ORIGINAL ARTICLE

Expert Systems WILEY

Multi-label logo recognition and retrieval based on weighted fusion of neural features

Marisa Bernabeu[1] | Antonio Javier Gallego[2] | Antonio Pertusa[2]

[1]Instituto Superior de Enseñanzas Artísticas, Comunidad Valenciana (ISEACV), Alicante, Spain

[2]University Institute for Computing Research, University of Alicante, Alicante, Spain

Correspondence
Antonio Javier Gallego, University Institute for Computing Research, University of Alicante, Carretera San Vicente del Raspeig s/n, 03690 San Vicente del Raspeig, Alicante, Spain.
Email: jgallego@dlsi.ua.es

Funding information
Conselleria d'Innovació, Universitats, Ciència i Societat Digital from Generalitat Valenciana and FEDER, Grant/Award Number: IDIFEDER/2020/003; Universidades y Empleo

Abstract

Classifying logo images is a challenging task as they contain elements such as text or shapes that can represent anything from known objects to abstract shapes. While the current state of the art for logo classification addresses the problem as a multiclass task focusing on a single characteristic, logos can have several simultaneous labels, such as different colours. This work proposes a method that allows visually similar logos to be classified and searched from a set of data according to their shape, colour, commercial sector, semantics, general characteristics, or a combination of features selected by the user. Unlike previous approaches, the proposal employs a series of multi-label deep neural networks specialized in specific attributes and combines the obtained features to perform the similarity search. To delve into the classification system, different existing logo topologies are compared and some of their problems are analysed, such as the incomplete labelling that trademark registration databases usually contain. The proposal is evaluated considering 76,000 logos (seven times more than previous approaches) from the European Union Trademarks dataset, which is organized hierarchically using the Vienna ontology. Overall, experimentation attains reliable quantitative and qualitative results, reducing the normalized average rank error of the state-of-the-art from 0.040 to 0.018 for the Trademark Image Retrieval task. Finally, given that the semantics of logos can often be subjective, graphic design students and professionals were surveyed. Results show that the proposed methodology provides better labelling than a human expert operator, improving the label ranking average precision from 0.53 to 0.68.

KEYWORDS
convolutional neural networks, logo image retrieval, multi-label classification, similarity search

1 | INTRODUCTION

The detection and recognition of logos is an important task given that companies need to detect the use of their logos in images (Bianco et al., 2017), social media (Orti et al., 2019) and sports events (Köstinger et al., 2010), or to discover unauthorized usages and plagiarism. Moreover, to register trademarks, it is necessary to verify that there are no similar logos within the same business sector. This is a relevant issue owing

to the volume of applications for trademark registration and the size of the databases containing existing trademarks, moreover when considering how costly it would be for humans to make these comparisons visually (Perez et al., 2018).

Most previous computer vision approaches for logos have focused on the trademark image retrieval (TIR) task, which consists of performing a similarity search to obtain the most similar logos given a query image. Schietse et al. (2007) describe the main challenges that TIR systems confront. Logo images differ from real pictures since they are artificially created and designed to have a visual impact. In addition, they may contain only text, images, or a combination of both. The most relevant feature for humans to characterize a logo is probably the shape. However, automatic shape classification is a challenging task. In addition to the structure of the elements that comprise a logo and its organization, semantic interpretation must also be considered to determine the objects present in logos. This is a very complex task related to how humans perceive and interpret images.

Colour also plays an essential role in designing and characterizing a logo. Brands within the same business area often use similar colours owing to their cultural and social connotations. However, this is not always the case since organizations may also use colour to differentiate themselves from the competition. For example, the authors of (Capsule, 2007) describe the case of the technology company Gear6, which uses the colour green to distinguish itself from its competitors. Colour is, therefore, important when performing TIR, but it is also necessary to consider that logos sometimes lose their colour and that we can also find versions in greyscale.

This work presents a method with a twofold purpose.[1] In addition to retrieving the most similar logos according to criteria provided by a user, it also allows performing multi-label classification (MLC). Figure 1 shows an overview of the method. First, a multi-label neural network architecture is proposed to classify different features of the logo, such as colour, shape, and figurative elements (semantics). This stage aims to facilitate the labelling of brands since the output contains a series of label options with an associated probability to assist the operator in this process, as seen in the bottom white box from the figure. Additionally, a similarity search module is added (top box), allowing the operator to adjust the search criteria by modifying the weight assigned to each characteristic. For instance, the user can tune the method to give a higher weight to the shape or colour when seeking the most similar logos. This is achieved through a module that conducts a weighted fusion of neural features according to the user-defined criteria, facilitating the identification of similar designs or the detection of plagiarism based on these preferences.

A preliminary study on multi-label logo classification and similarity search was proposed in (Gallego et al., 2019). We extend this method and its evaluation by making the following contributions:

1. The addition of a preprocessing stage for text detection, which includes an inpainting method to improve the retrieval results.
2. The inclusion of additional neural models and changes to the final stage of the similarity search using an alternative method that improves the results.
3. The analysis of existing topologies to understand the hierarchical Vienna classification system used for comparing and retrieving logos from a design perspective.
4. The proposal of multi-label tagging by grouping Vienna classification codes.
5. The use of a much larger corpus (logos from the European Union Intellectual Property Office, EUIPO, from 9 years rather than only 1 year) to train the networks.
6. A comparison of the proposal with 17 state-of-the-art methods, outperforming their results.
7. The presentation of a qualitative evaluation through surveys of graphic design students and expert designers.

The remainder of the paper is structured as follows: the background related to the topic is introduced in Section 2. Then, the proposed approach is developed thoroughly in Section 3. Next, Section 4 describes the experimental setup considered, while the results obtained and an analysis of them are included in Section 5. Finally, the general conclusions are discussed in Section 6.
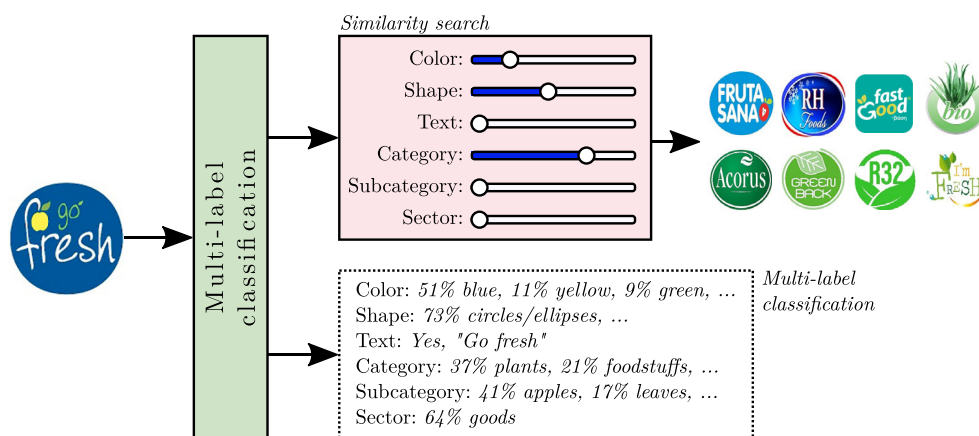

FIGURE 1   Overview of the proposed method.

2 | BACKGROUND

The state of the art of methods used for TIR and multi-label retrieval are reviewed in the first part of this section, after which the available datasets and the topologies used for their classification are detailed.

2.1 | Trademark image retrieval

Most traditional methods have addressed TIR by extracting a series of handcrafted features and using them to feed a k-nearest neighbour (kNN) (Duda et al., 2001) to obtain a ranking of the most similar logos. Some of the features used for this comparison include methods based on colour histograms (Ghosh & Parekh, 2015), shape (Qi et al., 2010), local descriptors such as SIFT (Chiam, 2015), or a combination of them (Guru & Kumar, 2018; Kumar et al., 2016). In some cases, the dimensionality of these features is reduced with Bags of Words (Iandola et al., 2015). In addition, distance metrics are generally employed for comparing the features processed, although more complex approaches based on template matching have also been proposed (Pornpanomchai et al., 2015).

Handcrafted features for this task are also found in more recent works, such as (Lourenço et al., 2019), which introduces Hierarchy of Visual Words. This TIR method decomposes images into simpler geometric shapes and defines a descriptor for binary logo image representation by encoding the hierarchical arrangement of component shapes. Nonetheless, most current TIR methods use deep learning (LeCun et al., 2015) architectures. For example, in Chiam (2015), an AlexNet network (Krizhevsky et al., 2012) with a sliding window is used to find logos in real images. The authors of Iandola et al. (2015) evaluated GoogleNet-based convolutional neural networks (CNNs) architectures for the brand classification of logos. More recently, Perez et al. (2018) proposed a combination of descriptors extracted from a VGG-19 network to find similarities using the cosine distance, and (Xia et al., 2019) used Transform-invariant Deep hashing for TIR by learning transformation-invariant features.

Our proposal is also based on deep learning. However, it employs a much more versatile method that combines the descriptors learned by a set of multi-label networks specialized in the classification of the different characteristics of logos. Most TIR methods reviewed rely on the brand to perform the similarity search or classification. However, a brand's image may change over time, in addition to the fact that the generic comparison of a logo does not allow its classification. For this, it is essential to consider distinctive characteristics of the logo, such as the use of colours, the semantic meaning of shapes, etc. The proposed approach makes it possible to perform similarity search based on different criteria while simultaneously taking advantage of these characteristics to perform multi-label logo retrieval.

2.2 | Multi-label logo image retrieval

As shown in the previous section, many TIR works exist in the literature. However, only a few approaches aim to classify logos using features other than brands. In this case, samples may have more than one simultaneous label (e.g., several colours, shapes, or figurative elements annotated for the same logo), making this problem a MLC task.

In traditional multi-class classification, each sample is assigned a single label from a set of mutually exclusive labels. In contrast, MLC allows each sample to be associated with multiple labels independently. Thus, a sample can have from no label to all possible labels. Formally, in MLC, the samples are associated with a subset of labels $Y \subseteq L$, where L represents the set of possible labels.

Several studies, as highlighted by Liu et al. (2021), have proposed exploiting label correlations in MLC problems. However, their applicability depends on the data type. In datasets originating from natural sources, it is feasible to rely on the identified relationships to improve results. However, for datasets from artificial sources, like logos, relying on such dependencies is less advisable, as the characteristics of a logo are limited only by the designer's creativity. A notable example is the case described previously about the technology company Gear6, which intentionally chose an atypical colour for its sector to stand out from competitors.

MLC has received significant attention in recent machine learning literature owing to its interesting applications. During the past decade, great strides have been made in this emerging paradigm. A review on this area emphasizing state-of-the-art multi-label learning algorithms can be found in Liu et al. (2021), Zhang and Zhou (2014). MLC approaches can roughly be categorized into problem transformation methods and algorithm adaptation methods. The first approach tackles the multi-label learning problem by transforming it into other well-established learning scenarios. This includes transforming the MLC problem into a binary classification task (Boutell et al., 2004a), into label ranking (LR) (Fürnkranz et al., 2008), or into multi-class classification (Tsoumakas & Vlahavas, 2007). The second category of MLC methods is based on algorithm adaptation, modifying standard learning techniques such as kNN (Zhang & Zhou, 2007), decision trees (Clare & King, 2001), or SVM (Elisseeff & Weston, 2001) to deal with multi-label data directly. In the case of neural networks, they can be adapted for multi-label problems by changing the output activation function and using appropriate loss functions, such as binary cross-entropy loss computed on all label predictions.

