Extending Feature Selection Strategies in VGG16: Convolutional Feature Aggregation for Content-Based Image Retrieval

1st Saha Kuljit Shantanu
Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
shantanu.cse@iubat.edu

1st Alina Zaman
Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
zmalina98@gmail.com

1st Md. Monirul Islam
Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka, Bangladesh
mmislam@cse.buet.ac.bd

***Abstract*—Content-Based Image Retrieval (CBIR) has achieved significant advances through deep learning, with VGG16 serving as a widely adopted architecture for robust feature extraction. Conventional feature aggregation strategies, such as AdCoW+I, typically rely on a single dominant feature channel, potentially overlooking complementary semantic information from other channels. This paper proposes an adaptive multi-channel fusion framework that selects multiple high-ranking channels based on a dataset-dependent threshold (γ), determined through an iterative process to optimize representation. The selected channels are fused using weighted averaging, and spatial as well as channel-wise weighting schemes are applied to enhance discriminative capability while reducing visual burstiness. Experiments on the Oxford5k and Paris6k datasets demonstrate consistent improvements over AdCoW+I, with Mean Average Precision (MAP) increasing from 39.7% to 42.04% on Oxford5k and from 43.78% to 46.12% on Paris6k. Grayscale feature was also evaluated, yielding dataset-dependent effects—beneficial for Oxford5k but detrimental for Paris6k. The results highlight the potential of adaptive multi-channel fusion to improve CBIR accuracy while maintaining computational efficiency.**

***Index Terms*—Content-Based Image Retrieval, VGG16, CNN, Feature Aggregation, Multi-channel Fusion**

I. Introduction

Content-Based Image Retrieval (CBIR) retrieves images from large datasets using visual features rather than textual metadata [1]. By exploiting color, texture, shape, and spatial cues [2], CBIR is useful when annotations are missing or unreliable. Applications include landmark recognition, augmented reality, surveillance, e-commerce, medical imaging and so on.

Deep learning, especially Convolutional Neural Networks (CNNs) [3], has enhanced CBIR by automatically extracting high-level semantic features. Architectures like AlexNet, ResNet, and VGG16 [4], [5] provide robust visual representations, with VGG16 popular for its simplicity and discriminative power. Removing fully connected layers allows it to serve as an effective feature extractor [6], [7].

Feature aggregation condenses convolutional maps into compact descriptors. Early methods like Bag-of-Words [8] and later approaches addressing visual burstiness [9] laid the groundwork. Recent techniques such as R-MAC and AdCoW+I [10], [11] improve weighting but may discard complementary channel information.

We propose an *adaptive multi-channel fusion strategy* for VGG16-based CBIR. Instead of a single dominant channel, our method selects a dataset-dependent proportion of top channels (controlled by γ) and fuses them via weighted averaging. Spatial and channel-wise weighting enhance discriminative cues and reduce burstiness [7], [9], while monochromatic features help suppress background noise.

Our contributions are:

- **Adaptive multi-channel fusion:** leverages multiple top-ranking channels.
- **Iterative γ-based selection:** adapts to dataset characteristics.
- **Monochromatic feature analysis:** reveals dataset-dependent effects.

The rest of the paper is organized as follows. In Section II we review the literature. Then we propose our methodology in Section III. We report our setup and findings in Section IV, followed by a brief conclusion in Section V.

II. Related Work

In CBIR, features are firstly extracted using deep CNN models. Then several methods are applied to enhance feature-based improvement. Finally, the processed features of a query image are matched against a database of stored features, typically using similarity metrics to retrieve the images based on their rank.

A. Feature Extraction

The traditional feature extraction methods mainly focus on low-level descriptors, which generally deal with color, texture, shape or spatial layout [12].

Modern CBIR methods, on the other hand, use pre-trained CNNs as a powerful means of extracting high-level contextual and semantic information from images [13], [14].

B. Feature Aggregation

Traditional aggregation strategies typically pool features from a predefined Region of Interest (ROI). For example, SPoC [6] pools from the last convolutional layer of VGG16 [14], assuming the geometric center as the ROI—a choice often invalid for many images.

To improve robustness, advanced methods incorporate spatial and channel-wise weighting. CroW [7] applies both, but treats all features equally despite their varying contributions. Building on this, AdCoW [11] selects a single dominant feature to guide spatial weighting, while CWAH [15] further refines this with contrastive channel weighting before pooling.

While these approaches enhance feature discrimination, their reliance on a single dominant feature can be limiting. In contrast, we propose weighting multiple discriminative channels, allowing richer and more adaptive feature representation.

