Multimedia Tools and Applications (2023) 82:11619–11661
https://doi.org/10.1007/s11042-022-13119-0

ExpertosLF: dynamic late fusion of CBIR systems using online learning with relevance feedback

Soraia M. Alarcão[1] · Vânia Mendonça[2] · Carolina Maruta[3] · Manuel J. Fonseca[1]

Received: 21 June 2021 / Revised: 11 January 2022 / Accepted: 10 April 2022 /
Published online: 20 August 2022

Abstract

One of the main challenges in CBIR systems is to choose discriminative and compact features, among dozens, to represent the images under comparison. Over the years, a great effort has been made to combine multiple features, mainly using early, late, and hierarchical fusion techniques. Unveiling the perfect combination of features is highly domain-specific and dependent on the type of image. Thus, the process of designing a CBIR system for new datasets or domains involves a huge experimentation overhead, leading to multiple fine-tuned CBIR systems. It would be desirable to dynamically find the best combination of CBIR systems without needing to go through such extensive experimentation and without requiring previous domain knowledge. In this paper, we propose ExpertosLF, a model-agnostic interpretable late fusion technique based on online learning with expert advice, which dynamically combines CBIR systems without knowing a priori which ones are the best for a given domain. At each query, ExpertosLF takes advantage of user's feedback to determine each CBIR contribution in the ensemble for the following queries. ExpertosLF produces an interpretable ensemble that is independent of the dataset and domain. Moreover, ExpertosLF is designed to be modular, and scalable. Experiments on 13 benchmark datasets from the Biomedical, Real, and Sketch domains revealed that: (i) ExpertosLF surpasses the performance of state of the art late-fusion techniques; (ii) it successfully and quickly converges to the performance of the best CBIR sets across domains without any previous domain knowledge (in most cases, fewer than 25 queries need to receive human feedback).

Keywords Content-based image retrieval · Late fusion · Prediction with expert advice · Online learning · Relevance feedback

1 Introduction

Nowadays, many image repositories are available for almost every domain, such as architecture, astronomy, education, geology, medicine, multimedia, and remote sensing. With such

✉ Soraia M. Alarcão
smalarcao@ciencias.ulisboa.pt

Extended author information available on the last page of the article.

a massive growth in the number of images available, the need to store and retrieve images in an efficient manner arises, leading to an increase in the importance of Content-Based Image Retrieval (CBIR) systems. Such systems have a multitude of real-life applications concerning crime prevention, digital libraries, medical diagnostic, textile industry, traffic congestion analysis, and so on.

One of the main challenges in CBIR is to choose features that are sufficiently discriminative to infer how similar images are, while keeping them compact to ensure that the system is timely and computationally efficient. Furthermore, human perception of image similarity, which is subjective, semantic, and task-dependent, may not be captured by commonly used low-level features (e.g., color, shape, texture). This phenomenon is known as the *semantic gap* between high-level concepts conveyed by an image (e.g., emotions, events, or objects) and the limited descriptive power of low-level visual features [2, 60, 86, 115]. For example, consider Fig. 1a (yellow car close to a green wall) as the query image of a retrieval system that only uses low-level features. Both Fig. 1b (lady with a yellow dress on green grass) and Fig. 1d (yellow car close to a tree) would be returned as a result (due to the color similarity). However, it would be desirable that Fig. 1c (red car in the wilderness) would be returned instead of Fig. 1b, since Fig. 1a and c are very similar semantically (both depict a car).

Over the years, multiple approaches have been proposed to mitigate the semantic gap: (i) generation of high-level features that mimic human perception using deep learning; (ii) multi-feature early, hierarchical, and late fusion methods to combine low-level features (and to a lower extent, low- and high-level features, and multiple high-level features); (iii) incorporation of human expertise, through relevance feedback, in the retrieval process, leading to perceptually and semantically more meaningful results.

Besides the semantic gap, another issue for CBIR systems concerns the choice of the best features for a certain domain. If we consider pictures of our everyday life, one might expect that, the more information available, the better a retrieval system's performance will be. However, in certain domains, such as art or medicine, this is not necessarily true. Not only images from those domains have characteristics that are quite different from everyday pictures, but also their characteristics within the same domain may differ significantly. For example, many medical images are only gray-level (e.g., radiography, computed tomography, Magnetic Resonance Imaging (MRI)), leading shape and texture to acquire increased relevance, when compared to color or semantic features [26, 103].

Unveiling the perfect combination of different features to design novel CBIR systems for new datasets or domains usually involves a huge experimentation overhead, and is highly task and domain-specific, leading to the fine-tuning of CBIR systems for each domain.

(a) Query Image (b) Visually similar (c) Semantically similar (d) Visually and semantically similar

Fig. 1 Example of a query image and three related images that illustrate the semantic gap between high-level concepts and low-level features: only Fig. 1d is similar according to both visual and semantic features. (figure best seen in color)

Consider, for instance, a scenario where multiple biomedical CBIR systems are available, each of them fine-tuned to a specific disease and diagnostic method (e.g., a shape-based CBIR tuned for brain cancer MRI scans, a color and texture-based CBIR for breast cancer histopathological images, etc.). How can one leverage on the existing fine-tuned CBIR systems to accommodate other diseases or diagnostic methods, or even other domains? Early fusion and hierarchical approaches would require even further experimentation to combine the different types of features, resulting in more fine-tuned systems. Although late fusion approaches are more robust and well-suited for such task, they lack interpretability, as it is unclear which CBIR system performs best for the task in hand. Relevance feedback techniques could also be used to tune the results and adapt the CBIR system, but they usually require constant feedback from the users. In the biomedical CBIR example above, it would be unrealistic and costly to expect medical specialists to give feedback on every result retrieved by a CBIR system.

To address these challenges, we present *ExpertosLF*, an interpretable late fusion technique that takes advantage of human feedback (requiring minimum effort and interaction). For that, we propose a novel application of online learning to late fuse multiple CBIR systems, under the framework of prediction with expert advice. To each CBIR in the ensemble is assigned a weight that determines how much it contributes to the final set of images to be retrieved for a given query. The systems' weights are updated in an online fashion, based on the quality of each system's results, assessed by one or more human evaluators at each query. The resulting ensemble will be independent of the dataset and domain, while being able to take advantage of previous experiments to create the individual CBIR systems. *ExpertosLF* is designed to be interpretable, model-agnostic, modular, and scalable.

With this work, we aim to address the following research questions:

- RQ1) Does our late fusion technique improve retrieval performance?
- RQ2) Does the resulting ensemble perform as well as the best individual CBIR?
- RQ3) Can we use the ensemble learned in an online setting in an offline setting?
- RQ4) Are the CBIR experts in the resulting ensemble plausible considering the domain in hand?

Our contribution is threefold:

1. A model-agnostic interpretable late fusion technique based on online learning with expert advice, which dynamically combines CBIR systems without knowing a priori which ones are the best for a given domain;
2. Mitigation of the semantic gap between the low-level information of an image and its high-level semantic concepts, by studying the impact of combining both kinds of descriptors in CBIR in different domains;
3. A set of extensive experiments on 13 benchmark datasets focusing on three different domains: Biomedical, Real, and Sketch.

ExpertosLF surpasses the performance of state of the art late fusion techniques for the majority of the datasets. It quickly converges to the performance of the best CBIR systems across domains, without any previous domain knowledge (in most cases, fewer than 25 queries need to receive human feedback). Moreover, the ensemble learned using our weighted late-fusion technique can be successfully applied to an offline scenario (i.e., in which there is no feedback available).

2 Related work

The typical flow of a CBIR system is depicted in Fig. 2. The first step consists of generating a set of features to accurately represent the content of each image in the database. These sets of features, also called *descriptors*, are used to compute the distance between the query image and each candidate image in the database, in order to retrieve the most similar images to the query image.

Ascertaining the most discriminative descriptors is highly dependent on both the type of images the CBIR system will handle (colored, black and white, or grey-level) and the domain in hand (e.g., art, medical, textile, remote sensing). For example, an image of a sunset will have more semantic and color information than an image from a medical examination (consider an X-ray or Computed Tomography scan whose shape of the organ under analysis is more prominent). Moreover, there is a *semantic gap* between high-level concepts, such as emotions, events, objects or activities conveyed by an image, and the limited descriptive power of low-level visual descriptors, as exemplified earlier in Section 1.

Here, we analyse how the semantic gap has been addressed in several CBIR works focusing on computational methods that: (i) propose novel low- and high-level descriptors (Sections 2.1 and 2.2), and combinations among them (Section 2.3); (ii) improve the retrieval process using human *relevance feedback* at each query (Section 2.4). Our analysis is focused on the last five years. For a more complete review, see [52, 77, 122, 123].

2.1 Low-level descriptors

Hand-crafted global and local low-level descriptors representing color, shape, and texture are widely used in current CBIR systems. Color is extensively used since it is the basic constituent of images, relatively robust to background complexity and independent of orientation and image size. Shape is useful for matching objects based on their physical structure and profile. Texture is used to look for visual patterns with properties of homogeneity that are not achieved by the presence of a single color, and how those patterns are spatially defined.

Global descriptors extracted from the whole image, are easy to compute, and have lower dimensionality. Multiple descriptors have been proposed: color (e.g., Auto Color Correlogram (ACC) [38], Color Coherence Vectors (CCV) [89], Color Histogram (CH) [89],

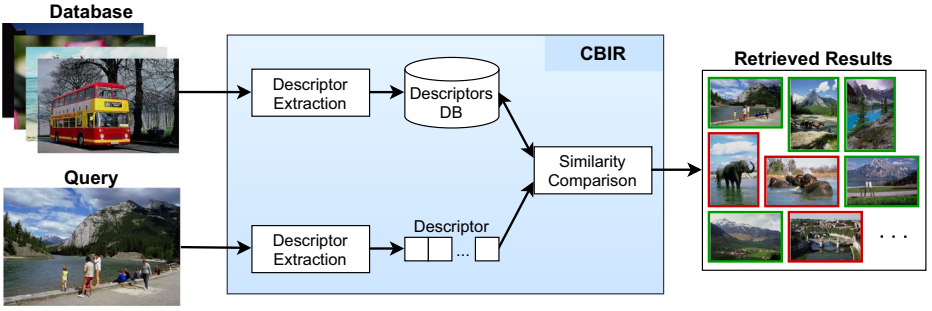


Fig. 2 Architecture of the typical CBIR setting. The retrieved images with a green border represent relevant images, while a red border represents non-relevant images for the given query. (figure best seen in color)

Color Moments (CM) [28], Opponent Histogram (OH) [105], and Reference Color Similarity (RCS) [48]), shape (e.g., Edge Histogram (EH) [19], and Zernike Moment Descriptor (ZMD) [47]), and texture (e.g., Gabor [64], Haralick [37], and Hybrid Directional Extrema Pattern (HDEP)).

Local descriptors are extracted from sub-images of a given image. They are robust to occlusion, changes in illumination and background, and geometric transformations; they are usually complex and produce high-dimensional vectors [113]. The Local Binary Patterns (LBP) descriptor is widely used in color and texture retrieval since it reflects the correlation among pixels within a local area [32, 71]. Other binary descriptors are Binary Robust Independent Elementary Features (BRIEF) [16], Binary Robust Invariant Scalable Keypoints (BRISK) [53], Fast Retina KeyPoint (FREAK) [4], Scale-Invariant Feature Transform (SIFT) [66], and Speeded Up Robust Features (SURF) [12].

Singular Value Decomposition (SVD)-based descriptors take advantage of the local spatial relationship of non-overlapping images' sub-regions [32, 63, 106]. Radon transforms are useful to reconstruct objects, and attain special attention on the medical domain [9, 100, 101].

2.2 High-level descriptors

Low-level information is useful to discriminate images, but it often fails at capturing high-level semantic concepts perceived by humans. To model high-level abstractions present in images, deep learning approaches have been proposed in the latest years. Deep approaches are able to learn complex representations from large amounts of data, in a supervised manner. An example of such representations are Convolutional Neural Networks (CNNs) [49], which have been widely adopted for multiple tasks, such as classification, image segmentation, or object recognition. Recently, CNNs have also been explored in retrieval tasks.

The most common approach is to extract feature representations from a pre-trained CNN model by feeding images in the input layer of a model, and taking activation values either from fully connected layers (to capture semantic information), or from convolutional layers using pooling techniques (to exploit spatial information). Pre-trained CNNs with ImageNet dataset are commonly used in CBIR systems [29]: AlexNet [88, 104, 110], Fast Convolutional Neural Network (FCNN) [118], VGG-19 [119]. Some authors have also proposed novel deep approaches: a CNN to retrieve images of different body organs [79], a CNN scalable face CBIR [98], a Convolutional Sparse Kernel Network for the medical domain [3], a Deep Belief Network for object-based retrieval [85], and a Fuzzy Neural Network to learn effective binary codes, while enhancing interpretability [60].

In some domains, the amount of images available is not sufficient to train a robust deep model, i.e., the model is prone to overfit. Transfer learning is beneficial in such situations, since features can be learnt in a resource-rich domain and then applied to a resource-scarce domain. Several works adopted pre-trained CNNs with natural images, and apply it to their target domain: VGG-m and VGG-16 for landmarks/monuments [5], ResNet-50 for diabetic retinopathy [26], Capsule Networks with 3D CNNs to detect Alzheimer disease using MRI [50], VGG19 for brain tumors [93], Inception-ResNet-V2 for otoscope images [17], and DenseNet121 for chest X-ray images [94]. Most models were pre-trained using ImageNet dataset [5, 26, 93, 94] (for the remaining ones, the information is missing/unclear).

2.3 Multi-feature fusion methods

Over the years, different approaches have been proposed to combine global and local descriptors (color and texture [13, 33, 44, 45, 107], color and shape [2, 25, 75, 113, 116], shape and texture [92, 103], and color, shape, and texture [7, 8, 14, 73, 81, 83, 86, 121]). To a lesser extent, authors have also proposed combinations of low- and high-level descriptors [57, 58], and an ensemble of high-level descriptors [36].

The most common approach is to extract multiple descriptors, and combine them using an early fusion approach [2, 7, 14, 25, 45, 81, 103, 113, 116, 121]. In the early fusion approach, the descriptors are extracted and combined into a single feature vector. The resulting vector is used to index all the images in the CBIR and search for the most similar ones. Usually, it is assumed that all descriptors have the same importance. However, that is not necessarily true: descriptors may not yield the same results for different categories of images. An alternative approach is to use weights to early fuse the descriptors using Particle Swarm Optimization (PSO) algorithm [33], genetic algorithms [75], or weighted functions [10, 83]. The fusion of descriptors at different levels can benefit CBIR systems, but it requires mechanisms for the selection of appropriate weights which are usually highly dependent of the dataset in use, and still involves a lot of experimentation (since the parameter tuning of the descriptors is carried out from the analysis of their performance in the proposed CBIR). Moreover, although a large number of features may better represent the discriminative properties of images, it may lead to the dimensionality curse problem.

Previous works focused mainly on single resolution processing of an image, however it may not be sufficient to gather varying level of details in an image since an image consists of both high and low resolution objects, and both large and small size objects [92]. Another approach is to combine descriptors in a hierarchical way by processing an image at multiple resolutions [8, 44, 45, 92]. This way, features that were not detected at a certain resolution, will be detected at another one. Wavelets offer a good energy compression and multi-resolution capability [8]. As such, LBP, Legendre moments, Gabor (or similar descriptors) are combined using Discrete Wavelet Transform (DWT) to extract shape information from texture features from an image at multiple resolutions.

A considerable body of work has been proposed to combine descriptors to create a single CBIR. Another possibility is to combine multiple CBIR systems, which may be less dependent on the task or type of images, while being able to take advantage of experiments already performed to create the individual CBIR. Some authors have proposed a hierarchical approach to combine low-level CBIR [73, 107], and low- and high-level CBIR [57, 58]. When combining multiple CBIR hierarchically, the main idea is to use a single CBIR to find the most relevant images [58, 73] or discard irrelevant images [57, 107], and then apply a second CBIR to refine the search.

