https://doi.org/10.1016/j.inffus.2017.01.003

Information Fusion in Content Based Image Retrieval: A Comprehensive Overview

Luca Piras, Giorgio Giacinto

Department of Electrical and Electronic Engineering
University of Cagliari
09123 Piazza D'armi, Cagliari, Italy
luca.piras@diee.unica.it, giacinto@diee.unica.it

Abstract

An ever increasing part of communication between persons involve the use of pictures, thanks to the cheap availability of powerful cameras on smartphones, and the cheap availability of storage space. The rising popularity of social networking applications such as Facebook, Twitter, Instagram, and of instant messaging applications, such as WhatsApp, WeChat, is the clear evidence of this phenomenon, thanks to the opportunity of sharing in real-time a pictorial representation of the context each individual is living in. The media rapidly exploited this phenomenon, using the same channel, either to publish their reports, or to gather additional information on an event through the community of users. While the real-time use of images is managed through metadata associated with the image (i.e., the timestamp, the geolocation, tags, etc.), their retrieval from an archive might be far from trivial, as an image bears a rich semantic content that goes beyond the description provided by its metadata. It turns out that after more than 20 years of research on Content-Based Image Retrieval (CBIR), the giant increase in the number and variety of images available in digital format is challenging the research community. It is quite easy to see that any approach aiming at facing such challenges must rely on different image representations that need to be conveniently fused in order to adapt to the subjectivity of image semantics. This paper offers a journey through the main information fusion ingredients that a recipe for the design of a CBIR system

should include to meet the demanding needs of users.

Keywords: Information Fusion, Content Based Image Retrieval

1. Introduction

The availability of a large variety of personal devices, the prominent being the smartphone, that allows capturing pictures, videos, and audio clips, and uploading them on different social sharing services, fosters the steep rise of the volume of digital data stored in many different archives [1]. To accurately extract information related to some specific topic from this vast amount of data, retrieval tools and algorithms need to quickly discard irrelevant information, and focus on the items of interests by evaluating a variety of multiple diverse features. Whereas the performances of textual data search methods have reached a good level of maturity, the same can not be said for visual search or multimedia data, due to the richness of content and the subjectivity of its interpretation [2].

So far, the most common method for retrieving multimedia content from an archive consists of using meta-data associated to the images such as the timestamp, the geolocation, keywords, tags, labels or short descriptions, and perform the retrieval task through a text-based search. Manual cataloguing a large image archive, even though it requires expensive work, and a large amount of time, often turns out not to be so effective, due to the subjectivity of the task compared to the richness of its semantic content. In addition, it is necessary to take into account that the use of a limited number of words for image labeling and tagging does not always allow clearly and completely describing what an image represents, and is prone to confusion, as sometimes the same word can have several meanings in different contexts.

Moreover, the manual annotation of images typically provides a general description that could be loosely connected to its specific visual content, as it might contain the author, the main subject depicted, the place where the image was taken, etc., rather than the details, and, consequently, a textual query can produce a multiplicity of results with different semantic meanings.

To overcome these drawbacks, several methods based on the automatic analysis of the image content from a computer perspective have been proposed over the years. They are predicated on the approach of content-based indexing by leveraging on the use of low- and mid-level features such as color, texture, shape, etc. a very active research field since more than 20 years [3, 4, 5, 6]. Retrieval systems that are grounded on these approaches are called *content-based image retrieval* (CBIR) systems.

Of course, the description of images through such low- and mid-level features is not always directly related to the common perception that the user has of an image. For a human being, indeed, an image can be seen as the representation of different concepts, either related to physical characteristics such as shapes, colors, textures, etc., or related to emotions and memories. For a computer perspective, an image is simply a set of pixels with different "colors" and different intensities.

The early papers on CBIR are, by now, almost twenty-five years old [7], but while the results attained so far allowed achieving some relevant milestones [5], we are still facing issues for which an acceptable solution is far to be devised. One of the most relevant issues is related to the type of features used to handle the "content" of an image, that is usually represented through low-level features that represent colors, shapes, edges that can be found in an image with numerical values. This means that, when two images are compared to find similarities, actually the similarities with respect to the intrinsic features of the images, such as the presence of objects with a given shape, and/or the dominance of a given color, etc., are computed. It can be easily seen however that the effectiveness of the search in this case is limited to a small subsets of semantic concepts.

For example, when the similarity between an image of an orange and an image of a lemon has to be measured, you might not always be satisfied by the result [8, 9]. A retrieval system based on this level of description of an image content, may respond either with a very high or very low value of similarity. It is not so difficult to see that a shape-based retrieval system would evaluate the two images as being similar, while a retrieval system based on color does not.

These early findings, paved the way for exploring CBIR systems based on the fusion of different features, either by employing weighted similarity measures, where different features are weighted according to their relevance to the task at hand, or by combining different similarity functions, where image similarity is first computed separately for each feature, and then their values are combined.

In the past years, there have been many attempts to bridge the gap between the high level features, those perceived by human beings which identify the semantic information related to the images, and the low level ones that are used in the searches. This difference in perception is widely known in the CBIR field as the *semantic gap*. In order to capture such a subjectivity, image retrieval tools may employ *Relevance Feedback* [10, 11] mechanisms. Relevance Feedback (RF) techniques involve the user in the process of refining the search. In a CBIR task in which RF is employed, the user submits to the system a query image which is an example of the pictures of interest. The system then assigns to each image in the database a score according to a similarity measure between each image and the query. The top k best scored images are returned to the user that labels them as being relevant or not, so that the system can consider all relevant images as additional examples to better specify the query, and the non-relevant ones as examples of images that the user is not interested in. With the availability of this new additional information, the system can improve the quality of the search results, by providing a larger number of relevant images in the next iteration.

It can be easily seen that this iterative and interactive procedure can benefit from the availability of multiple image representations, as the retrieval system can exploit the different similarity concepts embedded in the available representations, and adapt the search towards the user's interests. Consequently, the fusion of multiple image representations for content-based image retrieval tasks has been mainly addressed within the relevance feedback paradigm, as it provides a way to estimate the relevance of each feature and similarity measure with respect to the task at hand. Moreover, the relevance feedback paradigm can be employed to enable browse-to-search mechanisms, where the user has not

a specific target in mind, and the feedback captures the most relevant features that drive the browsing experience towards the images of interest [12, 13].

In addition to the use of different sets of visual features, it can be easily seen that the effectiveness of a visual retrieval system can be improved by combining information from different modalities, i.e., from different types of content. For example, if we consider Web pages, they usually contain both images and text. Even if the relationship between the surrounding text and images varies greatly, with much of the text being redundant and/or unrelated to the visual content, a large amount of information about an image can be found in the textual context of the Web pages [14]. Several works indeed proved that such data can be effectively combined within the traditional CBIR systems to improve the quality of the retrieval results [15, 16, 17].

This paper will introduce the reader to the major approaches proposed in the literature for fusing information in visual retrieval tasks. Section 2 describes the basic concepts behind the techniques proposed in the content based image retrieval field. Section 3 summarizes the main categories in which fusion approaches can be classified, each fusion approach being extensively addressed in Sections 4, 5, 6, and 7. Conclusion and future research perspectives are drawn in Section 9.

2. Architecture of a CBIR system

The design of a content-based image retrieval system requires a clear planning of the goal of the system [5]. As much as the images in an archive are of different types, are obtained by different acquisition techniques, and exhibit different content, the search for specific concepts is definitely a hard task. It is easy to see that as much as the scope of the archive is limited, and the content to be searched is clearly defined, than the task can be more easily managed [3]. On the other hand, the design of a general purpose multimedia retrieval engine is a challenging task, as the system should be capable of adapting to different semantic contents, and different intents of the users.

A number of content-based retrieval systems tailored to specific applications, usually referred to as narrow domain systems, have been proposed to date. Some of them are related to sport events, as the aspect of the scene is fixed, camera positions are known in advance, and the movements of the players and other objects (e.g., a ball) can be modeled [5]. Other applications are related to medical analysis, as the type of images, and the objects to look for can be precisely defined [18].