Multi-label methods are used in applications as diverse as text categorization (Dong et al., 2020), music categorization (Trohidis et al., 2011), or semantic scene classification (Boutell et al., 2004b). However, to the best of our knowledge, the literature contains no examples of MLC-related

works applied to logos, except for Gallego et al. (2019). As argued in the introduction section, features such as colour, shape, or semantic meaning play a key role in logo classification. Given the particular characteristics of this task, it is of particular interest to developing multi-label systems that make it possible to classify and search for logos based on different criteria that the user can configure.

This paper proposes an MLC approach applied to logos that, in addition to extending the methodology proposed in Gallego et al. (2019) and obtaining better results, also broadens the set of labels considered and considerably expands both the experiments and the analysis of the results obtained.

2.3 | Datasets and topologies

Reliable image datasets are crucial for tackling this task, although the corpora used in former works are not generally publicly available for copyright reasons (Ghosh & Parekh, 2015; Kalantidis et al., 2011). As a result, it has not been until recently that some free logo datasets appeared. Some examples are the large logo dataset (LLD) (Sage et al., 2018), which consists of more than 600,000 logos obtained from the Internet, METU (Tursun et al., 2017; Tursun & Kalkan, 2015), which contains 923,343 trademark images, and Logos in the Wild (Tüzkö et al., 2018), in which 11,054 images are labelled within 871 brands.

All the datasets mentioned above, and most of those used by the state-of-the-art methods, are labelled only by brand, as it is assumed that logos from the same brand tend to be similar. However, brands may evolve different versions of their logos (e.g., Disney has changed its logo more than 30 times; Iandola et al., 2015). These differences may include changes in the background, colour, texture, or shape, thus making the different versions of a logo very different in appearance and signifying that relying on visual similarity is not always a suitable means to classify logos from the same brand.

It is, therefore, complicated to establish a categorization method for logos. One of the topologies accepted as a standard by the different agencies for trademark registration worldwide is the Vienna classification, which was developed by the World Intellectual Property Organization (WIPO) (World Intellectual Property Organization, 2002). It is used by the European Union Intellectual Property Office (EUIPO) and the United States Patent and Trademark Office (UPSTO), among others, to classify their datasets.

Vienna classification (which will be described in detail in Section 2.3.1) is an international system used to label different characteristics of trademarks by employing a hierarchical topology ordered from the most general to the most specific concepts. It allows images to be labelled with metadata indicating their figurative meaning (semantics), colour, shape, and whether or not they contain text. Several patent and trademark agencies have recently released datasets along with their metadata, thus making possible works such as Rusiñol et al. (2011) or Gallego et al. (2019), which use this information for the classification or comparison of logos.

In addition to this labelling, there are classifications that professionals frequently use, such as the topologies proposed by Wheeler (2013) and Chaves and Belluccia (2003), which are based on other kinds of criteria.

Wheeler classifies logos into three general categories: Wordmarks (a freestanding acronym, company name, or product name), emblems (logos in which the company name is inextricably connected to a pictorial element), and only symbols (which are subdivided into letterforms, pictorial marks, and abstract/symbolic marks). However, the boundaries among these categories are pliant, and many logos may combine elements from more than one category.

The alternative categorization proposed by Chaves and Belluccia (2003) is similar but is more detailed and based on formal aspects. In this case, there are four main categories: logotypes (which is equivalent to Wheeler's 'Wordmarks' category, but with three subtypes: pure logotype, logotype with background, and logotype with accessory), logo-symbol (equivalent to emblems), logotypes with symbols, and only symbols, which, as occurs in Wheeler's version, is divided into three subtypes.

Given the relevance of the labelling method for the proposed methodology, the following section describes the Vienna classification and its relationship with the Wheeler and Chaves topologies that are oriented to designers.

2.3.1 | Vienna classification

The Vienna classification (World Intellectual Property Organization, 2002) proposes a hierarchical topology of logos, in which each image can be labelled with a series of codes indicating its semantics, shape, colour, and so on. It defines a set of 29 main categories, which are in turn divided into 2nd and 3rd level categories, creating a classification with hundreds of possible labels. The complete list of top-level categories can be seen at Appendix. Each code is indicated using the XX.YY.ZZ pattern. For example, the 5.9.1 code would assign the tag 'carrots' to a logo. The hierarchy of this code indicates that it belongs to the category of 2nd level 5.9 'vegetables' and the main category 5 'plants'.

This hierarchical organization makes it possible to group logos by different levels of labels and use higher hierarchical levels when the 3rd or 2nd levels have too much detail, are not very representative, contain few samples, or are ambiguous.

It is also necessary to consider that the labelling from trademark agencies is usually not exhaustive since only the most distinctive characteristics of brands are typically annotated. This means that incomplete or contradictory labelling can sometimes be found (e.g., a logo that has three colours and only one of them is labelled).

In this work, we propose solving some problems by grouping these codes according to their characteristics and semantics. The intention is not to replace Vienna but to carry out a selection of labels to keep those most useful for their application in machine learning tasks. The following four categories are, therefore, defined, in which the Vienna codes that uniquely describe these characteristics are selected or grouped:

- *Figurative*. This includes the Vienna codes from 1 to 25, which indicate categories related to the figurative or semantic meaning of the logo. For this category, we differentiate between the 1st and 2nd levels of the Vienna hierarchy, which we respectively denominate as the main category (which contains the 25 codes from the 1st level) and the sub-category (with 123 possible classes).
- *Colours*. Vienna category 29 refers to colours, although many codes indicate their number (e.g., 29.01.12 means that there are two predominant colours). It is, therefore, proposed that they should not be considered since they do not provide relevant information. After performing this filtering, the set of selected colour codes is reduced to 13 (included in Appendix).
- *Shapes*. In category 26, different types of shapes are labelled, including circles, triangles, quadrilaterals, and so on. In this case, the 3rd level of labelling is very specific and sometimes ambiguous (e.g., curved lines versus wavy lines, or dotted lines versus broken lines). We, therefore, propose using only up to the 2nd level. Moreover, codes 26.07 and 26.13 are grouped in category 26.5 (Other polygons) since a defined shape is not visually identified. After this grouping, a list of seven possible shape categories was eventually obtained (see Appendix).
- *Text*. Category 27 defines the text and its characteristics. This category is also too detailed (e.g., there are 20 different codes to indicate the appearance or shape of the text and as many to indicate the style of the font). Since the specific text in the logo is often made up of acronyms, monograms, or brand names that do not contribute much to the calculation of the similarity between logos, we, therefore, propose to label only the presence or absence of text in the image.

Table 1 shows a summary of the equivalence among the topologies proposed by Wheeler and Chaves and their relation to the proposed Vienna code groups. These equivalences allow us to determine the most relevant characteristics of logos when preparing or analysing their design. For example, colour and shape are features that appear in all types of designs and can, therefore, help the most to distinguish them. This is not the case with the presence of text or figurative elements, although they are very useful in determining some of the logo features. In summary, there is a relationship among the different topologies, signifying that the Vienna codes can describe the remaining classifications.

In this work, the modified Vienna classification is used to perform MLC and similarity search. We will specifically use the dataset provided by EUIPO (described in Section 4.1), which, in addition to Vienna, uses the alternative Nice classification[2] to label goods and services. This categorization organizes the sector into 45 subcategories. The labels used for goods include chemicals, medicines, metals, materials, machines, tools, vehicles, instruments, and so on, while the labels used for services include advertising, insurance, telecommunications, transport, and education.

3 | METHOD

Figure 2 shows the scheme of the proposed approach, which is divided into three main steps: a preprocessing of the input images, a MLC, and a similarity search step based on the features learned in the previous stage. Detailed explanations of each of these steps are provided in the following sections.

TABLE 1 Relationship between the Vienna classification and the topologies proposed by Wheeler and Chaves.

	Wheeler	Chaves	Vienna			
			Figurative	Colour	Shape	Text
Nominal Identifier	*Wordmark*	*Logotype*				
		• Logotype with background	–	✓	✓	✓
		• Pure logotype	–	✓	–	✓
		• Logotype with accessory	✓	✓	✓	✓
Symbolic identifier	*Emblem*	*Logo-symbol*	✓	✓	✓	✓
	–	*Logotype with symbol*	✓	✓	✓	✓
	Only symbol	*Only symbol*				
	• Pictorial mark	• Iconic symbol	✓	✓	–	–
	• Abstract/symbolic mark	• Abstract symbol	–	✓	✓	–
	• Letterform	• Alphabetic symbol	–	✓	✓	–

3.1 | Data preprocessing

Data preprocessing is performed to prepare the image for the next steps. The logo is first cropped to eliminate the borders containing a uniform background. Logo images used for trademark registration or similarity search (i.e., as long as it is not for the task of searching logos in the wild[3]) generally tend to have a uniform background. Therefore, the images are cropped by eliminating colour-uniform borders so that the logo will occupy all the available space in the image. This makes it possible to homogenize their size and facilitate the comparison process.

The second preprocessing step consists of detecting whether the input logo contains text and, if so, generating a new version of it without text. Many image brands include text. However, this information may be irrelevant or even confuse the detection of some logo characteristics. During experimentation, it was observed that shape classification improved notably when the text was eliminated. This was not the case with the other characteristics, such as colour or figurative elements. This process, therefore, was carried out only for the shapes network, using the full logo for the rest of the characteristics, as shown in Figure 2.

To remove the text, the image is first processed using the CRAFT text detector (Baek et al., 2019), which efficiently identifies the text area of an image by exploring each region and the affinity between text characters. As a result, if any text is found, a mask is obtained. Together with the original image, this mask is processed by an inpainting network (Wang et al., 2018) to fill the detected gaps with a background colour. For optimization, when the detected mask is surrounded by white pixels—which is quite common in these kinds of images—the gap is filled directly to white. Figure 3 shows some examples of the steps followed in this process.

3.2 | Multi-label classification

In the second step of the proposed method, a set of neural networks is used for the classification of different characteristics of the input image. Specifically, each of these networks specializes in the MLC of one of the proposed label groups (see Section 2.3.1), such as shape, colour, text, category, subcategory, and sector.

Figure 2 shows a diagram of the integration of these networks into the proposed methodology. The input used is the preprocessed result of the previous stage (in the case of the network specialized in shape, the version of the image without text is employed). The fact that they are independent allows the networks to be run in parallel, signifying that the algorithm's performance is not affected.

As discussed in the introduction, the current methods that obtain the best results as regards processing logos, or images in general, are those based on CNN (LeCun et al., 2010, 2015), and it is for this reason that we also use this type of architecture, but adapting it to an MLC configuration. The specific definition of the networks used is shown in Figure 4 (upper diagram). The proposed architecture consists of five layers alternately arranged into convolutions, batch normalization (Ioffe & Szegedy, 2015), max-pooling and dropout (Srivastava et al., 2014), plus two final fully-connected layers, also with dropout. Batch normalization and dropout (Srivastava et al., 2014) were included to reduce overfitting, help perform faster training, and improve accuracy.
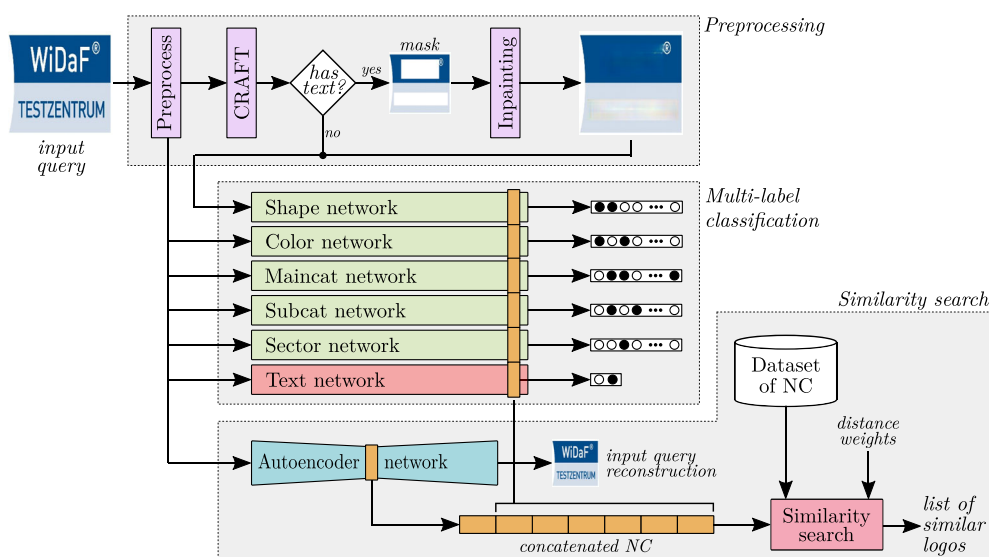

FIGURE 2 Scheme of the proposed method.

FIGURE 3 Examples of how selected text is removed from the image using CRAFT and how an inpainting neural network fills gaps.
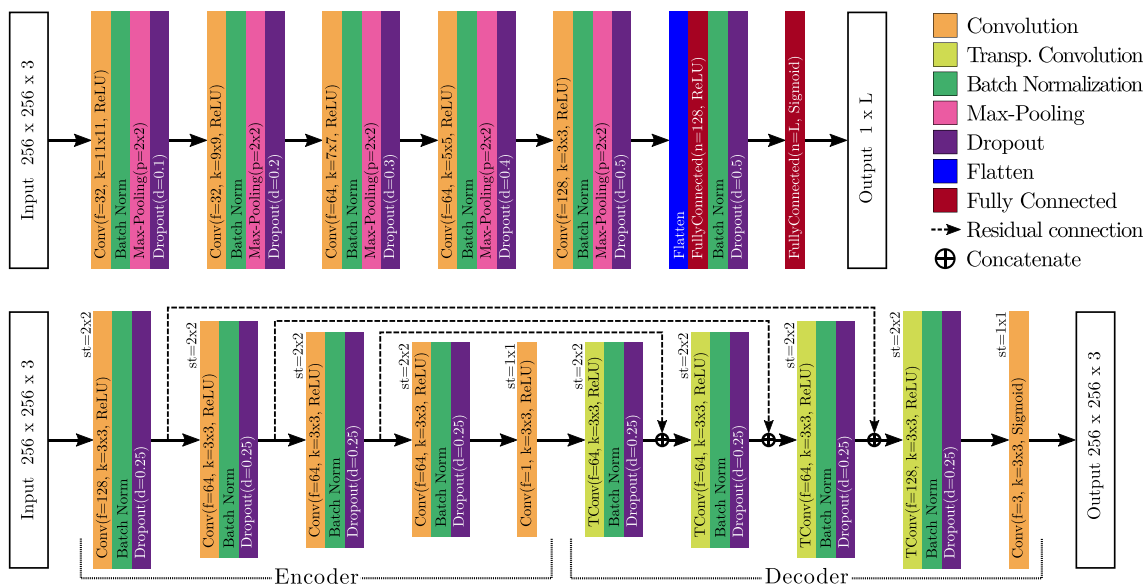

FIGURE 4 Schemes of the specialized convolutional neural networks (CNNs) (top, used in the multi-label classification stage) and the Auto-Encoder (bottom, used for the similarity search). In this figure, the layer type is labelled with colours according to the side legend. Each layer configuration is shown in the scheme, including the activation function, the number of filters (*f*) and kernel size (*k*) for convolutions and transposed convolutions, the pool size (*p*) for max-pooling, the ratio *d* used for dropout, the stride *st* applied to each layer of the auto-encoder, and the number of neurons *n* used for the fully-connected layers. Note that the output size of the top CNN depends on the *L* labels considered.

ReLU (Glorot et al., 2011) was used as the activation function for all layers except the output, which has a sigmoid activation function. The sigmoid function models the probability of each class as a Bernoulli distribution, in which each class is independent of the others, unlike that which occurs with Softmax. Therefore, the output is a MLC for each of the *L* labels considered, which depends on the number of classes of each particular network. This way, the MLC methodology proposed is based on algorithm adaptation, changing the Softmax by a sigmoid to deal with the multi-label scenario while learning the implicit correlations between labels.

In the case of the network specialized in text detection, only one output is necessary since it detects only the presence of text in the image. Unlike CRAFT, which searches for individual characters, this network seeks global features that allow this binary classification to be carried out. As we will see in the experimentation section, this difference has the advantage of allowing a generic comparison of the presence of text in the image (and not of the specific text that appears in it).

3.3 | Similarity search

The last step of the method takes advantage of the intermediate representation learned by the networks described in the previous section to perform the logo-similarity search. In this respect, it is possible to use the CNN as a feature extractor to obtain a suitable mid-level vector representation (also called a Neural Code or NC; Babenko et al., 2014) that is later used as the input for a search algorithm such as *k*NN (Gallego et al., 2020). This is done by feeding the trained networks with the raw data and extracting the NC from one of the last layers of the network (Huang & LeCun, 2006; Razavian et al., 2014), in our case, from the penultimate layer.

In addition to the six specialized networks used in the previous step, an auto-encoder architecture is added to capture generic characteristics that define logos. These networks were proposed decades ago by Hinton and Zemel (1994) and have since been actively researched (Baldi, 2012). They consist of feed-forward neural networks trained to reconstruct their input. They are usually divided into two stages: the first part (denominated as the *encoder*) receives the input and creates a meaningful intermediate or latent representation of it, and the second part (the *decoder*) takes this intermediate representation and attempts to reconstruct the input.

Figure 4 (bottom) depicts the topology of the auto-encoder used. The encoder consists of four convolutional layers combined with batch normalization and dropout. Down-sampling is performed by convolutions using strides rather than resorting to pooling layers. Four mirror layers are then followed to reconstruct the image to the same input size. Up-sampling is achieved through transposed convolution layers, which perform the inverse operation to a convolution to increase rather than decrease the resolution of the output. Residual connections were also added from each encoding layer to its analogous decoding layer, thus facilitating convergence and improving the results.

The size of the neural codes (NC) obtained from these networks is 128 for the CNNs and 256 for the auto-encoder. In preliminary experiments, it was observed that the accuracy decreased with smaller sizes and that larger sizes did not lead to any improvement. As shown in Figure 2, these NC are combined into a single feature vector, which is then used to perform the similarity search. An ℓ_2 normalization (Zheng et al., 2016) is applied for the regularization of this vector since this technique usually improves the results (Gallego et al., 2018).

During the training stage, the NCs from the training set are extracted and stored following the process described above. Then, in the inference phase, the NC representation of the query is obtained and compared with the stored NCs. In this process, in addition to using *k*NN (Duda et al., 2001), the result obtained was also compared with the following two multi-label similarity search methods:

- **Binary Relevance *k*NN (BR*k*NN)** (Eleftherios Spyromitros, 2008): This is a multi-label classifier based on the *k*NN method and the binary relevance (BR) problem transformation. It learns one binary classifier for each different label by checking whether samples are labelled with the label under consideration, thus following a one-against-all strategy.
- **LabelPowerset** (Boutell et al., 2004b): This also follows a problem transformation approach in which the multi-label set is transformed into a multi-class set. Then, a classifier (Random Forest, in this case) is trained on all the unique label combinations found in the training data.

Moreover, we use a weighted distance to search for the nearest neighbours. The advantage of this distance is that it allows the user to adjust the search criteria by modifying the weight assigned to each characteristic (e.g., the user can tune the method to give a higher weight to the shape or colour when seeking the most similar logos). The following weighted dissimilarity metric d_w was used to calculate the distance between two vectors A and B:

$$d_w(A,B) = \frac{\sum\limits_{c \in C} w^c d(A^c, B^c)}{\sum\limits_{c \in C} w^c} \tag{1}$$

where C is the set of all possible characteristics (i.e., colour, shape, etc.), A^c and B^c represent the subset of features corresponding to the characteristic c, w^c is the weight assigned to that characteristic, $\forall c \in C : w^c \in [0,1]$, $\sum_{c \in C} w^c = 1$, and $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_0^+$ is the dissimilarity metric used to compare the two vectors. We have employed the Euclidean distance since, as described above, the NC vectors obtained are numerical feature representations.

3.4 | Training process

The training of the networks was made using standard back-propagation, Stochastic Gradient Descent (Bottou, 2010), and considering the adaptive learning rate method proposed in Zeiler (2012). The *binary crossentropy* loss function was used to calculate the error between the CNN output and the expected result. The training lasted a maximum of 100 epochs with a mini-batch size of 32 samples and *early stopping* when the loss did not decrease during 15 epochs.

A model pretrained during 25,000 iterations with more than 10,000 images was used for the CRAFT (Baek et al., 2019) network. The inpainting network (Wang et al., 2018) was initialized with a model pretrained with ImageNet and fine-tuned with our dataset during 30,000 iterations.

4 | EXPERIMENTAL SETUP

4.1 | Dataset

The experimentation was carried out using the European Union Trademark (EUTM) dataset provided by EUIPO[4]. This dataset is labelled using the Vienna classification as depicted in Section 2.3. However, since the available labelling is not exhaustive, a filtering process was performed to select only those logos whose semantics, colour, and shape were labelled. We, therefore, eventually chose a subset of 76,000 logos corresponding to the 2010–2018 period.

It is important to state that even if this filtering is performed, the collected labelling is still not complete. This is owing to the subjectivity of some labels and the fact that operators usually indicate only the most representative characteristics of logos, that is, those that are distinctive of that brand. For example, in the first image in Figure 5, it will be noted that only the colour red was labelled, although it also contains black and blue. The same happens with the third, fourth, and fifth logos, in which only one colour is labelled although they contain more. In the case of the shape labelling, only circles were annotated in the first and fifth images, lines for the second logo, and triangles for the third, although they also contain other shapes.