Late-fusion techniques are also used to combine CBIR systems (mostly based on low-level features). In the late fusion approaches, multiple CBIR systems are created (where each one uses one or more early-fused descriptors to index and search for the images), and the results of each one are combined. They are usually split into two major groups: (i) similarity score-based rank list fusion and (ii) order based rank list fusion [6]. For the first group, the similarity scores of each image (of each retrieved list) are merged using an aggregation function (e.g., minimum, median, or maximum) to form the final search result. In the second one, a revised retrieval list is created as a function of the position in which images appear in different rank lists. Such fusion techniques tend to be more robust and efficient than early fusion techniques.

Finally, Hamreras et. al [36] proposed to take advantage of ensemble learning to combine different CNNs. However, its scope is very limited; the main focus was the identification of good parameters to form the ensemble (the number of neural networks to be used, and number of hidden neurons in each network).

All these methods, to some extent, require a huge experimental overload to find out which are the best combinations of descriptors or CBIR systems for each possible domain/dataset. Thus, in this work, we extend the late-fusion method so that it dynamically assigns a greater weight to the best CBIR systems for the domain/dataset in hand, taking advantage of relevance feedback provided by the user (when available).

2.4 Relevance feedback

Relevance feedback has been used in CBIR systems to modify the retrieval process in order to generate perceptually and semantically more meaningful results by involving the user in the retrieval process [65, 104]. The main idea is to present to the user the results from a given query, collect feedback about whether or not those results are relevant, and perform a new query based on that information; these steps are carried out iteratively until the user is satisfied with the results.

The most common types of feedback are explicit and implicit. In the explicit feedback, the user explicitly informs which images are relevant/not relevant (binary relevance feedback) or how relevant each image is (graded relevance feedback). In the implicit feedback, the system automatically infers user's feedback from their behavior.

Different approaches have been proposed to reformulate the query according to the feedback received: finding an optimized query feature vector using Rocchio's algorithm [11, 46, 67, 114], modifying the similarity measure so that relevant images have a high similarity value [1, 124], exploiting images' geometrical and discriminant structures to learn a semantic subspace [39, 120], or separating relevant and non-relevant images using Bayesian Networks [87], CNN [56, 76, 78, 104], Clustering [27], Logistic Regression [30], Optimum Forest algorithm [54], and Support Vector Machine [82, 97, 112].

Active learning has also been used to reduce the annotation effort, by selecting which images should be annotated by the users [82, 97]. Moreover, Tang *et.al* combined different active learning relevance feedback approaches carried out simultaneously, and then fused their results to improve the initial query [97].

All the aforementioned methods to mitigate the semantic gap helped furthering the development of CBIR systems, but they come with a number of drawbacks. Early and hierarchical fusion involve a huge experimentation overload to choose the best set of descriptors among descriptors of the same category or across categories. Furthermore, early approaches are prone to suffer from the so called *curse of dimensionality*. With late-fusion approaches, the merge of most similar images is done at a query-level, i.e., no knowledge of which CBIR is better is acquired over time. Moreover, regardless of the fusion technique in use, most CBIR solutions in the literature suffer from a lack of interpretability regarding which descriptors or CBIR systems are the best for a given domain or type of images. Finally, relevance feedback relies on user feedback (which is not always available), and the retrieval process is repeated multiple times until the user is satisfied.

To address some of these drawbacks, we propose to take advantage of human feedback (when available), not to improve the results for the current query, but instead to improve the late-fusion process for the next queries. Thus, our focus is to reward the CBIR systems that made the best contributions to the final set of images retrieved to the user, giving them a greater weight in the late-fusion for future queries.

3 Dynamic late fusion of CBIR using online learning

Given several existing CBIR systems (each one encompassing different descriptors or combinations of descriptors), how can we combine them in order to dynamically reach at least the performance of the best CBIR system (without knowing which are the best ones for a given domain a priori)? To tackle this question, we frame our late fusion technique as a problem of prediction with expert advice, using online learning to dynamically find which are the best CBIR in the ensemble, making the most of minimal human interaction.

We start by providing some background on the prediction with expert advice online learning framework (Section 3.1), and then we describe how we adapt it to late fuse multiple CBIR systems (Section 3.2).

3.1 Prediction with expert advice

A problem of prediction with expert advice can be seen as a repeated game between a *forecaster* and the *environment*, in which the forecaster resorts to a set of weighted *experts* to provide the best forecast [18]. At each round t, the forecaster F consults the predictions p_k^t in the decision space $\mathcal{A}$ made by each expert k. Considering the experts' predictions, the forecaster makes its own prediction, $\hat{p}_F^t \in \mathcal{A}$. At the same time, the environment reveals an outcome y^t in the decision space $\mathcal{Y}$.

In order to learn the experts' weights, an online learning algorithm can be used. A well-established algorithm for prediction with expert advice is the Exponentially Weighted Average Forecaster (EWAF) [18]. In EWAF, the prediction $\hat{p}_F^t$ made by the forecaster is given by (1):

$$\hat{p}_F^t = \frac{\sum_{k=1}^{K} \omega_k^{t-1} p_k^t}{\sum_{k=1}^{K} \omega_k^{t-1}}. \tag{1}$$

At the end of each round, the forecaster and each of the experts receive a non-negative loss based on the outcome y^t revealed by the environment (ℓ_F^t and ℓ_k^t respectively):

$$\ell_F^t, \ell_k^t : \mathcal{A} \times \mathcal{Y} \to \mathbb{R} \tag{2}$$

The weights $\omega_1^t, \ldots, \omega_K^t$ of each expert k are then updated according to the loss incurred by each expert as shown in (3).

$$\omega_k^t = \omega_k^{t-1} e^{-\eta \ell_k^t} \tag{3}$$

After T rounds, by setting:

$$\eta = \sqrt{8 \log \frac{K}{T}} \tag{4}$$

it can be shown that the forecaster's *regret* for not following the best expert's advice is bounded as follows:

$$\sum_{t=1}^{T} \ell_F^t - \min_{k=1,\ldots,K} \sum_{t=1}^{T} \ell_k^t \leq \sqrt{\frac{T}{2} \log K} \tag{5}$$

i.e., the forecaster quickly converges to the performance of the best expert [18].

3.2 CBIR late fusion with expert advice

We frame the late fusion of multiple CBIR systems as a problem of prediction with expert advice: given an ensemble of K CBIR systems, each system corresponds to an expert $k = 1, \ldots, K$, associated with a weight ω_k (all experts start with the same weight, $\omega_k = \frac{1}{K}$); all

the possible sets of images that can be retrieved (from the database of images of each expert) correspond to the decision space $\mathcal{A}$; the late fusion of the CBIR systems thus corresponds to the forecaster, i.e., the forecaster's decision is the final set of images to be retrieved, combining images from multiple systems in the ensemble.

An overview of the learning process is depicted in Fig. 3 and Algorithm 1, and can be summed up as follows. At each round t, a query image q_t is given as input to all the experts $m_1, \ldots, m_K$, and each returns the most similar ones to the query according to its descriptor(s), $retrieved_k^t$ (line 5). Based on the experts' selections, the forecaster selects the final set of retrieved images $queryResult^t$ (line 7). Both the forecaster's and each expert's set of images are then evaluated with a quality score, reflecting how similar the query image and the ones retrieved by each expert are according to one (or more) human evaluators (e.g., a user searching for similar images of their dog in a searching engine, or a doctor using a system to obtain medical examinations similar to those of their patient or a specific diagnosis) (line 9). This quality score is used at the end of the round to update the experts' weights (lines 11–12).

Algorithm 1 CBIR with expert advice.

Input: CBIR systems $\mathcal{M} = m_1, \ldots, m_K$, stream of query images $\mathcal{Q} = q_1, \ldots, q_T$, number of images to retrieve N, image database $\mathcal{D}$

1: $\omega_1^1, \ldots, \omega_K^1 \leftarrow 1/K$
2: **for each** $q_t \in \mathcal{Q}$ **do**
3: $\quad queryResult^t \leftarrow \emptyset$
4: $\quad$ **for each** $m_k \in \mathcal{M}$ **do**
5: $\quad\quad retrieved_k^t \leftarrow m_k.retrieveRelevant(q_t, \mathcal{D}, N)$
6: $\quad\quad I_k^t = \lfloor \omega_k^t * N \rceil$
7: $\quad\quad queryResult^t \leftarrow queryResult^t \cup head(retrieved_k^t, I_k^t)$
8: $\quad$ **end for**
9: $\quad score^t, score_1^t, \ldots, score_K^t \leftarrow askHumanForFeedback(queryResult^t)$
10: $\quad$ **for each** $m_k \in \mathcal{M}$ **do**
11: $\quad\quad \ell_k^t \leftarrow 1 - score_k^t$
12: $\quad\quad \omega_k^{t+1} \leftarrow m_k.updateWeight(\ell_k^t)$
13: $\quad$ **end for**
14: $\quad t \leftarrow t + 1$
15: **end for**

In this paper, we propose to late fuse the images retrieved by the experts in the ensemble by taking into account the weights learned using EWAF. Each expert contributes with a number of images I_k^t proportional to its weight (line 6), as shown in (6). Note that merging the results from multiple experts deviates from the traditional EWAF formulation, in which only a single expert is selected by the forecaster at a time (since prediction with expert advice typically deals with single-value predictions).

$$I_k^t = \lfloor \omega_k * N \rceil \tag{6}$$

We sort the images in each retrieved set (either the forecaster's or each of the experts' sets) in ascending order according to their distance to the query image (the smaller the distance, the more similar the images are; a perfect match corresponds to a distance of 0). In order to avoid duplicates, we skip images that have already been added to the set to be retrieved by some expert. We consider first the images retrieved by the experts with a lower weight (i.e.,

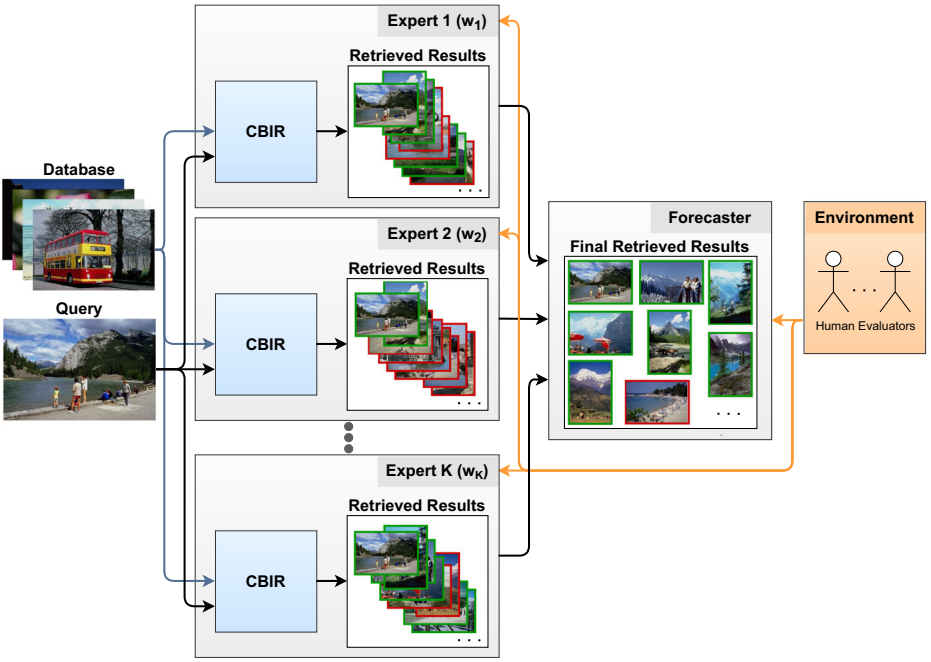


Fig. 3 Overview of our late fusion of multiple CBIR systems with expert advice. (figure best seen in color)

the one that performed worse). This way, we ensure that each expert k contributes with I_k^t images.

A key condition for applying online learning is the availability of feedback (which, in the case of EWAF, is based on the outcome of the environment). To simulate the feedback from human evaluators in a real-world scenario, we used the images' category present in the datasets, curated by human annotators, as a feedback source to compute the loss and update the weight of each CBIR system. In other words, images belonging to the same category are considered as relevant for the remaining images within that category. We thus compute the loss for each expert k at a round t as:

$$\ell_k^t = 1 - sim(relevant^t, retrieved_k^t) \tag{7}$$

where sim is computed using a set similarity measure that allows us to quantify how similar the sets of relevant images $relevant^t$ and retrieved images $retrieved_k^t$ are for each expert k (the set similarity measures experimented are listed in Section 5.2). The weights of all the CBIR systems are then updated based on the loss received, according to (3).

Note that although we rely on the notion of relevance feedback from the user to learn the weights, we did not follow the traditional setup, i.e., we did not refine the query by iteratively asking which images are relevant until the user is satisfied with the results retrieved. We only need input once.

In Fig. 4, we present a snapshot of the learning process, in which we consider the same query image as in Fig. 3, three CBIR experts, $numResults^t = 8$, and the Jaccard index to

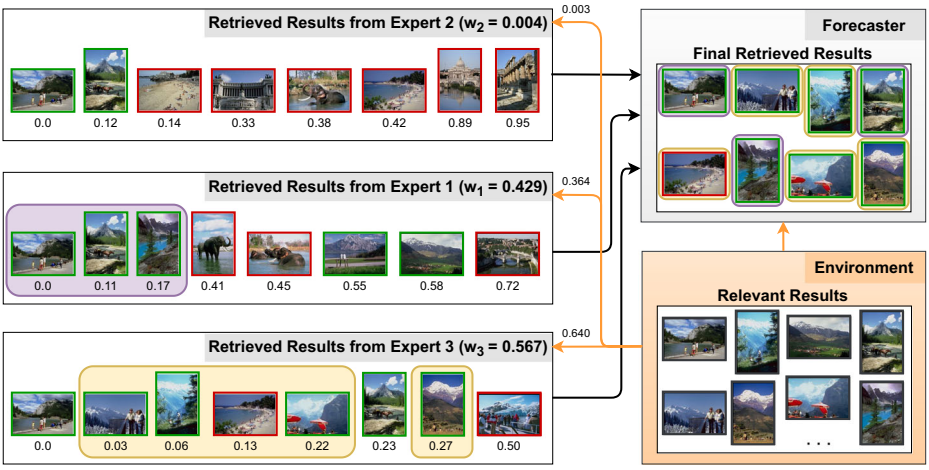


Fig. 4 Example of an iteration using our online CBIR setting. The orange arrows represent the new weight for each expert. (figure best seen in color)

compute the similarity between the sets (see Section. 5.2). Following (6) and considering the weights assigned to each expert, Expert 3 will contribute with five images to the final set, Expert 1 with three images, and Expert 2 with none, excluding possible duplicates. If we had considered the traditional EWAF setting in which only one expert is selected by the forecaster, we would have a precision of 75% (6 out of 8 successfully relevant images retrieved). By late fusing the retrieved images from several CBIR experts, it increases to 87.5% (7 out of 8). Moreover, if there is an expert that clearly outperforms the others, its weight would converge to 1, leading to EWAF's traditional behavior of choosing only one expert.

4 Implementation details

We created two late fusion CBIR solutions based on expert advice. In the first solution, `ExpertosLF_V`, we considered four CBIR systems as experts representing low-level information (color, shape, texture, and joint). The first three ones represent the early fusion of either color, shape, or texture descriptors alone. The joint one represents the early fusion of three existing descriptors that already encompass multiple visual characteristics: color, shape, and texture. In the second one, `ExpertosLF_VS`, we added a fifth CBIR expert that represents the semantic information.

One of our goals is to evaluate whether the experts in the resulting ensemble are plausible for the domain in hand. We focused on low- and high-level descriptors in order to study how the use of different kinds of information (visual or semantic) varies across the different domains, and whether the resulting ensemble reflects it.

Each CBIR follows the typical architecture of a retrieval system with a **Database** to store the images, a **Descriptor Extraction** module, a **Descriptors DB** (indexing structure), and a **Similarity Comparison** algorithm. Following, we present the list of descriptors under consideration (Section 4.1), and detail each component of the CBIR architecture (Sections 4.2 and 4.3).

