Content-based Image Retrieval and the Semantic Gap in the Deep Learning Era

Björn Barz and Joachim Denzler

Computer Vision Group, Friedrich Schiller University Jena, Jena, Germany
{bjoern.barz,joachim.denzler}@uni-jena.de

Abstract. Content-based image retrieval has seen astonishing progress over the past decade, especially for the task of retrieving images of the same object that is depicted in the query image. This scenario is called instance or object retrieval and requires matching fine-grained visual patterns between images. Semantics, however, do not play a crucial role. This brings rise to the question: Do the recent advances in instance retrieval transfer to more generic image retrieval scenarios?
To answer this question, we first provide a brief overview of the most relevant milestones of instance retrieval. We then apply them to a semantic image retrieval task and find that they perform inferior to much less sophisticated and more generic methods in a setting that requires image understanding. Following this, we review existing approaches to closing this so-called semantic gap by integrating prior world knowledge. We conclude that the key problem for the further advancement of semantic image retrieval lies in the lack of a standardized task definition and an appropriate benchmark dataset.

Keywords: Content-based Image Retrieval · Instance Retrieval · Object Retrieval · Semantic Image Retrieval · Semantic Gap.

1 Introduction

One sees well only with the heart. The essential is invisible to the eyes.

This famous quote from the French writer Antoine de Saint Exupéry applies to life as well as to computer vision. The human perception of images greatly exceeds the visual surface of pixels, colors, and objects. The *meaning* of an image cannot simply be described by enumerating all objects contained therein and defining their spatial layout. We as humans are able to grasp a plethora of diverse and complex information contained in an image at first glance, such as events happening in the depicted scene, activities performed by persons, the relationships between them, the atmosphere and mood of the image, and emotions transported by it. Many of these concepts elude textual description and are best illustrated by providing an example image.

The example in Fig. 1 illustrates this variety of information conveyed by images. The image depicted there can be described from several perspectives:

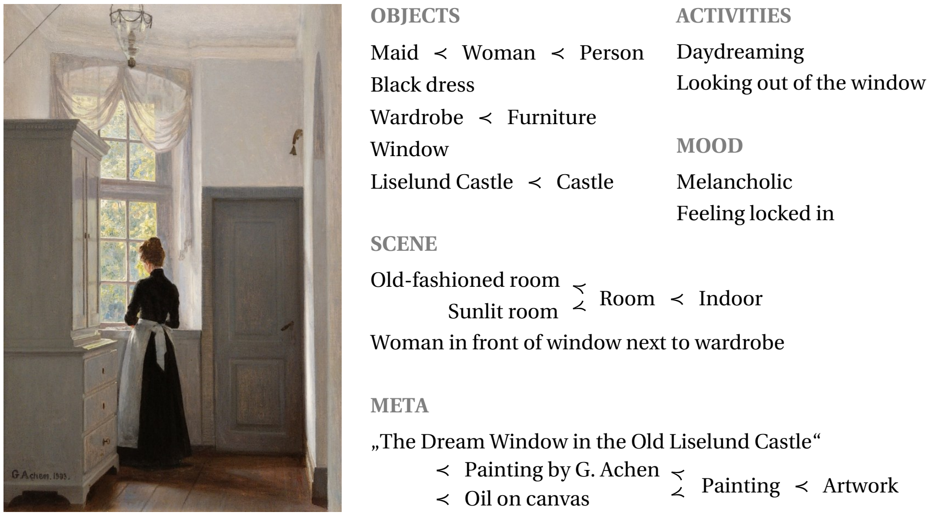

Fig. 1. An example for the ambiguity and semantic richness of images. All concepts listed on the right-hand side could be used to describe the image on the left, while different observers will pay attention to different subsets of these aspects. Moreover, some concepts can be organized hierarchically, indicated by the "≺" sign, which designates the hyponomy ("is-a") relationship.

its semantic content, artistic style, the emotions it evokes in the observer, or meta-information about the image itself. Depending on their background and the situational context, different observers will perceive and interpret this image differently. Searching for images on the web by means of textual descriptions or keywords is hence destined to fail, because most images are not exhaustively described in their surrounding text, for mainly two reasons: First, it is often difficult, if not impossible, to enumerate all aspects of an image explicitly, due to the potentially infinite amount of possible interpretations. Secondly, it is not necessary to do so, since most facets of an image are directly available to the viewer by simply looking at it. The textual description therefore focuses most often on the meta-information that is not encoded in the image itself, such as its author. The image shown in Fig. 1, for example, would probably be described as a photographic reproduction of the painting "The Dream Window in the Old Liselund Castle" by Georg Achen. This would prevent this image from being found by users searching for images of a woman looking out of a window, images showing the activity "daydreaming", or images with a melancholic atmosphere.

Searching through a large database of images not with textual keywords but using a representative example as query is hence the most natural, direct, and expressive way of finding images with a particular content, which might be complex and difficult to define. This approach is known as *content-based image retrieval (CBIR)* [49] and has been an active area of research since 1992 [31,36].

"Pictures have to be seen and searched as pictures", wrote Smeulders et al. [49] in their extensive survey at the end of the "early years" of CBIR in 2000. During the two decades that have passed since then, the field of content-based image retrieval has undergone at least two major revolutions (more on that in Section 2). However, most of the main challenges and directions had already been identified back then. One of these challenges is the *semantic gap*, as Smeulder et al. call it:

> *"The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation."* [49, sec. 2.4]

Phrased with the words of de Saint Exupéry, the semantic gap is the difference between perceiving an image with the *eyes*—objectively, as a depiction of objects, shapes, textures—and perceiving an image with the *heart*—subjectively, including world-knowledge and emotions, reading "between the pixels".

The size of the semantic gap depends on the level of abstraction of the search objective pursued by the user. Smeulders et al. [49] define this level of abstraction on a continuous scale between the two poles of a *narrow* and a *broad domain*. This terminology is best explained on the basis of the three currently most relevant CBIR tasks, depicted in Fig. 2:

Duplicate retrieval searches for images with exactly the same content. These are variants that originated from the same photo but might have been post-processed differently with regard to cropping, scaling, adjustments to color, brightness, contrast etc.