C. Image Query

The output of the Retrieval for the Query is a ranked arrangement of the database images based on a metric. The distance between two images can be measured using l_2-normalization as it ensures unit norm. Euclidean distance and cosine similarity rankings become equivalent, making dot product-based retrieval efficient [16].

III. Methodology

Let a tensor $X \in \mathcal{R}^{(W \times H \times K)}$ is extracted for an image from the last convolutional layer of VGG16. This tensor has K channels containing $W \times H$ pixel values. From the tensor, our task is to build a feature map adaptively and convert it into a spatial mask $I \in \mathcal{R}^{(W \times H)}$.

Conveniently, the adaptive Coweighting strategy with a spatial Mask ensures an adaptive Weighting yet somehow Discriminative Feature Selection [11]. Now that we decide to construct a Feature mask with a less discriminative feature selection process, the strategy can be coined as ExtAdCoW+I. We further explore the grayscaling of images before selecting features with a view to prioritizing texture over color.

The framework of our approach is presented in Fig. 1. We will discuss the methods in detail.

A. Selection of the Region of Interest

After a tensor is extracted from a pretrained CNN model, the selection of the Region of Interest is a crucial part of the feature aggregation to enhance the performance of the retrieval system.

We introduce a dataset-specific parameter γ that defines the percentage of images used for feature selection. Features are ranked by dominance (Fig. 2), and a weighted average is computed over the selected set. A feature k is retained only if its coverage f_k, added to the cumulative sum of previous features, remains within γN; otherwise, it is discarded. This yields an optimal subset of features.

$$n = \max \left\{ n \mid \sum_{l=1}^{n} f_{k_l} \leq \gamma N \right\} \tag{1}$$

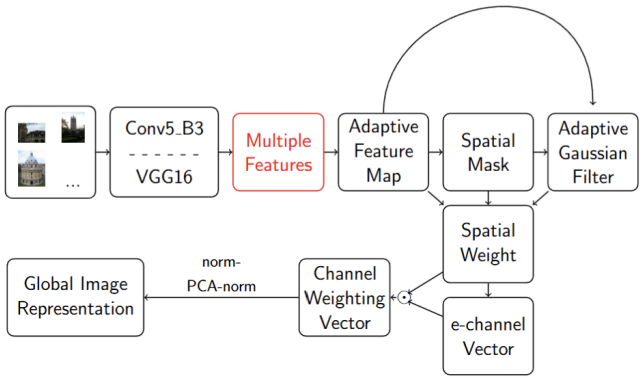


Fig. 1. The Framework of our Aggregation Method. Our contribution lies in Multiple Feature Selection to generate an adaptive Feature Map

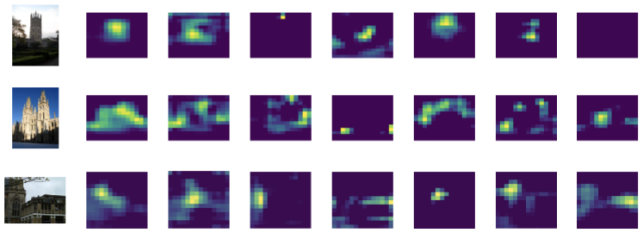


Fig. 2. The Top 7 Features by rank for three sample images selected from Oxford5K. The Sequence of channels is 360, 486, 359, 89, 220, 12, 413.

The number of features selected for aggregation obtained from Equation 1, n may vary based on different values assigned to γ as demonstrated in Fig. 3. The new Tensor is the weighed average of the selected top n features.

$$X_{(i,j)} = \frac{\sum_{l=1}^{n} X_{(i,j,k_l)} \cdot f_{k_l}}{\sum_{l=1}^{n} f_{k_l}} \tag{2}$$

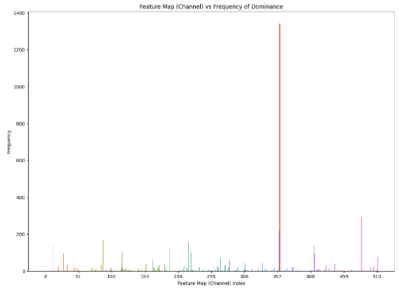
Experimentally, the value of γ for Oxford5K and Paris5K Dataset to perform the best on the proposed pipeline is 40% and 56% respectively.