4.1 Low and High-level descriptors

We selected a diverse set of low-level descriptors representing color (`Auto Color Correlogram (ACC)` [38], `Color Histogram (CH)` [89], `Itten Contrasts (IC)` [41], `Opponent Histogram (OH)` [105], and `Reference Color Similarity (RCS)` [48]), shape (`Edge Histogram (EH)` [19] and `Edges`), texture (`Tamura` [96] and `Haralick` [37]), joint color, shape, and texture information (`Color and Edge Directive Descriptor (CEDD)` [20], `Joint Composite Descriptor (JCD)` [22], and `Fuzzy Color and Texture Histogram (FCTH)` [21]), as well as high-level descriptors that represent images' semantic content using tags (`Adjective-Noun Pairs (ANP)`, `Adjectives`, `Nouns`, and `General Concepts (GC)`) (see Tables 1 and 2).

The majority of the aforementioned descriptors were computed using jFeatureLib [35] and LIRE [61]. The remaining descriptors were implemented by us. The `Edges` descriptor can be seen as a simplified version of `EH`, where the number of edges of an image along vertical, horizontal, 35°, 135°, non-directional, and all directions are counted and represented as a descriptor. Our implementation of the `IC` follows the details presented in [41].

We used SentiBank [15] to extract ANPs. Each image was annotated with the 10 ANPs with the highest probability. The `ANPs` descriptor has a dimensionality of 2089 (i.e., the number of pairs that can be identified by SentiBank), and the probability for each ANP was used as a feature. For each image, we divided each of the 10 ANP into adjective and noun, and computed the average of the probabilities (considering how many times each adjective or noun occurs in the image) to create the `Adjectives` and `Nouns` descriptors. Finally, the semantic tags were obtained automatically, using the Clarify API deep learning pre-trained General model [117], to avoid relying on human-generated tags. The General model computes the probability of the presence of relevant general concepts in the image. It is able to identify over 11,000 concepts within an image, but the set of possible concepts is not known a priori. Furthermore, each image can only be annotated with at most 200 concepts. To devise the final set of most relevant concepts among the possible 11,000 general concepts identified by the model, we annotated each image from all the datasets used in our study (with 200 concepts). The probability given to each concept for each image is used as a feature to create the `GC` descriptor.

4.2 Descriptor Extraction

Before computing any descriptor, we first resized each image to a maximum of 400 pixels on their larger dimension (width or height), keeping the original aspect ratio. Additionally, images of Digital Imaging and Communications in Medicine (DICOM) format were converted to the RGB color space (PNG or JPEG format).

Since we did not know a priori which were the best descriptors within each category, we conducted some preliminary tests. We started by testing each descriptor individually. All visual descriptors proved to be relevant for at least one dataset used, whereas the `GC` descriptor was consistently better than the remaining high-level descriptors (even when we combined it with the remaining high-level descriptors, those descriptors did not improve the system performance comparing to using solely `GC` for any of the datasets).

As mentioned earlier, each type of descriptor usually captures only one aspect of an image property. Thus, there is no single "best" descriptor that leads to accurate results

regardless of the setting, which means that a combination of descriptors is usually needed to provide adequate retrieval results [89]. As such, we tested multiple early fused combinations of descriptors within each category (using the min-max normalization of each feature before fusing them). Given the high dimensionality of some descriptors, we applied Principal Component Analysis (PCA) to reduce their dimensionality in all the tests performed. We tested a different number of principal components to be able to account from 90 to 100% of the variance. The best result was achieved for 90%. Besides ensuring that the system

Table 1 Summary of the color, shape, texture, and joint descriptors selected to study. The column '#' indicates the feature vector length

	Feature	#	Short Description
Color	Auto Color Correlogram	1024	Color histogram combined with spatial correlation between identical colors
	Color Histogram	128	HSB Color histogram with 8 bins for hue, 4 for saturation and 4 for brightness.
	Itten Contrasts	14	Histogram for saturation (avg., low, middle, and high), lightness (avg., very light, light, middle, dark, and very dark), hue (avg., warm and cold), and contrast.
	Opponent Histogram	64	Combination of 1D histograms based on the channels of the opponent color space, where O_1 and O_2 represent color information, while O_3 represents intensity.
	Reference Color Similarity	77	Average pixel color similarity of 77 colors spaced evenly in HSV color space (18 hues with 100% and 50% each in saturation and brightness, plus 5 gray values).
Shape	Edge Histogram	80	5-bin histogram counting edges in vertical, horizontal, 35°, 135°, and non-directional directions (image divided into 16-equal-sized, non-overlapping blocks).
	Edges	6	Number of edges along vertical, horizontal, 35°, 135°, non-directional, and all.
Texture	Haralick	14	Relative frequency distribution that describes how often one gray tone will appear in a specific spatial relationship to another gray tone on the image.
	Tamura	18	Histogram for coarseness, contrast, and directionality.
Joint	Color and Edge Directive Descriptor	144	Histogram that is constituted by 6 regions determined by texture, where each region is constituted by 24 individual HSV fuzzy color regions.
	Fuzzy Color and Texture Histogram	192	Histogram constituted by 8 regions determined by texture (Haar Wavelet), where which region is constituted by 24 individual regions resulting from the combination of YIQ and HSV color fuzzy systems.
	Joint Composite Descriptor	168	Combines CEDD and FCTH. It is made up of 7 texture areas, with each area made up of 24 color regions.

Table 2 Summary of the semantic descriptors selected to study. The column '#' indicates the feature vector length

	Feature	#	Short Description
Tags	Adjective-Noun Pairs	2089	Computes the probability for each possible adjective-noun pair.
	Adjectives	231	Computes the probability of the most relevant adjectives from adjective-noun pairs.
	Nouns	424	Computes the probability of the most relevant nouns from adjective-noun pairs.
	General Concepts	7786	Probability of the presence of relevant general concepts (objects, moods, etc.)

was usable in a timely manner, we also improved the discriminative power of the resulting vector.

4.3 Descriptors DB and similarity comparison

Indexing and Searching plays a fundamental role in Information Retrieval when dealing efficiently with large collections of data (in our case, image descriptors). Thus, we used NB-Tree [34], an efficient indexing structure for high–dimensional data points, which exhibits low insertion and searching times. For the similarity comparison between the descriptor(s) computed for the query image and those available at the descriptors database, we used the k-Nearest Neighbors (kNN) query provided by the NB-Tree. kNN is commonly used in content–based retrieval, and we chose this implementation since it takes advantage of the indexing structure, which was optimized for high–dimensional data points. As expected, the smaller the distance between the descriptor(s) of the query image and the descriptor(s) of each retrieved image, the more similar the images are. Note that the NB-Tree was used both in our dynamic late-fusion CBIR solutions and in the state of the art early and late-fusion techniques (used for comparison in the experimental tests performed), as the kNN algorithm and indexing structure.

5 Experimental setup

In this section, we present our experimental setup, in terms of: 1) the datasets in which we performed our experiments; 2) the similarity measures tested as a loss function for the update of experts' weights; 3) the state of the art early and late fusion techniques under comparison; 4) the evaluation metrics used to report and analyze the performance of each retrieval system; 5) the computing infrastructure in which we ran our experiments.

5.1 Datasets

Experiments were conducted on 13 benchmark datasets divided into three main categories: Biomedical, Real, and Sketch (see Table 3).

In the Biomedical category, we used BRAINCE-MRI, BREAKHIS, COVID19-RX, HAM10000, IRMA, and PLANTPATHOLOGY datasets. They depict brain and breast tumors, COVID-19, bone fractures, pneumonia, pigmented skin lesions, and leaf diseases (see Fig. 5). BRAINCE-MRI contains 3064 T1-weighted contrast-inhanced images of three types

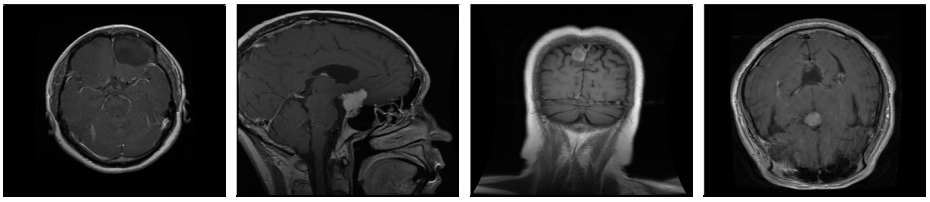
of brain tumor: glioma, meningioma, and pituitary. BREAKHIS contains 7909 Histopathological images of benign breast tumors (adenosis, fibroadenoma, phyllodes tumor, and tubular adenona), and breast cancer (carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma). COVID19-RX contains 3886 chest X-ray images for COVID-19 positive cases along with normal and viral Pneumonia images. HAM10000 contains 10015 multi-source dermatoscopic images of pigmented skin lesions: actinic keratoses and intraepithelial carcinoma/bowen's disease, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanocytic nevi, melanoma, and vascular lesions. IRMA, from ImageCLEF initiative, contains 14410 images of scanned X-rays of various human body parts. PLANTPATHOLOGY contains 1821 high-quality, real-life symptom images of apple foliar diseases, with variable illumination, angles, surfaces, and noise.

In the Real category, we used COPYDAYS, COREL1K, COREL10K, and GHIM10K datasets, which depict realistic images of diverse aspects of everyday life (see Fig. 6). COPYDAYS contains 3212 personal holidays photos that were artificially manipulated (cropped, scaled, and strongly attacked). COREL1K contains 1000 images depicting African people, beaches, buildings, buses, dinosaurs, elephants, flowers, foods, horses, and mountains. COREL10K contains 10000 images representing buildings, sunsets, fish, flowers, cars, mountains, tigers, etc. GHIM10K contains 10000 images depicting cars, insects, mountains, ships, sunsets, etc.

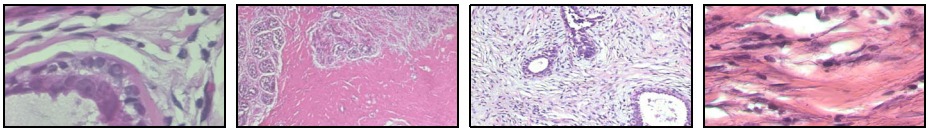
In the Sketch category, we used \$P, IMISKETCHS, and MCALI (see Fig. 7). These datasets have a large variety in the types of symbols represented (e.g. digits, furniture, mathematical, smiles), and the way they were drawn. \$P contains 4802 images, drawn by 10 users, with gestures (multi-stroke without rotation) representing geometric shapes, letters or symbols; IMISKETCHS contains 1871 images of furniture symbols (e.g., doors, or tables) drawn with multi-stroke and rotation; MCALI contains 8159 symbols, drawn by 17 users, with gestures (with multi-stroke and rotation) representing geometric shapes, smiles, generic symbols and letters.

Table 3 Summary of the datasets

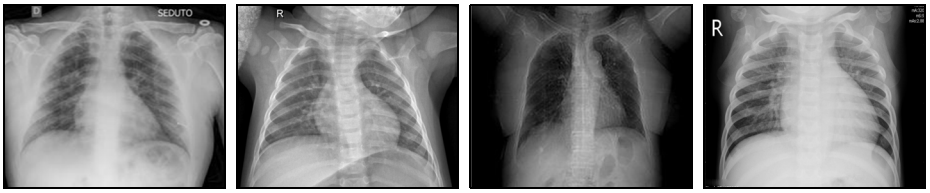
	Ref.	Dataset	#Images	#Categories	#Relevant
Biomedical	[23]	BRAINCE-MRI	3064	3	708 - 1426
	[90, 91]	BREAKHIS	7909	8	444 - 3451
	[24]	COVID19-RX	3886	3	1200 - 1345
	[102]	HAM10000	10015	7	115 - 6705
	[68]	IRMA	14410	193	1 - 2343
	[99]	PLANTPATHOLOGY	1821	4	91 - 622
Real	[43]	COPYDAYS	3212	157	20 - 24
	[55, 111]	COREL1K	1000	10	100
	[59]	COREL10K	10000	100	97 - 103
	[59]	GHIM10K	10000	20	500
Sketch	[108]	\$P	4802	16	299 - 301
	[40]	IMISKETCHS	1871	13	43 - 372
	[109]	MCALI	8159	24	339 - 340



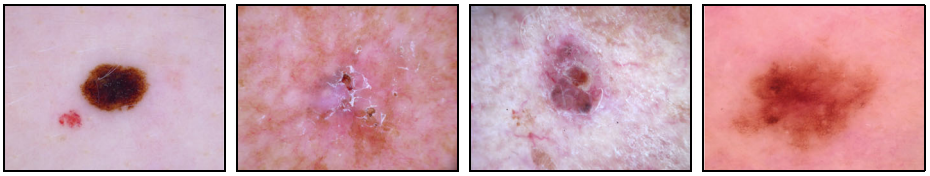
(a) BrainCE-MRI.



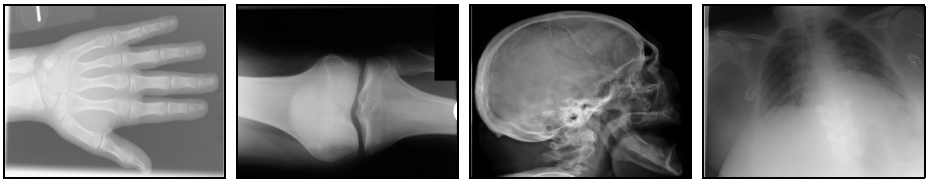
(b) BreakHis



(c) COVID19-Rx



(d) HAM10000



(e) IRMA

(f) PlantPathology

Fig. 5 Example images from the Biomedical datasets. (figure best seen in color)

5.2 Similarity Measures

In order to compute the loss for each expert in our late fusion solution with expert advice, we tested four measures to quantify how similar the sets of relevant and retrieved images

are: Jaccard index [42] (8), Otsuka-Ochiai coefficient [70, 72] (9), Overlap coefficient [74] (10), and Sørensen-Dice index [31, 95] (11).

$$jaccard = \frac{|relevant \cap retrieved|}{|relevant| + |retrieved| - |relevant \cap retrieved|} \tag{8}$$

$$otsukaOchiai = \frac{|relevant \cap retrieved|}{\sqrt{\Big(|relevant| * |retrieved|\Big)}} \tag{9}$$

$$overlap = \frac{|relevant \cap retrieved|}{min\Big(|relevant|, |retrieved|\Big)} \tag{10}$$

$$sorensenDice = \frac{2\,|relevant \cap retrieved|}{|relevant| + |retrieved|} \tag{11}$$

Sørensen-Dice and Jaccard are more rigid measures since they penalize the existence of more retrieved images than relevant ones, and vice-versa. Otsuka-Ochiai is a less rigid measure than the aforementioned ones since it penalizes less the existence of different cardinalities between the sets of retrieved and relevant images, so it can be seen as a more balanced measure. Finally, Overlap prioritizes the existence of retrieved images that are

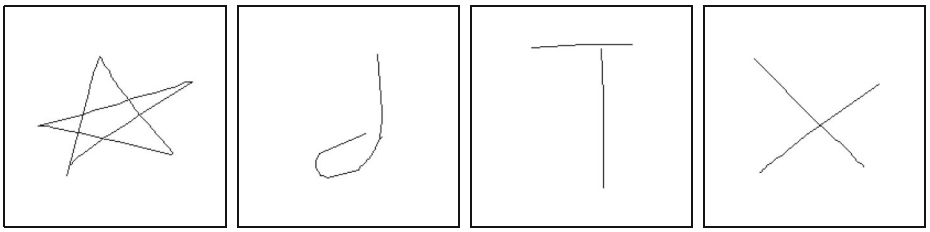
(a) CopyDays

(b) COREL1K

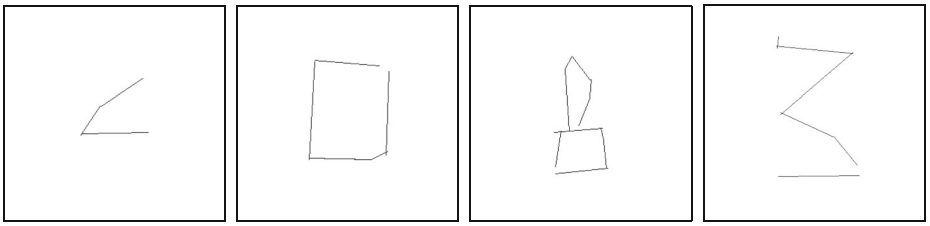
(c) COREL10K

(d) GHIM10K

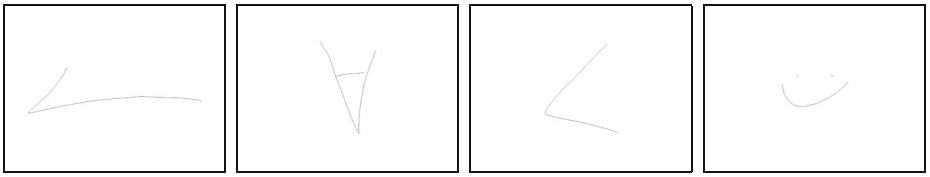
Fig. 6 Example images from the Real datasets. (figure best seen in color)



(a) $P



(b) IMISKETCHS



(c) MCALI

Fig. 7 Example images from the Sketch datasets. (figure best seen in color)

relevant, even though the cardinality of the sets differs, thus being the least restrictive measure.