Instance retrieval searches for images that depict the same instance of an object, i.e., a person or a certain building. Thanks to its nature as a well-defined but non-trivial task with a clear ground-truth, this is the most extensively studied CBIR sub-task [48,38,29,30,25,4,3,50,20,42,46,9]. A handful of established datasets are available for this task [28,39,40,43] and significant progress has been made during the past few years, which we will outline in Section 2.

Semantic retrieval covers most of the remaining spectrum broader than instance retrieval and aims for finding images belonging to the same category as the query. It is important to note that *category* does not necessarily mean *object class* in this context. In practice, the set of possible categories is limited by nothing but the imagination of the user and a single image usually belongs to a remarkably high number of categories at once (see Fig. 1). Thus, the exact search objective of the user can rarely be determined based on the query image alone and will almost certainly also vary between users, even for the same query. Therefore, approaches to this problem often comprise interaction with the user to adapt the similarity measure used by the system to that in the user's mind [55,12,15,5,7].

Learning meaningful image representations that capture fine semantic distinctions and the various facets of an image's meaning is hence of paramount importance. Despite its practical relevance, this CBIR sub-task has received

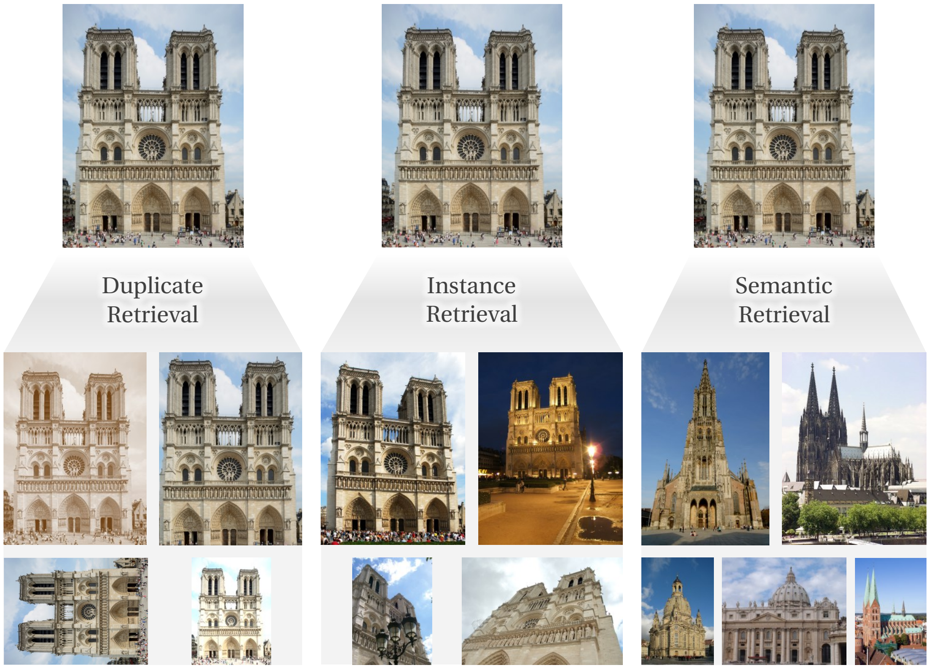

Fig. 2. Examples for three different sets of images to be retrieved given the same query depending on the type of the CBIR task.

> substantially less attention than instance retrieval, mainly due to the less well-defined notion of "relevance" and "similarity" and, as a result, the lack of a suitable benchmark. In this work, we will review recent approaches to semantic image retrieval (see Section 4) and assess the current state of the semantic gap, twenty years after the end of the "early years" of CBIR.

Duplicate retrieval marks one end of the spectrum, as it is the narrowest domain possible. In this case, the semantic gap is almost non-existent and all that is needed to overcome it is a list of invariances regarding the image's content (e.g., rotation, cropping etc.). The broader the domain, the larger the semantic gap.

While it is more challenging than duplicate retrieval, instance retrieval can still be handled by matching fine-grained distinctive visual patterns and their geometric layout. Content-based image retrieval has made substantial progress in this area in the past two decades, which we outline in Section 2. However, the applicability of such techniques is limited with respect to the much more generic broad domain of semantic retrieval, as we see in Section 3. One way to overcome this semantic gap, according to Smeulders et al. [49], lies in integrating sources of semantic information from outside the image. In Section 4, we review recent approaches in this direction, followed by a discussion of what is still missing for advancing CBIR in the broad domain further (Section 5).

2 The Evolution of Instance Retrieval

Between 2000 and 2020, CBIR—with a particular focus on instance retrieval—has undergone two major paradigm shifts: The first began in 2003 [48] and was initiated by the adaptation and subsequent improvement of techniques from text retrieval. The second wave of breakthrough achievements originated from the application of deep learning methods to CBIR, starting in 2014 [4,45]. We outline the major milestones of these two epochs of innovation in the following.

2.1 Hand-Crafted Features and Visual Words

Local Features as Visual Words In 2003, Sivic and Zisserman [48] sought to find occurrences of a certain object in videos and, to this end, adapted the *bag-of-words (BoW)* document descriptor, which is popular in the field of text retrieval, to image retrieval. As an analogy for words, they use local image features at distinctive keypoints and quantize them into a vocabulary of "visual words" using the k-Means clustering algorithm. Analogously to text retrieval, the occurrences of visual words per image are counted and the counts aggregated into a tf-idf vector representing the entire image. Since the Euclidean distance is not meaningful in high-dimensional spaces, the cosine similarity is then used to assess the similarity of two such image representations.

This process illustrates the general framework for extracting image representations that has been used in CBIR from that point on until today [30]: A local feature extractor computes features at keypoints in a given image. These local features are then embedded into a different space, such as quantized indices of visual words. Finally, they are aggregated into a global representation.