The description of the content of a specific image can be provided in multiple ways. First of all, an image can be described in term of its properties provided in textual form (e.g., creator, content type, keywords, etc.). This is the model used by Digital Libraries, where standard descriptors are defined, and guidelines for defining appropriate values are proposed. However, apart from descriptor such as the file format, the size of the image, etc., other keywords are typically provided by domain experts. In the case of very narrow-domain systems, it is possible to agree on an ontology that helps describing standard scenarios. On the other hand, when multimedia content is shared on the web, different users may assign the same keyword to different contents, as well as assign different keywords to the same content [19]. Thus, more complex ontologies, and reasoning systems are required to correctly assess the similarity between images [20].

Low-level and medium-level content-based features [5, 9] have been proposed in analogy with the possible way in which the human brain assesses the similarity between visual contents. While at present this analogy is not deemed valid, these features may provide some hints about the concept represented by the pictorial content. Currently, very sophisticated low-level features are defined that take into account multiple image characteristics such as color, edge, texture, etc. [21]. Indeed, as soon as the domain of the archive is narrow, very specific features that are directly linked with the semantic content are computed [22]. On the other hand, in a broad domain archive, these features may prove to be misleading, as the basic assumptions do not hold [9].

A different way of thinking [23] has recently attracted researchers in the computer vision community that dramatically advanced the state of the art in

image classification tasks through the use of Deep Convolutional Neural Networks (CNN) [24]. This approach has a long history [25], its basic concept stemming from artificial neural network research, where many layers of interconnected information processing units are exploited for pattern classification, or feature learning tasks [26, 27]. The more interesting characteristic of the deep learning paradigm is that features need not to be extracted from the raw data beforehand, but the raw data themselves are processed by the network that produces an internal feature representation of the data suited for the task at hand. The seminal paper, where CNN have been used in an image classification task [24], clearly showed that features emerging in the upper layers of the CNN can also serve as good descriptors for image retrieval. Subsequent works [28, 29, 30, 31] further investigated this aspect, and showed that the features learned for a given task, can be reused for other classification or retrieval tasks. It is worth to note that Babenko et al. [32] recently showed that the features extracted from a CNN trained for a classification task have the same performances when the retrieval dataset is quite different from the training dataset. Nevertheless, this performance can be further improved when the CNN is retrained by using images that have a stronger relationship with the retrieval dataset. However, in the same paper, the authors also point out that in some contexts these features do not outperform other state-of-the-art features, such as Fisher vectors [33, 34] or Triangulation embedding [35].

The past decade has witnessed many scientific advances in the CBIR field, mostly in vertical application areas [18] such as medical image retrieval[1], cultural heritage preservation [36], and a large variety of new class of applications that use the camera phone to search information about objects that are in visual proximity to the user [37], e.g., for identifying products, comparison shopping and so on [38]. Nevertheless, the use of content based image search engine in large and non-specific database produces results that are far from being satisfactory. In [39], for example, it easy to see how the recognition rate of the state-of-

[1] http://ganymed.imib.rwth-aachen.de/irma/

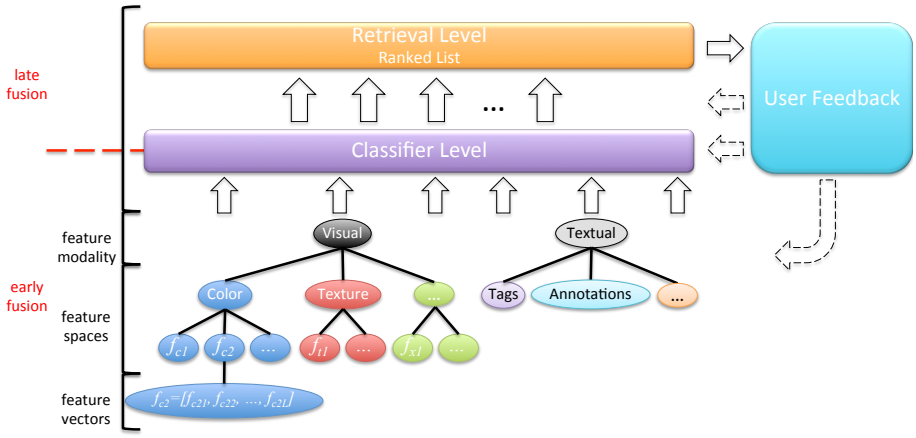

Figure 1: Schema of the structure of a CBIR system.

the-art scene classifiers employing state-of-the-art features decrease from 88.1% to 38.0% in two datasets with 15 and 397 categories respectively.

Summing up, the design of a content-based image retrieval system requires: i) the selection of a set of content-based features that are expected to capture the semantics of the images; ii) the selection of the technique used to extract and represent each feature; iii) the definition of suitable similarity measures in the feature representation of images that capture their semantic similarity; iv) the use of tailored fusion mechanisms to weigh the importance of each component in capturing the semantics.

3. Information Fusion in CBIR

Over the years, several techniques for combining and fusing different image descriptors have been proposed in the literature [40], each technique being tailored to the selected descriptors, the strategy of combination, and the application targeted. Fusion approaches are usually categorized into two classes, namely *early*, and *late* fusion approaches [41, 42], which refers to their relative position in the learning chain that goes from the feature extraction step up to the classification/retrieval step (See Figure 1).

Early fusion usually refers to the combination of the features into a single representation before the computation of similarity between images[41]. This kind of approaches is very common in the image retrieval field, and the simplest and well known solution is based on the concatenation of the feature vectors into a single vector such as in [43], where the authors propose two ways of integrating the SIFT [44] and LBP [45] descriptors, and the HOG [46] and LBP descriptors, respectively. Other approaches proposed for image retrieval are based on the early fusion of different feature spaces (See Section 5), such as color and shape [47], or texture and color [48].

Late fusion refers either to the combination of the outputs produced by different retrieval systems, or to the combination of the similarity rankings, the outputs and the rankings referring to different feature representations [49]. In an image retrieval task, the goal is to aggregate multiple ranked outputs to generate another ranked output. This kind of fusion can be implemented either according to a score-based approach, where the different similarities or distances from the query are combined, or by following a ranked-based paradigm that combines the different ranks obtained by the classifiers. The outputs to combine are usually weighted to give more importance to particular descriptors either using weights whose values are fixed *a priori*, or by learning them for a given image content [50].

In an image classification task, instead, late fusion usually involves a weighted voting strategy for the outputs of the classifiers associated to the individual descriptors [51, 52]. Over the years, more sophisticated strategies that take place at different levels of the learning chain, such as multiple kernels, have been proposed in the literature, so they are sometimes categorized as *intermediate* fusion strategies [53]. However, as kernels are employed to provide for a feature transformation where patterns from different classes can be linearly separated, these techniques can be regarded as a special case of *early* fusion strategies.

A quite different approach to use the available descriptors is to give each of them a different *role*, so that some of them are used to filter out a subset of images, while the rest of descriptors are used on the remaining subset of images, or

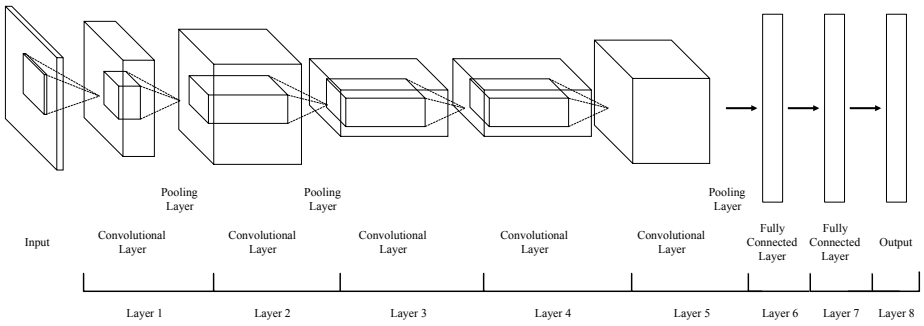

Figure 2: Diagram of a Convolutional Neural Networks (CNN) architecture as proposed in [24].

regions in the images. This kind of approaches could be categorized as *sequential fusion*. This paradigm can be used for combining different feature modalities [54], or simply different visual feature sets [55, 56]. In other approaches, global and local image descriptors are used sequentially, the first ones performing a coarse similarity search, the latter ones, to refine the search [57, 58]. In this light, *deep learning* approaches implicitly perform feature fusion. In fact, the CNN model consists of several convolutional layers and pooling layers stacked up with one on top of another, where the convolutional layer performs a weighted combination of the input values, while the pooling layer reduces the output of the convolutional layer (See Figure 2). At each level the *raw features* of the images are weighed and refined with respect to the previous level, to produce a better representation of the images. The hierarchical architecture of CNNs could be seen as a 'natural' way of combining more feature modalities taking advantage of the training algorithms embedded in its hidden layers [59, 60]. Unfortunately, despite the great attention paid by researchers in using deep learning approaches for image classification and recognition, there is still a limited amount of works that specifically focus on CBIR applications [61, 32], where the goal is not to retrieve the images of the most probable class(es), but to retrieve the most similar images. This perspective affects both the learning phase, and the output processing phase, as outlined by the few seminal works to date.