5.3 State of the art fusion techniques

To assess the quality of the ensemble of CBIR systems produced by our technique, we compared it to well-known state of the art fusion techniques.

The first technique is the widely used early fusion of the descriptors, followed by PCA to reduce the high dimensionality of the resulting feature vector (*EF*). Since we wanted to ensure that our technique performs as well as early fusion, we used *EF* as a baseline in our work. Following, we considered two late fusion techniques to combine the results of multiple CBIR systems. Let D be the set of all images in the database, and i an image from this set ($i \in D$). Each CBIR j returns a list of the most similar images to the query one (L_j), where each image i contains its normalized similarity score (representing how similar it is to the query image) denoted as $S_j(i)$. The goal of each late fusion method is to produce a final ranked list (L_f).

Late fusion techniques are usually split into two major groups: (i) order based rank list fusion and (ii) similarity score-based rank list fusion. For the first group, we implemented a method based on the frequency of occurrence of each image in L_j (*FreqRankLF*) [62]. L_f is sorted by descending order of the frequency of images. For the second one, we implemented a method based on the similarity score (*SimRankLF*) [69]. The scores $S_j(i)$ are arranged in

ascending order and the final list L_f is generated. If an image is present in more than one L_j, the lowest score is considered in the merging process.

5.4 Retrieval metrics

To evaluate the performance of our approach, we used precision (12), recall (13), F_1 score (14), and AveP (15). Precision allows us to identify the percentage of retrieved images that are relevant, while recall allows us to obtain the percentage of relevant images that are successfully retrieved. F_1 score combines both precision and recall measures in a balanced way (i.e., both metrics are evenly weighted). AveP evaluates whether all of the relevant images retrieved are ranked higher (or not).

All the results are reported in terms of the average precision at the top-10 retrieved images ($avgP@10$), average F_1 score ($avgF_1$), and the mean Average Precision (mAP). For each query, the number of retrieved results is set to be equal to the number of relevant images for that query. Images belonging to the same category within the same dataset are considered as relevant. With this setup, precision and recall are equal to the F_1 score, thus we did not report them individually.

$$precision = \frac{|relevant \cap retrieved|}{|retrieved|} \tag{12}$$

$$recall = \frac{|relevant \cap retrieved|}{|relevant|} \tag{13}$$

$$avg\,\mathrm{F}_1 = \frac{\sum_{q=1}^{Q} \mathrm{F}_1(q)}{Q}, \mathrm{F}_1 = \frac{2 * precision * recall}{precision + recall} \tag{14}$$

$$mAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}, AveP = \frac{\sum_{k=1}^{n} \left(precision(k) * relevant(k)\right)}{|relevant|} \tag{15}$$

5.5 Computing infrastructure

All the experiments were carried out on the same PC, running Arch Linux 5.12.7, with an Intel Core i7-8700 3.20GHz CPU, 64GB of memory, and two GeForce RTX 2080 8GB GDDR6.

6 Experimental results

In this section, we report the performance of our expert-based solutions. Recall that the `ExpertosLF_V` encompasses four CBIR systems as experts: `color` (`ACC`, `CH`, `IC`, `OH`, and `RCS`), `shape` (`EH` and `Edges`), `texture` (`Tamura` and `Haralick`), and `joint` (`CEDD`, `FCTH`, and `JCD` - combines color, shape, and texture information into a single expert). in `ExpertosLF_VS`, a fifth CBIR expert was added to represent the `semantic` information (`GC`). We considered the similarity measures presented in Section 5.2 to update the experts' weights. We report the best one for each solution, although overall the difference between the different measures is negligible.

We compared the performance of each solution to the following state of the art fusion techniques: a) early fusion of all the descriptors that compose our experts (*EF*); b) late fusion of the experts' results using frequency (*FreqRankLF*); c) late fusion of the experts'

results using similarity (*SimRankLF*). we analysed our expert-based solutions in an online setting, in which we assumed user feedback is always available (Section 6.1), and in an offline setting, in which we assumed that feedback is no longer available (Section 6.2).

6.1 Online setting

We started by shuffling each dataset and randomly select 1000 images to be used as queries. After each query, the weights of each CBIR expert in each expert-based CBIR solution were updated according to how relevant the retrieved images were.

6.1.1 Biomedical

In Tables 4 and 5, we present, respectively, the results for the `ExpertosLF_V` and `ExpertosLF_VS` solutions, and the best CBIR expert in each ensemble, and the aforementioned state of the art techniques, for comparison purposes.

For `ExpertosLF_V`, the best individual CBIR systems are `shape` and `color`. For the majority of the datasets, `ExpertosLF_V` performs as well as the best individual CBIR, outperforming it in the HAM10000 dataset. When we include the `semantic` CBIR in the ensemble, the results achieved by `ExpertosLF_VS` are very similar to the ones obtained by `ExpertosLF_V`. The only exception is the PLANTPATHOLOGY dataset, which benefits from using the `semantic` expert. SorensenDice metric shows better results for `ExpertosLF_V`, while Overlap is equally useful when we add semantic information.

In Fig. 8, for each dataset, we present the evolution of the weights for each expert in the `ExpertosLF_V` and `ExpertosLF_VS` solutions, and the evolution of the F_1 over the queries. Overall, both solutions converge quickly to the best individual expert. An interesting exception is COVID19-RX, for which `ExpertosLF_VS` converges to a combination of three experts: `shape`, `joint`, and `semantic`. With this, we are able to achieve better results than the ones provided by the best expert individually (an increase of ≈ 0.02 for $avgF_1$ and mAP).

To compare the different fusion techniques per domain, we report the difference between our technique and the remaining ones (Δ), averaged across the datasets of each domain. `ExpertosLF_V` and `ExpertosLF_VS` usually return more relevant images at the top of the retrieved set of images. This is also supported by an $avgP@10$ better than the $avgF_1$[1], with an increase varying from 0.14 (BRAINCE-MRI) to 0.41 (BREAKHIS). The early fusion of all experts (*EF*) is slightly better than ours when focusing on the top ten retrieved images ($\Delta avgP@10 = -0.04 \pm 0.07$); while for the remaining metrics, they perform similarly ($\Delta avgF_1 = 0.01 \pm 0.02$, $\Delta mAP = 0.00 \pm 0.02$). The performance of our proposed technique surpasses both *FreqRankLF* ($\Delta avgP@10 = 0.48 \pm 0.09$, $\Delta avgF_1 = 0.05 \pm 0.04$, $\Delta mAP = 0.07 \pm 0.06$) and *SimRankLF* ($\Delta avgP@10 = 0.11 \pm 0.09$, $\Delta avgF_1 = 0.06 \pm 0.04$, $\Delta mAP = 0.07 \pm 0.06$).

6.1.2 Real

The results for `ExpertosLF_V` and `ExpertosLF_VS` solutions, best CBIR expert in each ensemble, and fusion techniques are presented in Tables 6 and 7.

[1]Notice that, in our scenario, $avgF_1$ would be equal to the $avgPrecision$ and $avgRecall$.

Table 4 Results for Biomedical datasets using only visual descriptors. Bold values indicate the best results

Dataset	Method	Metric	avgP@10	avgF_1	mAP
BrainCE-MRI	BestExpert (texture)	-	0.67	**0.53**	**0.35**
	EF	-	**0.83**	0.49	0.33
	FreqRankLF	-	0.54	0.50	0.28
	SimRankLF	-	0.60	0.46	0.26
	ExpertosLF_V	Jaccard	0.67	**0.53**	**0.35**
BreakHis	BestExpert (color)	-	**0.78**	**0.36**	**0.21**
	EF	-	0.75	**0.36**	0.20
	FreqRankLF	-	0.07	0.35	0.17
	SimRankLF	-	0.53	0.27	0.11
	ExpertosLF_V	SorensenDice	0.77	**0.36**	**0.21**
COVID19-Rx	BestExpert (shape)	-	0.87	**0.61**	**0.46**
	EF	-	**0.90**	0.59	0.45
	FreqRankLF	-	0.34	0.54	0.40
	SimRankLF	-	0.81	0.55	0.37
	ExpertosLF_V	SorensenDice	0.87	**0.61**	**0.46**
HAM10000	BestExpert (color)	-	0.68	0.52	0.38
	EF	-	**0.71**	**0.53**	0.38
	FreqRankLF	-	0.04	0.50	0.23
	SimRankLF	-	0.68	0.52	**0.39**
	ExpertosLF_V	Overlap	0.70	0.52	0.37
IRMA	BestExpert (shape)	-	0.68	**0.39**	0.28
	EF	-	**0.69**	**0.39**	**0.30**
	FreqRankLF	-	0.05	0.22	0.09
	SimRankLF	-	0.46	0.28	0.16
	ExpertosLF_V	SorensenDice	0.67	**0.39**	0.28
Plant Pathology	BestExpert (shape)	-	0.44	**0.33**	0.13
	EF	-	**0.57**	**0.33**	**0.14**
	FreqRankLF	-	0.31	0.32	0.11
	SimRankLF	-	0.47	0.32	0.12
	ExpertosLF_V	OtsukaOchiai	0.54	**0.33**	0.13

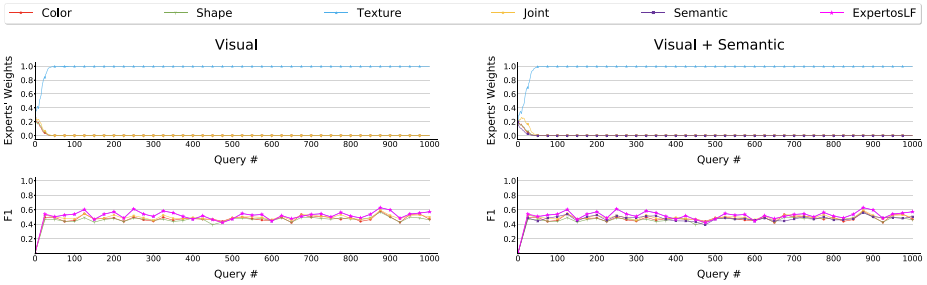
For `ExpertosLF_V`, the best individual CBIR systems are `color` and `joint`. `ExpertosLF_V` performs as well as the best individual CBIR for almost all datasets, and outperforms the individual expert for COREL1K dataset (except for $avgP@10$). When considering semantic information in the ensemble (`ExpertosLF_VS`), `semantic` expert becomes the best CBIR system for COREL1K, COREL10K, and GHIM10K. Overall, Sørensen-Dice and Otsuka-Ochiai similarity measures yield the best results for the majority of the datasets.

Figure 9 depicts the evolution of the weights for each expert, as well as the evolution of the F_1 over queries. `ExpertosLF_V` and `ExpertosLF_VS` quickly converge to the best

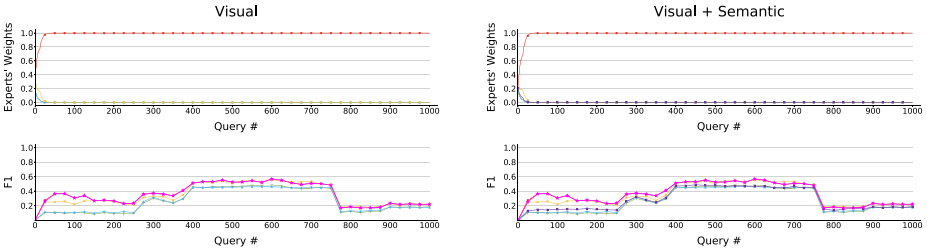
Table 5 Results for Biomedical datasets using visual and semantic descriptors. Bold values indicate the best results

Dataset	Method	Metric	avgP@10	avgF_1	mAP
BRAINCE-MRI	BestExpert (texture)	-	0.67	**0.53**	**0.35**
	EF	-	**0.82**	0.49	0.33
	FreqRankLF	-	0.47	0.48	0.26
	SimRankLF	-	0.60	0.47	0.28
	ExpertosLF_VS	Jaccard	0.67	**0.53**	**0.35**
BREAKHIS	BestExpert (color)	-	**0.78**	**0.36**	**0.21**
	EF	-	0.77	0.34	0.18
	FreqRankLF	-	0.06	0.29	0.11
	SimRankLF	-	0.55	0.27	0.11
	ExpertosLF_VS	SorensenDice	0.77	**0.36**	**0.21**
COVID19-RX	BestExpert (shape)	-	0.87	0.61	0.46
	EF	-	**0.93**	0.62	**0.50**
	FreqRankLF	-	0.48	0.58	0.46
	SimRankLF	-	0.86	0.59	0.44
	ExpertosLF_VS	Overlap	0.85	**0.63**	0.48
HAM10000	BestExpert (color)	-	0.68	**0.53**	0.37
	EF	-	**0.71**	**0.53**	**0.38**
	FreqRankLF	-	0.04	0.50	0.24
	SimRankLF	-	0.67	0.51	0.35
	ExpertosLF_VS	Overlap	0.70	0.52	0.37
IRMA	BestExpert (shape)	-	0.68	0.39	0.28
	EF	-	**0.70**	**0.41**	**0.31**
	FreqRankLF	-	0.09	0.30	0.14
	SimRankLF	-	0.49	0.28	0.17
	ExpertosLF_VS	SorensenDice	0.67	0.39	0.28
PLANT PATHOLOGY	BestExpert (semantic)	-	**0.67**	**0.39**	**0.21**
	EF	-	0.63	0.35	0.17
	FreqRankLF	-	0.36	**0.39**	0.17
	SimRankLF	-	0.58	0.37	0.17
	ExpertosLF_VS	OtsukaOchiai	**0.67**	**0.39**	**0.21**

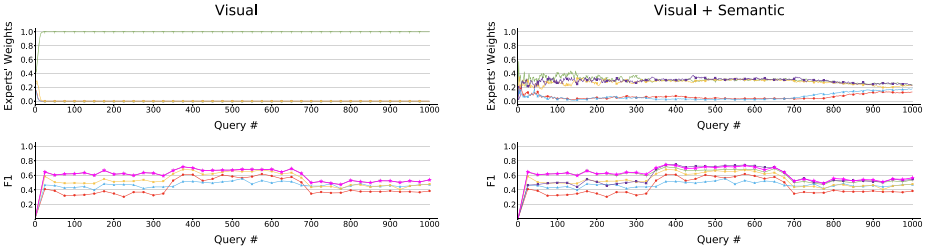
individual expert. The only exception is the COPYDAYS dataset, for which `ExpertosLF_V` converge to a combination of two experts: `color` and `joint`. The result achieved by the combination slightly outperforms the result of the best expert individually (an increase of ≈ 0.04 for both $avgF_1$ and mAP). Overall, our `ExpertosLF_V` and `ExpertosLF_VS` solutions have a very good retrieval performance, in particular, with the inclusion of the `semantic` expert. Many relevant images are successfully retrieved, with more relevant images at the top of the retrieved set ($avgP@10$ is slightly better than the $avgF_1$ for all datasets).