The global representation allows for efficient retrieval of an initial list of candidate images. In addition, the local features are often used to perform a spatial verification and re-ranking step for the top-ranking candidates to eliminate false matches [48,39]. This technique is quite specific to instance retrieval and matches local feature vectors between the query and a retrieved image to verify that the local features have a matching geometric layout.

Towards More Complex Embeddings Subsequent works of this epoch focused mainly on improving the embedding and aggregation step, while using the same local feature extractor over the course of a decade. The Hessian-affine detector [34] is typically used for finding keypoints at which local features should be extracted. This detector finds point of interest that are invariant to affine transformations as well as robust to limited changes of illumination and viewpoint. These keypoints are then described using SIFT [33] or RootSIFT [1] features. The latter is a simple transformation of SIFT, which consists in L^1-normalizing the SIFT vector and taking the element-wise square root. In the resulting space, the Euclidean distance between RootSIFT vectors corresponds to a histogram matching kernel in SIFT space.

In the case of Sivic and Zisserman [48], the embedding transforms each local feature vector into a space of one-hot vocabulary index vectors with tf-idf weights.

The aggregation then simply consists in a sum operation. However, representing local feature vectors by a single integer (the cluster index), incurs a severe loss of information and does not capture the actual distribution of the local features well. Hard assignment to a single cluster is furthermore not robust against small variations of local descriptors close to cluster boundaries. To overcome these issues, Perronnin et al. [38] propose the use of Fisher vectors for CBIR. The training data is quantized into visual words by fitting a Gaussian mixture model. Each local feature vector is then transformed into the gradient of its log-likelihood with respect to the means of the Gaussians. This realizes a weighted soft assignment to clusters and results in a dense, more informative, but also high-dimensional descriptor. In fact, the authors show that a Fisher vector with a single visual word achieves comparable performance to a BoW descriptor with 4,000 words.

A simplification with comparable and sometimes even superior performance are *vectors of locally aggregated descriptors (VLADs)*, proposed by Jégou et al. [29]. VLAD still uses hard-assignment of local descriptors to the nearest cluster, but captures the element-wise residuals of all local features from the center of their cluster. That means, the embedding feature vector is partitioned into k segments, where k is the number of clusters. The segment corresponding to the closest cluster center equals the difference between the local descriptor and that center and all other segments are 0. The dimensionality of the embedding space is hence the number of clusters times the local feature dimensionality. The aggregation consists in taking the sum over all transformed local feature vectors, L^2-normalizing the result, and applying PCA with whitening to reduce the high dimensionality of the global descriptor to something more manageable (usually in the order of a few hundred dimensions).

VLAD is, by definition, sensitive to the distance between a local feature vector and its cluster center. However, the Euclidean distance is of limited meaning in high-dimensional spaces. In a follow-up work, Jégou and Zisserman [30] account for this fact by L^2-normalizing the residuals, thus encoding their angle instead of their magnitude, which gives rise to the name *triangulation embedding*. Because distance is not meaningful, hard assignments to single clusters are not reasonable either. Triangulation embedding hence encodes the angles between the local feature vector and *all* visual words. This representation is subsequently whitened and has been found to outperform fisher vectors and VLAD.

However, Husain and Bober [25] find that comparing each local feature vector with all visual words does not scale to large datasets. Soft cluster assignment, on the other hand, often behaves unstable and degrades to single assignment in practice. To overcome this, they propose a middle ground by assigning the local descriptors to the few cluster centers that are closest and base the weights on their ranks among the nearest neighbors instead of their actual distances. These *robust visual descriptors (RVDs)* are furthermore not whitened globally but on a per-cluster level. The authors found that RVD performs competitively to triangulation embedding, while being faster to compute and more robust to dimensionality reduction.

The Role of Datasets While the paradigm of using aggregated local features for CBIR dates back to 2003 [48], research in this area has been most active between 2010 and 2016. One likely reason for this delay is the lack of suitable and established benchmark datasets. In the years 2007 and 2008, the Oxford Buildings [39], Paris Buildings [40], and INRIA Holidays [28] datasets were published, which quickly emerged as the standard benchmarks for instance retrieval and gave new impetus to the field by providing a proper ground for evaluation and comparison of methods.

The two building datasets comprise different photos of various landmark buildings in Oxford and Paris, with a large variety of perspectives, scales, and occlusions. The Holidays dataset, on the other hand, contains a collection of personal holiday photos with on average three different perspectives per scene. While these datasets are challenging, the task of retrieving images showing the same object or scene as the query is well-defined with a clear ground truth.

2.2 Off-the-shelf CNN Features

After hand-crafted local features had remained unquestioned in CBIR for over a decade, the renaissance of deep learning finally led to a substantial change regarding image representations. The independent works of Babenko et al. [4] and Razavian et al. [45] first showed that surprisingly good results can be achieved by simply extracting global image descriptors, so-called *neural codes*, from the first fully-connected layer of an off-the-shelf CNN pre-trained on ImageNet [14]. Given the extreme simplicity of this approach, requiring close to zero engineering effort compared to detecting keypoints, extracting local features, and aggregating them, this was a remarkable result. Just a year later, Babenko and Lempitsky [3] considerably improved the performance of this approach by extracting image features not from a fully-connected but from the last convolutional layer, which still has a spatial resolution. The result is, thus, a set of feature vectors, which can roughly be associated with different regions in the image. These are summed up for aggregation, L^2-normalized, reduced in dimensionality using PCA, and L^2-normalized again, leading to the speaking name *sum-pooled convolutional features (SPoC)* for these descriptors.

In the following years, research mainly adhered to using such pre-trained neural feature extractors and focused on designing sophisticated aggregation functions. Many of them try to find a middle ground between sum and maximum pooling, e.g., by averaging activations over the top few responses only as in *partial mean pooling (PMP)* [54], or by smoothly interpolating between the two extremes as in *generalized-mean pooling (GeM)* [42].

Aggregated convolutional features have one drawback, though: As opposed to traditional local features, they do not allow for precise localization of the matching object and, thus, are not compatible with techniques such as spatial verification and re-ranking, which depend on geometric information. To this end, Tolias et al. [50] propose the *regional maximum activation of convolutions (R-MAC)* aggregation, which follows a two-step approach: The convolutional feature map is divided into overlapping regions of different sizes and the local feature vectors in each region are aggregated using maximum pooling. These so-called MAC vectors are then whitened and aggregated by sum pooling into a global R-MAC image descriptor. For spatial re-ranking, the similarity of the query's MAC vector and the individual regional MAC vectors of the top few retrieval results can be used to localize the query object in the retrieved images and refine the ranking.

These techniques took CBIR based on features extracted from pre-trained CNNs quite far, but the hand-crafted RVD descriptor [25] is still able to compete with them on instance retrieval benchmarks.

2.3 End-to-end Learning for Image Retrieval

Deep learning finally became undeniably superior to traditional CBIR techniques based on hand-crafted features when researchers began to adapt the CNN used for feature extraction to the task of image retrieval instead of using a pre-trained one. We regard this shift of focus from feature transformation and aggregation to actual feature learning as the second important paradigm shift in CBIR.

Global Features Gordo et al. [20] were among the first to be successful in this endeavor and set the state of the art in instance retrieval for at least two years. They build upon R-MAC [50] and implement it as differentiable layers on top of a VGG16 CNN architecture, which can then be trained end-to-end. To this end, they employ the triplet loss [47], a training objective from the field of deep metric learning. By training on a curated dataset of famous landmarks, they learn a feature representation where images of the same landmark are closer together by a certain margin than two images of different landmarks, which supports the objective of instance retrieval.

This approach has later been extended by extracting R-MAC features from multiple layers of a CNN and weighting individual features of each region by the Kullback-Leibler divergence between the distributions of the Euclidean distance between matching and non-matching descriptors, so that more discriminative regional features obtain a higher weight [26]. The motivation for combining features from multiple layers lies in the different degrees of visual abstraction: features from earlier layers are more indicative of visual properties, while later layers provide a semantically more abstract representation.

As opposed to the triplet loss, Radenović et al. [42] find the contrastive loss to provide better final performance, while furthermore requiring only pairs instead of triplets of images for training. More importantly, they propose an unsupervised technique for generating training data consisting of matching and non-matching image pairs for instance retrieval without human annotation: Images in the training dataset are clustered based on their BoW representation using local RootSIFT features and spatial verification is applied to ensure that all images in a cluster show the same object. A 3-D model is then constructed for each cluster using structure-from-motion (SfM) techniques, so that it can be determined from these models whether two images depict the same object or not. This also allows images of the same landmark but captured from different and disjoint viewpoints to be considered as non-matching. The information about camera positions obtained from SfM furthermore enables mining of challenging positive image pairs that exhibit a non-trivial amount of overlap.

These metric learning approaches have led to an impressive improvement of instance retrieval performance in terms of average precision (AP), even though they do not optimize it directly but a proxy objective based on distances in the learned feature space. Since AP is the most important metric for evaluating retrieval methods, it seems desirable to optimize it directly instead of a proxy-task. However, that entails taking into account not only a single sample, a pair, or a triplet as before, but the entire list of ranked results. One apparent benefit is that such listwise objectives are position-sensitive: The impact of a single pair or triplet involving images at the top of the ranking should be higher than at the end of the list. However, average precision is not differentiable, because it involves sorting images by their similarity to the query. For being able to optimize AP in an end-to-end learning context nevertheless, He et al. [22] proposed a differentiable approximation of AP using histogram binning, which has been adopted by Revaud et al. [46] for CBIR and improved the state of the art. Since the cosine similarity, which is usually employed for retrieval, is bounded in $[-1, 1]$, the range of possible similarity scores can easily be divided into a fixed number of equally sized bins. Images are then soft-assigned to the bins whose centers are closest to the image's retrieval score to obtain histograms of positive and negative match counts in each bin. Instead of computing precision and recall for each possible position in the ranking, these metrics can now be computed for each bin and combined to approximate AP.

However, the quantization of similarity scores into bins ignores variations of the ranking within each bin, which can have particularly large impacts on AP at the top positions of the ranking. This deficiency has recently been overcome by a different approach to approximating AP: Instead of quantized sorting by binning, the sorting operation itself is relaxed by replacing the Heaviside step function indicating whether one element of the list precedes another with a sigmoid function to avoid vanishing gradients [41]. This allows for differentiable sorting and computation of a relaxed version of AP, called Smooth-AP [9].

With these listwise approaches, global representations for CBIR can finally be learned end-to-end without hand-crafted intermediate steps or proxy objectives.

Local Features While global image descriptors are convenient for retrieval applications, they are neither robust in the presence of occlusion or background clutter nor suitable for spatial verification, which is an important technique for instance retrieval. Other works hence aimed at learning local feature detectors and descriptors in an end-to-end manner.

Deep Local Features (DELF) [37], for example, uses coarse regional features extracted from a convolutional layer of a pre-trained CNN and then trains another small CNN to assess the importance of these densely sampled keypoints. For training, these predicted weights are used for weighted sum pooling of the local descriptors into a global feature vector, which allows for fine-tuning of the local features using image-level supervision.

Most instance retrieval systems using local features adopt a two-stage approach: First, a set of candidate images is retrieved by comparing global features and then re-ranked using spatial verification based on local features. Cao et al. [10] unified the learning of both types of features into a single model with two branches: One branch aggregates all feature vectors of the last convolutional layer of a CNN as global feature vectors and is trained with a metric learning loss. The other branch learns an attention module to identify distinctive local features and is trained using categorical cross-entropy.

The Need for More Challenging Benchmarks Besides plenty of computing capacity, deep learning techniques require one thing most of all: data. The existing instance retrieval datasets were too small for training deep neural networks, wherefore Babenko et al. [4] created a novel landmarks dataset with over 200,000 images for training purposes, which was later used by other works on deep image retrieval as well [20]. Nowadays, the large-scale Google-Landmarks dataset [37] proposed in 2017 is often used for training. It comprises over a million images of 12,894 landmarks from all over the world.