In the next sections, the different fusion strategies will be extensively addressed, and, for each strategy, several approaches will be discussed in order to provide an overview of the different techniques proposed in the past years for fusing multiple information in image retrieval tasks.

4. Feature weighting for early fusion

If the retrieval task is modelled as a classification task, where a pool of images described by a set of low-level features are assigned to a set of labels, so that images with the same labels are considered to be similar to each other, then a new feature space that captures the semantic similarity can be extracted by techniques such as PCA. However, to attain reliable results, the number of training images should be large, and labels should be reliable. This is hardly the case in image retrieval from large datasets. Conversely, instead of formulating the problem in terms of the transformation of the feature space to discover the hidden relationship between relevant images, an alternative solution that has been widely investigated consists in using feature selection strategies, or feature weighting approaches, that can be considered to perform a soft selection strategy. In fact, by following this paradigm, non-discriminative features will receive a weight close to zero. As above mentioned, the idea comes from the observation that the effectiveness of CBIR techniques strongly depends on the choice of the set of visual features.

However, no matter how suitably for the task at hand the features have been designed, the set of retrieved images often fits the users needs only partly. This is because, in general, the exact intent of the user's query cannot be fully captured even when multiple images are used for querying the archive. As a consequence, it is not possible to choose "a priori" the subset of features that is best suited to a user's query. The basic idea behind weighting mechanisms is that the exploitation of relevance feedback from the user implicitly defines which images should be considered similar to each other. For example, in a metric space, relevance feedback information can be exploited by modify the

similarity measure so that similar images are represented as neighbors of each other (i.e., relevant images), and non-relevant images do not fall within the neighborhood of relevant images.

More formally, feature weighting mechanisms can be formulated as follows. An image I is represented as $I = I(F)$, where F is a set of low level feature spaces f_i, such as color, texture, etc. Each feature space f_i can be modeled by several representations f_{ij}, e.g. color histogram, color moments, etc.. Each representation f_{ij} is itself a vector with multiple components

$$f_{ij} = [f_{ij1}, \ldots, f_{ijh}, \ldots, f_{ijk}, \ldots, f_{ijL}], \tag{1}$$

where L is the vector length. For each level f_i, f_{ij} and f_{ijk} it is possible to associate a set of weights denoted with w_i, w_{ij} and w_{ijk}, aimed at representing the effectiveness of each feature to the query at hand. For example, for a given feature representation f_{ij}, the similarity between the images can be computed by the "weighted" Minkowski metric [10]:

$$S(f_{ij}) = \left(\sum_{k=1}^{L} w_{ijk} |I_A(f_{ijk}) - I_B(f_{ijk})|^p \right)^{1/p} \tag{2}$$

with $p \geq 1$. The majority of papers that addressed the problem of weight estimation, followed a "probabilistic" approach. In [62] the authors proposed to estimate the weights using the inverse of the standard deviation of the values of a feature component computed over a "class" of relevant images. The rationale behind this proposal is that if a certain component of the feature vector takes similar values for all the relevant images, it means that the component is relevant to the query; on the contrary, if all relevant images have different values for that component, then the component is not relevant. In [63], a "local" measure of relevance has been proposed to estimate features' weights (*Probabilistic Feature Relevance Learning*, PFRL). The estimation followed a least-squares approach, that is, a certain feature is more relevant to the query if it contributes more to the reduction in terms of the prediction error. A different approach is used in [64], where the features with maximum *balanced information gain* obtained from

the entropy of the set of labelled images have been selected. In [8], the weights have been estimated with the goal of privileging those feature spaces where the set of relevant images form a compact cluster or, in terms of probabilities, by assigning more importance to features for which the relevant examples have a high likelihood, and less importance to features for which the non-relevant examples have a low likelihood. Finally, in [65] the authors proposed a dynamic feature weighting approach by exploiting intra-cluster and inter-cluster information for representing the descriptive and discriminative properties of the features according to the labels given by the user.

The same estimation procedure could be used to weigh not only the components of each individual feature space, but also to estimate the weights to be associated to subsets of components. The idea stems from the fact that the feature vectors can be decomposed into "sub-vectors", each sub-vector describing a different part of the image or a specific characteristic. Therefore, by assigning a greater or a lower weight to one of them, it is possible to better adapt the search to the concept the user is looking for.

In [66] a different point of view with respect to the usual probabilistic approach has been proposed. The weights associated to a given feature have estimated so that they reflect the capability of representing nearest-neighbor relationships according to the user's choices. This method is tailor-made for retrieval techniques based on the nearest neighbor paradigm, and the same algorithm can be used either to weight each component of one feature space, or to weigh different subsets of feature values within the same feature space, or to weight different feature spaces.

Another approach that exploits the feedback from the user to assign a larger importance to features related to similar images, and less importance to other features, has been proposed in [67]. The rationale behind this approach can be explained by observing that, if the variance of the images relevant to the query is large along a given axis of the feature space, any value on this axis could be acceptable by the user, i.e., the value of the corresponding feature is irrelevant with respect to the user's needs, and therefore this axis should

be given a low weight, and vice-versa. In that paper, the authors formulated the relevance feedback approach as a minimization problem whose solutions are the optimal query and a weight matrix used to define the distance metric between images. In [68], Rui et al. improved the algorithm described in [67] by proposing a hierarchical model in which each image is represented by a set of different features. In that work, in addition to estimating weights related to each feature representation (inter-feature weighting), the different components of each feature representation are also weighted (intra-feature weighting).

Recently, the relevance feedback paradigm has been also exploited to improve the retrieval capabilities of CNNs by modifying the weights of the convolutional layers according to the feedback of the user, by following an *early* fusion approach, where the internal layers of the networks are seen as implicit feature representations of the input images [69].

5. Representation by multi-feature spaces for late fusion

The previous section showed that the combination of multiple image representations (colors, shapes, textures, etc.) by early fusion approaches can effectively cope with the reduced inter-class variation that is experienced by resorting to just one type of features. As a drawback, the use of multiple image representations with a high number of components increases the computational cost of retrieval techniques. As a consequence, the response time of the system might become an issue for interactive applications (e.g., web searching). Over the years, the pattern recognition community proposed a number of solutions for fusing the information from different feature spaces through the combination of the output of different pattern classifiers [70]. The most popular and effective techniques for output combination are based on late fusion techniques, such as the mean rule, the maximum rule, the minimum rule, and weighted means.

In the field of content-based image retrieval, similar approaches can be employed by considering the value of similarity between images as the output of a classifier. In particular, combination approaches have been proposed for fus-

ing different feature representations, where the appropriate similarity metric is computed in each feature space, and then all the similarities are fused through a weighted sum [10].

In the same spirit, in [71], a large set of *highly selective visual features* has been used, where each feature was highly selective for a small percentage of images, and, at the same time, only a few features were selective for the set of relevant images. In this way, after the choice of the most selective features for a given query, each image in the archive can be evaluated very rapidly, by discarding all other features.

Artificial neural networks have been used in [72], where self-organizing maps (SOMs) are employed to measure the similarity between images. This approach aims at mapping the sequence of the queries based on the user's responses during the retrieval process. A separate SOM is trained for each feature vector type, then the system adapts to the user's preferences by returning more images from those SOMs where the responses of the user have been most densely mapped.

More recently, [73] proposed a different probabilistic strategy to combine similarity measures. The authors considered a subjective similarity judgement given by users on a fixed set of image and related it to a measure of similarity, then combined the different values of similarity evaluated in different feature spaces. The different feature representations can be combined by fusing all the similarity metrics through a weighted sum [10, 66]. The main issue for this kind of approach is to increase the performances in terms of *Precision*, *Recall*, and *Average Precision* [74], by limiting the increase of the processing time.