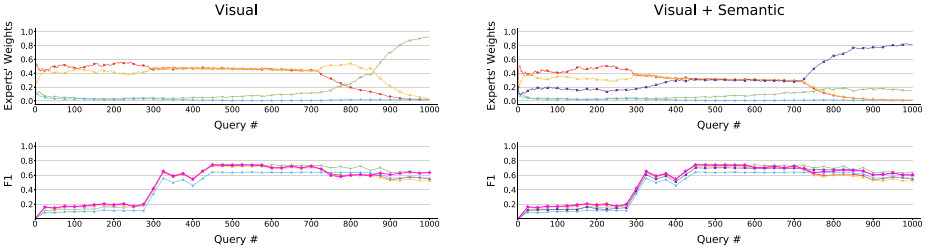
(a) BrainCE-MRI



(b) BreakHis



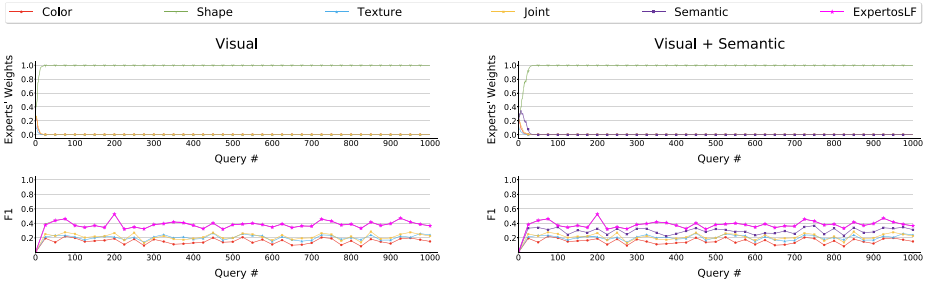
(c) COVID-19Rx



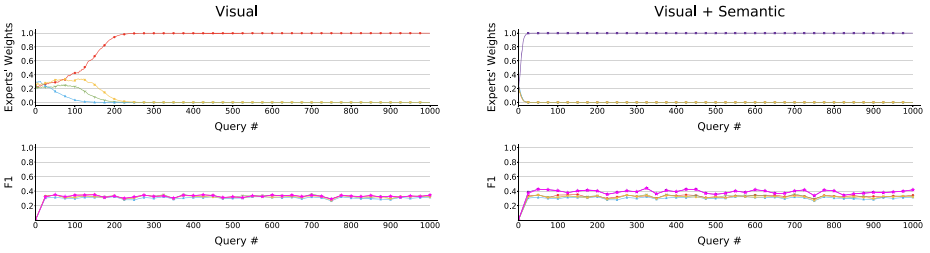
(d) HAM10000

Fig. 8 Evolution of the weights for each expert, and evolution of F_1 over queries. (figure best seen in color)

When considering only visual experts, *EF* performs slightly better than our technique ($\Delta avgP@10 = -0.07 \pm 0.07$, $\Delta avgF_1 = -0.01 \pm 0.03$, $\Delta mAP = -0.03 \pm 0.06$). Considering both visual and semantic experts, on average, our technique surpasses all the remaining techniques for all the metrics. In particular, it is considerably better than *SimRankLF* ($\Delta avgP@10 = 0.39 \pm 0.32$, $\Delta avgF_1 = 0.38 \pm 0.22$, $\Delta mAP = 0.43 \pm 0.25$).



(e) IRMA



(f) PlantPathology

Fig. 8 (continued)

6.1.3 Sketch

In Tables 8 and 9, the results for the `ExpertosLF_V` and `ExpertosLF_VS` solutions, the best CBIR expert in each ensemble, and the state of the art fusion techniques are presented.

Considering only visual descriptors, the best individual CBIR systems are `shape` and `joint`. For \$P and MCALI datasets, `ExpertosLF_V` performs as well as the best individual CBIR, while for IMISKETCHS it outperforms the best individual expert. All datasets benefit from the inclusion of `semantic` information in the `ExpertosLF_VS`. Overall, SorensenDice similarity metric shows better results.

In Fig. 10, we present the evolution of the weights for each expert, and the evolution of the F_1 over queries. For the \$P dataset, we can see that the `ExpertosLF_VS` solution converges to a combination of the best experts: `shape` and `semantic`. For IMISKETCHS dataset, `ExpertosLF_V` solution also converges to a combination of the best experts: `shape` and `joint`. Our solutions return more relevant images at the top of the retrieved set of images ($avgP@10$ is always considerably better than the $avgF_1$, with increases varying from 0.147 (IMISKETCHS) to 0.375 (MCALI)).

The early fusion of all experts, *EF*, achieves slightly better results when compared to ours, in particular when considering only visual experts ($\Delta avgP@10 = -0.09 \pm 0.07$, $\Delta avgF_1 = -0.02 \pm 0.02$, $\Delta mAP = -0.03 \pm 0.01$). Compared with the late fusion techniques, our techniques achieves overall better results, being particularly better than *SimRankLF* ($\Delta avgP@10 = 0.38 \pm 0.11$, $\Delta avgF_1 = 0.21 \pm 0.06$, $\Delta mAP = 0.26 \pm 0.10$) when considering both visual and semantic experts.

For all the domains under study, when considering both visual and semantic experts, our late fusion technique achieved, as desirable, similar performance to the baseline ($\Delta avgP@10 = -0.03 \pm 0.05$, $\Delta avgF_1 = 0.01 \pm 0.03$, $\Delta mAP = 0.01 \pm 0.04$).

Table 6 Results for Real datasets using only visual descriptors. Bold values indicate the best results

Dataset	Method	Metric	avgP@10	avgF_1	mAP
	BestExpert (joint)	-	0.88	**0.74**	**0.73**
	EF	-	**0.89**	0.70	0.69
CopyDays	FreqRankLF	-	0.75	**0.74**	0.62
	SimRankLF	-	0.82	0.64	0.60
	ExpertosLF_V	SorensenDice	0.88	**0.74**	**0.73**
	BestExpert (color)	-	0.78	0.49	0.40
	EF	-	**0.83**	**0.55**	**0.46**
COREL1K	FreqRankLF	-	0.27	0.50	0.32
	SimRankLF	-	0.40	0.27	0.14
	ExpertosLF_V	Overlap	0.67	0.53	0.40
	BestExpert (color)	-	0.51	0.22	0.15
	EF	-	**0.59**	**0.27**	**0.20**
COREL10K	FreqRankLF	-	0.12	0.21	0.09
	SimRankLF	-	0.17	0.07	0.03
	ExpertosLF_V	OtsukaOchiai	0.57	0.25	0.11
	BestExpert (joint)	-	0.58	0.25	0.12
	EF	-	**0.66**	**0.27**	**0.14**
GHIM10K	FreqRankLF	-	0.12	0.25	0.10
	SimRankLF	-	0.24	0.12	0.02
	ExpertosLF_V	SorensenDice	0.57	0.25	0.11

We would like to emphasize that our technique is not expected to always achieve better results than the baseline, since it needs a few queries to learn the best ensemble. Yet, our technique achieved better results than the two late fusion techniques tested: *SimRankLF* ($\Delta avgP@10 = 0.25 \pm 0.23$, $\Delta avgF_1 = 0.19 \pm 0.19$, $\Delta mAP = 0.22 \pm 0.21$) and *FreqRankLF* ($\Delta avgP@10 = 0.44 \pm 0.19$, $\Delta avgF_1 = 0.03 \pm 0.04$, $\Delta mAP = 0.09 \pm 0.06$).

6.2 Offline setting

Our late fusion technique relies on the existence of feedback from its users to gauge how well the experts are behaving, but such feedback may not be always available. Thus, we studied how many queries need to receive feedback from users in order to successfully learn the best set of weights to apply them in an offline setting (for a given dataset of images).

For each dataset, we used the set of weights learned in the online setting until learning round X to create the `ExpertosLF_V` and `ExpertosLF_VS` solutions (using our weighted late fusion technique). We tested multiple X values: 25, 50, 75, 100, 125, 250, 500 and 1000. Contrary to what happens in the online setting, the weights used are always the same throughout the offline queries, i.e., they are never updated.

In the offline setting, only the remaining images for each dataset are used as queries (i.e., we did not consider the images used as learning rounds in the online setting): 800 images for CopyDays, ImiSketchS, and PlantPathology, 2000 for BrainCE-MRI, COVID19-Rx, and $P, 6000 for BreakHis and mCali, and 9000 for COREL10K, GHIM10K,

Table 7 Results for Real datasets using visual and semantic descriptors. Bold values indicate the best results

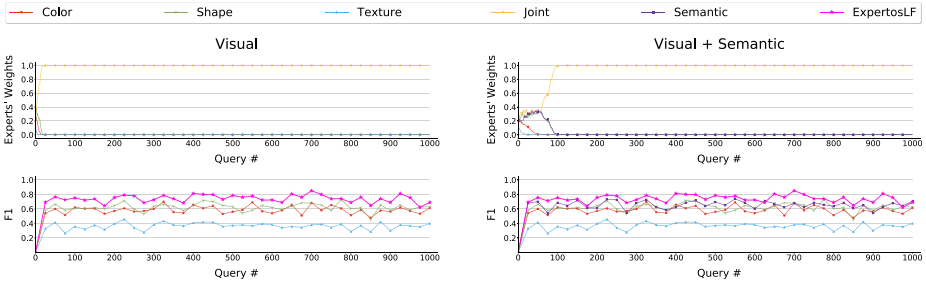
Dataset	Method	Metric	avgP@10	avgF_1	mAP
	BestExpert (joint)	-	0.88	0.74	0.73
	EF	-	**0.91**	**0.77**	**0.75**
CopyDays	FreqRankLF	-	0.65	0.66	0.51
	SimRankLF	-	0.85	0.66	0.63
	ExpertosLF_VS	Jaccard	0.88	0.74	0.73
	BestExpert (semantic)	-	**0.98**	**0.93**	**0.92**
	EF	-	**0.98**	0.92	0.91
COREL1K	FreqRankLF	-	0.86	**0.93**	0.89
	SimRankLF	-	0.78	0.59	0.52
	ExpertosLF_VS	OtsukaOchiai	**0.98**	**0.93**	**0.92**
	BestExpert (semantic)	-	0.87	**0.67**	**0.61**
	EF	-	**0.89**	**0.67**	**0.61**
COREL10K	FreqRankLF	-	0.41	**0.67**	0.52
	SimRankLF	-	0.25	0.17	0.09
	ExpertosLF_VS	SorensenDice	0.86	**0.67**	**0.61**
	BestExpert (semantic)	-	**0.98**	**0.84**	**0.81**
	EF	-	0.97	0.76	0.70
GHIM10K	FreqRankLF	-	0.45	**0.84**	0.72
	SimRankLF	-	0.27	0.25	0.10
	ExpertosLF_VS	SorensenDice	**0.98**	**0.84**	**0.81**

HAM10000, and IRMA datasets. We did not use COREL1K because it only has 1000 images.

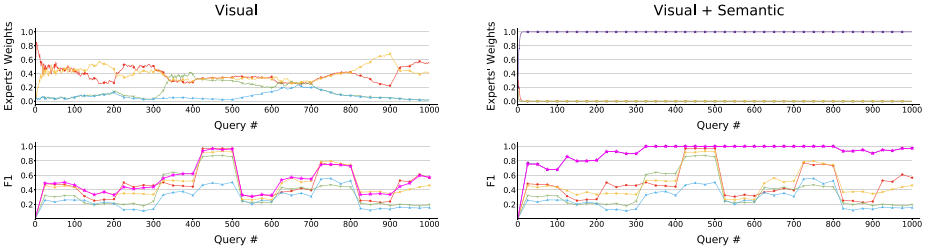
This way, we ensure that the weights learned in the online setting are independent of the queries used in the offline setting. Moreover, it allows us to evaluate the quality of the ensembles learned using different query sizes, in particular, whether the performance deteriorates with the increase of the number of unseen queries.

6.2.1 Biomedical

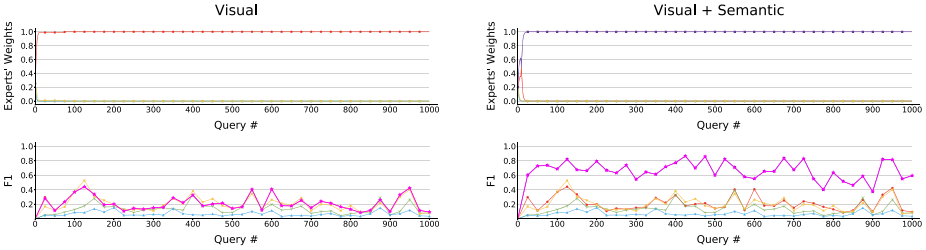
Figure 11 depicts the evolution of the F_1 for each expert individually and both expert-based `ExpertosLF_V` and `ExpertosLF_VS` solutions. For BrainCE-MRI (Fig. 11a), both `ExpertosLF_V` and `ExpertosLF_VS` surpass the best expert performance using the weights learnt up to $X = 25$ and 50, keeping a similar performance after that. For BreakHis (Fig. 11b) and IRMA (see Fig. 11d), both `ExpertosLF_V` and `ExpertosLF_VS` achieve a similar performance as the best expert for $X = 50$. For COVID19-Rx (Fig. 11c), `ExpertosLF_V` performs similarly to the performance of the best expert for $X = 25$. `ExpertosLF_VS` also performs similarly to the best expert for $X = 25$, and it ends up surpassing it (best performance achieved for $X = 100$ with an $avgF_1$ of 0.627 against 0.605 of that expert `shape`). HAM10000 (Fig. 11e) and PlantPathology (Fig. 11f) are the ones for which our solution takes the longest to converge. In HAM10000, `ExpertosLF_VS`



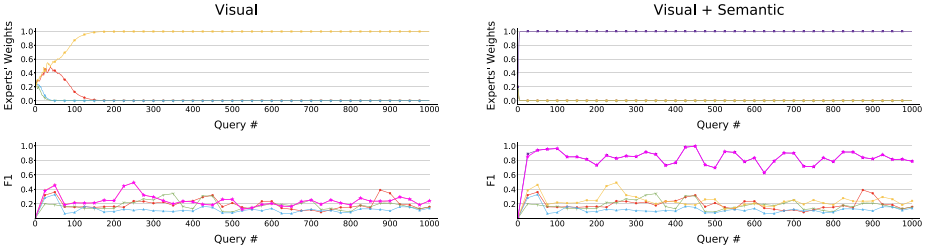
(a) CopyDays



(b) COREL1K



(c) COREL10K



(d) GHIM10K

Fig. 9 Evolution of the weights for each expert, and evolution of F_1 over queries. (figure best seen in color)

does not converge for the best expert performance until $X = 1000$, while for PlantPathology, `ExpertosLF_V` converges at $X = 500$. For the remaining datasets, our solutions converge at $X = 50$ (PlantPathology) and 75 (HAM10000).