These datasets are orders of magnitudes larger than the Oxford and Paris Buildings dataset, but the latter were still relevant for evaluating and comparing novel methods. The rapid advances in deep learning for CBIR, however, quickly resulted in a saturation of performance on these benchmarks. Therefore, Radenović et al. [43] revisited these two datasets in 2018 by improving the ground-truth annotations, finding more difficult queries, adding challenging distractor images, and defining three different evaluation protocols of varying difficulty.

These developments demonstrate the importance of suitable training and benchmark datasets for the advancement of content-based image retrieval.

3 Impact on the Semantic Gap

The previous section outlined the impressive advances of instance retrieval in the deep learning era. However, instance retrieval is a rather narrow domain, where a broad understanding of the scene semantics are not required to solve the task satisfactorily. The interesting question is, therefore: Do these advances transfer to the broader domain of semantic retrieval?

To answer this question, we evaluate several seminal methods and models on an instance retrieval and a semantic retrieval task. For instance retrieval, we use the Revisited Oxford Buildings dataset [39,43] (see above), on which these methods have originally been evaluated. As an indicator for their performance in a broader domain, we evaluate them on the MIRFLICKR-25K dataset [24], which comprises 25,000 images from Flickr, each annotated with a subset of 25 concepts such as “sky”, “lake”, “sunset”, “woman”, “portrait” etc. While most images in the dataset are annotated with more than one concept, 3,054 of

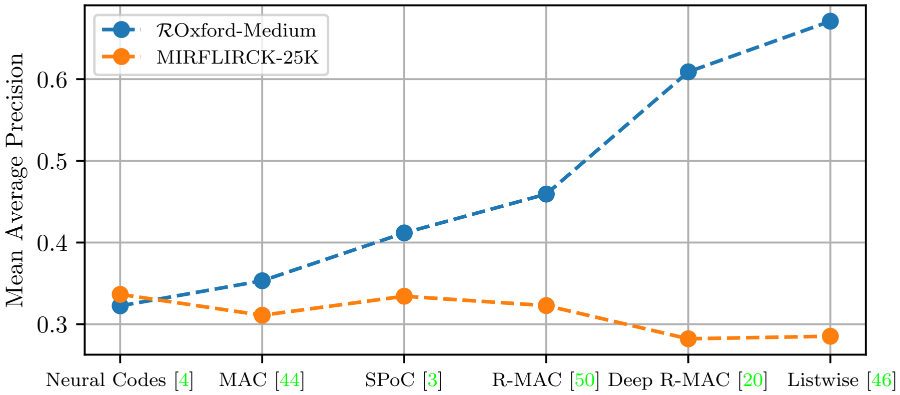

Fig. 3. Milestones of CNN-based instance retrieval, evaluated on an instance retrieval ($\mathcal{R}$Oxford [43]) and a semantic retrieval dataset (MIRFLICKR-25K [24]).

them exhibit only a single label. We use these images as queries to avoid query ambiguity. We consider a retrieved image as relevant if it shares this concept.

Figure 3 depicts the mean average precision of several milestones of CBIR research in the deep learning era on both tasks. While the performance on instance retrieval tasks increased steadily, the semantic retrieval performance did not only not improve, but even deteriorated slightly. The majority of developments in the past years have focused on instance retrieval and hence tuned feature representations towards this tasks, for which fine-grained visual features are important. This, however, degraded their performance on broader-domain tasks, for which a different set of features is necessary.

While instance retrieval has reached a very advanced level of maturity during the past 20 years, content-based image retrieval in general is still facing the challenges of the semantic gap.

4 Knowledge Integration for Semantic Image Retrieval

One way to overcome the semantic gap lies in incorporating additional sources of information outside the image, as Smeulders et al. [49] already stated back in 2000. In the following, we briefly review the most common sources of such external information as well as approaches for leveraging them to improve image representations for semantic image retrieval.

4.1 Class Labels

Image-level class labels are one of the most frequently available and cheapest types of semantic information about images. To provide robust performance in an open world, however, a huge number of classes or sophisticated methodology beyond training a simple classifier is required.

OASIS [11] combines both: Method-wise, OASIS learns a bilinear similarity metric using the triplet loss for comparing hand-crafted features with respect to semantic image similarity. The training dataset consisted of over two million images sourced from Google Image Search using about 150,000 textual queries entered by real users. Working at Google, the authors did not only have access to these queries, but also to relevance ratings based on click statistics, which allowed them to collect this large-scale but non-public dataset.

With the advent of deep learning, Yu et al. [53] exploit the intrinsic hierarchical representation generated by CNNs by combining features from shallow and deep layers. While the former capture rather visual patterns, features from deeper layers are expected to be more abstract and carry semantic information. Despite this, they only evaluate their approach on instance retrieval benchmarks.

More recent approaches optimize CNNs directly for multiple tasks to learn diverse representations. *MultiGrain* [8], for instance, aims for learning features that are useful for class-level, instance-level, and identity-level recognition by combining a classification and a metric learning objective. Evaluation, however, is conducted separately for each task in terms of classification accuracy on ImageNet [14] and retrieval accuracy on instance retrieval benchmarks. This evaluation protocol does not provide information about semantic retrieval performance.

To deploy CBIR at production-level within the *Microsoft Bing* search engine, Hu et al. [23] employ a large ensemble of different network architectures trained for various tasks: for classification with cross-entropy loss, with a metric learning objective such as the contrastive or triplet loss, for face recogmition, or for object detection. This ensemble is intended to capture a broad variety of both visual and semantic properties of images and, hence, cover most objectives a user of the visual search engine could pursue. The training data for this system is non-public and was collected by human annotators in an expensive data collection and annotation effort. The evaluation was conducted using human relevance judgments as well. For these two reasons, this work is neither publicly reproducible nor directly comparable with other works.

