay attention to a more general method for selecting the RoI. Besides, since the channels working as semantic detectors in CSBA play a significant role in the performance of CSBA, we will attempt to get a more efficient but simple way to obtain the detector channels.

Acknowledgements This research was funded by National Natural Science Foundation of China Grant 61603289, China Postdoctoral Science Foundation Grant 2016 M602823, and Fundamental Research Funds for the Central Universities xjj2017118.

References

1. Azizpour H, Razavian AS, Sullivan J, Maki A, Carlsson S (2015) From generic to specific deep representations for visual recognition. In: Computer vision and pattern recognition workshops. pp 36–45
2. Babenko A, Lempitsky V (2015) Aggregating deep convolutional features for image retrieval. Computer Science
3. Babenko A, Slesarev A, Chigorin A, Lempitsky V (2014) Neural Codes for Image Retrieval 8689:584–599
4. Cao X, Wang P, Meng C, Bai X, Gong G, Liu M, Qi J (2018) Region based CNN for foreign object debris detection on airfield pavement. Sensors 18(3):737
5. Chen Z, Kuang Z, Wong KYK, Zhang W (2017) Aggregated deep feature from activation clusters for particular object retrieval. In: Thematic workshops of ACM multimedia. pp 44–51
6. Chu WT, Wu YL (2018) Image style classification based on learnt deep correlation features. IEEE Trans Multimed (99):1–1
7. Chum O, Philbin J, Sivic J, Isard M, Zisserman A (2007) Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval. 1–8
8. Do TT, Hoang T, Tan DKL, Cheung NM (2018) From Selective Deep Convolutional Features to Compact Binary Representations for Image Retrieval

9. Gao L, Guo Z, Zhang H, Xu X, Shen HT (2017) Video captioning with attention-based LSTM and semantic consistency. IEEE Trans Multimed 19(9):2045–2055
10. Gong Y, Wang L, Guo R, Lazebnik S (2014) Multi-scale Orderless Pooling of Deep Convolutional Activation Features. 8695:392–407
11. Gordo A, Almazán J, Revaud J, Larlus D (2016) Deep image retrieval: learning global representations for image search. In: European conference on computer vision. pp 241–257
12. Gordo A, Almazán J, Revaud J, Larlus D (2016) End-to-end learning of deep visual representations for image retrieval. Int J Comput Vis:1–18
13. He L, Xu X, Lu H, Yang Y, Shen F, Shen HT (2017) Unsupervised cross-modal retrieval through adversarial learning. IEEE Int Conf Multimed Expo: 1153–1158
14. Jégou H, Chum O (2012) Negative evidences and co-occurences in image retrieval: the benefit of PCA and whitening. Eur Conf Comput Vision: 774–787
15. Jégou H, Zisserman A (2014) Triangulation Embedding and Democratic aggregation for image search. In: Computer vision and pattern recognition. pp 3310–3317
16. Jegou H, Douze M, Schmid C (2009) On the burstiness of visual elements. Computer Vision Pattern Recogn 2009. CVPR 2009. IEEE Conf: 1169–1176
17. Jian X, Chunheng W, Chengzuo Q, Cunzhao S, Baihua X (2018) Unsupervised Semantic-based Aggregation of Deep Convolutional Features. arXiv:10
18. Kalantidis Y, Mellina C, Osindero S (2016) Cross-dimensional weighting for aggregated deep convolutional features. In: European conference on computer vision. 685–701
19. Kim DS, Arsalan M, Park KR (2018) Convolutional neural network-based shadow detection in images using visible light camera sensor. Sensors 18 (4)
20. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Int Conf Neural Inform Process Syst: 1097–1105
21. Lowe DG (2004) Distinctive image features from scale-invariant Keypoints. Int J Comput Vis 60(2):91–110
22. Lu H, Li Y, Chen M, Kim H, Serikawa S (2017) Brain intelligence: go beyond artificial intelligence. Mobile Netw Appl 23(2):368–375
23. Lu H, Li Y, Mu S, Wang D, Kim H, Serikawa S (2017) Motor anomaly detection for unmanned aerial vehicles using reinforcement learning. IEEE Int Things J PP (99):1–1
24. Lu H, Li Y, Uemura T, Ge Z, Xu X, He L, Serikawa S, Kim H (2017) FDCNet: filtering deep convolutional network for marine organism classification. Multimed Tools Appl (2):1–14
25. Lu H, Li B, Zhu J, Li Y, Li Y, Xu X, He L, Li X, Li J, Serikawa S (2017) Wound intensity correction and segmentation with convolutional neural networks. Concurr Comput Pract Exper 29 (6)
26. Lu H, Li Y, Uemura T, Kim H, Serikawa S (2018) Low illumination underwater light field images reconstruction using deep convolutional neural networks. Futur Gener Comput Syst 82
27. Mao XJ, Shen C, Yang YB (2016) Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections
28. Pang S, Ma J, Zhu J, Xue J, Tian Q Improving object retrieval quality by integration of similarity propagation and query expansion. IEEE Trans Multimed (99):1–1
29. Philbin J, Chum O, Isard M, Sivic J, Zisserman A (2007) Object retrieval with large vocabularies and fast spatial matching. In: Computer vision and pattern recognition, 2007. CVPR 2007. IEEE conference on. pp 1–8
30. Philbin J, Chum O, Isard M, Sivic J (2008) Lost in quantization: improving particular object retrieval in large scale image databases. In: Computer vision and pattern recognition, 2008. CVPR 2008. IEEE conference on. pp 1–8
31. Radenović F, Tolias G, Chum O CNN (2016) Image retrieval learns from BoW: unsupervised fine-tuning with hard examples. In: European conference on computer vision. 3–20
32. Ran L, Zhang Y, Wei W, Zhang Q (2017) A hyperspectral image classification framework with spatial pixel pair features. Sensors 17(10):2421
33. Razavian AS, Azizpour H, Sullivan J, Carlsson S CNN (2014) Features off-the-shelf: an astounding baseline for recognition. In: IEEE conference on computer vision and pattern recognition workshops. 512–519
34. Razavian AS, Sullivan J, Maki A, Carlsson S (2014) A baseline for visual instance retrieval with deep convolutional networks. Computer Science
35. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE transactions on Pattern Analysis & Machine. Intelligence 39(6):1137–1149
36. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
37. Scott BL, Hardesty LH (2018) Method and apparatus for speech recognition. J Acoust Soc Am 109(3):864
38. Serikawa S, Lu H (2014) Underwater image dehazing using joint trilateral filter. Pergamon press, Inc
39. Shelhamer E, Long J, Darrell T (2017) Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 39(4):640–651

40. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. Computer Science
41. Tolias G, Sicre R, Jégou H (2015) Particular object retrieval with integral max-pooling of CNN activations. Computer Science
42. Tollari S, Detyniecki M, Marsala C, Fakeri-Tabrizi A, Amini MR, Gallinari P (2009) Exploiting visual concepts to improve text-based image retrieval. Eur Conf Ir Res Adv Inform Retriev: 701–705
43. Tuan H, Thanh-Toan D, Dang-Khoa Le T, Ngai-Man C (2017) Selective deep convolutional features for image retrieval arXiv:9 pp.-9 pp
44. Wang L, Xu X, Dong H, Gui R, Pu F (2018) Multi-pixel simultaneous classification of PolSAR image using convolutional neural networks. Sensors 18(3):769
45. Wang X, Pang S, Zhu J, Wang J, Wang L (2018) An efficient aggregation method of convolutional features for image retrieval. In: International symposium on artificial intelligence and robotics, Nanjing, China
46. Wang J, Zhu J, Pang S, Li Z, Li Y, Qian X (2018) Adaptive Co-weighting Deep Convolutional Features For Object Retrieval
47. Wei XS, Luo JH, Wu J, Zhou ZH (2016) Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Trans Image Proc 99:1–1
48. Xiu-Shen W, Jian-Hao L, Jianxin W (2016) Selective convolutional descriptor aggregation for fine-grained image retrieval. arXiv:16 pp.-16 pp.
49. Xu X, He L, Shimada A, Taniguchi RI, Lu H (2016) Learning unified binary codes for cross-modal retrieval via latent semantic hashing. Neurocomputing 213:191–203
50. Xu X, Shen F, Yang Y, Shen HT, Li X (2017) Learning discriminative binary codes for large-scale cross-modal retrieval. IEEE Trans Image Process (99):1–1
51. Xu J, Shi C, Qi C, Wang C, Xiao B (2017) Unsupervised Part-based Weighting Aggregation of Deep Convolutional Features for Image Retrieval
52. Xu X, He L, Lu H, Gao L, Ji Y (2018) Deep adversarial metric learning for cross-modal retrieval. World Wide web-internet & web Inf Syst:1–16
53. Yandex AB, Lempitsky V (2016) Aggregating local deep features for image retrieval. In: IEEE international conference on computer vision. 1269–1277
54. Yang J, She D, Sun M, Cheng MM, Rosin P, Wang L (2018) Visual sentiment prediction based on automatic discovery of affective regions. IEEE Trans Multimed (99):1–1
55. Zeiler MD, Fergus R (2013) Visualizing and Understanding Convolutional Networks 8689:818–833
56. Zhang X, Xiong H, Zhou W, Lin W, Tian Q (2016) Picking deep filter responses for fine-grained image recognition. In: Computer vision and pattern recognition, –1142
57. Zhang Y, Wei XS, Wu J, Cai J, Lu J, Nguyen VA, Do MN (2016) Weakly supervised fine-grained categorization with part-based image representation. IEEE Trans Image Process 25(4):1713–1725

Xinsheng Wang received his B.S. degree from Northwest University, China, in 2015. He is currently a doctoral candidate of Xi'an Jiaotong University. His research interests include computer vision and machine learning.

Shanmin Pang is currently an assistant professor in the School of Software Engineering at Xi'an Jiaotong University. His research interests include pattern recognition, computer vision and image processing. He won the Best Application Paper Award at the ACCV 2012 conference.

Jihua Zhu received his Ph.D. degree in pattern recognition and intelligence system from Xi'an Jiaotong University, China, in 2011. He is an Associate Professor in the School of Software Engineering at Xi'an Jiaotong University. His research interests include mobile robot and computer vision.

Jiaxing Wang received the B.E. degree from Central South University, Changsha, China, in 2016. He is currently a graduate student in the School of Software Engineering at Xi'an Jiaotong University, Xi'an, China. His research interests include machine learning, multimedia information retrieval and computer vison.

Lin Wang received his Ph.D. degree in Control Science and Engineering from Xi'an Jiaotong University, China, in 2012. He is an Associate Professor in the School of Information Science and Technology at Northwest University, China. His research interests include point cloud processing, and swarm intelligence optimization.