Table 8 Results for Sketch datasets using only visual descriptors. Bold values indicate the best results

Dataset	Method	Metric	avgP@10	avgF_1	mAP
	BestExpert (shape)	-	0.94	0.58	0.50
	EF	-	**0.95**	**0.62**	**0.54**
$P	FreqRankLF	-	0.24	0.45	0.28
	SimRankLF	-	0.88	0.44	0.31
	ExpertosLF_V	OtsukaOchiai	0.94	0.58	0.50
	BestExpert (joint)	-	0.43	0.33	0.16
	EF	-	**0.57**	**0.34**	0.19
IMISKETCHS	FreqRankLF	-	0.27	0.33	**0.20**
	SimRankLF	-	0.50	0.29	0.14
	ExpertosLF_V	Overlap	0.48	**0.34**	0.17
	BestExpert (shape)	-	0.52	0.21	0.10
	EF	-	**0.59**	**0.24**	**0.13**
MCALI	FreqRankLF	-	0.06	0.18	0.06
	SimRankLF	-	0.48	0.16	0.06
	ExpertosLF_V	SorensenDice	0.43	0.21	0.10

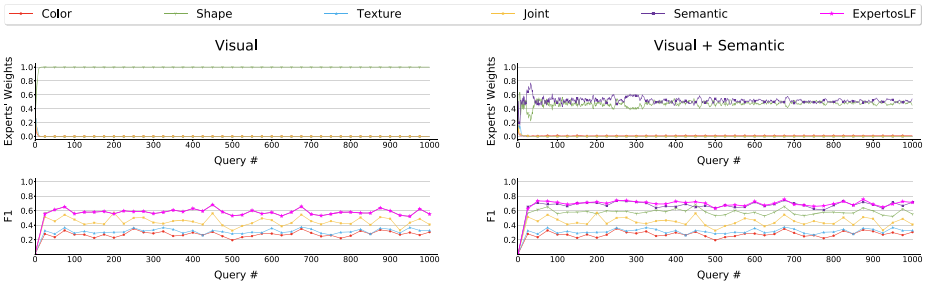
Table 9 Results for Sketch datasets using visual and semantic descriptors. Bold values indicate the best results

Dataset	Method	Metric	avgP@10	avgF_1	mAP
	BestExpert (semantic)	-	**0.97**	0.68	0.61
	EF	-	**0.97**	**0.70**	**0.63**
$P	FreqRankLF	-	0.36	0.68	0.51
	SimRankLF	-	0.62	0.42	0.24
	ExpertosLF_VS	Overlap	0.89	**0.70**	0.61
	BestExpert (semantic)	-	0.74	**0.42**	0.28
	EF	-	**0.75**	**0.42**	**0.29**
IMISKETCHS	FreqRankLF	-	0.33	**0.42**	0.26
	SimRankLF	-	0.36	0.26	0.10
	ExpertosLF_VS	SorensenDice	0.74	**0.42**	0.28
	BestExpert (semantic)	-	0.82	**0.44**	0.33
	EF	-	**0.83**	**0.44**	**0.34**
MCALI	FreqRankLF	-	0.16	**0.44**	0.26
	SimRankLF	-	0.33	0.25	0.11
	ExpertosLF_VS	SorensenDice	0.82	**0.44**	0.33

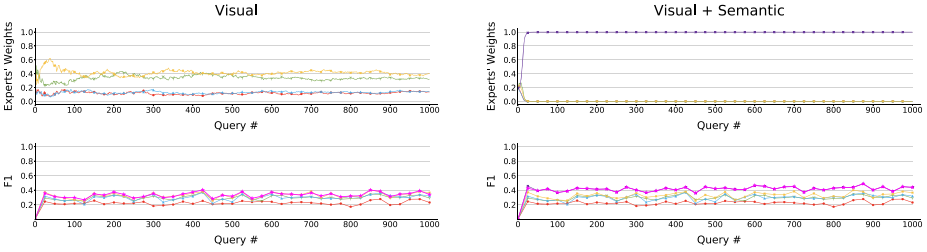
6.2.2 Real

In Fig. 12, we present the evolution of the F_1 for each expert individually, and both expert-based solutions (ExpertosLF_V and ExpertosLF_VS). For all datasets, the weighted ensembles obtained from our expert-based solutions very quickly achieve the performance of the best individual expert.

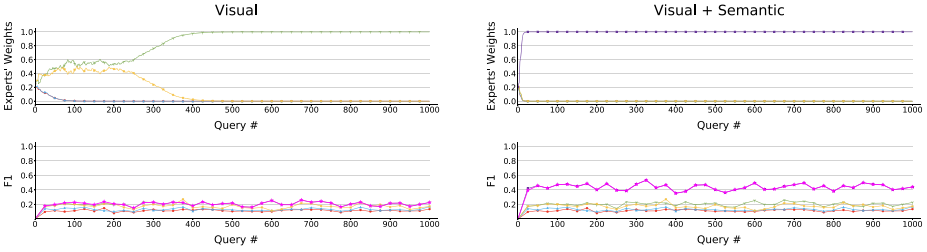
In CopyDays (Fig. 12a) and COREL10K (Fig. 12b), ExpertosLF_V achieves a performance similar to that of the best expert for $X = 25$; for COREL10K, ExpertosLF_VS also achieves a similar performance for $X = 25$, while for CopyDays, it does so for $X = 100$, and it performs even better for $X = 75$. In GHIM10K (Fig. 12c), ExpertosLF_V achieves the best performance for $X = 75$, being marginally better at $X = 150$, while ExpertosLF_VS quickly converges to the best expert at $X = 25$.



(a) $P

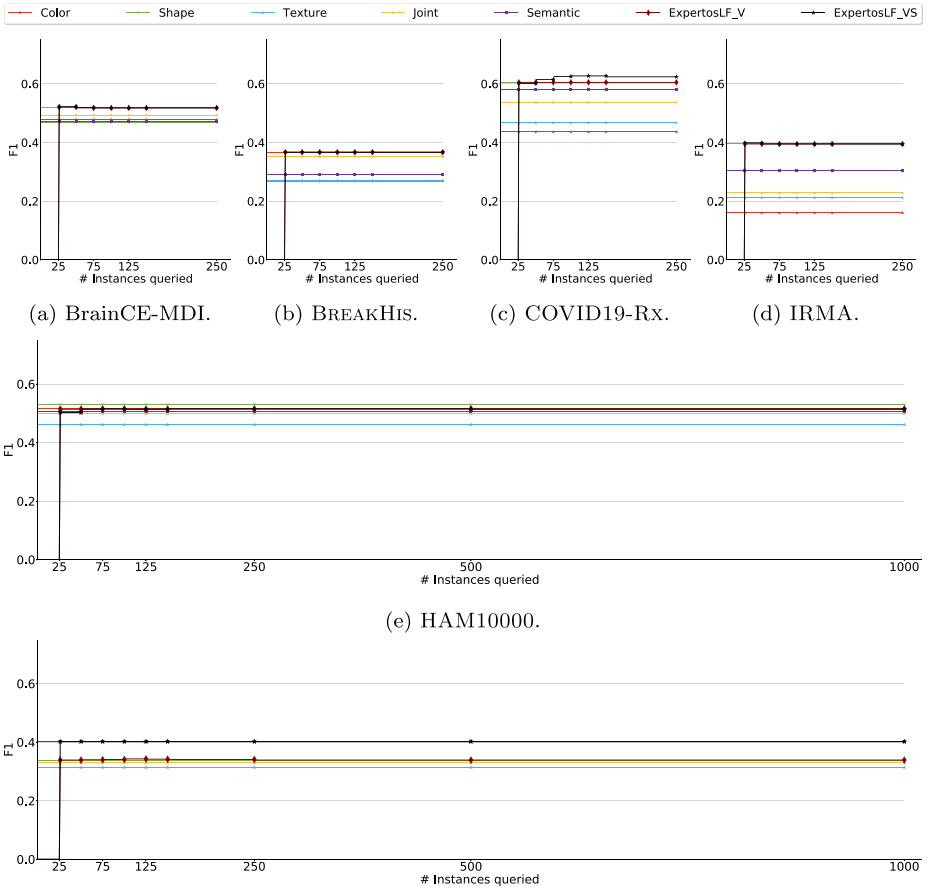


(b) ImiSketchS



(c) mCali

Fig. 10 Evolution of the weights for each expert, and evolution of F_1 over queries. (figure best seen in color)

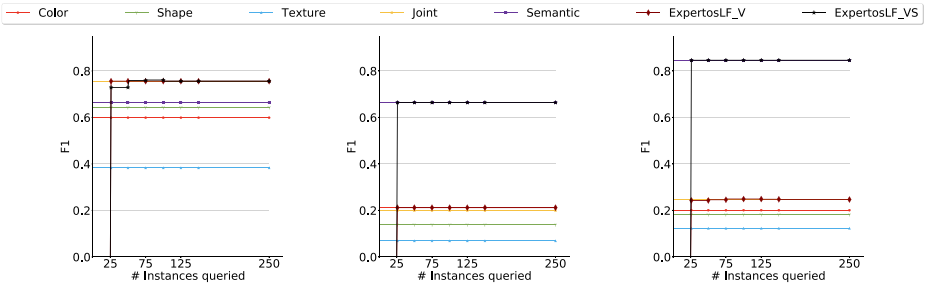


(a) BrainCE-MDI. (b) BreakHis. (c) COVID19-Rx. (d) IRMA.

(e) HAM10000.

(f) PlantPathology.

Fig. 11 Biomedical: PlantPathology (Q = 800), BrainCE-MRI, COVID19-Rx (Q = 2000), BreakHis (Q = 6000), and IRMA and HAM10000 (Q = 9000). (figure best seen in color)



(a) CopyDays. (b) COREL10K. (c) GHIM10K.

Fig. 12 Real. CopyDays (Q = 800), COREL10K, and GHIM10K (Q = 9000). (figure best seen in color)

6.2.3 Sketch

Figure 13 depicts the evolution of the F_1. Once again, for all datasets, the ensembles learnt with our solutions very quickly achieve (or surpass) the performance of the best individual expert. For \$P (Fig. 13a), ImiSketchS (Fig. 13b), and mCali (Fig. 13c), `ExpertosLF_V` achieves the performance of the best expert for $X = 25$, and surpasses it for $X = 50$ (for ImiSketchS) and 75 (mCali). `ExpertosLF_VS` achieves the best expert's performance for $X = 50$ for ImiSketchS, $X = 25$ for \$P (surpassing the best expert performance for $X = 125$), and $X = 25$ for mCali.

To sum up these results: in the online setting, for all the domains under study, our late fusion technique achieves similar results to the early fusion of all the experts (with the exception of the $avgP@10$ metric), and surpasses the *FreqRankLF* and *SimRankLF* late fusion techniques. We also validated that our solutions, `ExpertosLF_V` and `ExpertosLF_VS`, converge to the performance of the best experts individually, surpassing them in some cases. For the offline setting, the weighted ensembles obtained from our solutions very quickly achieve (or surpass) the performance of the best individual expert for almost all datasets.

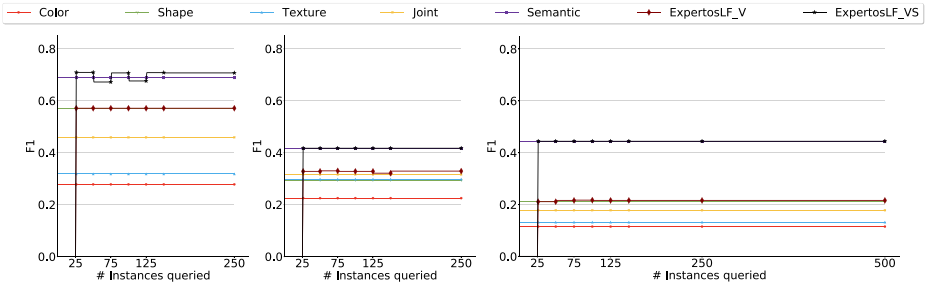
7 Discussion

In this section, we discuss the results obtained in our experiments in the light of our research questions.

RQ1) Does our late fusion technique improves retrieval performance?

Our late fusion technique achieved similar or better results than the baseline (*EF*), and surpassed the *FreqRankLF* and *SimRankLF* late fusion techniques.

We believe that our technique achieves best results because we acknowledge the subjectivity of human perception of image similarity by including human annotators in the loop to learn which CBIR systems are more suitable for the different domains. Furthermore, the fact that the best performing CBIR systems vary across domains illustrates how task dependency affects the quality of the retrieval results: if, instead of using our dynamic ensemble of CBIR systems, one had committed to a single CBIR system, it would perform inconsistently across domains/tasks.



(a) \$P. (b) ImiSketchS. (c) mCali.

Fig. 13 Sketch. ImiSketchS (Q = 800), \$P (Q = 2000), and mCali (Q = 6000). (figure best seen in color)

Moreover, our technique allows to create ensembles adapted to each domain without introducing an overhead of time in the retrieval process. To assess this, we considered the larger datasets for each domain using visual and semantic experts. We performed five runs to collect the average elapsed time of each query for each dataset and fusion technique, which we report in Fig. 14.

Our technique is more efficient at performing a query than *FreqRankLF* and *SimRankLF*. It takes a little longer at computing results than the baseline (*EF*), but the creation (extraction and indexation of the descriptors) of the CBIR system for the baseline is slower than for any of the late fusion techniques (e.g., for the IRMA dataset, creation with the late fusion takes around 20 hours, and with early fusion 27 hours).

RQ2) Does the resulting ensemble perform as well as the best individual CBIR?

As we have seen across the different domains for all the datasets under evaluation, our late fusion solutions based on expert advice indeed quickly converged to the performance of the best CBIR expert. They usually needed fewer than 25 queries to converge to the most suitable combination of weights for the CBIR experts.

Our solutions were also very quick to re-adapt the weights distribution to follow the current best CBIR. This can be observed on the plots depicting the evolution of the F_1 over queries. For the majority of the datasets, our solutions achieved (or surpassed) the $precision$ of the best CBIR expert(s) at a $recall$ cut-off of 0.1 (BREAKHIS, PLANTPATHOLOGY, COPYDAYS, COREL1K, COREL10K, GHIM10K, IMISKETCHS, and MCALI), 0.2 (BRAINCE-MRI, and COVID19-RX), or 0.3 (IRMA).

We observed a possible limitation of our technique: it tends to converge to a single expert in the ensemble, even if a combination of multiple experts yielded better results in previous iterations. This may be explained by the exponential behavior of EWAF, which tends to favor the highest weighted expert over the remaining ones. We believe the same effect may happen with other late fusion techniques.

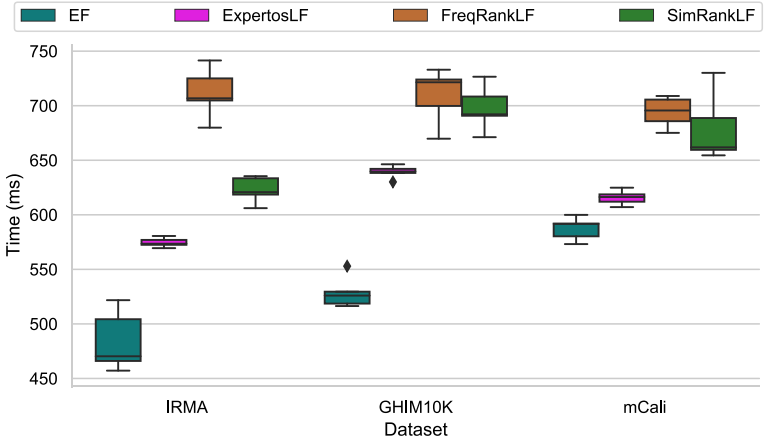


Fig. 14 Average elapsed time of performing a query on the datasets IRMA, GHIM10K, and MCALI, using different fusion approaches. (figure best seen in color)

RQ3) Can we use the ensemble learned in an online setting in an offline setting?

The ensemble learned using our weighted late fusion technique, for each dataset across the three domains under study, was successfully applied to an offline scenario, in which we did not have feedback available. The expert-based solutions (used to learn the weights) only needed to receive feedback on how good the retrieved results are for approximately 25 queries for the Real and Sketch domains, and around 50 queries for the Biomedical domain. After that, it is ready to be used in an offline scenario.

RQ4) Are the CBIR experts in the resulting ensemble plausible considering the domain in hand?

In Fig. 15, we present an overview of the CBIR experts weights' distribution for the `ExpertosLF_V` and `ExpertosLF_VS` solutions considering either only CBIR visual experts or the combinations of visual and semantic CBIR experts. We considered the combinations of weights of each CBIR expert at the first iteration for which they converged to (or surpassed) the performance of the best CBIR expert. Each horizontal bar encodes the distribution of the weights of each expert for each dataset. As we can see, the best experts vary across the different domains.

Leveraging on the interpretability associated with each ensemble, we provide a thorough and detailed analysis of each solution. Our aim is to demonstrate that the distribution of the weights for the experts across the domains is plausible and well-rooted on the type of images within each dataset.

In the Biomedical domain, and considering only the use of visual descriptors, BRAINCE-MRI is well described by the `texture` expert with a slight contribution of the remaining visual experts. This result can be explained by the fact that the images are gray-level with brain and tumors' shapes being roughly similar (they may vary in size), thus, there is little color, shape or joint information available. As a result, mainly texture information is able to successfully account for differences on tissue characteristics, such as calcifications, fat, cysts, contrast enhancement, or signal intensity.