4.2 Class Taxonomies

Plain class labels do not take into account the semantic relationships between classes. Despite their visual similarity, images of humans and apes, for example, are generally considered to be semantically much less similar than images of a caterpillar and a butterfly, although the latter are not particularly similar from a visual perspective. Taxonomies such as WordNet [16] are a popular tool for measuring the semantic similarity between classes. They organize concepts on different levels of abstraction in terms of is-a relationships ("a poodle is a dog is an animal etc."). Several works strive for integrating this prior knowledge about the world to improve the semantic consistency of CBIR results.

Deng et al. [13] construct a hand-crafted bilinear similarity measure from the class taxonomy of ImageNet [14] and use it for comparing vectors of class probabilities predicted by a classifier. Instead of a similarity measure, Barz and Denzler [6] construct a semantic feature space spanned by class embeddings, where the cosine similarity between two class embeddings equals their semantic similarity derived from the taxonomy. They then use a CNN to map images into the same semantic space. Arponen and Bishop [2] do not constrain the feature space in this explicit way, but instead integrate the same objective directly into the loss function, so that the layout of the semantic feature space is learned. They combine this with an additional term encouraging the individual features to be binary, which allows for compact and memory-efficient descriptors.

The aforementioned works evaluate their approaches on ImageNet using "hierarchical precision" [13], which replaces the binary relevance of the retrieval results used by ordinary precision with the semantic similarity of their class and the class of the query. This metric suits the task better, but is best plotted for several cut-off positions in the ranking and cannot easily be summarized in a single number to facilitate comparison.

Yang et al. [52] combine semantic and visual similarity by first ranking images according to semantic similarity and then ordering the images within the same class according to visual similarity to the query. To this end, they use the contrastive loss with an adaptive margin proportional to the dissimilarity. The evaluation, however, is limited to fine-grained classification datasets and conducted using binary relevance, which does not take semantics into account.

Long et al. [32] not only embed the classes but all concepts in the taxonomy into a hyperbolic space, so that sub-classes lie in their parent class' entailment cone. As before, a CNN is then used to map samples onto their class embeddings. Although their method could also be applied for content-based image retrieval, they focus on video retrieval and evaluate their approach on that task only.

4.3 Textual Descriptions

While taxonomies provide information about the semantic similarity between classes, their full semantic meaning goes far beyond that. Several works have aimed for extracting such rich semantics from textual descriptions of classes or images and leverage them for learning meaningful image features. *DeViSE* [17] and *HUSE* [35], for example, learn word embeddings on Wikipedia and use the embedding of a class' name as its semantic embedding. DeViSE [17] then maps images into that space by maximizing the dot-product similarity between their feature vector and the respective class embedding, while enforcing a certain minimum distance to any other class embedding. HUSE [35], in contrast, adopts a pair-wise optimization approach by forcing the distance of pairs of images to be equal to the dissimilarity of their class embeddings. This approach provides more flexibility regarding the learned image feature space since it is separate from the space of word embeddings. Like some of the hierarchy-based approaches described above, both methods were evaluated using hierarchical precision. Thus, the semantic information used for evaluation was not the same as that used for training, which incurs a disadvantage compared to hierarchy-based methods.

Instead of using texts associated with classes, other methods leverage texts belonging to individual images, such as titles and captions, and learn a multimodal embedding space. Gomez et al. [19] do so by training a CNN to regress the text embeddings generated by a separately trained language model. However, they evaluate their approach only with textual queries and not in a *content-based* image retrieval scenario. Wu et al. [51], in contrast, learn text and image embeddings jointly and additionally predict individual embeddings for components of the caption such as objects, object-attribute pairs, and object-relation phrases. These semantic components are automatically aligned with the local features of the corresponding image regions using contrastive learning. However, their experiments only investigate the cross-modal image-to-caption and caption-to-image retrieval scenarios, while semantic CBIR performance is not analyzed.

4.4 Artistic Style

An entirely different dimension of image semantics is opened up by stylistic concepts such as artistic style, mood, and atmosphere. Learning image features that respect such properties requires either specialized annotations or prior knowledge about their characteristics.

Ha et al. [21] define style in terms of color composition, i.e., the distribution and layout of colors in an image. They construct a dataset with subjective 5-star similarity ratings for pairs of images, which have been collected in a laborious crowd-sourcing process involving active learning. A siamese network is trained to predict the distribution of similarity ratings for a given pair of images.

To avoid the expensive collection of large-scale style datasets, Gairola et al. [18] draw on knowledge from the field of visual style transfer, where Gram matrix features have been found to capture the stylistic properties of images. They extract these features from a pre-trained CNN, cluster them, and use the cluster labels as ground-truth for training another CNN using the triplet loss. They evaluate their approach on numerous datasets annotated with artistic styles, photographic styles, historical art styles, moods, or genres.

5 The Missing Ingredient

The two lines of research on instance retrieval and semantic retrieval portrayed in Sections 2 and 4, respectively, exhibit one apparent difference: The research on instance retrieval shows measurable continuous progress thanks to the Oxford [39] and Paris [40] benchmark datasets, whose release was followed by a clear surge of research activity in the field. With the Google-Landmarks dataset [37], sufficient training data is available for modern deep learning methods. The recent revision of the two aforementioned benchmark datasets [43] maintains their usefulness as a benchmark despite the substantial performance improvements.

Existing works on semantic image retrieval, in contrast, vary widely with respect to their evaluation protocol, training data (some of which is closed-source), and even the problem definition, rendering a clear comparison between approaches impossible. This is perhaps the biggest obstacle for further progress in this field and the likely reason why research still focuses on instance retrieval.

A thoroughly curated benchmark dataset for semantic image retrieval would hence greatly contribute to advancing the field. However, constructing such a benchmark is highly non-trivial due to numerous aspects. This begins already with the evaluation metric. In a semantic CBIR scenario, precision is often more important than recall, since most users are not interested in all potentially relevant images from a large-scale database. Average precision is hence a sub-optimal measure, but also precision alone is insufficient, since it only considers binary relevance. In reality, however, relevance is a graded phenomenon [49]. A candidate for an evaluation metric is the normalized discounted cumulative gain (NDCG) [27], which is capable of taking into account the degree of relevance between two images. The dataset, however, also needs to provide such graded relevance ratings for each pair of query and retrieved image. Ideally, the relevance should be based on real user ratings, which poses a major annotation effort.