The issue of combining different feature representations is also relevant when relevance feedback mechanisms are used. In this case, at each iteration, similarities have to be computed by exploiting relevance feedback information, for example by resorting to Nearest-Neighbor or Support Vector Machine [75] techniques.

In this view, an approach that has been proposed in the pattern recognition field to classify patterns represented by a set of similarity measures is the so called "dissimilarity space". This approach is based on the creation of a new

feature space where patterns are represented in terms of their (dis)similarities to some reference prototypes. The dimension of this space does not depend on the dimensions of the low-level features employed, but it depends on the number of reference prototypes used to compute the dissimilarities, and on the number of dissimilarity measures employed. If we denote with $P = \{\mathbf{p}_1, \ldots, \mathbf{p}_P\}$ the set of prototypes, and the dissimilarity measure between an image $\mathbf{I}_i$ and one of the prototypes $\mathbf{p}_j$ as $d(\mathbf{I}_i, \mathbf{p}_j)$, then the image $\mathbf{I}_i$ can be represented in the dissimilarity space as follows:

$$\mathbf{I}_i^P = [d(\mathbf{I}_i, \mathbf{p}_1), \ldots, d(\mathbf{I}_i, \mathbf{p}_P)]. \quad (3)$$

This representation can be easily extended to take into account multiple dissimilarity measures by stacking the corresponding dissimilarity vectors.

This technique has been employed to exploit relevance feedback in content-based image retrieval field [76, 77], where the set of relevant images plays the role of reference prototypes. In addition, dissimilarity spaces have been also proposed for image retrieval to exploit information from different multi-modal characteristic [78].

In addition, [56] propose another use of the dissimilarity representation for improving the performances of relevance feedback approaches based on the Nearest-Neighbor approach [79]. Instead of computing (dis)similarities by using different prototypes (e.g., the relevant images) and a single feature space, the authors propose to compute similarities by using just one prototype, and multiple feature representations. Each image is thus represented by a very compact vector that summarizes different low-level characteristics, and allows images that are relevant to the user's goals to be represented as near points.

In the past years, the combination of multi-feature spaces for image retrieval tasks, has been proposed through the use of CNN [80] to learn both the metrics for each feature space, and the combination function for the different feature representations.

6. Fusing different relevance feedback approaches

The relevance feedback paradigm has been introduced to refine retrieval results, both to overcome inaccuracies in textual information, and to bridge the semantic gap between the low level image descriptors and the user semantics. The user is actively involved in the retrieval process as she is asked to label a set of retrieved images as being relevant or not [81] with respect to her interests. In general, the approaches proposed in the literature to exploit the RF paradigm can be divided into two groups. One group of techniques exploit relevance feedback by modifying some parameters of the search, either by computing a new query vector in the feature space [62], or by choosing a more suitable similarity measure, or by using a weighted distance [82, 66]. Another group of approaches are based on the formulation of RF in terms of a pattern classification task, by using popular learning algorithms such as SVMs [83], neural networks and self-organizing maps [75, 84, 72], and using the relevant and non-relevant image sets for training purposes.

One of the first techniques to be employed for RF in CBIR tasks, that is still used in a number of image retrieval applications, is based on the so-called query shifting paradigm [62]. This technique has been developed in the text retrieval field, and it is represented by the Rocchio formula [85]:

$$Q_{opt} = \frac{1}{N_R} \sum_{i \in D_R} D_i - \frac{1}{N_T - N_R} \sum_{i \in D_N} D_i \tag{4}$$

where D_R and D_N are the sets of relevant and non relevant images respectively, N_R is the number of images in D_R, N_T the number of the total documents, and D_i is the representation of an image in the feature space. This approach is motivated by the assumption that the query may lie in a region of the feature space that is in some way "far" from the images that are relevant to the user. On the contrary, according to Eq.(4), the optimal query should lie near to the euclidean center of the relevant images and 'far" from the non relevant images.

Relevance feedback has been also formulated in terms of a pattern classification task using neural networks, self-organizing maps (SOMs) [72], or approaches

based on SVM. The latter have been widely used to model the concepts behind the set of relevant images, and adjust the search accordingly [75, 84]. However, it is worth noting that in many practical CBIR settings it is usually difficult to produce a high-level generalization of a "class" of objects, as the number of available relevant and non-relevant samples cases may be too small, and the concept of "class" is variable, due to the subjectivity of the definition of similarity between images. This kind of problems has been partially mitigated thanks to the use of the *active learning* paradigm [86], where the system is trained not only with the most relevant images according to the user judgement, but also with the most informative images that allows driving the search into more promising regions of the feature space [87, 88].

For a given image database, and for different users, the best performances might be provided by different relevance feedback approaches. This behaviour can be easily seen if we model the set formed by each query image and the associated positive feedback samples as a "class" of a classification problem. For each "class" of query images, one relevance feedback technique might be better other RF approaches.

According to this behaviour, [89] proposed a combination of multiple relevance feedback strategies. The proposed combination integrates three relevance feedback techniques, namely *Query Vector Modification* [85], *Feature Relevance Estimation* [62, 63], and *Bayesian Inference* [90], and dynamically selects the most appropriate technique for a particular query, or even for a particular iteration, by evaluating the retrieval precision of each approach.

In [91], a different approach is followed. The authors proposed to employ the Support Vector Machine ensembles technique to construct a *group-based relevance feedback* algorithm, by assuming the data as coming from multiple positive classes and one negative class, i.e. the problem was modeled as a *(x+1)-class* classification problem. An SVM ensemble was also proposed in [92, 93] to address the unbalanced learning issue, whereas the authors of [94] suggested to use a set of *one-class* classifiers based on the *Information Bottleneck* framework [95].

Apart from the above mentioned papers, there have not been other significant investigations on the potentialities of combining different relevance feedback approaches. The vast majority of papers that propose the use of classifier ensembles for content based image retrieval tasks, are based on a single approach for relevance feedback, where different instances are created either by training on different "classes" of images, or on different bags of relevant/non relevant images in order to improve the performance of that particular approach.

7. Multimodal retrieval

While it has been more than 20 years since the first proposal of a system that allowed the user to combine the textual information contained in a HTML document along with the attached image, with the information in image metadata (i.e., its width, height, the file size, type, etc.), and with the number of faces in the image [96, 97], nevertheless the paradigm to combine multimodal feature has never ceased to arouse interest in researchers [98, 42, 99]. The roots of this approach lie in the fact that the performance of a content-based image retrieval (CBIR) system is inherently constrained by low-level features, and it cannot give satisfactory retrieval results when users' high-level concepts are not easily expressed by low-level features [100]. Keywords have been used to assist content-based image retrieval tasks according to two main approaches: their use as additional features, or their use to seed a text-based query [101].

The first approach combines keywords with low level features of the images in order to use a combined input space. Many works followed this path as [100], where the authors proposed an algorithm for learning the keyword similarity matrix during user interaction, namely word association via relevance feedback (WARF). They assume that the images in the database have textual annotations in terms of short phrases or keywords that can come from keywords spotting from surrounding HTML text on Web pages, manual annotation, and so forth. To combine the use of low-level visual features with keywords, they convert keyword annotations for each image into a vector, where each component is

related to the presence or to the probability of a certain keyword in a specific image.

Several other researchers have addressed this problem from different points of view. [102] proposed to combine textual and visual statistics in a single index vector, where textual statistics are captured in vector form using a latent semantic approach based on the text in the containing HTML document, while visual features are captured in vector form using color and orientation histograms. Barnard and Forsyth [103] proposed a method that organizes image databases using both image features and associated text by integrating the two types of information during model construction. The system learns the relationships between the image features and semantics by modelling the statistics of word and feature occurrence and co-occurrence. In [104], the authors proposed an approach based on associating a fuzzy membership function with the distribution of the features' distances, and assigning a degree of worthiness to each feature based on its relative average performance. The memberships and the feature weights are then aggregated to produce a confidence value that could be used to rank the retrieved images. The basic idea is to assign high membership values to distances that are relatively low and low membership values to relatively large distances. The membership functions is designed according to the distribution of the distances within each feature for a small set of training images. In particular, the features' memberships values and their relevance weights have been combined according to two distinct approaches: the first one is linear and is based on a simple weighted combination, the second one is non-linear and is based on the discrete Choquet integral [105].