BREAKHIS and PLANTPATHOLOGY datasets are best discriminated only by the `color` expert. These results can be explained, respectively, by the kind of differences observed in images regarding the tissue of different breast tumors (BREAKHIS) or how the leaves change with the foliar diseases (PLANTPATHOLOGY). BREAKHIS presents images in shades of pink, white, and purple to represent what percent of the tumor forms normal duct structures, how larger, darker, or irregular the cell nucleus is, and how many cells exist. PLANT-PATHOLOGY depicts images mainly in shades of green where the visual symptoms of a disease vary greatly between varieties, but color plays a major role in differentiating them. For example, in apple scab, the initial infection appears as black or olive-brown lesions, while in cedar apple rust, early symptoms of the disease are small, light yellow spots on leaves, that will expand and turn into bright orange ones.

COVID19-RX, IRMA, and HAM10000 are better described by the `shape` expert. The latter also slightly benefits from the inclusion of the remaining experts, in particular the `joint` one. Both COVID19-RX and IRMA depict gray-level images where mainly the size and shape of either the opacities and pleural abnormalities of the organs under study, respectively, differ significantly. Color could also be relevant to capture the density of the whiter pixels in the COVID19-RX dataset. Overall, our results are in line with the characteristics of the images. HAM10000 depicts images of skin lesions with different shapes in tones of pink (for the skin) and combinations of pink, brown, and black (for the lesion itself).

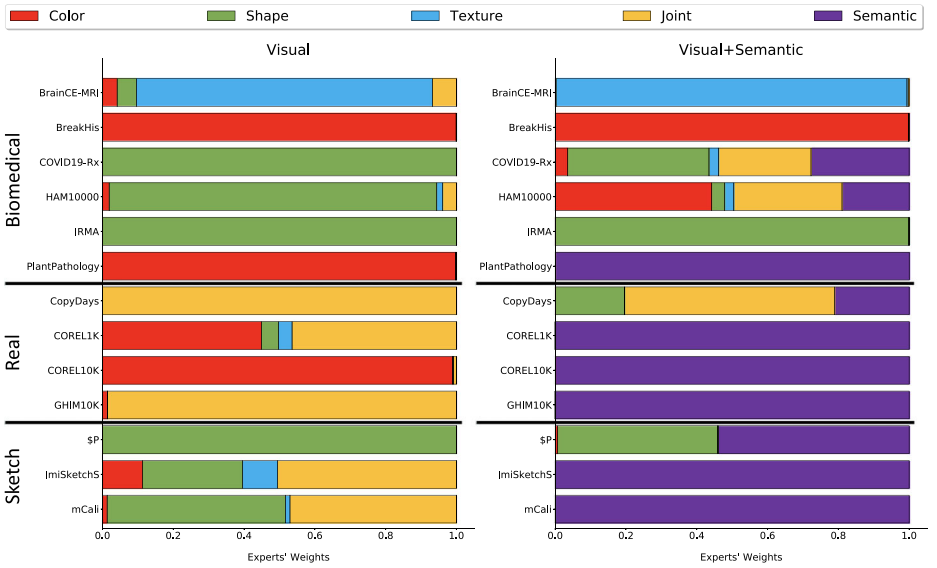


Fig. 15 Distribution of the experts weights per dataset. (figure best seen in color)

Moreover, there is little texture information available in the images. It is widely known that color in skin lesions provide important morphologic information (melanin is the most important chromophore in pigmented skin lesions), however, our solution failed to capture that when using only visual experts: only the shape of the lesion itself was useful to distinguish among types of lesions.

The inclusion of the `semantic` tags expert was useful to half of the Biomedical datasets: COVID19-Rx, HAM10000, and PLANTPATHOLOGY. We believe this is due to the semantic richness of the identified terms: although they may not be suited to the domain, they may be sufficiently distinct to discriminate the images and improve the performance of the retrieval task. Interestingly, the inclusion of semantic information made the system adjust itself in a way that, in addition to the `semantic` tags, it benefits from `color` and `joint` experts for the HAM10000, and `shape` and `joint` for COVID19-Rx. These results are in line with what we believe it would be expected considering the characteristics of the images of those datasets as described above.

In the Real domain, datasets are best described by either `color` (COREL10K), `joint` information (COPYDAYS), or the combination of both (COREL1K and GHIM10K). COREL1K also slightly benefits from the contribution of texture and shape experts. These results are not surprising, since all datasets depicts natural colored images with diverse colors, shapes, and textures. The inclusion of the semantic information majored almost all datasets. These results were also expected since the semantic richness of the tags (which describes objects, people, emotions, events, etc.) is much more powerful than just the visual content of the images. For the COPYDAYS dataset, it is useful to use both visual (`shape` and `joint`) and semantic information. We believe this is due to the fact that several images have been heavily manipulated (scaled, decreased image quality, and parts of the image painted or blurred). Thus, for many images, the semantic information is less discriminative.

In the Sketch domain, datasets are best discriminated by `shape` ($P), or a combination of `shape` and `joint` information (IMISKETCHS and MCALI). The latter datasets, in

particular the ImiSketchS, benefit slightly from the inclusion of the remaining visual experts. All datasets depict black and white images representing numbers, geometrical shapes, furniture, mathematical symbols, among others. As expected, shape plays an important role in our expert-based solution (demonstrated by the use of both the `shape` and `joint` expert). Similar to the Real domain, the inclusion of semantic information majored our solution for the ImiSketchS and mCALI datasets, and we believe it happens for the same reasons. The $P dataset benefits from the use of semantic information combined with shape.

Figure 16 depicts the difference between the performance achieved by our solution when using both visual and semantic experts and the performance achieved by using only visual experts. As we can see, the Biomedical domain is well described using mainly visual CBIR experts. while the Real and Sketch ones benefit from the use of semantic information.

8 Conclusions and future work

We presented a novel late fusion technique using online learning and prediction with expert advice to the problem of combining, in a dynamic fashion, the best types of descriptors to discriminate images in a CBIR scenario, regardless of the dataset or domain in hand. We did so by leveraging on relevance feedback that may be available in a realistic scenario.

Our late fusion solutions based on expert advice were indeed able to quickly learn the best descriptor sets in three distinct domains (Biomedical, Real, and Sketch), spanning a total of 13 benchmark datasets. The expert-based solutions achieved similar performance to that of the early fusion of all the experts, and surpassed existing state of the art late fusion techniques (*FreqRankLF* and *SimRankLF*). The expert-based solutions were also

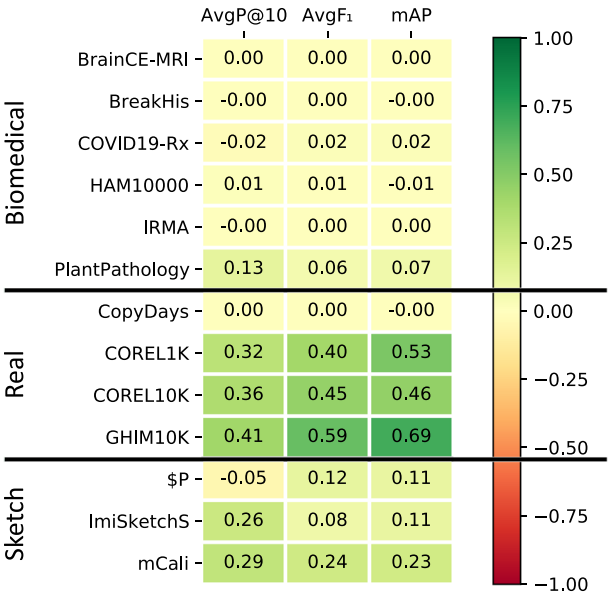


Fig. 16 Difference between the `ExpertosLF_VS` solution with visual and semantic experts and the `ExpertosLF_V` solution using only visual experts performance. Darker shades of green mean that the performance of `ExpertosLF_VS` is better than `ExpertosLF_V`. (figure best seen in color)

more efficient than state of the art techniques with similar or better results. Moreover, our solutions guaranteed that the retrieval performance was as good as the best CBIR system in the ensemble. Finally, the ensembles learnt through our approach also proved useful in an offline setting (i.e., when human feedback is no longer available).

In this work, we focused mainly on low- and high-level descriptors (instead of using, for instance, CNN layers), since we intended to 1) ensure that the resulting ensembles were interpretable, and 2) study how the use of different kinds of information (visual or semantic) varied across the different domains, and whether the resulting ensemble reflected it. For future work, neural descriptors could also be included in the ensemble, since our technique is model-agnostic, modular, and scalable.

Another line of future work to be explored is that of the online learning frameworks. The framework used in this work, prediction with expert advice, assumes that the forecaster learns both its own loss and the loss of each expert after the environment's outcome is revealed (i.e., that the set of the most relevant images for a given query is known). However, this may not always be the case in CBIR. Thus, we intend to explore a related class of problems, *multi-armed bandits* [51, 84], in which the environment's outcome is unknown, and only the forecaster learns its own loss , i.e., only the expert chosen by the forecaster receives feedback regarding its set of retrieved images.

Acknowledgements This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) under the LASIGE Research Unit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020, and the INESC-ID Research Unit, ref. UIDB/50021/2020. Soraia M. Alarcão is funded by an FCT grant, ref. SFRH/BD/138263/2018, and Vânia Mendonça was funded by an FCT grant, ref. SFRH/BD/12 1443/2016.

Declarations

Conflict of interests The authors declare that they have no conflict of interest.

References

1. Ahmed A (2020) Implementing relevance feedback for content-based medical image retrieval. IEEE Access 8:79969–79976
2. Ahmed KT, Ummesafi S, Iqbal A (2019) Content based image retrieval using image features information fusion. Inf Fusion 51:76–99
3. Ahn E, Kumar A, Fulham M, Feng D, Kim J (2019) Convolutional sparse kernel network for unsupervised medical image analysis. Med Image Anal 56:140–151
4. Alahi A, Ortiz R, Vandergheynst P (2012) Freak: fast retina keypoint. In: 2012 IEEE Conference on computer vision and pattern recognition. pp 510–517. Ieee
5. Alzu'bi A, Amira A, Ramzan N (2017) Content-based image retrieval with compact deep convolutional features. Neurocomputing 249:95–105
6. Arun K, Govindan V, Kumar SM (2017) On integrating re-ranking and rank list fusion techniques for image retrieval. Intl J of Data Sci and Anal 4(1):53–81
7. Ashraf R, Ahmed M, Ahmad U, Habib MA, Jabbar S, Naseer K (2020) Mdcbir-mf: multimedia data for content-based image retrieval by using multiple features. Multimed Tools Appl 79(13):8553–8579
8. Ashraf R, Ahmed M, Jabbar S, Khalid S, Ahmad A, Din S, Jeon G (2018) Content based image retrieval by using color descriptor and discrete wavelet transform. J of Med Syst 42(3):1–12
9. Babaie M, Tizhoosh HR, Khatami A, Shiri M (2017) Local radon descriptors for image search. In: 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA). pp 1–5. IEEE
10. Bai S, Zhou Z, Wang J, Bai X, Jan Latecki L, Tian Q (2017) Ensemble diffusion for retrieval. In: Proceedings of the IEEE international conference on computer vision. pp 774–783

11. Banerjee I, Kurtz C, Devorah AE, Do B, Rubin DL, Beaulieu CF (2018) Relevance feedback for enhancing content based image retrieval and automatic prediction of semantic image features: application to bone tumor radiographs. J Biomed Inform 84:123–135
12. Bay H, Tuytelaars T, Van Gool L (2006) Surf: speeded up robust features. In: European conference on computer vision. pp 404–417. Springer
13. Bella MIT, Vasuki A (2019) An efficient image retrieval framework using fused information feature. Comput Electr Eng 75:46–60
14. Bhardwaj S, Pandove G, Dahiya PK (2020) A futuristic hybrid image retrieval system based on an effective indexing approach for swift image retrieval. Int J Comput Inf Syst Ind Manag Appl 12:1–13
15. Borth D, Ji R, Chen T, Breuel T, Chang SF (2013) Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: ACM International Conference on Multimedia. pp 223–232
16. Calonder M, Lepetit V, Strecha C, Fua P (2010) Brief: binary robust independent elementary features. In: European conference on computer vision. pp 778–792. Springer
17. Camalan S, Niazi MKK, Moberly AC, Teknos T, Essig G, Elmaraghy C, Taj-Schaal N, Gurcan MN (2020) Otomatch: content-based eardrum image retrieval using deep learning. Plos One 15(5):e0232776
18. Cesa-Bianchi N, Lugosi G (2006) Prediction, learning and games. Cambridge University Press
19. Chang SF, Sikora T, Purl A (2001) Overview of the mpeg-7 standard. IEEE Transactions on Circuits and Systems for Video Technology 11(6):688–695
20. Chatzichristofis SA, Boutalis YS (2008) Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In: International conference on computer vision systems. pp 312–322. Springer
21. Chatzichristofis SA, Boutalis YS (2008) Fcth: fuzzy color and texture histogram-a low level feature for accurate image retrieval. In: 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services. pp 191–196. IEEE
22. Chatzichristofis S, Boutalis Y, Lux M (2009) Selection of the proper compact composite descriptor for improving content based image retrieval. In: Proc. of the 6th IASTED International Conference. vol 134643, pp 064
23. Cheng J, Yang W, Huang M, Huang W, Jiang J, Zhou Y, Yang R, Zhao J, Feng Y, Feng Q, Chen W (2016) Retrieval of brain tumors by adaptive spatial pooling and fisher vector representation. PLOS ONE 11(6):1–15
24. Chowdhury MEH, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, Islam KR, Khan MS, Iqbal A, Emadi NA, Reaz MBI, Islam MT (2020) Can ai help in screening viral and covid-19 pneumonia? IEEE Access 8:132665–132676
25. Chu K, Liu GH (2020) Image retrieval based on a multi-integration features model. Mathematical problems in engineering. vol 2020
26. Chung YA, Weng WH (2017) Learning deep representations of medical images using siamese cnns with application to content-based image retrieval. arXiv:1711.08490
27. Dang-Nguyen DT, Piras L, Giacinto G, Boato G, Natale FGD (2017) Multimodal retrieval with diversification and relevance feedback for tourist attraction images. ACM Trans Multimed Comput Commun Appl (TOMM) 13(4):1–24
28. Datta R, Joshi D, Li J, Wang JZ (2006) Studying aesthetics in photographic images using a computational approach. In: European Conference on Computer Vision. pp 288–301. Springer
29. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on computer vision and pattern recognition. pp 248–255. Ieee
30. de Ves E, Benavent X, Coma I, Ayala G (2016) A novel dynamic multi-model relevance feedback procedure for content-based image retrieval. Neurocomputing 208:99–107
31. Dice LR (1945) Measures of the amount of ecologic association between species. Ecol 26(3):297–302
32. Dubey SR, Singh SK, Singh RK (2017) Local svd based nir face retrieval. J of Vis Commun and Image Represent 49:141–152
33. Fadaei S, Amirfattahi R, Ahmadzadeh MR (2016) New content-based image retrieval system based on optimised integration of dcd, wavelet and curvelet features. IET Image Process 11(2):89–98
34. Fonseca MJ, Jorge JA (2003) Indexing high-dimensional data for content-based retrieval in large databases. In: Eighth International Conference on Database Systems for Advanced Applications, 2003.(DASFAA 2003). Proceedings. pp 267–274. IEEE
35. Graf F (2015) Jfeaturelib v1.6.3
36. Hamreras S, Boucheham B, Molina-Cabello MA, Benitez-Rochel R, Lopez-Rubio E (2020) Content based image retrieval by ensembles of deep learning object classifiers. Integrated computer-aided engineering. pp 1–15
37. Haralick RM, Shanmugam K, Dinstein IH (1973) Textural features for image classification. IEEE transactions on systems, man, and cybernetics. pp 610–621