Furthermore, the benchmark should define a diverse set of relevance criteria a user can have in mind when using a CBIR system, including instance identity, object category identity on different levels of abstraction, similarity regarding artistic style, mood, emotions, actions, and relationships portrayed in the image. Further complications are caused by the fact that a single query image can be interpreted differently with respect to each of these dimensions. The relevance of a retrieved image hence does not only depend on the query, but also on the search objective pursued by the user. This ambiguity can only be resolved by interaction with the user or by providing multiple query images sharing the relevant aspect. Therefore, the benchmark should ideally provide different evaluation protocols, an interactive one and a non-interactive one, which could be restricted to less ambiguous queries. The interactive scenario furthermore requires the definition of a feedback simulation protocol.

6 Conclusions

Content-based image retrieval has made astounding progress over the past two decades, especially in the area of instance retrieval, where a clearly defined objective and evaluation benchmarks exist. However, the methodological advances in this area do not translate to the more challenging task of semantic image retrieval. On the contrary, more advanced instance retrieval methods often perform worse than simpler ones in that domain. Despite the seeming advances, the semantic gap has rather become larger than smaller.

Due to the lack of an established benchmark, semantic image retrieval methods are often hardly comparable and vary widely regarding the task definition and the evaluation data and protocol. The history of instance retrieval shows that such a benchmark would be an invaluable catalyst for research on semantic image retrieval and a necessity for closing the semantic gap.

References

1. Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition

(CVPR). pp. 2911–2918 (June 2012). https://doi.org/10.1109/CVPR.2012.6248018

2. Arponen, H., Bishop, T.E.: SHREWD: Semantic hierarchy based relational embeddings for weakly-supervised deep hashing. In: ICLR 2019 Workshop on Learning from Limited Labeled Data (2019)
3. Babenko, A., Lempitsky, V.: Aggregating local deep features for image retrieval. In: IEEE International Conference on Computer Vision (ICCV). pp. 1269–1277 (Dec 2015). https://doi.org/10.1109/ICCV.2015.150
4. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) European Conference on Computer Vision (ECCV). pp. 584–599. Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_38
5. Barz, B., Denzler, J.: Automatic query image disambiguation for content-based image retrieval. In: International Conference on Computer Vision Theory and Applications (VISAPP). vol. 5, pp. 249–256. INSTICC, SciTePress (2018). https://doi.org/10.5220/0006593402490256
6. Barz, B., Denzler, J.: Hierarchy-based image embeddings for semantic image retrieval. In: IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 638–647 (2019). https://doi.org/10.1109/WACV.2019.00073
7. Barz, B., Käding, C., Denzler, J.: Information-theoretic active learning for content-based image retrieval. In: Brox, T., Bruhn, A., Fritz, M. (eds.) Pattern Recognition. GCPR 2018. Lecture Notes in Computer Science. vol. 11269, pp. 650–666. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-12939-2_45
8. Berman, M., Jégou, H., Vedaldi, A., Kokkinos, I., Douze, M.: MultiGrain: A unified image embedding for classes and instances. arXiv preprint arXiv:1902.05509 (2019)
9. Brown, A., Xie, W., Kalogeiton, V., Zisserman, A.: Smooth-AP: Smoothing the path towards large-scale image retrieval. In: European Conference on Computer Vision (ECCV). Springer Berlin Heidelberg, Berlin, Heidelberg (2020)
10. Cao, B., Araujo, A., Sim, J.: Unifying deep local and global features for image search. In: European Conference on Computer Vision (ECCV). Springer Berlin Heidelberg, Berlin, Heidelberg (2020)
11. Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large scale online learning of image similarity through ranking. Journal of Machine Learning Research (JMLR) **11**(36), 1109–1135 (2010)
12. Cox, I.J., Miller, M.L., Minka, T.P., Papathomas, T.V.: The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments. IEEE Transactions on Image Processing **9**(1), 20–37 (Jan 2000). https://doi.org/10.1109/83.817596
13. Deng, J., Berg, A.C., Fei-Fei, L.: Hierarchical semantic indexing for large scale image retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 785–792. IEEE (2011)
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255. IEEE (2009)
15. Deselaers, T., Paredes, R., Vidal, E., Ney, H.: Learning weighted distances for relevance feedback in image retrieval. In: International Conference on Pattern Recognition (ICPR). pp. 1–4. IEEE (2008). https://doi.org/10.1109/ICPR.2008.4761730
16. Fellbaum, C.: WordNet. Wiley Online Library (1998)
17. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: DeViSE: A deep visual-semantic embedding model. In: International Conference on

Neural Information Processing Systems (NIPS). pp. 2121–2129. NIPS'13, Curran Associates Inc., USA (2013)

18. Gairola, S., Shah, R., Narayanan, P.J.: Unsupervised image style embeddings for retrieval and recognition tasks. In: IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 3270–3278 (2020)
19. Gomez, R., Gomez, L., Gibert, J., Karatzas, D.: Learning to learn from web data through deep semantic embeddings. In: Leal-Taixé, L., Roth, S. (eds.) European Conference on Computer Vision (ECCV) Workshops. pp. 514–529. Springer International Publishing, Cham (2018)
20. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: End-to-end learning of deep visual representations for image retrieval. International Journal of Computer Vision (IJCV) **124**(2), 237–254 (Sep 2017). https://doi.org/10.1007/s11263-017-1016-8
21. Ha, M.L., Hosu, V., Blanz, V.: Color composition similarity and its application in fine-grained similarity. In: IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 2559–2568 (2020)
22. He, K., Lu, Y., Sclaroff, S.: Local descriptors optimized for average precision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 596–605 (June 2018). https://doi.org/10.1109/CVPR.2018.00069
23. Hu, H., Wang, Y., Yang, L., Komlev, P., Huang, L., Chen, X.S., Huang, J., Wu, Y., Merchant, M., Sacheti, A.: Web-scale responsive visual search at Bing. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). pp. 359–367. KDD '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3219819.3219843
24. Huiskes, M.J., Lew, M.S.: The MIR Flickr retrieval evaluation. In: ACM International Conference on Multimedia Information Retrieval. ACM, New York, NY, USA (2008), `http://press.liacs.nl/mirflickr/`
25. Husain, S.S., Bober, M.: Improving large-scale image retrieval through robust aggregation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **39**(9), 1783–1796 (Sep 2017). https://doi.org/10.1109/TPAMI.2016.2613873
26. Husain, S.S., Bober, M.: REMAP: Multi-layer entropy-guided pooling of dense CNN features for image retrieval. IEEE Transactions on Image Processing (2019). https://doi.org/10.1109/TIP.2019.2917234
27. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS) **20**(4), 422–446 (Oct 2002). https://doi.org/10.1145/582415.582418
28. Jégou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) European Conference on Computer Vision (ECCV). pp. 304–317. Springer Berlin Heidelberg, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88682-2_24
29. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3304–3311 (June 2010). https://doi.org/10.1109/CVPR.2010.5540039
30. Jégou, H., Zisserman, A.: Triangulation embedding and democratic aggregation for image search. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3310–3317 (June 2014). https://doi.org/10.1109/CVPR.2014.417
31. Kato, T., Kurita, T., Otsu, N., Hirata, K.: A sketch retrieval method for full color image database – query by visual example. In: IAPR Interna-

tional Conference on Pattern Recognition (ICPR). pp. 530–533 (Aug 1992). https://doi.org/10.1109/ICPR.1992.201616

32. Long, T., Mettes, P., Shen, H.T., Snoek, C.G.: Searching for actions on the hyperbole. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1141–1150 (2020)
33. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (Nov 2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94
34. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International Journal of Computer Vision (IJCV) **60**(1), 63–86 (Oct 2004). https://doi.org/10.1023/B:VISI.0000027790.02288.f2
35. Narayana, P., Pednekar, A., Krishnamoorthy, A., Sone, K., Basu, S.: HUSE: Hierarchical universal semantic embeddings. arXiv preprint arXiv:1911.05978 (2019)
36. Niblack, C.W., Barber, R., Equitz, W., Flickner, M.D., Glasman, E.H., Petkovic, D., Yanker, P., Faloutsos, C., Taubin, G.: QBIC project: querying images by content, using color, texture, and shape. In: Proc. SPIE, Storage and Retrieval for Image and Video Databases. vol. 1908, pp. 173–188. International Society for Optics and Photonics (1993). https://doi.org/10.1117/12.143648
37. Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B.: Large-scale image retrieval with attentive deep local features. In: IEEE International Conference on Computer Vision (ICCV). pp. 3476–3485 (2017)
38. Perronnin, F., Liu, Y., Sánchez, J., Poirier, H.: Large-scale image retrieval with compressed Fisher vectors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3384–3391 (June 2010). https://doi.org/10.1109/CVPR.2010.5540009
39. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–8 (June 2007). https://doi.org/10.1109/CVPR.2007.383172
40. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–8 (June 2008). https://doi.org/10.1109/CVPR.2008.4587635
41. Prillo, S., Eisenschlos, J.M.: SoftSort: A continuous relaxation for the argsort operator. In: International Conference on Machine Learning (ICML) (2020)
42. Radenović, F., Tolias, G., Chum, O.: Fine-tuning CNN image retrieval with no human annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2018). https://doi.org/10.1109/TPAMI.2018.2846566
43. Radenović, F., Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5706–5715 (June 2018). https://doi.org/10.1109/CVPR.2018.00598
44. Razavian, A.S., Sullivan, J., Carlsson, S., Maki, A.: Visual instance retrieval with deep convolutional networks. ITE Transactions on Media Technology and Applications **4**(3), 251–258 (2016). https://doi.org/10.3169/mta.4.251
45. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: An astounding baseline for recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-WS). pp. 512–519 (June 2014). https://doi.org/10.1109/CVPRW.2014.131

46. Revaud, J., Almazan, J., de Rezende, R.S., de Souza, C.R.d.: Learning with average precision: Training image retrieval with a listwise loss. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
47. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (June 2015). https://doi.org/10.1109/CVPR.2015.7298682
48. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: IEEE International Conference on Computer Vision (ICCV). vol. 2, pp. 1470–1477 (Oct 2003). https://doi.org/10.1109/ICCV.2003.1238663
49. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **22**, 1349–1380 (12 2000). https://doi.org/10.1109/34.895972
50. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of CNN activations. In: International Conference on Learning Representations (ICLR) (2016)
51. Wu, H., Mao, J., Zhang, Y., Jiang, Y., Li, L., Sun, W., Ma, W.Y.: Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6602–6611 (2019)
52. Yang, S., Yu, W., Zheng, Y., Yao, H., Mei, T.: Adaptive semantic-visual tree for hierarchical embeddings. In: ACM International Conference on Multimedia (ACMMM). pp. 2097–2105. MM '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3343031.3350995
53. Yu, W., Yang, K., Yao, H., Sun, X., Xu, P.: Exploiting the complementary strengths of multi-layer CNN features for image retrieval. Neurocomputing **237**, 235–241 (2017). https://doi.org/10.1016/j.neucom.2016.12.002
54. Zhi, T., Duan, L.Y., Wang, Y., Huang, T.: Two-stage pooling of deep convolutional features for image retrieval. In: IEEE International Conference on Image Processing (ICIP). pp. 2465–2469 (Sep 2016). https://doi.org/10.1109/ICIP.2016.7532802
55. Zhou, X.S., Huang, T.S.: Relevance feedback in image retrieval: A comprehensive review. Multimedia Systems **8**(6), 536–544 (Apr 2003). https://doi.org/10.1007/s00530-002-0070-3