In the case of text labelling, only 30% of the images had this information. Again, in this dataset, the presence of text is labelled only when it is a distinctive element. For example, Figure 5 shows that the text is only labelled in the first three images when all the logos contain text. For this reason, all the images were processed and reviewed using the CRAFT text detector to complete this labelling, thus obtaining much more complete ground truth (GT) for this feature.

The input images were scaled to a spatial resolution of 256 × 256 pixels, and their values were normalized into the range [0, 1] to feed the networks. Of the 76,000 logo images, 80% were selected for training, and the remaining samples were employed for testing. These partitions were maintained for all experiments, keeping the same train and test sets to ensure a fair comparison.

4.2 | Metrics

In multi-label learning, each sample may have more than one ground-truth label. To assess this problem quantitatively, better ranks are assigned as the method correctly predicts more GT labels. In this work, we considered the following two multi-label metrics (Tsoumakas et al., 2010).

4.2.1 | Label ranking average precision

This is a LR metric that is linked to the average precision score but based on the notion of LR rather than precision and recall. Label ranking average precision (LRAP) averages over the samples the answer to the following question: for each GT label, what fraction of higher-ranked labels were true labels? This performance measure will be higher if the method can give a better rank to the labels associated with each sample. The score is always strictly greater than 0, with 1 being the best score.

FIGURATIVE 01.15 - 18.03	FIGURATIVE 01.03 - 25.05	FIGURATIVE -	FIGURATIVE 01.01 - 25.12	FIGURATIVE 05.07
COLOR 29.01.01	COLOR 29.01.04/96/98	COLOR 29.01.01	COLOR 29.01.04	COLOR 29.01.03
SHAPE 26.01	SHAPE 26.11	SHAPE 26.03	SHAPE 26.15	SHAPE 26.01
SECTOR 12 - 35 - 40	SECTOR 05	SECTOR 07 - 09 - 37	SECTOR 07	SECTOR 43
TEXT 27.05.05	TEXT 27.05.21 - 27.99 .03	TEXT 27.05.05	TEXT -	TEXT -

FIGURE 5 Some examples of trademarks in the EUTM dataset. Note that some of them have only partial labelling of some characteristics, such as colour and shape, and that the text is not labelled although it is present, as it is not considered to be a characteristic element of the design.

Formally, given a binary indicator matrix of the GT labels, $y \in \{0,1\}^{N \times L}$, where N and L are the amount of samples and labels, respectively, and the score associated with each label $\widehat{f} \in \mathbb{R}^{N \times L}$, the LRAP is defined as:

$$\text{LRAP}\left(y,\widehat{f}\right) = \frac{1}{N} \sum_{i=0}^{N-1} \frac{1}{\|y_i\|_0} \sum_{j:y_{ij}=1} \frac{|\mathscr{L}_{ij}|}{\text{rank}_{ij}} \tag{2}$$

where $\mathscr{L}_{ij} = \left\{ k : y_{ik} = 1, \widehat{f}_{ik} \geq \widehat{f}_{ij} \right\}$, $\text{rank}_{ij} = \left| \left\{ k : \widehat{f}_{ik} \geq \widehat{f}_{ij} \right\} \right|$, $| \cdot |$ calculates the cardinality (number of elements) of the set, and $\| \cdot \|_0$ is the ℓ_0-norm that computes the number of nonzero elements in a vector. If there is exactly one relevant label per sample, LRAP is equivalent to the mean reciprocal rank.

4.2.2 | Label ranking loss

This LR metric computes the ranking, which averages the number of label pairs that are incorrectly ordered in the samples (true labels with a lower score than false labels), weighted by the inverse of the number of ordered pairs of false and true labels. The best performance is achieved with a label ranking loss (LRL) of zero. This metric is formally defined as:

$$\text{LRL}\left(y,\widehat{f}\right) = \frac{1}{N} \sum_{i=0}^{N-1} \frac{1}{\|y_i\|_0(L - \|y_i\|_0)} \left| \left\{ (k,l) : \widehat{f}_{ik} \leq \widehat{f}_{il}, y_{ik} = 1, y_{il} = 0 \right\} \right| \tag{3}$$

5 | EVALUATION

The proposed methodology is evaluated at different levels, starting with the MLC phase and continuing with the similarity search, also comparing it with other state-of-the-art approaches. In addition to quantitative results, a qualitative evaluation is carried out by analysing the response of each stage of the method and comparing the results with the classification made by experts and graphic design students.

5.1 | Multi-label classification

At this stage, the method returns a MLC for each characteristic considered, that is, colour, shape, main category, subcategory, and sector. Table 2 shows the results obtained for each of these characteristics in terms of LRAP and LRL. There is a consensus regarding the best and worst results for both metrics except for the sector. Overall, colour, main category, and sub-category are the best-detected characteristics. The worst-ranked feature using LRL is the sector, probably because no specific pattern, characteristic, or type of design is detected with which to determine it since the type of design applied to each sector is subjective. However, the sector is the best-ranked feature according to LRAP, but probably because usually there is a single sector label per image.

Intermediate precision was attained for shape classification, mainly owing to the labelling noise and the ambiguity of the possible classes. In these results, there is also an improvement produced by the proposed preprocessing to eliminate the text from the image ('Shape+' row) compared to using the original version of the logo that includes the text ('Shape' row).

TABLE 2 Results obtained with the proposed method for the multi-label classification stage. Two cases are shown for the Shape network: 'Shape+', which includes the preprocessing to remove the text, and 'Shape', which does not. Lower LRL values and higher LRAP values indicate better performance.

Model	LRL ↓	LRAP ↑
Colour	0.0561	0.8642
Sub-category	0.0561	0.7376
Main category	0.0635	0.7979
Shape+	0.1169	0.7699
Shape	0.1534	0.6899
Sector	0.2220	0.8890

The results for the text classification network are not included in this table since this is not a multi-label classifier (it discriminates only whether or not the image contains text). For this reason, the accuracy metric was chosen, obtaining 96.06% for this task.

Table 3 depicts some examples of the results obtained for MLC, including only the predictions made with a confidence greater than 2%. When the prediction is compared with the GT, the method succeeds in all cases, with a fairly high confidence percentage. The only error was in the main category of the second logo since the class 'plants' was selected as the first option. However, this error is understandable when the examples labelled with this class are analysed since they usually are green and define shapes with curves. For the shape labelling, it can be seen that in some examples, such as the first, third and fourth, other classes that were not labelled are proposed but that, nevertheless, describe characteristics present in the logos.

5.2 | Similarity search

In this section, we evaluate the similarity search results using the NCs learned by the neural networks in the MLC stage together with the NCs of the auto-encoder. In this case, the results are reported by considering only the LRAP metric since, as stated in the previous section, the tendency of both metrics is similar.

To establish the value of k used by the kNN and BRkNN while simultaneously analysing the labelling noise, we shall now evaluate the result obtained when performing the similarity search for the single-label case. For this, when processing each class, only the samples with a single label for that characteristic were considered. Table 4 depicts the results of this experiment (in terms of LRAP) when considering the kNN method and values of k in the range [1, 11]. As will be noted, the best results are obtained with high k values, between 7 and 11. The intermediate value of $k = 9$ was eventually chosen for the remaining experiments since this reports a higher average result. These results demonstrate that the labels provided contain noise since the method improves by considering more neighbours in the inference stage.

Since LabelPowerset is based on Random Forests, we also carried out a similar experiment by evaluating the number of trees considered in the range $t \in [100,500]$, eventually obtaining the best result with $t = 100$. These parameter settings were used to compare the three multi-label similarity search algorithms: kNN and BRkNN with $k = 9$, and LabelPowerset with $t = 100$. Table 5 shows the results of this experiment using the LRAP metric. As can be seen, a better result is obtained for almost all the characteristics when using LabelPowerset. The only exception is the sector, which, as previously argued, is a very subjective characteristic and may contain a higher level of noisy labels.

In the case of the auto-encoder, it is necessary to consider that it is trained in an unsupervised manner for the reconstruction of the input. The labelling of characteristics is, therefore, not used during training. For this reason, to assess its performance, its result for the characteristics

TABLE 3 Examples of multi-label classification of the EUTM dataset, including the ground truth (GT) and the prediction made by the main category, shape, colour, and text networks when the confidence percentage of the prediction exceeds 2%. Two examples include text, and two do not.

GT	Main-category	Ornamental motifs	Human beings	Plants	Ornamental motifs games, toys
	Shape	Quadrilaterals lines, bands	Circles, ellipses	Lines, bands	Quadrilaterals
	Colour	Black; Orange	Green	Blue	Black; White
	Text	Yes	Yes	No	No
Prediction	Main-category	100% Ornamental motifs	47.10% plants 17.70% human beings 4% arms, ammunition	46.47% plants 31% heraldry, coins 9.66% celestial bodies	100% Ornamental motifs
	Shape	94.85% Quadrilaterals 6.07% lines, bands 4.08% other polygons	99.81% circles, ellipses	63.62% circles, ellipses 61.87% lines, bands 10.44% quadrilaterals	99.96% Quadrilaterals 10.06% circles, ellipses
	Colour	48.44% black 94.41% orange 58.45% white	99.22% green	100% blue	87.54% black 78.36% white
	Text	Yes	Yes	No	No

TABLE 4 Similarity search results (in terms of LRAP) obtained with the kNN classifier for the single-label search task and different k values.

	CNN + kNN					
Model	k = 1	k = 3	k = 5	k = 7	k = 9	k = 11
Colour	0.8322	0.8366	0.8369	**0.8394**	0.8378	0.8331
Main Category	0.7673	0.7842	0.7875	0.7885	**0.7886**	0.7880
Subcategory	0.7409	0.7611	0.7660	0.7682	0.7695	**0.7716**
Sector	0.8020	0.8027	0.8054	0.8060	0.8065	**0.8067**
Shape	0.5489	0.5513	0.5503	0.5542	0.5544	**0.5552**
Shape+	0.6583	0.6717	0.6707	0.6689	0.6712	**0.6728**
Average*	0.7601	0.7713	0.7733	0.7742	**0.7747**	0.7744

Note: The best results are highlighted in bold type.
*The average is calculated excluding the result from the 'Shape' network (without preprocessing), due to the superior performance of the 'Shape+' network, which will be employed in further analyses.

TABLE 5 Results obtained for the different characteristics with the three multi-label classifiers using the LRAP metric.

Network	kNN	BRkNN	LabelPowerset
Colour	0.7042	0.7042	**0.7070**
Main category	0.7015	0.7015	**0.7396**
Subcategory	0.6589	0.6589	**0.6850**
Sector	**0.8434**	**0.8434**	0.8001
Shape	0.5333	0.5333	**0.5594**
Shape+	0.6242	0.6242	**0.6579**
Average*	0.7064	0.7064	**0.7179**

Note: The best results are shown in bold type.
*The average is calculated excluding the result from the 'Shape' network (without preprocessing), due to the superior performance of the 'Shape+' network, which will be employed in further analyses.
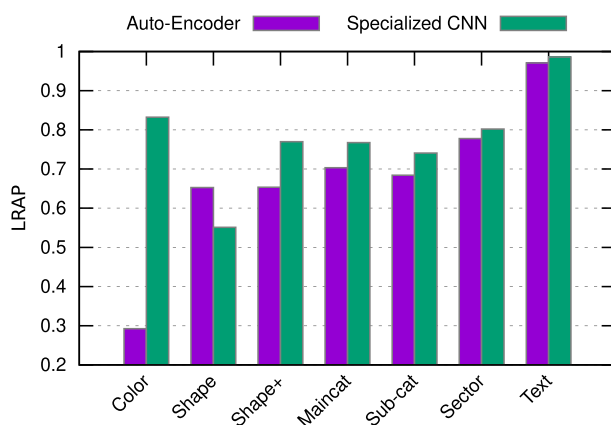
FIGURE 6 Similarity search results obtained for each of the characteristics considered when using the neural codes learned by the auto-encoder. The results obtained by the specialized networks for these characteristics are included as a reference.

considered is compared with that obtained by the specialized networks. Figure 6 shows this analysis. Good results are yielded for almost all characteristics except for colour. This indicates that the auto-encoder learns a generic representation of combined features, which primarily considers shape versus colour.

It is also evident that the auto-encoder works even better than the shape network when the text is not eliminated, but this is not the case when the proposed preprocessing (Shape+) is applied. The auto-encoder is not the best for any particular feature (except for shape without preprocessing). This method is, therefore, beneficial for searching for similarity generically, considering appearance without looking at any specific characteristic.

5.2.1 | Qualitative results

In this section, we qualitatively analyse the results obtained after the similarity search. Figure 7 includes a series of examples of the logos found when using each specialized network separately, assigning 100% of the search weight to a single characteristic. In this figure, the first logo in the row is the query, and the others are the 8-nearest neighbours retrieved.

In the case of colour (first row of the figure), it will be noted that the results retrieved are correctly matched, even when there are multiple colours, independently of other characteristics such as the shape. The second row depicts an example of shape, which is also perfectly detected without, in this case, taking into account colour.

The main category and sub-category (3rd and 4th rows) of figurative designs are more difficult to analyse visually since elements can often be represented creatively or abstractly. It is for this reason that 'Plants' has been selected for the main category and 'Leaves, needles, branches with leaves or needles' for the sub-category, as they contain easily recognizable designs. In both cases, it will be noted that similar logos, in which leaves or plants appear, have been retrieved. For the main category, the design appears to be a little more generic, including other elements such as people, while for the sub-category, the designs are more specific, and only logos that include leaves are shown.

The case of the sector (5th row) is even more difficult to analyse visually since the classification into goods and services is quite subjective and does not always depend on visual information. Nevertheless, this example shows a correct search result for a logo used for goods. In the case of the text (penultimate row), in addition to retrieving logos containing text, the model also considers the logo's composition since a similar design

FIGURE 7 Example of the 8-nearest neighbours obtained by using each of the specialized networks separately, that is, assigning 100% of the search weight to a single feature. The first logo is the query.

appears in all of them (with the text at the bottom). Finally, the auto-encoder (last row) focuses principally on the spatial distribution or the layout of the logo and, in some cases, also considers the colours.

We shall now analyse the effect of combining several characteristics using the proposed weighted distance (see Equation 1) and the capacity it gives users to refine the search. Figure 8 shows some examples of the results obtained by applying different weights to combine colour, shape, and figurative elements.

In the first row, the logo used previously in Figure 7 for the shape characteristic is evaluated, but shape and colour are combined in this example. As can be seen, when adding the colour, circular logos are again retrieved, but in this case, they have similar colours. When reducing the weight of the colour to 30%, other colours such as blue begin to appear, but red and black are always maintained. These results contrast those previously obtained in Figure 7, in which the colours changed completely.

In the second example, the same logo from Figure 7 (3rd row) is evaluated, but in this case, the figurative elements from the main category are combined with the shape. As will be noted, by giving some weight to shape, the recovered figurative elements keep the same shape, unlike the previous result in which this characteristic was not considered. By assigning 30% of the weight to the shape, only two logos that do not have a circular shape are obtained, and by giving more weight to the shape (70%), all the results obtained are 'circles, ellipses'.

To analyse the representations learned by the models, a visualization of the grouping formed by the NCs is included for the colour and shape characteristics using the t-Distributed Stochastic Neighbor Embedding technique (t-SNE; van der Maaten, 2008). Figure 9 shows that, despite being a multi-label task, the learned NCs tend to group similar characteristics. For example, in the case of the colour (top image), grey or silver are grouped in the upper right, blues on the right, yellows, browns, oranges on the left, and greens in the upper left part. Similar shapes are also grouped (see images below, in which two areas of the representation generated are shown zoomed in). In the left-hand image, there are circular shapes, and in the right-hand one, there are quadrilaterals. It should be noted that logos that include text next to these shapes are also grouped separately.

These results show how the networks transform the input images into a new dimensional space (the NCs extracted) in which the logos with similar characteristics are close. This makes it possible to perform a similarity search based on the distance between the representation of the logos in this dimensional space and thus analyse the neighbourhood space of a given query to retrieve similar images.

## 5.2.2 | Class activation maps

Figure 10 shows five examples of class activation maps (CAM) that highlight the areas of the logo that were most relevant for its classification. These maps were obtained using Grad-CAM (Gradient-weighted class activation mapping; Selvaraju et al., 2017), a technique to visually explain the decisions of a CNN by highlighting the regions of the input image that are more important for the prediction of a specific class. This method backpropagates the gradients obtained at the final convolutional layer of the CNN in order to produce a coarse localization map highlighting important regions for the prediction.

In Figure 10, it can be seen that the model focuses on the areas of the image where the most relevant characteristics are best visually appreciated. In the first, second, and last images, the model focuses on the areas where the colour, shape, and text are identified,


FIGURE 8 Results obtained using the weighted distance with two different characteristics. The first column shows the query and the weights applied. The second column includes the 8-nearest neighbours retrieved.
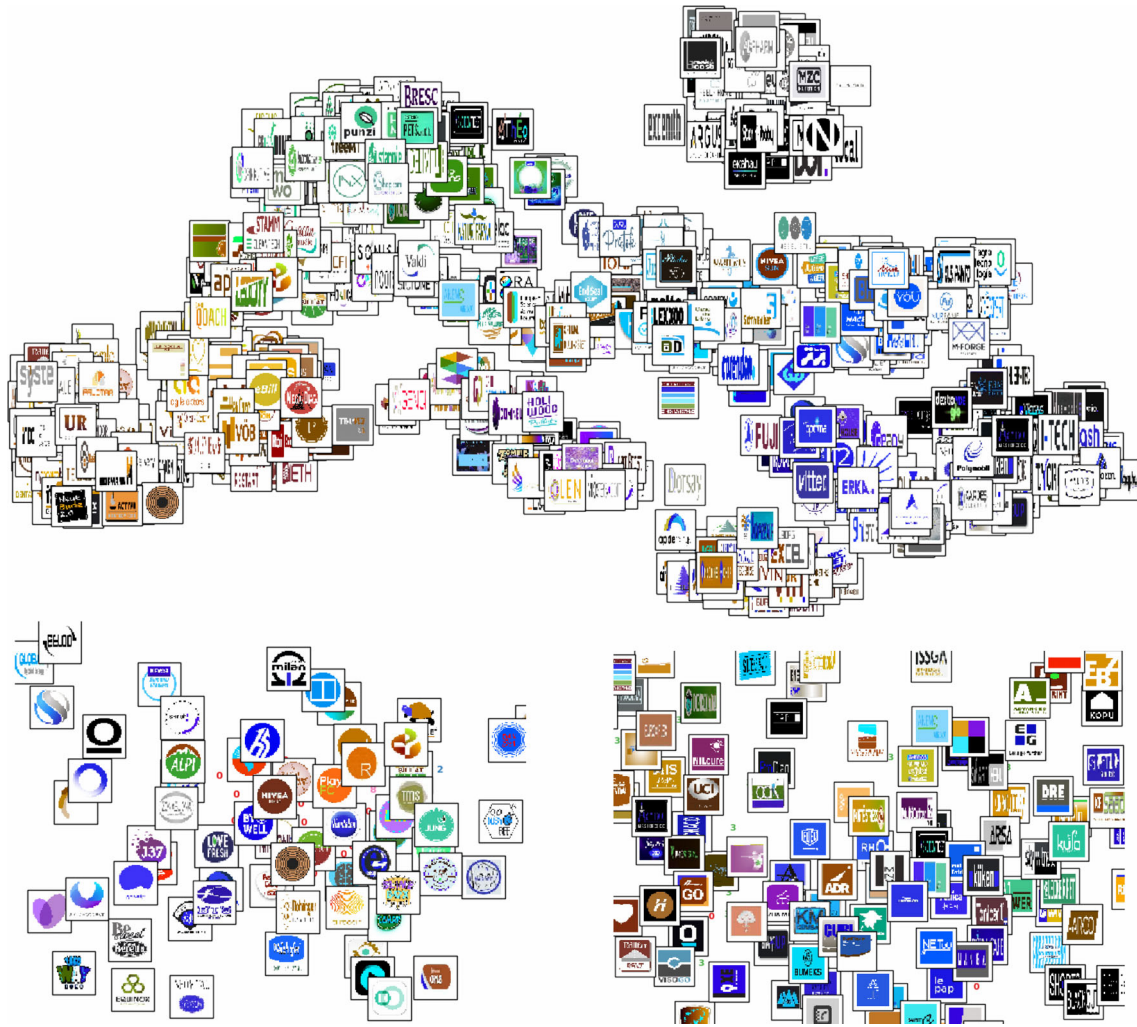
FIGURE 9 Clusters formed by the neural codes from the networks of colour (top) and shape (bottom) using the t-Distributed Stochastic Neighbor Embedding method. In the case of the shape, two images are included by zooming in on areas in which the circular (left) and quadrilateral (right) shapes are located.
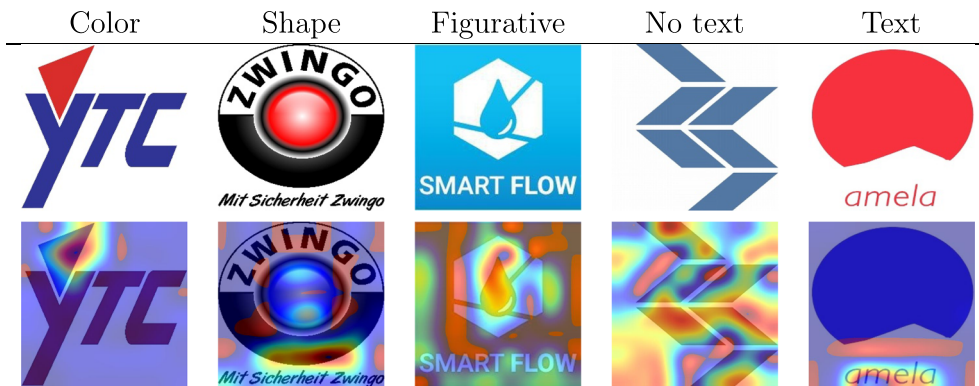
FIGURE 10 Grad-CAM activation maps for logo search using colour, shape, figurative elements, and text classifiers. These maps were obtained from the `dropout-4` layer of the network specialized for each characteristic.

respectively. In the figurative case, the CNN focuses on the centre of the image that contains the main identifiable element, in this case, a drop. In the image without text, the network does not identify any specific area, since the CNN is looking for text and the logo does not have any.

5.2.3 | Runtime analysis

We now include an additional benchmark that analyses the search cost taking into account the runtime. It is important to keep in mind that the execution time depends on factors that are not related to the method itself but to the programming language, the libraries used for the implementation, or the hardware settings, among others. In this regard, we clarify that all these experiments were carried out using the Python programming language (v. 3.7) with the TensorFlow (v. 2.1), Keras (v. 2.3), and the cuDNN libraries. The machine used consists of an Intel(R) Core(TM) i7-7700HQ CPU @ 2.80 GHz with 16 GB RAM and an NVIDIA GeForce GTX 1060 with 6 GB graphics processing unit (GPU). Note also that to obtain the execution times, 50 predictions were made and the average of the inference time and the deviation of the obtained times were calculated.

The pipeline of the proposed approach is divided into three main steps (see Figure 2): preprocessing, MLC, and similarity search. The first two stages take only 5.26 ± 0.9 ms and 2.19 ± 0.37 ms, respectively. It must be taken into account that in these stages the prediction consists of the use of trained networks, which basically implies matrix multiplication. The third stage takes a little longer since it performs the similarity search of the query and the entire training set. However, this process can be optimized by using a KD search tree, taking only 12.76 ± 1.2 ms.

5.3 | Comparison with state of the art

This section compares the proposed method with other state-of-the-art methods for TIR. Since, and as previously mentioned, there are, to the best of our knowledge, no other MLC approaches for logos that use the Vienna classification, we perform this comparison using METU v2.[5] This dataset is the largest public dataset for TIR and contains 922,926 trademark images belonging to approximately 410,000 companies. Its evaluation set is composed of 417 queries divided into 35 groups of about 10–15 trademarks, in which the logos within the same group are similar (Tursun et al., 2017; Tursun & Kalkan, 2015).

The evaluation was carried out using the normalized average rank (NAR) metric since it is the measure most commonly employed in reference state-of-the-art works. This metric is calculated by injecting the query set into the main dataset and, for each query logo, the rank obtained for the logos in the same group is calculated as follows:

$$\text{NAR} = \frac{1}{N \times N_{rel}} \sum_{i=1}^{N_{rel}} R_i - \frac{N_{rel}(N_{rel} + 1)}{2} \tag{4}$$

where N_{rel} is the number of relevant images for a particular query image (the number of injected images), N is the size of the image set, and R_i is the rank of the ith injected image. The value 0 corresponds to the best performance and 0.5 to a random order.

Table 6 shows the result of the comparison carried out. As can be seen, different types of approximations were considered, which were based on both hand-crafted features and neural networks. In the case of those based on hand-crafted features, the use of colour histograms (Lei et al., 1999), LBP (Ojala et al., 2002), SIFT (Lowe, 2004), SURF (Bay et al., 2008), TRI-SIFT and OR-SIFT (Kalantidis et al., 2011) was compared. We also considered two more elaborated proposals: the use of SIFT while excluding the features of the text areas (Perez et al., 2018), and an enhanced version of SIFT (Feng et al., 2018) in which reversal invariant features are extracted from edges of segmented blocks which are then aggregated to perform the similarity search.

The use of pre-trained neural network models was also compared. In particular, we evaluated GoogLeNet (Szegedy et al., 2015), AlexNet (Krizhevsky et al., 2012), and VGG16 (Simonyan & Zisserman, 2015), extracting the NCs from one of its layers (77S1, FC7, and Pool5, respectively). Specific proposals for this dataset were also considered, such as the work of Tursun et al. (2017), in which six hand-crafted features are combined with NCs extracted from three different CNN architectures. We also evaluated the proposal of Perez et al. (2018), which compares three solutions: the results of the VGG19 architecture trained in two ways (one to distinguish visual similarities and the other for conceptual similarities), and the result of merging the features of both. Finally, we included a work based on attention mechanisms (Tursun et al., 2020), which pays direct attention to critical information, such as figurative elements, and reduces the attention paid to non-informative elements, such as text and background. This process, denominated as ATRHA (automated text removal hard attention), is combined with two proposals for the elaboration of the features compared, one based on the regional maximum activations of convolutions (R-MAC) and the other based on the saliency of convolutional activations maps (CAM) that were detected through the use of soft attention mechanisms (CAMSA) and the aggregation of maximum activations of convolutions (MAC).

As noted in the results shown in Table 6, generally, the methods based on neural networks are significantly better than those based on hand-crafted features. There are, however, some exceptions: since the pre-trained networks have not been specifically prepared for this type of data, they do not achieve good results and are even surpassed by a method based on hand-crafted features ('Enhanced SIFT'; Feng et al., 2018). It is interesting to see how the combination of hand-crafted features with features extracted from a CNN (proposed in Tursun et al., 2017) achieves a

Expert Systems —WILEY—

TABLE 6 Comparison with the previous state-of-the-art results for METU dataset. NAR is the normalized average rank metric. Smaller NAR values indicate better results.

Approach	Method	NAR
Hand-crafted features	Colour histograms (Lei et al., 1999)	0.400
	SIFT (Lowe, 2004)	0.348
	TRI-SIFT (Kalantidis et al., 2011)	0.324
	LBP (Ojala et al., 2002)	0.276
	SURF (Bay et al., 2008)	0.207
	OR-SIFT (Kalantidis et al., 2011)	0.190
	SIFT without text (Perez et al., 2018)	0.154
	Enhanced SIFT (Feng et al., 2018)	0.083
Neural networks-based	GoogLeNet (Szegedy et al., 2015)	0.118
	AlexNet (Krizhevsky et al., 2012)	0.112
	VGG16 (Simonyan & Zisserman, 2015)	0.086
	Visual network (Perez et al., 2018)	0.066
	Conceptual network (Perez et al., 2018)	0.063
	ATRHA R-MAC (Tursun et al., 2020)	0.063
	Fusion of hand-crafted & CNN features (Tursun et al., 2017)	0.062
	Fusion of visual and conceptual networks (Perez et al., 2018)	0.047
	ATRHA CAMSA MAC (Tursun et al., 2020)	0.040
Our approach	Autoencoder	0.118
	Colour	0.090
	Shape	0.044
	Weighted features (70% colour, 30% shape)	0.034
	Weighted features (30% colour, 70% shape)	0.018

notable improvement. Of the methods based solely on neural networks, the proposal that uses attention mechanisms (Tursun et al., 2020) stands out.

Concerning the results obtained by our proposal, it can be seen that the auto-encoder obtains low results for this task. These are similar to those attained by the approximations based on hand-crafted features, possibly because it is trained unsupervised and learns overly generic features. Using the features learned by the networks specialized in colour and shape separately, the results improve, with the best being the result obtained for the shape. In particular, the shape classifier is better than all the state-of-the-art works except (Tursun et al., 2020). Finally, the results are further improved when our proposal combines the features in a weighted manner, surpassing the other state-of-the-art methods. When assigning more weight to shape (70%) than to colour (30%), our method outperforms previous works by a notable margin.

Figure 11 shows an example of the results obtained for the METU dataset when using our proposal combining the characteristics of shape (70%) and colour (30%). In this figure, the first logo is the query, and the rest are the ten most similar logos—an asterisk (*) marks the correct results. For the query in the first row, the GT contains thirteen similar logos. Our method found eight among the first ten results; the others are in positions 11, 16, 17, 20, and 23. For the query in the second row, the GT had nine similar logos. In this case, the method returned seven of them among the first ten results, and the other two were in positions 57 and 65 (out of a total of 923,340 possible logos).

5.4 | Surveys

Since the classification of brands can often be subjective, to assess the effectiveness of our proposal, we also evaluated (using the same metrics) the results that experts in this task would obtain.

This was achieved by surveying 107 graphic design students and professionals. In this survey, 3 logos with colour labels, 3 with shapes, and 6 with figurative elements were randomly selected for each participant, asking 12 questions per participant. A reduced set of possible answers to each question was provided, and the participants were asked to mark only the labels they considered to be present in the logo.
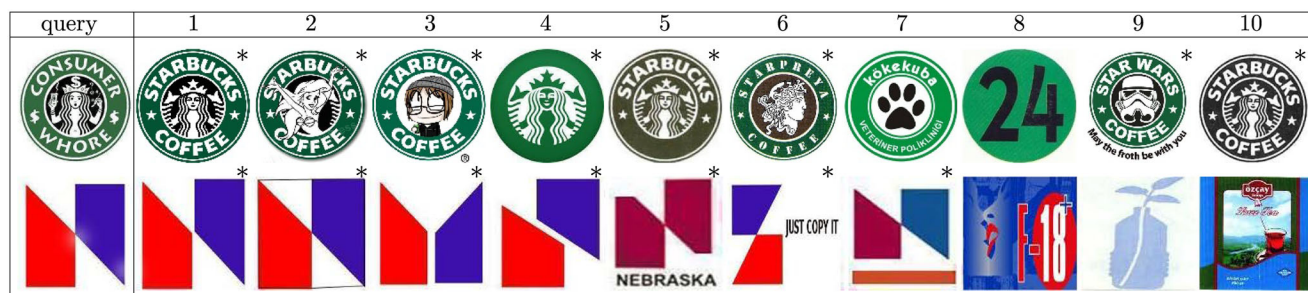
FIGURE 11 Two examples of the 10-nearest neighbours obtained in the METU dataset by assigning 30% of weight to colour and 70% to shape. The first logo is the query. The correct results found are marked with an asterisk (*).

TABLE 7 Results obtained in the survey of design students and professionals using the LRAP metric, compared with the result obtained by our proposal. Higher LRAP values indicate better results.

Criteria	Students and professionals of design	Our proposal
Colour	0.6735	0.7070
Shape	0.5467	0.6579
Sub-category	0.3673	0.6850
Average	0.5292	0.6833

In the colour questions, the participants were shown the following statement: 'Indicate whether you can see the following colors in this logo (the white background is not considered to be a color)'. The following 13 possible colours were then provided, and the respondent had to choose one or more: Red, Yellow, Green, Blue, Violet, White, Brown, Black, Grey, Silver, Gold, Orange, and Pink.

In the case of shape, the respondents were instructed to select the distinctive shapes when provided with 8 possible options: (1) Circles or ellipses; (2) segments or sectors of circles or ellipses; (3) triangles, lines forming an angle; (4) quadrilaterals; (5) other polygons or geometrical figures; (6) different geometrical figures, juxtaposed, or joined; (7) lines, bands; and (8) geometrical solids (3D objects: spheres, cubes, cylinders, pyramids, etc.).

In the case of figurative elements, since there are 123 possible labels, only the correct answers, along with another 4 or 5 incorrect answers, were given to the respondents rather than all the options.

Table 7 depicts the results obtained from the surveys using the same LRAP metric considered previously. These results are compared with those obtained using the CNN networks specialized in classifying these same characteristics (previously shown in Table 5). As can be seen, the proposed methodology improves the precision of the labelling carried out by the professionals and design students surveyed, especially in the case of the labelling of figurative elements. These results confirm the difficulty of this task owing to subjectivity when interpreting the meaning of the elements that appear in a logo or the characteristics that could be considered representative of the brand.

5.4.1 | Analysis of the survey responses

Figure 12 shows some examples of the questions asked in the survey, including the correct responses (based on the database labelling) and statistical data on the number of correct answers provided by the participants. For example, if there are two possible options, the number of participants who got both or only one correct is indicated. The cases in which, in addition to having one or two correct answers, they also answered other incorrect options are also detailed.

Figure 12a,b shows two examples of the colour questions. In the case of the first, no respondent attained the correct answers. Most appreciated blue and red in the image, although the image was not labelled blue but black, and the colour red was not labelled in the dataset. In Figure 12b, most respondents selected the correct answer (some included other options), and seven confused Yellow with Gold or Brown. As can be seen in these examples, people can appreciate colour differently, either by the nature of the individual, the tone assigned to the colour, or defects in the image related to the means of production. Another error source is subjectivity in the labelling process since sometimes only the colour considered representative of the brand is labelled.

Concerning semantic labels, in Figure 12c nine of the respondents answered correctly to both questions, selecting an additional label in two cases. When analysed individually, 95% of the respondents recognized one of the two labels. On the other hand, in Figure 12d, which is labelled with a single class, only 36% marked the correct answer, with the majority selecting other options such as 'Furniture', 'Electrical Equipment' or

(a) Color labels: White; Black

Answers	#	%
Two	0	0
One	0	0
Two (and others)	3	11.54
One (and others)	7	26.92
Others (red and/or blue)	16	61.54

(b) Color labels: Black; Gold

Answers	#	%
Two	19	67.86
One	0	0
Two (and others)	2	7.14
One (and others)	7	25.00
Others	0	0

(c) Figurative labels: Stars, Comets; Armillary Spheres, Planetaria, ...

Answers	#	%
Two	7	33.33
One	8	38.10
Two (and others)	2	9.52
One (and others)	3	14.29
Others	0	0
None	1	4.76

(d) Figurative labels: Lighting Wireless Valves

Answers	#	%
One	2	8.00
One (and others)	7	28.00
Others	14	56.00
None	2	8.00

(e) Shape labels: Circles or ellipses; Segments or sectors of circles or ellipses; Lines, bands.

Answers	#	%
Three	0	0
Two	6	15.38
One	4	10.26
Three (and others)	5	12.82
Two (and others)	13	33.33
One (and others)	8	20.51
Others	3	7.69

(f) Shape labels: Circles or ellipses; Quadrilaterals; Geometrical solids.

Answers	#	%
Three	0	0
Two	1	3.57
One	1	3.57
Three (and others)	3	10.71
Two (and others)	9	32.14
One (and others)	7	25.00
Others	7	25.00

FIGURE 12 Examples of the questions and answers in the surveys. The correct responses and a summary of the answers provided for each option are included for each question.

'Heating, Cooking Or Refrigerating Equipment, Washing Machines, Drying Equipment'. As will be noted, recognizing semantic elements in a logo is not a trivial task. In many cases, figures are oversimplified and may be confused with other representations. In addition, the interpretation often depends on the individuals who perceive it and their cultural, personal, or professional background.

Figure 12e is labelled with three shape classes. Of these, 12 of the respondents (out of a total of 39) got only one correct. The case of Figure 12f is similar since the answers are multiple and there are very different combinations. In this case, the image contains 'Geometrical solids

(3D objects)', and only 5 of the 28 people who evaluated this logo marked this answer. These examples illustrate the complexity of detecting all the shapes in an image. It is usually somewhat subjective if a shape is representative of a logo design. Moreover, a predominant shape can sometimes influence the observer to ignore other shapes in the image.

6 | CONCLUSIONS

This paper presents a methodology for the MLC of logos, considering main characteristics such as colour, shape, semantic elements, and text. Furthermore, the proposed method also allows obtaining a ranking of the most similar logos, in which users can select the characteristics to consider in the search process. To the best of our knowledge, no other methods in the literature address these two objectives. Therefore, a proposal of this kind is of great interest, both methodologically and practically, as regards assisting in multiple tasks, such as labelling logos, detecting plagiarism, or similarities between brands.

The proposed architecture combines, in a weighted fashion, the representation learned by a series of MLC networks that specialize in detecting the most distinctive characteristics of logos. Moreover, the method performs a preprocessing stage to remove uniform backgrounds and text from input images. The experiments showed that removing the text from the logo helps classify the shape, but not other types of characteristics. This may be because the text often includes representative characteristics of the logo, such as colour or figurative elements, and removing them worsens the result.

The experimental results show that the proposed approach is reliable for both classification and similarity search. Furthermore, the comparison made with 17 state-of-the-art TIR methods shows that our proposal is notably better than previous approaches, especially considering colour and shape.

This paper also studies the logo labelling issues in trademark registration databases since only the most distinctive characteristics of the brand are generally labelled by registration agencies, resulting in incomplete and often inconsistent labelling. Moreover, the semantics of trademarks can be subjective, which results in difficulties for operators. These problems are produced either by the labelling process itself or are motivated by the Vienna coding since it is a closed categorization and some characteristics are challenging to define. We also conclude that the provided labels contain noise since the method improves by considering more neighbours in the inference stage. For future work, an estimation of the noise rate should be performed to deal with this issue (Song et al., 2023), selecting the most suitable method for noise mitigation and making changes in the architecture, regularization, loss function, or sample selection.

One of the proposed methodology's advantages is aiding in this task, since it suggests an initial classification that follows homogeneous criteria, which, in addition to facilitating the work, is complete and exhaustive. Furthermore, given that many people label ground-truth data, an automatic classification method reduces the inconsistency of human subjectivity caused by the different perceptions of the same visual representation and the difficulty of expressing graphic qualities in words.

We also performed a qualitative evaluation, which was carried out with expert designers to assess labelling consistency. These experiments showed that the proposed methodology provides better labelling than a human operator would assign, even in the case of experts in this task. The labelling suggested by the system could be used as an initial proposal to be reviewed by the operator. In addition, students and design professionals could use the system as aid since they could check the labelling proposal for a new design, search for references, ideas, and styles, or detect similar marks and possible plagiarism.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID
Marisa Bernabeu https://orcid.org/0000-0003-3722-1379
Antonio Javier Gallego https://orcid.org/0000-0003-3148-6886

ENDNOTES
[1] This work is available in a non-commercial preprint server, in accordance with Wiley's Preprint Policy, and can be accessed at the following Arxiv address: https://arxiv.org/abs/2205.05419.

[2] https://euipo.europa.eu/ohimportal/en/nice-classification.

[3] This term refers to the task of finding the position of a logo from a generic image that may contain many other elements.

[4] https://euipo.europa.eu/ohimportal/en/open-data.

[5] https://github.com/neouyghur/METU-TRADEMARK-DATASET.

REFERENCES

Babenko, A., Slesarev, A., Chigorin, A., & Lempitsky, V. (2014). Neural codes for image retrieval. In *European Conference on Computer Vision (ECCV)* (pp. 584–599). Springer.

Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9365–9374). IEEE.

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning* (pp. 37–49). PMLR.

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding, 110*(3), 346–359. https://doi.org/10.1016/j.cviu.2007.09.014

Bianco, S., Buzzelli, M., Mazzini, D., & Schettini, R. (2017). Deep learning for logo recognition. *Neurocomputing, 245*, 23–30. https://doi.org/10.1016/j.neucom.2017.03.051

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of the 9th International Conference on Computational Statistics, COMPSTAT 2010* (pp. 177–186). Springer.

Boutell, M. R., Luo, J., Shen, X., & Brown, C. M. (2004). Learning multi-label scene classification. *Pattern Recognition, 37*(9), 1757–1771. Elsevier.

Capsule. (2007). *Design matters: Logos 01: An essential primer for today's competitive market*. Rockport Publishers.

Chaves, N., & Belluccia, R. (2003). *La Marca Corporativa: Gestión y Diseño de Símbolos y Logotipos*. Estudios de Comunicación Series.

Chiam, J.-H. (2015). *Brand logo classification*, Technical Report. Stanford University.

Clare, A., & King, R. D. (2001). Knowledge discovery in multi-label phenotype data. In *European conference on principles of data mining and knowledge discovery* (pp. 42–53). Springer.

Dong, H., Wang, W., Huang, K., & Coenen, F. (2020). Automated social text annotation with joint multi-label attention networks. *IEEE Transactions on Neural Networks and Learning Systems, 32*(5), 2224–2238. https://doi.org/10.1109/TNNLS.2020.3002798

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). Wiley.

Eleftherios Spyromitros, I. V. (2008). Grigorios Tsoumakas, An empirical study of lazy multilabel classification algorithms. In *5th Hellenic Conference on Artificial Intelligence (SETN 2008)* (pp. 401–406). Springer.

Elisseeff, A., & Weston, J. (2001). A kernel method for multi-labelled classification. In *Advances in neural information processing systems* (Vol. 14, pp. 681-687). MIT Press.

Feng, Y., Shi, C., Qi, C., Xu, J., Xiao, B., & Wang, C. (2018). Aggregation of reversal invariant features from edge images for large-scale trademark retrieval. In *2018 4th International Conference on Control, Automation and Robotics (ICCAR)* (pp. 384–388). IEEE. https://doi.org/10.1109/ICCAR.2018.8384705

Fürnkranz, J., Hüllermeier, E., Loza Mencía, E., & Brinker, K. (2008). Multilabel classification via calibrated label ranking. *Machine Learning, 73*, 133–153.

Gallego, A. J., Calvo-Zaragoza, J., & Rico-Juan, J. R. (2020). Insights into efficient k-nearest neighbor classification with convolutional neural codes. *IEEE Access, 8*, 99312–99326. https://doi.org/10.1109/ACCESS.2020.2997387

Gallego, A.-J., Pertusa, A., & Bernabeu, M. (2019). Multi-label logo classification using convolutional neural networks. In A. Morales, J. Fierrez, J. S. Sánchez, & B. Ribeiro (Eds.), *Pattern recognition and image analysis* (pp. 485–497). Springer International Publishing.

Gallego, A.-J., Pertusa, A., & Calvo-Zaragoza, J. (2018). Improving convolutional neural networks' accuracy in noisy environments using k-nearest neighbors. *Applied Sciences, 8*(11), 2086.

Ghosh, S., & Parekh, R. (2015). Automated color logo recognition system based on shape and color features. *International Journal of Computer Applications, 118*(12), 13–20.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In G. Gordon, D. Dunson, & M. Dudík (Eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Vol. 15 of Proceedings of Machine Learning Research* (pp. 315–323). PMLR.

Guru, D. S., & Kumar, N. V. (2018). Interval valued feature selection for classification of logo images. In A. Abraham, P. K. Muhuri, A. K. Muda, & N. Gandhi (Eds.), *Intelligent systems design and applications* (pp. 154–165). Springer International Publishing.

Hinton, G. E., & Zemel, R. S. (1994). Autoencoders, minimum description length and helmholtz free energy. In *Advances in Neural Information Processing Systems* (pp. 3–10). MIT Press.

Huang, F., & LeCun, Y. (2006). Large-scale learning with SVM and convolutional nets for generic object categorization. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR* (pp. 284–291). IEEE.

Iandola, F. N., Shen, A., Gao, P., & Keutzer, K. (2015). DeepLogo: Hitting logo recognition with the deep neural network hammer, CoRR abs/1510.02131.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)* (pp. 448–456). Curran Associates.

Kalantidis, Y., Pueyo, L. G., Trevisiol, M., van Zwol, R., & Avrithis, Y. (2011). Scalable triangulation-based logo recognition. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ICMR'11, Association for Computing Machinery*. ACM. https://doi.org/10.1145/1991996.1992016

Köstinger, M., Roth, P. M., & Bischof, H. (2010). *Planar trademark and logo retrieval*. Technical Report. Computer Graphics and Vision, Graz University of Technology.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 25). Curran Associates.

Kumar, N. V., Pratheek, V. V. K., Govindaraju, K., & Guru, D. (2016). Features fusion for classification of logos. In *International Conference on Computational Modelling and Security (CMS)* (Vol. 85, pp. 370–379). Elsevier.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010). Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS)* (pp. 253–256). IEEE.

Lei, Z., Fuzong, L., & Bo, Z. (1999). A cbir method based on color-spatial feature. In *TENCON, Proceedings of the IEEE Region 10 Conference* (pp. 166–169). IEEE. https://doi.org/10.1109/TENCON.1999.818376

Liu, W., Wang, H., Shen, X., & Tsang, I. W. (2021). The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(11), 7955–7974.

Lourenço, V. N., Silva, G. G., & Fernandes, L. A. F. (2019). Hierarchy-of-visual-words: a learning-based approach for trademark image retrieval.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 971–987.

Orti, O., Tous, R., Gomez, M., Poveda, J., Cruz, L., & Wust, O. (2019). Real-time logo detection in brand-related social media images. In I. Rojas, G. Joya, & A. Catala (Eds.), *Advances in computational intelligence* (pp. 125–136). Springer International Publishing.

Perez, C. A., Estévez, P. A., Galdames, F. J., Schulz, D. A., Perez, J. P., Bastías, D., & Vilar, D. R. (2018). Trademark image retrieval using a combination of deep convolutional neural networks. International Joint Conference on Neural Networks (IJCNN).

Pornpanomchai, C., Boonsriporchai, P., Puttong, P., & Rattananirundorn, C. (2015). Logo recognition system. In *2015 IEEE International Computer Science and Engineering Conference (ICSEC)* (pp. 1–6). IEEE.

Qi, H., Li, K., Shen, Y., & Qu, W. (2010). An effective solution for trademark image retrieval by combining shape description and feature matching. *Pattern Recognition*, *43*(6), 2017–2027.

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW'14* (pp. 512–519). IEEE Computer Society.

Rusiñol, M., Aldavert, D., Karatzas, D., Toledo, R., & Lladós, J. (2011). Interactive trademark image retrieval by fusing semantic and visual content. In *European Conference on Information Retrieval* (pp. 314–325). Springer.

Sage, A., Agustsson, E., Timofte, R., & Van Gool, L. (2018). Logo synthesis and manipulation with clustered generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5879–5888). IEEE. https://doi.org/10.1109/cvpr.2018.00616

Schietse, J., Eakins, J., & Veltkamp, R. (2007). Practice and challenges in trademark image retrieval. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR* (pp. 518–524). ACM. https://doi.org/10.1145/1282280.1282355

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). IEEE. https://doi.org/10.1109/ICCV.2017.74

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Y. Bengio & Y. LeCun (Eds.), *3rd International Conference on Learning Representations*. ICLR.

Song, H., Kim, M., Park, D., Shin, Y., & Lee, J.-G. (2023). Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(11), 8135–8153. https://doi.org/10.1109/TNNLS.2022.3152527

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–9). IEEE. https://doi.org/10.1109/CVPR.2015.7298594

Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. (2011). Multi-label classification of music by emotion. *EURASIP Journal on Audio, Speech, and Music Processing*, *2011*(1), 1–9. https://doi.org/10.1186/1687-4722-2011-426793

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook* (pp. 667–685). Springer.

Tsoumakas, G., & Vlahavas, I. (2007). Random k-labelsets: An ensemble method for multilabel classification. In *European Conference on Machine Learning* (pp. 406–417). Springer.

Tursun, O., Aker, C., & Kalkan, S. (2017). A large-scale dataset and benchmark for similar trademark retrieval. *arXiv:1701.05766*.

Tursun, O., Denman, S., Sivapalan, S., Sridharan, S., Fookes, C., & Mau, S. (2020). Component-based attention for large-scale trademark retrieval. *IEEE Transactions on Information Forensics and Security*, *17*, 2350–2363. https://doi.org/10.1109/tifs.2019.2959921

Tursun, O., & Kalkan, S. (2015). METU dataset: A big dataset for benchmarking trademark retrieval. In *2015 14th IAPR International Conference on Machine Vision Applications (MVA)* (pp. 514–517). IEEE.

Tüzkö, A., Herrmann, C., Manger, D., & Beyerer, J. (2018). Open set logo detection and retrieval. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*. SCITEPRESS.

van der Maaten, G. (2008). Laurens; Hinton, Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, *9*(11), 2579–2605.

Wang, Y., Tao, X., Qi, X., Shen, X., & Jia, J. (2018). Image inpainting via generative multi-column convolutional neural networks. *Advances in Neural Information Processing Systems* (pp 329–338). MIT Press. https://doi.org/10.1016/j.displa.2021.102028

Wheeler, A. (2013). *Designing brand identity: An essential guide for the whole branding team*. Wiley.

World Intellectual Property Organization. (2002). *International Classification of the Figurative Elements of Marks: (Vienna Classification)*. World Intellectual Property Organization.

Xia, Z., Lin, J., & Feng, X. (2019). Trademark image retrieval via transformation-invariant deep hashing. *Journal of Visual Communication and Image Representation*, *59*, 108–116. https://doi.org/10.1016/j.jvcir.2019.01.011

Zeiler, M. D. (2012). ADADELTA: An adaptive learning rate method. *arXiv:1212.5701*.

Zhang, M.-L., & Zhou, Z.-H. (2007). Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*(7), 2038–2048.

Zhang, M.-L., & Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *26*, 1819–1837. https://doi.org/10.1109/TKDE.2013.39

Zheng, L., Zhao, Y., Wang, S., Wang, J., & Tian, Q. (2016). Good practice in CNN feature transfer abs/1604.00133. https://doi.org/10.48550/ARXIV.1604.00133
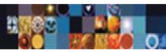
AUTHOR BIOGRAPHIES

Marisa Bernabeu received B.Sc. & M.Sc. degrees in Computer Science and a Ph.D. from the University of Alicante. She is a professor at the Escuela de Arte y Superior de Diseño de Alicante (EASDA), affiliated with the Instituto Superior de Enseñanzas Artísticas (ISEA), Comunidad Valenciana (ISEACV). She has been working for 12 years in a software development company from Valencia. She has published 6 papers, and her research interests include deep learning, computer vision, and arts.

Antonio Javier Gallego is an associate professor in the Department of Software and Computing Systems at the University of Alicante, Spain. He received B.Sc. & M.Sc. degrees in Computer Science from the University of Alicante in 2004, and a PhD in Computer Science and Artificial Intelligence from the same university in 2012. He has been principal investigator and collaborator in 18 research projects financed by the Spanish Government, the Generalitat Valenciana, and private companies. He has authored more than 80 works published in international journals, conferences and books. His research interests include Deep Learning, Pattern Recognition, Computer Vision, and Remote Sensing.

Antonio Pertusa is an associate professor at the Department of Software and Computing Systems (DLSI) at the University of Alicante and he belongs to the Pattern Recognition and Artificial Intelligence Group (PRAIG). He has authored more than 60 publications in international journals, conferences, and books. His research interests include machine learning methods applied to computer vision, music information retrieval, remote sensing, and medical knowledge extraction. He is currently the director of the University Institute of Computing Research (IUII) at the University of Alicante.

APPENDIX: VIENNA CLASSIFICATION

Next, we include the list of labels used in Vienna classification (World Intellectual Property Organization, 2002). In the scope of this work, figurative elements are those with codes from 1 to 25. Codes from 26 onwards are related to shape, text and colour, and were not used for figurative classification.

1	Celestial bodies, natural phenomena, geographical maps.
2	Human beings.
3	Animals.
4	Supernatural, fabulous, fantastic or unidentifiable beings.
5	Plants.
6	Landscapes.
7	Constructions, structures for advertisements, gates, or barriers.
8	Foodstuffs.
9	Textiles, clothing, sewing accessories, headwear, footwear.
10	Tobacco, smokers' requisites, matches, travel goods, fans, toilet articles.
11	Household utensils.
12	Furniture, sanitary installations.
13	Lighting, wireless valves, heating, cooking or refrigerating equipment, washing machines, drying equipment.
14	Ironmongery, tools, ladders.
15	Machinery, motors, engines
16	Telecommunications, sound recording or reproduction, computers, photography, cinematography, optics.
17	Horological instruments, jewellery, weights and measures.
18	Transport, equipment for animals.
19	Containers and packing, representations of miscellaneous products.
20	Writing, drawing or painting materials, office requisites, stationery, and booksellers' goods.
21	Games, toys, sporting articles, roundabouts.
22	Musical instruments and their accessories, music accessories, bells, pictures, sculptures.
23	Arms, ammunition, armour.
24	Heraldry, coins, emblems, symbols.
25	Ornamental motifs, surfaces or backgrounds with ornaments.
26	Geometrical figures and solids.
27	Forms of writing, numerals.
28	Inscriptions in various characters.
29	Colours.

Vienna codes used in this work for colours (code 29 of the main Vienna classification).

29.01.01	Red.
29.01.02	Yellow.
29.01.03	Green.
29.01.04	Blue.
29.01.05	Violet.
29.01.06	White.
29.01.07	Brown.
29.01.08	Black.
29.01.95	Silver.
29.01.96	Gray.
29.01.97	Gold.
29.01.98	Orange.
29.01.99	Pink.

Vienna codes used in this work for shapes (code 26 of the main Vienna classification).

26.1	Circles, ellipses.
26.2	Segments or sectors of circles or ellipses.
26.3	Triangles, lines forming an angle.
26.4	Quadrilaterals.
26.5	Other polygons. This category also groups 26.13 (Other geometrical figures, indefinable designs) and 26.7 (Different geometrical figures, juxtaposed, joined, or intersecting).
26.11	Lines, bands.
26.15	Geometrical solids.