After obtaining the representative tensor from Equation 2, the spatial feature mask is constructed as demonstrated in Fig. 4. The location where the activation value is greater than the median of the pixel values of the adaptive Feature Map in X represents the region of interest. Thus, we design $I' \in \mathcal{R}^{W \times H}$:

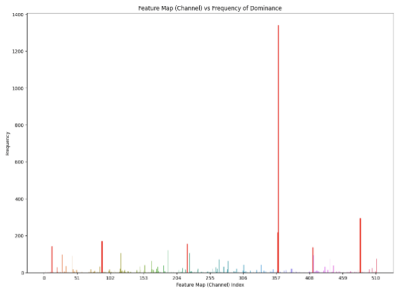
$$I'_{(i,j)} = \begin{cases} 0 & \text{if } X_{(i,j)} < t_m \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

Now from the mask I' obtained in Equation 3, we remove the noise, a floating area less than 4 pixels, to obtain mask I''. Finally we apply convex hull on I'' to obtain spatial mask I.

Fig. 5 shows the feature map when $\gamma = 40\%$ for two images in Oxford5K, then the corresponding values of I', I'' and I.



(a) AdCoW method



(b) $\gamma = 50\%$

Fig. 3. The features selected for Oxford5K dataset highlighted in red while (a) Following AdCoW Method (b) Selecting $\gamma = 50\%$

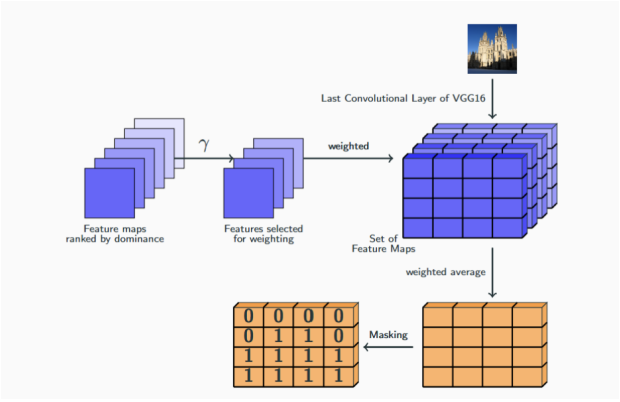


Fig. 4. Multiple features are selected Adaptively and are given priority based on the number of images they dominate in the dataset. Then a representative feature map is computed by weighted average of the features selected. The spatial mask is computed from the representative feature map.

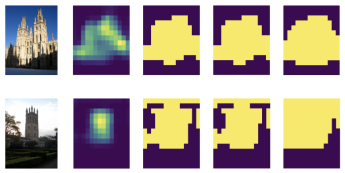


Fig. 5. The pipeline to obtain Spatial Matrix I. The five columns consecutively represent the original image, the adaptive feature map, I', I'' and I for the images when $\gamma = 40\%$

B. Adaptive Gaussian Filter

Applying an adaptive Gaussian Filter can further distinguish Region of Interest [11]. An Adaptive Gaussian filter introduces

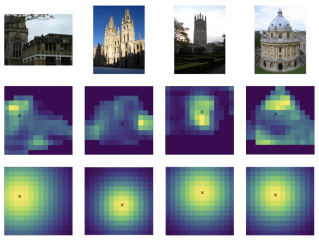


Fig. 6. The pipeline to obtain Adaptive Gaussian Filter. The three rows consecutively represent the original image, the adaptive feature map marked with the center of Gaussian Distribution, and the Weighting Vector S for the images when $\alpha = 10\%$

a new parameter α and experimentally from [11] $\alpha = 10\%$ yields to have the best retrieval performance on both the datasets. α is defined to be a parameter that ranks pixels on The matrix of aggregated response $S' \in \mathcal{R}^{(W \times H)}$ as obtained from Equation 4. From the ranked pixels, the centroid of the Gaussian Distribution (i_0, j_0) and the standard deviation of the Gaussian Distribution, σ is determined [11].

$$S'_{(i,j)} = \sum_{k=1}^{K} X_{(i,j,k)} I_{(i,j)} \qquad \forall i = 1, 2, ..., W, \;\; j = 1, 2, ..., H \tag{4}$$

$$S_{(i,j)} = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(i-i_0)^2 + (j-j_0)^2}{2\sigma^2}\right) \tag{5}$$

The weighting function, S is determined by Equation 5, and assigns higher values to the region of interest as demonstrated in Fig. 6. The spatial weighting vector is computed in the following manner.

$$\Omega_k = \sum_{i=1}^{W} \sum_{j=1}^{H} X_{(i,j,k)} I_{(i,j)} S_{(i,j)} \qquad \forall k = 1, 2, ..., K \tag{6}$$

C. Channel Weighting

Generally, different channels activate different semantic features of an image. To fully utilize the available information, an e-channel vector is first introduced in [11]. For each channel k, the Spatial Weighting vector scales itself in equation 6, then the enhanced visual burstiness [9] is balanced in Equation 7

$$b_k = \left(\frac{\Omega_K}{W \times H}\right)^p \tag{7}$$

A new scaling parameter p is proposed . By the Experiment of Zhu et al. [11], the ideal value is $p = 2$.

$$B_k = log\left(\frac{\varepsilon k + \sum_{c=1}^{K} b_c}{\varepsilon + b_k}\right) \tag{8}$$

ε is a negligible value added for numerical stability.

The final approach is to learn a PCA whitening matrix from the pipeline. The process is demonstrated in Fig. 7.

We present the entire aggregation strategy we are going to discuss in Algorithm 1. The symbol $\odot$ represents element-wise product between two vectors.

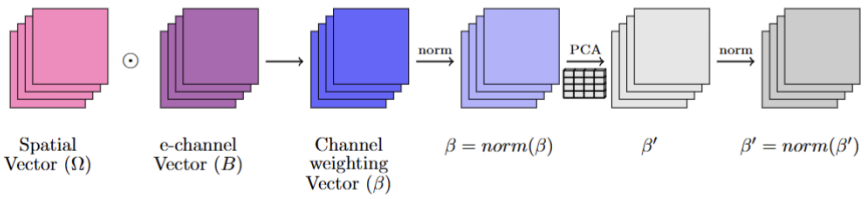


Fig. 7. The Framework to obtain Channel Weighting Vector β and applying PCA whitening to learn the rotation Matrix.

Algorithm 1 Deep Feature Aggregation Framework

Require: Tensor X, PCA whitening matrix $W \in \mathcal{R}^{K \times K'}$
Ensure: K'-dimensional global representation $\beta \in \mathcal{R}^{K'}$

1: Compute the representative feature map by Eq. 2
2: Compute spatial mask matrix I with the representative feature map
3: $X_{(i,j,k)} = X_{(i,j,k)} I_{(i,j)}$ //Spatial Mask
4: Compute adaptive Gaussian filter S by Eq. 5
5: $\Omega_k = \sum_{i=1}^{W} \sum_{j=1}^{H} X_{(i,j,k)} S_{(i,j)}$ //Spatial Weighting
6: Compute channel weighting vector B by Eq. 8
7: Perform norm-PCA-norm as per Fig. 7

IV. Experimental Results

A. Datasets and Evaluation Protocol

The proposed method is evaluated on two benchmark datasets widely adopted in image retrieval, Oxford5K [17] (5063 building images with 55 queries including 11 landmarks) and Paris6K [18] (6433 building images with 55 queries including 11 landmarks). We also evaluate the experiment on grayscale Oxford5K and grayscale Paris6k. For convenience, the grayscale versions of Oxford5K and Paris6K are dubbed as Oxford5K_GS and Paris6K_GS respectively.

The retrieval performance is measured by mean Average Precision (mAP), the mean of the average precision scores across multiple queries Q.

$$MAP = \frac{1}{Q} \sum_{q=1}^{Q} AP_q$$

Average Precision (AP) measures how well relevant items are ranked for a single query. It is the average of precision values at each relevant retrieved item. Our method of computing AP is 11-point-interpolation that evaluates retrieval performance by computing precision at 11 standard recall levels: $r \in \{0.0, 0.1, 0.2, \ldots, 1.0\}$

B. CNN model

As the title suggests, we have used VGG16 [14] as it is the most widely used CNN model for retrieval and classification tasks. As there is a really negligible gap between conv5_3 layer and pool5 layer of CNN, we can work with either of those. For our work, we have selected conv5_3 for feature extraction. We have studied the mAP for the AdCoW+I method [11] for three different dimensions on the Oxford5K documents, as demonstrated in Table. I do this for this too

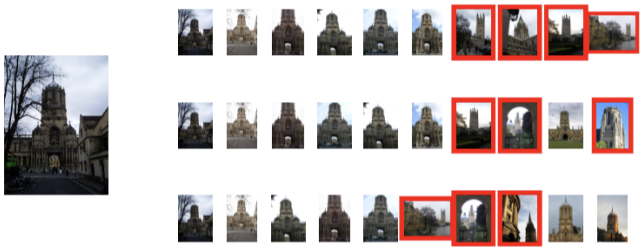


Fig. 8. Top 10 results of AdCoW+I, ExtAdCoW+I and ExtAdCoW+I after grayscaling for a randomly selected query of Oxford5K, using a 100-dimensional global vector to represent each image. The query image is displayed on the leftmost place. The false results are marked with red borders.

TABLE I
MAP OF ADCOW+I WHEN VGG16 FEATURES REDUCED TO DIFFERENT DIMENSION

No. of reduced features	mAP(%)	Remarks
256	37.92	Not good performance
128	40.22	Better, still space complex
100	39.70	Better and Space efficient

TABLE II
PERFORMANCE OF EXTENDED SELECTION WITH DIFFERENT VALUES OF γ ON OXFORD5K

γ	No. of Features selected	mAP(%)
AdCoW + I [11]	1	39.70
40	**4**	**42.04**
50	7	41.56
70	21	41.39

TABLE III
SELECTED VALUES OF γ ACROSS THE AFOREMENTIONED DATASETS OXFORD5K, OXFORD5K_GS, PARIS6K AND PARIS6K_GS

Dataset	γ	No. of Features selected
Oxford5K	40	4
Oxford5K_GS	44	5
Paris6K	56	10
Paris6k_GS	56	11

TABLE IV
SELECTED VALUES OF γ ACROSS THE AFOREMENTIONED DATASETS OXFORD5K, OXFORD5K_GS, PARIS6K AND PARIS6K_GS

Method	Dim	Paris6K	Paris6K_GS	Oxford5K	Oxford5K_GS
AdCoW+I [11]	100	43.78	41.36	39.70	51.57
ExtAdCoW+I	100	46.12	43.18	42.04	53.81

C. Impact of the Parameters

The proposed Algorithm ExtAdCow + I introduces a new parameter γ. The parameter γ is dataset-specific and regulates the number of features to be prioritized for Aggregation.

The parameter γ varies over datasets and aims to select the best number of features to be prioritized for feature aggregation upon co-weighing them. Table II demonstrates the retrieval performance on Oxford5K respective to different values of γ. Table III shows the best γ values chosen for datasets Oxford5K, Oxford5K_GS, Paris5K and Paris6K_GS.

D. Comparison with the state-of-the-art

To demonstrate the extent of our proposed method ExtAdCoW+ I, we compare the retrieval performance of our method with that of AdCoW+I [11] method across the four datasets OXford5K, Oxford5K_GS, Paris6K and Paris6K_GS. Table IV demonstrates how the extended feature selection outperforms the state-of-the-art method across the mentioned datasets. The fig. 8 demonstrates the first 10 relevant images for a selected Query experimented with the state-of-the-art method and proposed method to justify our comparison.

V. CONCLUSION

Experiments show that the proposed adaptive multi-channel fusion consistently outperforms single-channel AdCoW+I on Oxford5k and Paris6k datasets. By leveraging multiple semantically relevant channels, it captures broader discriminative cues, yielding higher Mean Average Precision (MAP).

The choice of γ strongly affects performance, varying across datasets, highlighting the framework's adaptability and the potential need for automated γ selection. Results also indicate that dataset characteristics, such as irrelevant image ratio and reliance on structural versus color features, influence retrieval.

Overall, adaptive multi-channel fusion enhances deep feature aggregation for CBIR, improving accuracy with minimal computational overhead and offering flexibility for large-scale or domain-specific retrieval tasks.

ACKNOWLEDGMENT

We acknowledge Bangladesh University of Engineering Technology for allowing us to conduct this experiment as our undergraduate thesis.

REFERENCES

[1] G.-H. Liu, J.-Y. Yang, and Z. Li, "Content-based image retrieval using computational visual attention model," *Pattern Recognition*, vol. 48, no. 8, pp. 2554–2566, 2015.

[2] E. Saykol, U. Gudukbay, and O. Ulusoy, "A histogram-based approach for object-based query-by-shape-and-color in image and video databases," *Image and Vision Computing*, vol. 23, no. 13, pp. 1170–1180, 2005.

[3] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 512–519, 2014.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems (NeurIPS)*, 2012, pp. 1097–1105.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[6] B. Babenko and M. Berezhnoy, "Aggregating deep convolutional features for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1791–1796, 2015.

[7] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 685–701.

[8] G. Csurka, C. Dance, L. Fan *et al.*, "Visual categorization with bags of keypoints," in *European Conference on Computer Vision*, vol. 1, 2004, pp. 1–22.

[9] H. Jegou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1169–1176.

[10] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

[11] J. Zhu, J. Wang, S. Pang, W. Guan, Z. Li, Y. Li, and X. Qian, "Co-weighting semantic convolutional features for object retrieval," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 368–380, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320319301798

[12] A. Smeulders, M. Worring, S. Santini *et al.*, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[13] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 157–166.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[15] F. Lu and G.-H. Liu, "Image retrieval using contrastive weight aggregation histograms," *Digital Signal Processing*, vol. 123, p. 103457, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051200422000744

[16] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3304–3311.

[17] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[18] ——, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.