Linear combination models have been widely used in multimedia information retrieval for combining textual and visual features, even if obtaining an effective system is not straightforward due to the difficulty to estimate the weights of the different modalities. In [106] the authors proposed an approach based on Fisher Linear Discriminant Analysis, aimed to learn the weights for multimedia documents composed of text and images. In particular, the authors reformulate the task of learning the combination parameters as a dimensionality reduction

problem in a binary classification context, i.e., to find the linear combination which best separate relevant and non-relevant documents.

More recently, it has been also proposed an alternative way to combine different modalities. In [54] the authors use the Balanced Iterative Reducing and Clustering (BIRCH) algorithm [107] on textual and visual descriptors to diversify the results obtained by the search. In order to combine textual and visual information together, they first build a clustering tree based on textual information, and then the resulting tree is refined by replacing the text features with the visual ones. In particular, for each node of the tree, its center and radius are recomputed based on the visual feature vectors instead of the former textual feature vectors.

In [108], a novel scheme for online multi-modal distance metric learning (OMDML) is investigated, which learns distance metrics from multi-modal data or multiple types of features via an online learning scheme. The key idea of OMDML is to learn to optimize a separate distance metric for each individual modality, and to find an optimal combination of diverse distance metrics on multiple modalities.

Another kind of approach is based on the use of keywords to seed a query, and then employs both keywords and visual features to conduct query refinement [109, 110]. In [111] the authors proposed a framework that performed relevance feedback both on keywords representing the images' semantic contents through a semantic network, and the low-level feature vectors such as color, texture, and shape. In [101] the authors proposed a multimodal learning approach that uses images' semantic labels to guide a concept-dependent, active-learning process. The system is based on the definition of complexity of a concept, and then it proposed making adjustments to the sampling strategy from which images are to be selected and labeled, to improve the capability of the concept learning. The idea behind this Concept Dependent Active Learning approach (CDAL) is to address the scarcity problem by using keywords to seed a query. According to this approach the user can use a keyword to describe the target concept and the images that are annotated with that keyword are added to the initial pool. If

the number of images with matching keywords is small, the system can perform query expansion using a thesaurus to obtain related words that have matching images in the dataset.

Some approaches propose to perform separate visual/textual queries in parallel, and then take the union/intersection of the two retrieved lists as in [112], where the results are combined using a weighted sum of the scores given by each retrieval system. A linear weighted combination is also used in [113], where a relevance feedback system that refines its results after each iteration, using late fusion methods is proposed. It allows also the user to dynamically tune the amount of textual and visual information to be used for retrieving similar images. Other systems perform the two queries sequentially and use one modality to filter the search space for the other modality as in [114]. In that paper, the authors proposed an asymmetric multimedia fusion strategy, which exploits the complementarity of the text and the visual features. The schema consists in a prefiltering textual step, which reduces the collection for the visual retrieval step.

In [115], the authors presented iLike, an image search engine that integrates both textual and visual features to improve the retrieval performance. The system claims to bridge the semantic gap by capturing the meaning of each text term in the visual feature space, and re-weight visual features according to their significance to the query terms. The system is able to infer the "visual meanings" behind the textual queries and provide a visual thesaurus, which is generated from the statistical similarity between the visual space representation of textual terms.

The paper by Ngiam et al. [59] can be considered one of the first attempts to learn and combine features over multiple modalities exploiting the deep learning paradigm. That paper presents a series of tasks for multimodal learning, and shows how to train deep networks that learn features to address each of the proposed tasks. For the sake of clarity, it is worth noting that the authors did not focus their proposal on the *multimodal* domain, properly speaking, but rather on the so-called *cross-modal* domain. In the multimodal fusion setting, data from

all modalities is available at feature learning, for training the system, and for each test pattern. In the cross-modality setting, data from multiple modalities is available only during feature learning, while, during the training and test phases, only data from a single modality is provided [59]. Accordingly, cross-modal retrieval refers to the search paradigm where information in one modality can be retrieved using the other available modalities, such as searching images using text and vice-versa [116]. In [60], the authors propose a Deep Boltzmann Machine (DBM) [117] model for learning multimodal data representations where the key idea is to learn a joint density model over the space of multimodal inputs. More recently Wang et al. [118] proposed a 5-layer neural network to learn a joint model for semantically correlating multiple features from different modalities, where deep learning feature as image representation, and topic feature as text representation are fused.

8. Discussion

The above sections clearly showed that for image retrieval tasks, the use of different representations is essential for capturing the multiple concepts that each image can be associated to by different persons. How to model the fusion of multiple representations according to the context in which the retrieval system is expected to be used is a hard task, as the increase in the number of parameters controlling the fusion mechanisms goes along with the availability of large training datasets, and with solid assumptions on their ranges and relationships, to ensure that robust estimations are produced. While nowadays a large quantity of data is available for training purposes, still the way in which they are processed to produce personalized results, and avoid biases, requires a careful design of the learning architecture and algorithm [23]. So, while the deep learning paradigm is now regarded as an effective solution for many classification and similarity retrieval tasks, previous works on information fusion for image retrieval provide a set of guidelines on the design of the learning architecture and the learning function to fully exploit the potentialities of this popular

paradigm.

- *early fusion by Feature weighting.* Feature weighting allows building retrieval systems where the designer has a full control of the processing steps, and the reasons behind the output of the system can be traced back to the importance given to different image representations. Moreover, relevance feedback mechanisms can be implemented to modify the weights according to the users' needs. The algorithm to estimate the weights should be as simple as possible, as the choice of the objective function, and the related estimation procedure could drive the system to produce biased and unexpected results. On the other hand, for narrow domain applications, this mechanism could prove to be winning, as the weights can be tailored to the domain of the images at hand. If no automatic estimation algorithm is used, then for the casual users it could be a difficult task to understand the effect of the weight values on the final results.

- *Late fusion by multi-feature spaces.* From a conceptual point of view, this is one of the more promising approaches, as it avoids dealing directly with the fusion of different image representations, as fusion is performed at a later stage, where the outputs of different systems can be regarded as new features to be combined. This approach is actually taken as a paradigm to implement multi-modal and cross-modal retrieval system employing deep learning architectures [116, 118], as different independent retrieval systems are seen as feature transformation functions, producing a new feature space for a second layer retrieval function. Again, the main research issues in the implementation of such an approach is the training of different systems, and how changes in one level propagates to the next levels and the output.

- *Fusing different relevance feedback approaches.* The exploitation of relevance feedback information can be carried out according to different approaches, in terms of the assumptions on the underlying distribution of relevant images. As the different relevance feedback approaches proposed

in the literature reflect the richness in the way the similarity between images can be modelled, their fusion might provide the system with an additional level of flexibility in capturing the user's needs, Again, the estimation of the parameters of the fusion mechanism should be kept as simple as possible, as the amount of information produced during the feedback iterations is quite limited, and, consequently, the number of parameters to be estimated should be small.

9. Conclusion

The cheap availability of cameras embedded in portable devices, and the availability of almost unlimited storage space, unleashed the natural tendency of people to communicate through images, for the richness of semantic content, and the immediacy of the message conveyed. This vast amount of visual information can be searched through textual queries related to the geolocation, timestamp, tags and labels provided by the users, with the shortcoming that these textual descriptors capture the semantics of images only partly, due to the richness of the semantic content of an image compared to the subjectivity of image tagging and labeling. To this end, the query-by-content paradigm allows searching for images beyond the purpose of their first use, by leveraging on the extraction of multiple descriptors to allow associating each image to different semantic concepts. This paper aimed at providing an overview of the techniques that could be used to fuse different descriptors, both in terms of the components in the processing pipeline in which the fusion takes place, and in terms of the techniques that can be used to estimate the parameters of the query mechanism to adapt to the needs and goals of the target application, and to the interests of the users involved. While the past 20+ years of research in the field allowed to reach a number of milestones, so that image classification and retrieval functions are now available in consumer products, the steep increase in the number of images stored, and the consequent

requests for more advanced functionalities by different categories of users, are making the old challenges even harder:

- Labeled data is needed in order to design the system, and estimate the parameters. However, the reliability of the labeling process clearly affects the quality of the performance of the system. For this reason benchmarks that help researchers to develop new approaches and evaluate the related performances are, now more than ever, crucial. In this line of reasoning it is worth to note that evaluation campaigns such as ImageCLEF[2] and ImageNET[3], since 2003 and 2010 respectively, created a number of publicly-accessible evaluation resources. Anyway, the creation of public datasets suited to test the performances of CBIR system in real settings still remains one of the main issues in this field. How can labeling be improved, and how can unlabeled data as well as partially labeled data being exploited to incrementally improve the system performances?
- How can the implicit feedback provided by the user be exploited when browsing the private archive, or searching through the media content shared within his/her social network?
- How to design easy-to-use interfaces that allows users to interact with the system in an intuitive way so that labels and feedbacks are provided in non ambiguous way?
- How to design fusion approaches tailored to vertical applications or scenarios, e.g., the media industry, fashion, design, forensics, etc.?
- More in general, principled approaches providing design guidelines are still needed, as the vast majority of papers support the proposal of new techniques and algorithms through experimental evaluation, and trial and error procedures.

[2] http://imageclef.org/2016

[3] http://image-net.org/challenges/LSVRC/2016/

The availability of more computing power, especially through the 'cloud computing' paradigm, the large popularity of deep learning approaches, as well as the interest from a variety of actors, both from the research community, and from an increasing number of companies, will allow addressing the above issues with novel approaches that will leverage on cooperation and knowledge sharing.

References

[1] V. Cisco, The zettabyte era: Trends and analysis, Whitepaper, WhitePaper (2015).

[2] C. Kofler, M. Larson, A. Hanjalic, User intent in multimedia search: A survey of the state of the art and future challenges, ACM Comput. Surv. 49 (2) (2016) 36:1–36:37.

[3] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1349–1380.

[4] M. S. Lew, N. Sebe, C. Djeraba, R. Jain, Content-based multimedia information retrieval: State of the art and challenges, ACM Trans. Multimedia Comput. Commun. Appl. 2 (1) (2006) 1–19.

[5] R. Datta, D. Joshi, J. Li, J. Z. Wang, Image retrieval: Ideas, influences, and trends of the new age, ACM Computing Surveys 40 (2) (2008) 1–60.

[6] B. Thomee, A picture is worth a thousand words: content-based image retrieval techniques, Ph.D. thesis, Leiden University, The Netherlands (November 2010).

[7] T. Kato, Database architecture for content-based image retrieval, in: Image Storage and Retrieval Systems (SPIE), Vol. 1662, 1992, pp. 112–123.

[8] M. L. Kherfi, D. Ziou, Relevance feedback for cbir: a new approach based on probabilistic feature weighting with positive and negative examples, IEEE Transactions on Image Processing 15 (4) (2006) 1017–1030.

[9] T. Pavlidis, Limitations of content-based image retrieval, Technical report, Stony Brook University (2008).

[10] Y. Rui, T. S. Huang, Relevance feedback techniques in image retrieval, in: Lew M.S. (ed.): Principles of Visual Information Retrieval, Springer-Verlag, London, 2001, pp. 219–258.

[11] X. S. Zhou, T. S. Huang, Relevance feedback in image retrieval: A comprehensive review, Multimedia Syst. 8 (6) (2003) 536–544.

[12] S. Lu, T. Mei, J. Wang, J. Zhang, Z. Wang, S. Li, Browse-to-search: Interactive exploratory search with visual entities, ACM Trans. Inf. Syst. 32 (4) (2014) 18:1–18:27.

[13] R. Tronci, L. Piras, G. Giacinto, Performance evaluation of relevance feedback for image retrieval by "real-world" multi-tagged image datasets, IJMDEM 3 (1) (2012) 1–16.

[14] A. Gilbert, L. Piras, J. Wang, F. Yan, E. Dellandréa, R. J. Gaizauskas, M. Villegas, K. Mikolajczyk, Overview of the ImageCLEF 2015 scalable image annotation, localization and sentence generation task, in: L. Cappellato, N. Ferro, G. J. F. Jones, E. SanJuan (Eds.), Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015., Vol. 1391 of CEUR Workshop Proceedings, CEUR-WS.org, 2015.

[15] A. Torralba, R. Fergus, W. T. Freeman, 80 million tiny images: A large data set for nonparametric object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 30 (11) (2008) 1958–1970.

[16] J. Weston, S. Bengio, N. Usunier, Large scale image annotation: learning to rank with joint word-image embeddings, Machine Learning 81 (1) (2010) 21–35.

[17] X. Wang, L. Zhang, M. Liu, Y. Li, W. Ma, ARISTA - image search to annotation on billions of web photos, in: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, pp. 2987–2994.

[18] H. Müller, P. D. Clough, T. Deselaers, B. Caputo (Eds.), ImageCLEF, Experimental Evaluation in Visual Information Retrieval, Springer, 2010.

[19] T. Li, T. Mei, S. Yan, I.-S. Kweon, C. Lee, Contextual decomposition of multi-label images, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 0 (2009) 2270–2277.

[20] M. Bertini, A. D. Bimbo, G. Serra, C. Torniai, R. Cucchiara, C. Grana, R. Vezzani, Dynamic pictorially enriched ontologies for digital video libraries, IEEE MultiMedia 16 (2) (2009) 42–51.

[21] S. A. Chatzichristofis, Y. S. Boutalis, Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval, in: A. Gasteratos, M. Vincze, J. K. Tsotsos (Eds.), ICVS, Vol. 5008 of Lecture Notes in Computer Science, Springer, 2008, pp. 312–322.

[22] J. Sivic, A. Zisserman, Efficient visual search for objects in videos, Proceedings of the IEEE 96 (4) (2008) 548 –566.

[23] N. Cristianini, A different way of thinking, New Scientist 232 (3101) (2016) 39 – 43.

[24] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: P. L. Bartlett, F. C. N.

Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States., 2012, pp. 1106–1114.

[25] Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Handwritten digit recognition with a back-propagation network, in: Proceedings of the 2Nd International Conference on Neural Information Processing Systems, NIPS'89, MIT Press, Cambridge, MA, USA, 1989, pp. 396–404.

[26] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.

[27] L. Deng, A tutorial survey of architectures, algorithms, and applications for deep learning, APSIPA Transactions on Signal and Information Processing 3.

[28] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, in: Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, Vol. 32 of JMLR Workshop and Conference Proceedings, JMLR.org, 2014, pp. 647–655.

[29] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Fleet et al., pp. 818–833.

[30] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '14, IEEE Computer Society, Washington, DC, USA, 2014, pp. 1717–1724.

[31] A. S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, Cnn features off-the-shelf: An astounding baseline for recognition, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW '14, IEEE Computer Society, Washington, DC, USA, 2014, pp. 512–519.

[32] A. Babenko, A. Slesarev, A. Chigorin, V. S. Lempitsky, Neural codes for image retrieval, in: Fleet et al., pp. 584–599.

[33] F. Perronnin, J. Sánchez, T. Mensink, Improving the fisher kernel for large-scale image classification, in: Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV'10, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 143–156.

[34] E. Valveny, Leveraging category-level labels for instance-level image retrieval, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), CVPR '12, IEEE Computer Society, Washington, DC, USA, 2012, pp. 3045–3052.

[35] H. Jégou, A. Zisserman, Triangulation embedding and democratic aggregation for image search, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, IEEE Computer Society, 2014, pp. 3310–3317.

[36] H. Chen, A socio-technical perspective of museum practitioners' image-using behaviors, The Electronic Library 25 (1) (2007) 18–35.

[37] B. Girod, V. Chandrasekhar, D. M. Chen, N. M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. S. Tsai, R. Vedantham, Mobile visual search, IEEE Signal Processing Magazine 28 (4) (2011) 61–76.

[38] O. Marques, Visual information retrieval: The state of the art, IT Professional 18 (4) (2016) 7–9.

[39] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, A. Torralba, SUN database: Large-scale scene recognition from abbey to zoo, in: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, pp. 3485–3492.

[40] N. Bhowmik, V. R. Gonzlez, V. Gouet-Brunet, H. Pedrini, G. Bloch, Efficient fusion of multidimensional descriptors for image retrieval, in: 2014 IEEE International Conference on Image Processing (ICIP), 2014, pp. 5766–5770.

[41] C. Snoek, M. Worring, A. W. M. Smeulders, Early versus late fusion in semantic video analysis, in: H. Zhang, T. Chua, R. Steinmetz, M. S. Kankanhalli, L. Wilcox (Eds.), Proceedings of the 13th ACM International Conference on Multimedia, Singapore, November 6-11, 2005, ACM, 2005, pp. 399–402.

[42] P. K. Atrey, M. A. Hossain, A. El-Saddik, M. S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, Multimedia Syst. 16 (6) (2010) 345–379.

[43] J. Yu, Z. Qin, T. Wan, X. Zhang, Feature integration analysis of bag-of-features model for image retrieval, Neurocomputing 120 (2013) 355 – 364, image Feature Detection and Description.

[44] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[45] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[46] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA, IEEE Computer Society, 2005, pp. 886–893.

[47] P. A. S. Kimura, J. M. B. Cavalcanti, P. C. Saraiva, R. da Silva Torres, M. A. Gonçalves, Evaluating retrieval effectiveness of descriptors for searching in large image databases, JIDM 2 (3) (2011) 305–320.

[48] J. Yue, Z. Li, L. Liu, Z. Fu, Content-based image retrieval using color and texture fused features, Mathematical and Computer Modelling 54 (34) (2011) 1121 – 1127, mathematical and Computer Modeling in agriculture (CCTA 2010).

[49] H. J. Escalante, C. A. Hérnadez, L. E. Sucar, M. Montes, Late fusion of heterogeneous methods for multimedia image retrieval, in: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, MIR '08, ACM, New York, NY, USA, 2008, pp. 172–179.

[50] R. da S. Torres, A. X. Falco, M. A. Gonalves, J. P. Papa, B. Zhang, W. Fan, E. A. Fox, A genetic programming framework for content-based image retrieval, Pattern Recognition 42 (2) (2009) 283 – 292, learning Semantics from Multimedia Content.

[51] W. Zhang, Z. Qin, T. Wan, Image scene categorization using multi-bag-of-features, in: Machine Learning and Cybernetics (ICMLC), 2011 International Conference on, Vol. 4, 2011, pp. 1804–1808.

[52] L. Piras, R. Tronci, G. Giacinto, Diversity in ensembles of codebooks for visual concept detection, in: A. Petrosino (Ed.), ICIAP (2), Vol. 8157 of Lecture Notes in Computer Science, Springer, 2013, pp. 399–408.

[53] D. Picard, N. Thome, M. Cord, An efficient system for combining complementary kernels in complex visual categorization tasks, in: 2010 IEEE International Conference on Image Processing, 2010, pp. 3877–3880.

[54] D. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, F. G. B. D. Natale, A hybrid approach for retrieving diverse social images of landmarks, in: 2015 IEEE International Conference on Multimedia and Expo, ICME 2015, Turin, Italy, June 29 - July 3, 2015, IEEE, 2015, pp. 1–6.

[55] Y. Cao, H. Zhang, Y. Gao, X. Xu, J. Guo, Matching image with multiple local features, in: Pattern Recognition (ICPR), 2010 20th International Conference on, 2010, pp. 519–522.

[56] L. Piras, G. Giacinto, Dissimilarity representation in multi-feature spaces for image retrieval, in: G. Maino, G. L. Foresti (Eds.), Image Analysis and Processing - ICIAP 2011 - 16th International Conference, Ravenna, Italy, September 14-16, 2011, Proceedings, Part I, Vol. 6978 of Lecture Notes in Computer Science, Springer, 2011, pp. 139–148.

[57] D. A. Lisin, M. A. Mattar, M. B. Blaschko, E. G. Learned-Miller, M. C. Benfield, Combining local and global image features for object class recognition, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops, 2005, pp. 47–47.

[58] V. Risojevi, Z. Babi, Fusion of global and local descriptors for remote sensing image classification, IEEE Geoscience and Remote Sensing Letters 10 (4) (2013) 836–840.

[59] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng, Multimodal deep learning, in: L. Getoor, T. Scheffer (Eds.), Proceedings of the

28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011, Omnipress, 2011, pp. 689–696.

[60] N. Srivastava, R. Salakhutdinov, Multimodal learning with deep boltzmann machines, J. Mach. Learn. Res. 15 (1) (2014) 2949–2980.

[61] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, J. Li, Deep learning for content-based image retrieval: A comprehensive study, in: Proceedings of the 22Nd ACM International Conference on Multimedia, MM '14, ACM, New York, NY, USA, 2014, pp. 157–166.

[62] Y. Rui, T. S. Huang, S. Mehrotra, Content-Based image retrieval with relevance feedback in MARS, in: International Conference on Image Processing Proceedings, 1997, pp. 815–818.

[63] J. Peng, B. Bhanu, S. Qing, Probabilistic feature relevance learning for content-based image retrieval, Computer Vision and Image Understanding 75 (1/2) (1999) 150–164.

[64] Y. Wu, A. Zhang, Interactive pattern analysis for relevance feedback in multimedia information retrieval, Multimedia Syst. 10 (1) (2004) 41–55.

[65] E. Guldogan, M. Gabbouj, Feature selection for content-based image retrieval, Signal, Image and Video Processing 2 (3) (2008) 241–250.

[66] L. Piras, G. Giacinto, Neighborhood-based feature weighting for relevance feedback in content-based retrieval, in: WIAMIS, IEEE Computer Society, 2009, pp. 238–241.

[67] Y. Ishikawa, R. Subramanya, C. Faloutsos, MindReader: Querying databases through multiple examples, in: Proceedings of the 24th VLDB Conference, 1998, pp. 433–438.

[68] Y. Rui, T. Huang, Optimizing learning in image retrieval, in: Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on, Vol. 1, 2000, pp. 236–243 vol.1.

[69] M. Tzelepi, A. Tefas, Relevance feedback in deep convolutional neural networks for content based image retrieval, in: Proceedings of the 9th Hellenic Conference on Artificial Intelligence, SETN '16, ACM, New York, NY, USA, 2016, pp. 27:1–27:7.

[70] L. I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, WILEY, 2004.

[71] K. Tieu, P. A. Viola, Boosting image retrieval, in: CVPR, IEEE Computer Society, 2000, pp. 1228–1235.

[72] J. Laaksonen, M. Koskela, E. Oja, Picsom-self-organizing image retrieval with mpeg-7 content descriptors, IEEE Transactions on Neural Networks 13 (4) (2002) 841–853.

[73] M. Arevalillo-Herráez, J. Domingo, F. J. Ferri, Combining similarity measures in content-based image retrieval, Pattern Recognition Letters 29 (16) (2008) 2174–2181.

[74] H. Müller, W. M. 0002, D. Squire, S. Marchand-Maillet, T. Pun, Performance evaluation in content-based image retrieval: overview and proposals, Pattern Recognition Letters 22 (5) (2001) 593–601.

[75] L. Zhang, F. Lin, B. Zhang, Support vector machine learning for image retrieval, in: ICIP (2), 2001, pp. 721–724.

[76] G. Giacinto, F. Roli, Dissimilarity representation of images for relevance feedback in content-based image retrieval, in: P. Perner, A. Rosenfeld (Eds.), MLDM, Vol. 2734 of Lecture Notes in Computer Science, Springer, 2003, pp. 202–214.

[77] G. P. Nguyen, M. Worring, A. W. M. Smeulders, Similarity learning via dissimilarity space in cbir, in: J. Z. Wang, N. Boujemaa, Y. Chen (Eds.), Multimedia Information Retrieval, ACM, 2006, pp. 107–116.

[78] E. Bruno, N. Moënne-Loccoz, S. Marchand-Maillet, Learning user queries in multimodal dissimilarity spaces, in: M. Detyniecki, J. M. Jose, A. Nürnberger, C. J. van Rijsbergen (Eds.), Adaptive Multimedia Retrieval, Vol. 3877 of Lecture Notes in Computer Science, Springer, 2005, pp. 168–179.

[79] G. Giacinto, A nearest-neighbor approach to relevance feedback in content based image retrieval, in: CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval, ACM, New York, NY, USA, 2007, pp. 456–463.

[80] P. Wu, S. C. Hoi, H. Xia, P. Zhao, D. Wang, C. Miao, Online multimodal deep similarity learning with application to image retrieval, in: Proceedings of the 21st ACM International Conference on Multimedia, MM '13, ACM, New York, NY, USA, 2013, pp. 153–162.

[81] B. Thomee, M. S. Lew, Interactive search in image retrieval: a survey, IJMIR 1 (1) (2012) 71–86.

[82] Y. Rui, T. S. Huang, S. Mehrotra, Relevance feedback: A power tool in interactive content-based image retrieval, IEEE Trans. on Circuits and Systems for Video Technology 8 (5) (1998) 644–655.

[83] S. Liang, Z. Sun, Sketch retrieval and relevance feedback with biased svm classification, Pattern Recognition Letters 29 (12) (2008) 1733 – 1741.

[84] Y. Chen, X. S. Zhou, T. Huang, One-class svm for learning in image retrieval, in: ICIP, Vol. 1, 2001, pp. 34 –37.

[85] J. J. Rocchio, Relevance feedback in information retrieval, in: G. Salton (Ed.), The SMART Retrieval System - Experiments in

Automatic Document Processing, Prentice Hall, Englewood, Cliffs, New Jersey, 1971, pp. 313–323.

[86] D. A. Cohn, L. E. Atlas, R. E. Ladner, Improving generalization with active learning, Machine Learning 15 (2) (1994) 201–221.

[87] S. C. H. Hoi, R. Jin, J. Zhu, M. R. Lyu, Semisupervised svm batch mode active learning with applications to image retrieval, ACM Trans. Inf. Syst. 27 (3) (2009) 16:1–16:29.

[88] S. Tong, E. Y. Chang, Support vector machine active learning for image retrieval, in: ACM Multimedia, 2001, pp. 107–118.

[89] P.-Y. Yin, B. Bhanu, K.-C. Chang, A. Dong, Integrating relevance feedback techniques for image retrieval using reinforcement learning, IEEE Trans. Pattern Anal. Mach. Intell. 27 (10) (2005) 1536–1551.

[90] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, P. N. Yianilos, The bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments, IEEE Transactions on Image Processing 9 (1) (2000) 20–37.

[91] C.-H. Hoi, M. R. Lyu, Group-based relevance feedback with support vector machine ensembles, in: ICPR (3), 2004, pp. 874–877.

[92] D. Tao, X. Tang, X. Li, X. Wu, Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval, IEEE Trans. on Pattern Analysis and Machine Intelligence 28 (2006) 1088–1099.

[93] Y. Rao, P. Mundur, Y. Yesha, Fuzzy svm ensembles for relevance feedback in image retrieval, in: H. Sundaram, M. R. Naphade, J. R. Smith, Y. Rui (Eds.), CIVR, Vol. 4071 of Lecture Notes in Computer Science, Springer, 2006, pp. 350–359.

[94] Y. Tu, G. Li, H. Dai, Integrating local one-class classifiers for image retrieval, in: X. Li, O. R. Zaïane, Z. Li (Eds.), ADMA, Vol. 4093 of Lecture Notes in Computer Science, Springer, 2006, pp. 213–222.

[95] K. Crammer, G. Chechik, A needle in a haystack: local one-class optimization, in: C. E. Brodley (Ed.), ICML, Vol. 69 of ACM International Conference Proceeding Series, ACM, 2004.

[96] R. K. Srihari, Automatic indexing and content-based retrieval of captioned images, Computer 28 (9) (1995) 49–56.

[97] C. Frankel, M. J. Swain, V. Athitsos, Webseer: An image search engine for the world wide web, Tech. rep., Chicago, IL, USA (1996).

[98] A. Depeursinge, H. Müller, Fusion Techniques for Combining Textual and Visual Information Retrieval, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 95–114.

[99] D. Rafailidis, S. Manolopoulou, P. Daras, A unified framework for multimodal retrieval, Pattern Recognition 46 (12) (2013) 3358 – 3370.

[100] X. S. Zhou, T. S. Huang, Unifying keywords and visual contents in image retrieval, IEEE MultiMedia 9 (2) (2002) 23–33.

[101] K.-S. Goh, E. Y. Chang, W.-C. Lai, Multimodal concept-dependent active learning for image retrieval, in: Proceedings of the 12th Annual ACM International Conference on Multimedia, MULTIMEDIA '04, ACM, New York, NY, USA, 2004, pp. 564–571.

[102] S. Sclaroff, M. La Cascia, S. Sethi, Unifying textual and visual cues for content-based image retrieval on the world wide web, Comput. Vis. Image Underst. 75 (1-2) (1999) 86–98.

[103] K. Barnard, D. Forsyth, Learning the semantics of words and pictures, in: Computer Vision, 2001. ICCV 2001. Proceedings. Eighth

IEEE International Conference on, Vol. 2, 2001, pp. 408–415 vol.2. doi:10.1109/ICCV.2001.937654.

[104] H. Frigui, J. Caudill, A. C. B. Abdallah, Fusion of multi-modal features for efficient content-based image retrieval, in: Fuzzy Systems, 2008. FUZZ-IEEE 2008. (IEEE World Congress on Computational Intelligence). IEEE International Conference on, 2008, pp. 1992–1998.

[105] H. Tahani, J. M. Keller, Information fusion in computer vision using the fuzzy integral, IEEE Transactions on Systems, Man, and Cybernetics 20 (3) (1990) 733–741.

[106] C. Moulin, C. Largeron, C. Ducottet, M. Gry, C. Barat, Fisher linear discriminant analysis for text-image combination in multimedia information retrieval, Pattern Recognition 47 (1) (2014) 260 – 269.

[107] T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: An Efficient Data Clustering Method for Very Large Databases, in: ACM SIGMOD International Conference on Management of Data, 1996, pp. 103–114.

[108] P. Wu, S. C. H. Hoi, P. Zhao, C. Miao, Z. Y. Liu, Online multi-modal distance metric learning with application to image retrieval, IEEE Transactions on Knowledge and Data Engineering 28 (2) (2016) 454–467.

[109] L. Zhu, A. B. Rao, A. Zhang, Theory of keyblock-based image retrieval, ACM Trans. Inf. Syst. 20 (2) (2002) 224–257.

[110] H. Zhang, Z. Chen, M. Li, Z. Su, Relevance feedback and learning in content-based image search, World Wide Web 6 (2) (2003) 131–155.

[111] Y. Lu, C. Hu, X. Zhu, H. Zhang, Q. Yang, A unified framework for semantics and feature based relevance feedback in image retrieval

systems, in: Proceedings of the Eighth ACM International Conference on Multimedia, MULTIMEDIA '00, ACM, New York, NY, USA, 2000, pp. 31–37.

[112] R. Besançon, P. Hède, P.-A. Moellic, C. Fluhr, Cross-Media Feedback Strategies: Merging Text and Image Information to Improve Image Retrieval, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 709–717.

[113] F. M. Segarra, L. A. Leiva, R. Paredes, A relevant image search engine with late fusion: mixing the roles of textual and visual descriptors, in: P. Pu, M. J. Pazzani, E. André, D. Riecken (Eds.), IUI, ACM, 2011, pp. 455–456.

[114] X. Benavent, A. Garcia-Serrano, R. Granados, J. Benavent, E. de Ves, Multimedia information retrieval based on late semantic fusion approaches: Experiments on a wikipedia image collection, IEEE Transactions on Multimedia 15 (8) (2013) 2009–2021.

[115] Y. Chen, H. Sampathkumar, B. Luo, X. wen Chen, ilike: Bridging the semantic gap in vertical image search by integrating text and visual features, IEEE Transactions on Knowledge and Data Engineering 25 (10) (2013) 2257–2270.

[116] Y. He, S. Xiang, C. Kang, J. Wang, C. Pan, Cross-modal retrieval via deep and bidirectional representation learning, IEEE Transactions on Multimedia 18 (7) (2016) 1363–1377.

[117] R. Salakhutdinov, G. E. Hinton, Deep boltzmann machines, in: D. A. V. Dyk, M. Welling (Eds.), Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009, Vol. 5 of JMLR Proceedings, JMLR.org, 2009, pp. 448–455.

[118] C. Wang, H. Yang, C. Meinel, A deep semantic framework for multimodal representation learning, Multimedia Tools Appl. 75 (15) (2016) 9255–9276.