38. Huang J, Kumar SR, Mitra M, Zhu WJ (2001) Image indexing using color correlograms. US patent 6,246,790
39. Huu QN, Viet DC, Thuy QDT (2021) Semantic class discriminant projection for image retrieval with relevance feedback. Multimedia tools and applications. pp 1–26
40. Imisketchsdb (2012) http://www.irisa.fr/intuidoc/IMIsketchSDB.html. Accessed date June 2021
41. Itten J (1973) The art of color: the subjective experience and objective rationale of color; translated by ernst van haagen. van nostrand reinhold
42. Jaccard P (1912) The distribution of the flora in the alpine zone. New Phytologist 11(2):37–50
43. Jégou H, Douze M, Schmid C (2008) Hamming embedding and weak geometry consistency for large scale image search-extended version. In: Proceedings of the 10th European Conference on Computer Vision
44. Jian M, Yin Y, Dong J, Lam KM (2018) Content-based image retrieval via a hierarchical-local-feature extraction scheme. Multimed Tools and Appl 77(21):29099–29117
45. Kanaparthi SK, Raju U, Shanmukhi P, Aneesha GK, Rahman MEU (2019) Image retrieval by integrating global correlation of color and intensity histograms with local texture features. Multimedia Tools and Applications. pp 1–37
46. Karamti H, Tmar M, Visani M, Urruty T, Gargouri F (2018) Vector space model adaptation and pseudo relevance feedback for content-based image retrieval. Multimed Tools and Appl 77(5):5475–5501
47. Kim WY, Kim YS (2000) A region-based shape descriptor using zernike moments. Signal Process Image Commun 16(1-2):95–102
48. Kriegel HP, Schubert E, Zimek A (2011) Evaluation of multiple clustering solutions. In: Multiclust@ ECML/PKDD. pp 55–66
49. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
50. Kruthika K, Maheshappa H, Initiative ADN et al (2019) Cbir system using capsule networks and 3d cnn for alzheimer's disease diagnosis. Inform Med Unlocked 14:59–68
51. Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. Adv Appl Math 6(1):4–22. https://doi.org/10.1016/0196-8858(85)90002-8
52. Latif A, Rasheed A, Sajid U, Ahmed J, Ali N, Ratyal NI, Zafar B, Dar SH, Sajid M, Khalil T (2019) Content-based image retrieval and feature extraction: a comprehensive review. Mathematical problems in engineering. vol 2019
53. Leutenegger S, Chli M, Siegwart RY (2011) Brisk: binary robust invariant scalable keypoints. In: 2011 International Conference on Computer Vision. pp 2548–2555. Ieee
54. Li H, Toyoura M, Shimizu K, Yang W, Mao X (2016) Retrieval of clothing images based on relevance feedback with focus on collar designs. The Vis Comput 32(10):1351–1363
55. Li J, Wang JZ (2003) Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans Pattern Anal Mach Intell 25(9):1075–1088
56. Liu S, Feng L, Liu Y, Wu J, Sun M, Wang W (2017) Robust discriminative extreme learning machine for relevance feedback in image retrieval. Multidim Syst Sign Process 28(3):1071–1089
57. Liu P, Guo JM, Wu CY, Cai D (2017) Fusion of deep learning and compressed domain features for content-based image retrieval. IEEE Trans Image Process 26(12):5706–5717
58. Liu X, Tizhoosh HR, Kofman J (2016) Generating binary tags for fast medical image retrieval based on convolutional nets and radon transform. In: 2016 International Joint Conference on Neural Networks (IJCNN). pp 2872–2878. IEEE
59. Liu GH, Yang JY, Li Z (2015) Content-based image retrieval using computational visual attention model. Pattern Recogn 48(8):2554–2566
60. Lu H, Zhang M, Xu X, Li Y, Shen HT (2020) Deep fuzzy hashing network for efficient image retrieval. IEEE transactions on fuzzy systems
61. Lux M, Marques O (2013) Visual information retrieval using java and LIRE. Synthesis lectures on information concepts, retrieval, and services. Morgan & Claypool Publishers
62. Mahmoud AM, Karamti H, Hadjouni M (2020) A hybrid late fusion-genetic algorithm approach for enhancing cbir performance. Multimed Tools Appl 79(27):20281–20298
63. Majhi M, Pal AK (2021) An image retrieval scheme based on block level hybrid dct-svd fused features. Multimed Tools and Appl 80(5):7271–7312
64. Marĉelja S (1980) Mathematical description of the responses of simple cortical cells. JOSA 70(11):1297–1300
65. Markonis D, Schaer R, de Herrera AGS, Müller H (2017) The parallel distributed image search engine (paradise). arXiv:1701.05596

66. Mortensen EN, Deng H, Shapiro L (2005) A sift descriptor with global context. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol 1, pp 184–190. IEEE
67. Mosbah M, Boucheham B (2017) Pseudo relevance feedback based on majority voting mechanism. International Journal of Web Science 3(1):58–81
68. Müller H, Clough P, Deselaers T, Caputo B (2010) Image CLEF: experimental evaluation in visual information retrieval. vol 32. Springer science & business media
69. Neshov NN (2013) Comparison on late fusion methods of low level features for content based image retrieval. In: International Conference on Artificial Neural Networks. pp 619–627. Springer
70. Ochiai A (1957) Zoogeographical studies on the soleoid fishes found in Japan and its neighbouring regions-i. Bull Jpn Soc Scient Fish 22:522–525
71. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern analysis and Machine Intelligence 24(7):971–987
72. Otsuka Y (1936) The faunal character of the japanese pleistocene marine mollusca, as evidence of climate having become colder during the pleistocene in Japan. Biogeograph Soc Japan 6:165–170
73. Pavithra L, Sharmila TS (2018) An efficient framework for image retrieval using color, texture and edge features. Comput Electr Eng 70:580–593
74. Pesenko YA (1982) Principles and methods of quantitative analysis in Faunistical researches. Moscow (Nauka) [in Russian]
75. Phadikar BS, Phadikar A, Maity GK (2018) Content-based image retrieval in dct compressed domain with mpeg-7 edge descriptor and genetic algorithm. Pattern Anal and Applic 21(2):469–489
76. Pinjarkar L, Sharma M, Selot S (2020) Deep cnn combined with relevance feedback for trademark image retrieval. J Intell Syst 29(1):894–909
77. Piras L, Giacinto G (2017) Information fusion in content based image retrieval: a comprehensive overview. Information Fusion 37:50–60
78. Putzu L, Piras L, Giacinto G (2020) Convolutional neural networks for relevance feedback in content based image retrieval. Multimed Tools and Appl 79(37):26995–27021
79. Qayyum A, Anwar SM, Awais M, Majid M (2017) Medical image retrieval using deep convolutional neural network. Neurocomputing 266:8–20
80. Raghuwanshi G, Tyagi V (2020) Texture image retrieval using hybrid directional extrema pattern. Multimedia Tools and Applications. pp 1–23
81. Rana SP, Dey M, Siarry P (2019) Boosting content based image retrieval performance through integration of parametric & nonparametric approaches. J Vis Commun and Image Represent 58:205–219
82. Rao Y, Liu W, Fan B, Song J, Yang Y (2018) A novel relevance feedback method for cbir. World Wide Web 21(6):1505–1522
83. Reta C, Solis-Moreno I, Cantoral-Ceballos JA, Alvarez-Vargas R, Townend P (2018) Improving content-based image retrieval for heterogeneous datasets using histogram-based descriptors. Multimed Tools and Appl 77(7):8163–8193
84. Robbins H (1952) Some aspects of the sequential design of experiments. Bull of the Am Math Soc 58(5):527–535. https://doi.org/10.1090/S0002-9904-1952-09620-8
85. Saritha RR, Paul V, Kumar PG (2019) Content based image retrieval using deep learning process. Clust Comput 22(2):4187–4200
86. Sathiamoorthy S, Natarajan M (2020) An efficient content based image retrieval using enhanced multitrend structure descriptor. SN Applied Sciences 2(2):1–20
87. Satish B, Supreethi K (2017) Content based medical image retrieval using relevance feedback bayesian network. In: 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT). pp 424–430. IEEE
88. Sezavar A, Farsi H, Mohamadzadeh S (2019) Content-based image retrieval by combining convolutional neural networks and sparse representation. Multimed Tools and Appl 78(15):20895–20912
89. Shete D, Chavan M (2012) Content based image retrieval: review. Int J Emerg Technol Adv Eng 2(9):85–90
90. Spanhol FA, Oliveira LS, Cavalin PR, Petitjean C, Heutte L (2017) Deep features for breast cancer histopathological image classification. In: 2017 IEEE International Conference on Systems, man, and Cybernetics (SMC). pp 1868–1873. IEEE
91. Spanhol FA, Oliveira LS, Petitjean C, Heutte L (2016) Breast cancer histopathological image classification using convolutional neural networks. In: 2016 International Joint Conference on Neural Networks (IJCNN). pp 2560–2567. IEEE
92. Srivastava P, Khare A (2017) Integration of wavelet transform, local binary patterns and moments for content-based image retrieval. J Vis Commun and Image Represent 42:78–103

93. Swati ZNK, Zhao Q, Kabir M, Ali F, Ali Z, Ahmed S, Lu J (2019) Content-based brain tumor retrieval for mr images using transfer learning. IEEE Access 7:17809–17822
94. Sze-To A, Tizhoosh H (2020) Searching for pneumothorax in half a million chest x-ray images. In: International Conference on Artificial Intelligence in Medicine. pp 453–462. Springer
95. Sørensen TA (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, and its application to analyses of the vegetation on {Danish} commons. Biologiske Skrifter 15:1–34
96. Tamura H, Mori S, Yamawaki T (1978) Textural features corresponding to visual perception. IEEE Trans Syst Man Cybern 8(6):460–473
97. Tang X, Jiao L, Emery WJ (2017) Sar image content retrieval based on fuzzy similarity and relevance feedback. IEEE J Sel Top Appl Earth Obs Remote Sens 10(5):1824–1842
98. Tang J, Li Z, Zhu X (2018) Supervised deep hashing for scalable face image retrieval. Pattern Recogn 75:25–32
99. Thapa R, Snavely N, Belongie S, Khan A (2020) The plant pathology 2020 challenge dataset to classify foliar disease of apples. arXiv:2004.11958
100. Tizhoosh HR, Mitcheltree C, Zhu S, Dutta S (2016) Barcodes for medical image retrieval using autoencoded radon transform. In: 2016 23Rd International Conference on Pattern Recognition (ICPR). pp 3150–3155. IEEE
101. Tizhoosh HR, Zhu S, Lo H, Chaudhari V, Mehdi T (2016) Minmax radon barcodes for medical image retrieval. In: International Symposium on Visual Computing. pp 617–627. Springer
102. Tschandl P, Rosendahl C, Kittler H (2018) The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data 5(1):1–9
103. Tsochatzidis L, Zagoris K, Arikidis N, Karahaliou A, Costaridou L, Pratikakis I (2017) Computer-aided diagnosis of mammographic masses based on a supervised content-based image retrieval approach. Pattern Recogn 71:106–117
104. Tzelepi M, Tefas A (2018) Deep convolutional learning for content based image retrieval. Neurocomputing 275:2467–2478
105. Van De Sande K, Gevers T, Snoek C (2009) Evaluating color descriptors for object and scene recognition. IEEE Trans Pattern Anal Mach Intell 32(9):1582–1596
106. Varish N, Pal AK (2016) Content-based image retrieval using svd-based eigen images. International journal of image mining
107. Varish N, Pradhan J, Pal AK (2017) Image retrieval based on non-uniform bins of color histogram and dual tree complex wavelet transform. Multimed Tools and Appl 76(14):15885–15921
108. Vatavu RD, Anthony L, Wobbrock JO (2012) Gestures as point clouds: a $P recognizer for user interface prototypes. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction. pp 273–280
109. Vieira J (2014) mCALI: reconhecedor de esboços multiuso. Master's thesis, IST/ULisboa
110. Wang X, Lee F, Chen Q (2019) Similarity-preserving hashing based on deep neural networks for large-scale image retrieval. J Vis Commun and Image Represent 61:260–271
111. Wang JZ, Li J, Wiederhold G (2001) Simplicity: semantics-sensitive integrated matching for picture libraries. IEEE Trans Pattern Anal Mach Intell 23(9):947–963
112. Wang XY, Liang LL, Li WY, Li DM, Yang HY (2016) A new svm-based relevance feedback image retrieval using probabilistic feature and weighted kernel function. J Vis Commun and Image Represent 38:256–275
113. Wei Z, Liu GH (2020) Image retrieval using the intensity variation descriptor. Mathematical problems in engineering. vol 2020
114. Xu H, Wang JY, Mao L (2017) Relevance feedback for content-based image retrieval using deep learning. In: 2017 2Nd International Conference on Image, Vision and computing (ICIVC). pp 629–633. IEEE
115. Yan C, Li L, Zhang C, Liu B, Zhang Y, Dai Q (2019) Cross-modality bridging and knowledge transferring for image understanding. IEEE Trans Multimed 21(10):2675–2685
116. Yuan BH, Liu GH (2020) Image retrieval based on gradient-structures histogram. Neural Comput and Applic 32(15):11717–11727
117. Zeiler M (2013) Clarifai. http://www.image-net.org/challenges/LSVRC/2013/results.php. Accessed date June 2021
118. Zhang J, Peng Y (2017) Ssdh: semi-supervised deep hashing for large scale image retrieval. IEEE Trans on Circuits and Syst for Video Technol 29(1):212–225
119. Zhang J, Peng Y (2018) Query-adaptive image retrieval by deep-weighted hashing. IEEE Trans Multimed 20(9):2400–2414

120. Zhang L, Shum HP, Shao L (2016) Discriminative semantic subspace analysis for relevance feedback. IEEE Trans Image Process 25(3):1275–1287
121. Zhao M, Zhang H, Sun J (2016) A novel image retrieval method based on multi-trend structure descriptor. J of Vis Commun and Image Represent 38:73–81
122. Zheng L, Yang Y, Tian Q (2017) Sift meets cnn: a decade survey of instance retrieval. IEEE Trans Pattern Anal Mach Intell 40(5):1224–1244
123. Zhou W, Li H, Tian Q (2017) Recent advance in content-based image retrieval: a literature survey. arXiv:1706.06064
124. Zhu Y, Jiang J, Han W, Ding Y, Tian Q (2017) Interpretation of users' feedback via swarmed particles for content-based image retrieval. Inf Sci 375:246–257

Soraia M. Alarcão holds a master's degree in Information Systems and Computer Engineering from IST, University of Lisbon (2014), and is currently a PhD student at the Informatics Department at Faculty of Sciences, University of Lisbon, Portugal. She is a researcher at LASIGE since 2014. Her research interests include Accessibility, Emotions Recognition, Human-Computer Interaction, Health Systems, and Multimedia Information Retrieval.

Vânia Mendonça holds a master's degree in Information Systems and Computer Engineering from IST, University of Lisbon (2015), and is currently a PhD student at the same institution, as well as a Junior Researcher at INESC-ID Lisbon. Her current research interests include low-resource Natural Language Processing and Accessibility.

Carolina Maruta is a Clinical neuropsychologist, holds a PhD (2015) in Biomedical Sciences (Neurosciences) from Faculty of Medicine, University of Lisbon, maintains clinical and research activities in the Laboratory of Language Research of FMUL and in CENC - Sleep Medicine Centre, and is an invited assistant professor in the Faculty of Human Sciences of Universidade Católica Portuguesa and Escola Superior de Saúde de Alcoitão. Current research interests focus on normal and pathological aging (dementia) and the relationship between sleep and cognition. Member of the executive committee of the Behavioural Neurology Section of the Portuguese Neurology Society.

Manuel J. Fonseca holds a PhD (2004) in Information Systems and Computer Engineering from IST/ULisbon, is an Associated Professor at Faculty of Sciences, University of Lisbon, and a senior researcher at LASIGE. His main research areas include Human-Computer Interaction, Emotions Recognition, Brain-Computer Interfaces, Multimedia Information Retrieval, Sketch Recognition, and Health Systems. He is a senior member of IEEE and ACM, and a member of Eurographics.

Affiliations

Soraia M. Alarcão[1] · Vânia Mendonça[2] · Carolina Maruta[3] · Manuel J. Fonseca[1]

Vânia Mendonça
vania.mendonca@tecnico.ulisboa.pt

Carolina Maruta
carolmaruta@gmail.com

Manuel J. Fonseca
mjfonseca@ciencias.ulisboa.pt

1 LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

2 INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal

3 Laboratório de Estudos de Linguagem, Centro de Estudos Egas Moniz, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal