Hierarchy-of-Visual-Words: a Learning-based Approach for Trademark Image Retrieval

Vítor N. Lourenço, Gabriela G. Silva, Leandro A. F. Fernandes
Instituto de Computação, Universidade Federal Fluminense
Niterói, Rio de Janeiro, Brazil
{vitorlourenco, gabrielagomessilva}@id.uff.br, laffernandes@ic.uff.br

***Abstract*—In this paper, we present the Hierarchy-of-Visual-Words (HoVW), a novel trademark image retrieval (TIR) method that decomposes images into simpler geometric shapes and defines a descriptor for binary trademark image representation by encoding the hierarchical arrangement of component shapes. The proposed hierarchical organization of visual data stores each component shape as a visual word. It is capable of representing the geometry of individual elements and the topology of the trademark image, making the descriptor robust against linear as well as to some level of nonlinear transformation. Experiments show that HoVW outperforms previous TIR methods on the MPEG-7 CE-1 and MPEG-7 CE-2 image databases.**

I. Introduction

Trademark images are complex patterns consisting of graphical or figurative shape patterns (device-mark), text words or phrases (word-in-mark), or both. Trademark images carry not only the identification meaning but, also, the reputation and the quality meanings of the associated product or service. Thus, it is of the intrinsic interest of companies to ensure the ownership and exclusive use of their trademark images. The design of automatic trademark image retrieval (TIR) systems has been an active research topic [1]–[11] due to the complexity of manual trademark image matching analysis.

All these artificially-produced images are designed to have a visual impact and consisting of multiple elements, which may be closed regions, lines, or areas of texture. Existing TIR systems, however, typically treat trademark images as indivisible structures by computing descriptors integrating global and local image features [1]–[6] or by partitioning the image [7]–[9] without considering the distribution of their component shapes. Such a practice has been successful in retrieving near-duplicated images but may fail in detecting similar instances that preserve the topology of their components without conserving the relative location of their elements.

This paper proposes a novel TIR method called Hierarchy-of-Visual-Words (HoVW). The key insights of our solution is that shape is probably the single most important feature used by human observers to characterize an image [12]. Also, image structure and the layout of individual image elements are essential when judging similarity [12]. Therefore, we have designed the HoVW approach as a method that takes component shapes, image structure, and layout of individual image elements into account while computing descriptors of trademark images. Fig. 1 shows the main steps of the HoVW. In the training stage, our approach decomposes the set of training trademark images (a) into simple component shapes (b-c) and learns a codebook of visual words for those shapes (d-e). The hierarchical arrangement of components within each image leads to the representation of trademarks as trees of visual words (f). Then, our approach learns a codebook of visual hierarchies (g), which defines a labeling system for trademark image representation. In the evaluation stage, the HoVW uses the visual words codebook to encode the simple shapes extracted from the query image (h-k). Next, the hierarchical relationship of its components is enconded (l), and the visual hierarchies codebook is used to accelerates the retrieval of similar images from the database (m). The feature vectors in (d) and (k) are comprised of Zernike moments (ZM), circularity, average bending energy, eccentricity, and convexity. The dissimilarity between two hierarchies produced in (f) and (l) is computed using an efficient tree edit distance algorithm [13]. Clustering in (e) and (g) are performed using, respectively, point-based k-means [14] and mean shift for distance matrices [15].

The main contributions of this paper include:

i) A new learning-based framework for the hierarchical representation of elements in binary images; and
ii) Its application on trademark image description and retrieval from image databases.

We have performed experiments using the popular MPEG-7 CE-1 and MPEG-7 CE-2 image databases to compare the efficiency of our method against state-of-the-art solutions in TIR tasks. Precision-recall curves show that the proposed hierarchical decomposition leads to smaller dissimilarity measures between trademarks of the same class than other solutions. In other words, the query image is expected to be closer to similar images in the feature space while performing TIR for judging trademark infringement.

II. Related Work

The first systems to implement content-based retrieval for trademark images were developed by Kato *et al.* [1], Wu *et al.* [2] and Eakins *et al.* [3]. The TRADEMARK system [1] uses the spatial distribution, spatial frequency, local correlation, and local contrast of pixel blocks as visual features. The STAR system [2] represents trademarks as structural patterns consisting of word-in-mark and device-mark information. The former is manually informed by the user. The later includes Fourier descriptor of the shape, moment invariants, and grey

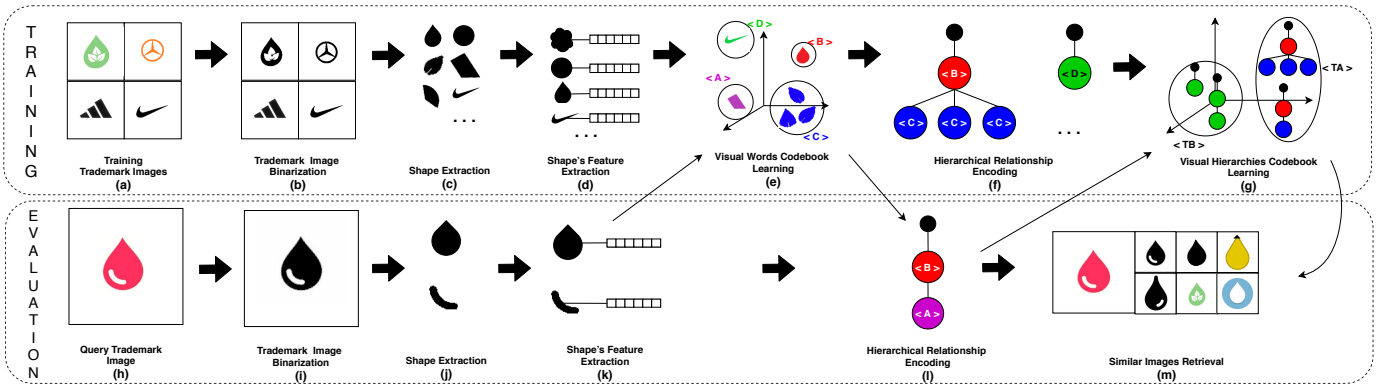


Fig. 1: Overview of the training and evaluation stages of the Hierarchy-of-Visual-Words (HoVW) approach.

level projections. ARTISAN [3] applies a structural pattern similar to STAR's and incorporates principles derived from Gestalt psychology to cope with device-only marks.

Some solutions integrate global features to capture the gross essence of the shapes and local features to describe the interior details of binary trademark images. For instance, Wei *et al.* [4] employ ZM to extract global features. Curvature and distance to the centroid are used as local features. Anuar *et al.* [5] also use ZM, but they employ the edge-gradient co-occurrence matrix derived from the contour information as the local descriptor. Qi *et al.* [6] use the histogram of centroid distances and a region descriptor based on improved feature points matching and the spatial distribution of feature points to avoid the calculation of ZM. It is important to emphasize that, in contrast to those approaches, the final descriptor produced by HoVW does not encode local and global information separately. The multilevel organization of the data gives a global representation at some level of detail, while each node of the structure represents local information.

In recent work, Liu *et al.* [7] proposed a hierarchical region feature descriptor where the binary trademark image is iteratively partitioned into progressively smaller ones along with various directions. Density, compactness, rectangularity, and eccentricity are computed for each partition. The binary image retrieval methods developed by Sidiropoulos *et al.* [8] and Yang *et al.* [9] differ from Liu's *et al.* work because they consider a single direction (the most descriptive one). Also, they split the image region recursively regarding the four sub-regions for the next partition in each iteration instead of only one. HoVW does not rely on image partitions. Our approach decomposes binary trademark images into sets of simpler component shapes and builds the hierarchical arrangement of those components. As a result, our descriptor encodes the topology of the basic image components using the hierarchy and their geometry using feature values invariant to rotation, translation, and scale.

The image retrieval system proposed by Alajlan *et al.* [10], [11] uses a structured representation called Curvature Tree to encode both shape and topology of objects and holes comprising a binary image. Our approach uses a different tree structure to represent topology. Also, our tree dissimilarity measure is based on tree edit distance [13] instead of maximum similarity subtree isomorphism [16].

The use of codebooks and structured representation of images was also considered by Silva *et al.* [17], [18]. However, in this case, local information of grayscale images is extracted by Hessian Affine [19] and Scale Invariant Feature Transform (SIFT) [20] detectors. As a result, Silva's *et al.* approach is not suitable for low-textured trademarks. Also, the visual-word arrangement is defined by a planar graph derived from the Delaunay triangulation of feature points instead of the tree structure applied by our approach to organizing the component shapes of the trademark image.

Recent approaches based on neural network advances are presented by Perez *et al.* [21] and Lan *et al.* [22]. Lan's *et al.* introduces the use of deep Convolutional Neural Networks (CNN) features aided by constraint theory. Perez's *et al.* approach resides on the use of pretrained VGG19 models for learning visual and conceptual similarities. Both approaches require a set of classes and the original trademark images with modifications to train the CNN models and learn concepts of similarity. Perez *et al.*, for instance, was aided by human experts to define the set of classes on each trained model, which could introduce human bias and possibly limit the model capability to learn unknown concepts. Our approach learns synthetic classes, *i.e.*, a codebook, without human interference, through the similarity between the tree-structured shapes.

III. Hierarchy-of-Visual-Words (HoVW)

As a learning-based approach, the HoVW framework is comprised of training and evaluation stages. Fig. 1 illustrates the main steps of each one of them.

Trademark image binarization is performed at both stages of HoVW (Figs. 1 (b) and (i)). It consists of converting digital trademark images into binary images from which component shapes will be extracted. In this step, we first convert a given color image to grayscales. Then, we apply a median filter [23] to reduce impulsive noise and a bilateral filter [24] to remove texture without losing overall shapes since sharp edges are

preserved. The final binary image is obtained by applying Otsu's method [25] on the textureless grayscale image.

Shape extraction aims to split binary images into objects and holes, *i.e.,* their component shapes (Figs. 1 (c) and (j)).

Definition 1 *(Object and hole): Objects and holes are connected sets of, respectively, foreground and background pixels.*

For instance, in Fig. 2 (top), all the pixels that are set to white are foreground pixels, while the background pixels are set to black. Thus, this image includes two objects (the circle and the cloud) and two holes (the outer rectangle and the square inside the circle).

In this work, we extract shapes by using the border-following method proposed by Suzuki and Abe [26]. Their approach applies a border labeling mechanism capable of describing the relationship among the outer borders and the hole borders, capturing the topological structure of a given binary image.

Fig. 3 presents the algorithm used to extract the component shapes of a given trademark image. The algorithm, first, associates the input binary image $\mathcal{I}$ with a decomposing list M (line 2). After, the main iterative process (lines 3 to 10) segments each element m_i on the decomposing list M concerning their foreground and background pixels (lines 5 and 6). From the background, the procedure extracts the holes' shapes and associate them with the component shapes' list (lines 7 and 8). The foreground shapes are used in the next iterations (lines 5 and 9) until all component shapes have been extracted from the initial binary trademark image.

Shape's feature extraction consists of building a feature vector for each component shape of a given trademark image (Figs. 1 (d) and (k)). These 29-dimension feature vectors combine region-based and contour-based descriptors.

Shape's region is described by the 25 moments of the Zernike polynomials (ZM) of order p from 0 to 8:

$$Z_{p,q} = \frac{p+1}{\pi} \sum_{\rho} \sum_{\theta} V_{p,q}\left(\rho,\theta\right)^{*} \mathcal{I}\left(\rho,\theta\right), \tag{1}$$

where $\rho = \sqrt{x^2+y^2}$ is the length of vector from origin to pixel (x, y), θ is the angle between the vector defining ρ and the x-axis in the counter clockwise direction and $V_{p,q}\left(\rho,\theta\right)$ is a Zernike polynomial of order p with repetition q that forms a complete set over the interior of the unit disk inscribing the component shape:

$$V_{p,q}\left(\rho,\theta\right) = R_{p,q}\left(\rho\right) \exp\left(-\mathrm{i}\, q\, \theta\right). \tag{2}$$

In Equation (2), $\mathrm{i} = \sqrt{-1}$, p is a non-negative integer value and q is a positive integer subject to $p - |q| =$ even and $q \leq p$. The radial polynomial $R_{p,q}\left(\rho\right)$ is defined as:

$$R_{p,q}\left(\rho\right) = \sum_{s=0}^{(p-|q|)/2} \frac{\left(-1\right)\left(p-s\right)!}{s!\left(p+|q|-s\right)!\left(p-|q|-s\right)} \rho^{p-2s}. \tag{3}$$

ZM are rotation invariant by construction. We use Khotanzad and Hong's approach [27] to obtain invariance to translation and scale, too.

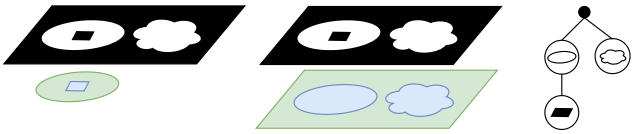


Fig. 2: Examples of inclusion and exclusion relationships: (top, left and center) parallel orthographic projection of a binary image; (bottom, left) the small square hole is in included in the circular object; (bottom, center) the circle and the cloud are included in the rectangle and exclude each other; and (right) the hierarchical relationship of the component shapes.

We have used four measures to describe the shape's contour [28]:

- *Circularity* defines the relation between the perimeter of the shape $\mathcal{S}$ and its area:

$$d_c = \frac{\mathrm{perimeter}\left(\mathcal{S}\right)^2}{4\pi\ \mathrm{area}\left(\mathcal{S}\right)}; \tag{4}$$

- *Average bending energy* defines the mean sum of the shape's curvature:

$$d_b = \frac{1}{N} \sum_{t=0}^{N-1} K(t)^2, \tag{5}$$

  where $K(t)$ is the curvature function, t is the arc length parameter and N is the number of points on the contour [29];
- *Eccentricity* characterizes the statistical distribution of contour points around the principal axes of shape's contour:

$$d_e = \frac{\lambda_2}{\lambda_1}, \tag{6}$$

  where λ_1 and λ_2 are the eigenvalues of the covariance matrix of the set of contour points of $\mathcal{S}$, for $\lambda_1 > \lambda_2$;
- *Convexity* defines the relation between the perimeter of the convex hull and the perimeter of the shape:

$$d_x = \frac{\mathrm{hull}\left(\mathcal{S}\right)}{\mathrm{perimeter}\left(\mathcal{S}\right)}. \tag{7}$$

We have chosen the descriptors mentioned above because they showed to be robust and have low computational cost. Also, they are invariant to rotation, translation, and scale.

Visual words codebook learning is performed only during the training stage of the HoVW approach (Fig. 1 (e)). In this step, we apply k-means clustering [14] on the feature vectors computed for the component shapes of the training images. The resulting clusters are a general representation of the shapes, in which each cluster acts as a word in the codebook Λ that assigns a given shape to a learned visual word by using the Euclidean distance between the shape's feature vector and the cluster's centroid.

The use of k-means or similar clustering techniques for defining codebooks of visual words is not new. The original contribution of our approach is the way the simple component shapes of the binary image is arranged hierarchically, combining topological and invariant geometrical information in the same representation.

Hierarchical relationship encoding is a key step of the HoVW framework (Fig. 1 (f) and (l)). For each binary image, this step produces a tree structure induced by:

Definition 2 *(Shape inclusion): A shape $\mathcal{A}$ is said to be included in a shape $\mathcal{B}$ if and only if $\mathcal{A}$ is a hole surrounded by object $\mathcal{B}$ or $\mathcal{A}$ is an object surrounded by hole $\mathcal{B}$.*

Definition 3 *(Shape exclusion): Shapes $\mathcal{A}$ and $\mathcal{B}$ exclude each other if and only if they are included in shape $\mathcal{C}$.*

Corollary 1 *(Visual hierarchy): The recursive inclusion and exclusion relationship of objects and holes yields the hierarchical organization of visual data into a tree structure where each node corresponds to one shape, and it is related to its ancestor node by inclusion and to its siblings by exclusion.*

For instance, in Fig. 2 (left), the small square hole is included in the circular object, while in Fig. 2 (center) the circle and cloud exclude each other but are included in the black rectangle. According to Corollary 1, the hierarchical relationship of those shapes leads to the tree in Fig. 2 (right).

In the HoVW framework, each node of the visual hierarchy stores a reference to the word that better describes the respective node's shape in the codebook of visual words. This way, the hierarchy encodes topology while nodes encode the geometry of component shapes.

Visual hierarchies codebook learning is the last step of the training stage (Fig. 1 (g)). It aims to discover the most suitable set of labels to represent similar visual hierarchies within a database. Those labels are used to accelerate TIR queries during the evaluation stage of the HoVW framework (Fig. 1 (m)).

In this step, we compute the dissimilarity matrix of visual hierarchies representing the training images and use this matrix as the input of the mean shift clustering procedure [15] with RBF kernel. Ideally, the dissimilarity measure between two visual hierarchies has to be robust against changes on the compared trademark images. Those changes include linear and non-linear transformations and the addition and removal

Input: Binary trademark image $\mathcal{I}$, where the background pixels are black, *i.e.*, 0 bit, and the foreground pixels are white, *i.e.* 1 bit.
Output: Component shapes of the trademark image.
1: $S \leftarrow \{\}$
2: $M \leftarrow \{\mathcal{I}\}$
3: **while** $M \neq \varnothing$ **do**
4: $m_i \leftarrow$ The first element in M
5: $F \leftarrow$ Foreground pixels in m_i
6: $B \leftarrow$ Background pixels in m_i
7: $H \leftarrow$ Hole shapes extracted from B
8: $S \leftarrow S \cup H$
9: $M \leftarrow (M \setminus \{m_i\}) \cup F$
10: **end while**
11: **return** S

Fig. 3: Algorithm for extracting the components shapes of a given trademark image.

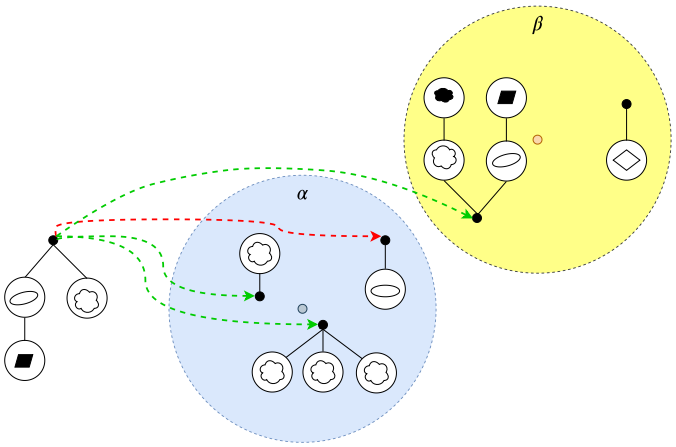


Fig. 4: Example of our similar image search strategy. The query hierarchy one the left was related to hierarchies included in clusters α and β.

of elements. Also, it must be computationally and memory-efficient. We satisfy those requirements by using the All Path Tree Edit Distance (AP-TED) algorithm developed by Pawlik and Augsten [13], an approach that counts the minimal-cost sequence of editing operations needed to transform a tree into another while keeping low computational cost and memory footprint.

We have modeled the costs of the *rename*, *insert* and *remove* editing operations performed by AP-TED as follows:

Rename a node is similar to change the visual word stored by it. Thus, the cost of this operation is the Euclidean distance between the centroids of the clusters corresponding to the actual and the desired words in the codebook Λ:

$$\delta_r(n_a, n_b) = \mathrm{dist}_E(\lambda_a, \lambda_b), \tag{8}$$

where dist_E denotes the Euclidean distance, n_a and n_b are tree nodes, and λ_a and λ_b are their respective visual words.

Insert and *remove* costs are calculated as functions of the mean distance between all pairs of visual words in Λ, modulated by a factor α. This factor is proportional to the most influential value between the deep $\mathcal{D}$ of the node n in the tree and the number of siblings of n at same level $\mathcal{L}$ of the tree:

$$\delta_x(n) = \alpha \frac{2}{m\,(m-1)} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \mathrm{dist}_E(\lambda_i, \lambda_j), \tag{9}$$

where $\alpha = \min\{\log_2^{-1} \mathcal{L}, \log_2^{-1} \mathcal{D}\}$. In Equation (9), m is the number of visual words in Λ and λ_i and λ_j are visual words in Λ.

Both cost heuristics were chosen given the spatial characteristics of the shapes' representation through visual words and the hierarchical semantics encoded in the hierarchies.

The cost of rename a node n_a to a node n_b is straightforward from Equation (8) as the Euclidean distances between their visual words. The heuristic of the cost of a node n insertion or removal, *i.e.*, $\delta_x(n)$, was defined based on the characteristics of the tree-structured trademark images' shape.

These characteristics ensure that the deeper a shape on the hierarchy, the smaller its size compared to the whole trademark image. Also, it assures the more shapes on a tree level, the greater the amount of information on it. Then, in both cases, the relevance of a single shape is ensured to be lower. The semantics of the presented characteristics are embedded on factor α, while the other part from Equation (9) performs the average distance between all visual words.

Similar images retrieval step receives the hierarchical representation of the query image as input (Fig. 1 (m)) and returns a set of related images. Recall that the proposed representation is a tree structure. As a result, conventional point-based searching strategy for retrieving the k-nearest neighbors in feature space cannot be used with our approach. We overcome this issue by using the codebook Θ of visual hierarchies and a k-d tree to speed up database search. Our searching strategy first compares the query image representation with the set of database entries having the same label in Θ. It only looks within other sets of entries having close labels if the user wants to retrieve more images. Retrieved images are naturally presented in ascending order of dissimilarity.

An example of the used similar image search strategy is illustrated in Fig. 4. In this example, the hierarchical representation of the query image (right) is associated with the cluster having label α. To retrieve all similar images of the query image, first, all images within the same α cluster are retrieved according to the measured dissimilarity. Since not all images related to the query image are within label α, the search strategy looks for other similar images in the cluster closest α, *i.e.,* cluster β. On β cluster, the similar missing image is found. Then, it joins the retrieved set, again, taking into account the dissimilarity between the query image and the image found.

Fig. 5 presents two examples of TIR performed by HoVW in the MPEG-7 CE-1 database. In the first example, the first 20 images retrieved are of the same class as the query image. In the second examples, the 14 first images retrieved are butterfly images. This database includes 20 images of each class.

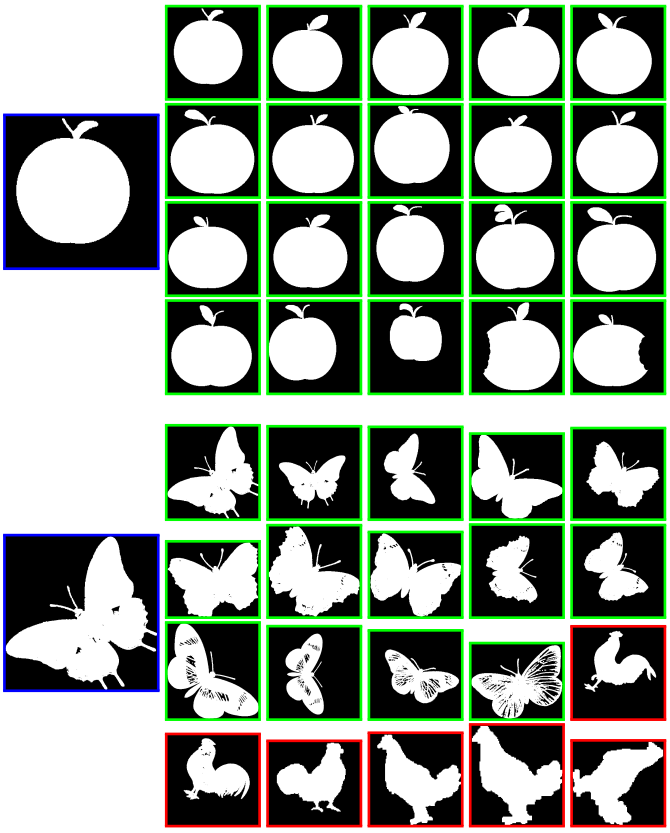


Fig. 5: The top-20 retrieval set of the HoVW for `apple-1.png` and `butterfly-1.png` images of the MPEG-7 CE-1 database. Green denotes images of the same class as the query entry (blue). Red denotes different classes.

IV. Experiments and Results

Sections IV-A and IV-B present the materials and methods of our experiments. In Section IV-C, we compare the performance of our approach to the traditional use of ZM [27] and to the state-of-the-art TIR techniques proposed by Liu *et al.* [7] and Anuar *et al.* [5]. In this analysis, we use precision-recall curves [30] to observe the quality of retrieval simulating the gradual detection of related samples in the database. In Section IV-D, we analyze the harmonic mean between precision and recall, *i.e.,* the F_1 score, of the aforementioned techniques and a CNN-based approach inspired by Perez's *et al.* [21] work and implemented by us. It is important to comment that, to the best of our knowledge, there is no available implementation of Perez's *et al.* technique. Also, the CNN-based approach was not included in the analysis of precision-recall curves because its outcome does not have the same semantic distance than the distance functions used by other approaches, which allows the simulation of gradual detection of related samples. Section IV-E discusses the limitations of our approach.

A. Databases

We have used two image databases in our experiments: *MPEG-7 Core Experiment CE-Shape-1* and *MPEG-7 Region Shape Dataset CE-2* [31]. The MPEG-7 CE-1 database is comprised of 1,400 binary images organized into 70 classes having 20 similar images each. The MPEG-7 CE-2 database includes 871 binary images organized into 51 classes having from 11 to 21 images each and 2,750 images that do not belong to any category. Following Anuar *et al.* [5] and Liu *et al.* [7], we have used only the categorized images in our experiments because uncategorized images do not follow an overall shape pattern. Fig. 6 shows some samples from both databases.

B. Implementation, Training and Evaluation

Liu *et al.*, Anuar *et al.*, and ZM Data Acquisition. We have used the results reported by the authors in their original papers [4], [5], [7] since those approaches do not require a training phase.

HoVW Implementation and Parameterization. We have implemented our algorithms using Python. Our system[1] performs codebook learning and efficient closest visual word searching using the implementations of k-means, mean shift and k-d tree provided by the `scikit-learn`[2]. Image filtering and shape extraction are performed using `OpenCV`[3]. The dissimilarity of hierarchies is computed using the AP-TED's implementation provided by the authors[4].

We have performed median filtering using a 5×5 window during the first step of the HoVW procedure. We skipped the application of bilateral filtering since the databases have only binary images.

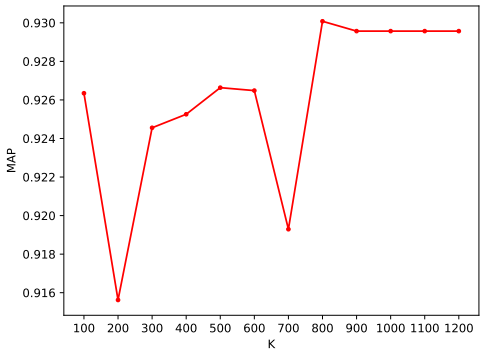
The size of the visual words codebook was set to $k = 800$ for MPEG-7 CE-1 and to $k = 600$ for MPEG-7 CE-2 after looking for the maximum among the mean average precision (MAP) metric values [30] computed as function of the number of clusters in k-means, for $k \in \{100, 200, \cdots, 1200\}$. Results on MPEG-7 CE-1 are presented in Fig. 7 (a). The k vs. MAP chart for MPEG-7 CE-2 is equivalent.

The bandwidth h of the mean shift clustering procedure was chosen after assuming $h \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ for MPEG-7 CE-1 and $h \in \{1.1, 1.3, 1.5, 1.7, 1.9\}$ for MPEG-7 CE-2 and then analyzing the respective MAP values. A small bandwidth value produces a codebook with an increased number of labels, which makes it more detailed in terms of nuances between visual hierarchies. On the other hand, large h values produce codebooks with fewer labels, which makes them more resilient to small changes. The parameter h was set to 0.7 and 1.7 for, respectively, MPEG-7 CE-1 and MPEG-7 CE-2 databases. Both values maximized the MAP value (see Fig. 7 (b) for the former) and produced codebooks having 25 and 21 labels, respectively. According to our experience, the advantage of using mean shift with the RBF kernel instead of k-means for visual hierarchies codebook learning is that it is quite more difficult to set the expected number of clusters (k) for visual words representing trademark images than for simpler shapes.

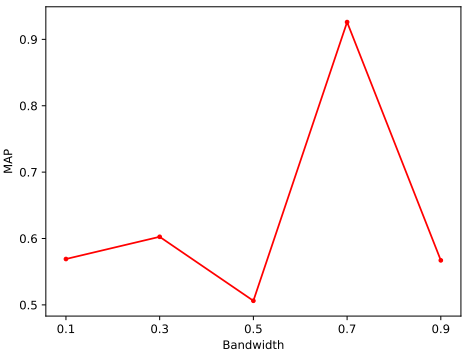
CNN-based Implementation and Training. The CNN-based approach inspired by Perez's *et al.* [21] work consists of a VGG16 network [32] in which the last fully connected layer was replaced by a 70 neurons layer and a 51 neurons layer, each corresponding to the number of classes included in, respectively, the MPEG-7 CE-1 and MPEG-7 CE-2 databases. The networks were implemented using `TensorFlow`[5] and `Keras`[6] frameworks.

We have used the same optimization hyperparameters for both networks. They were trained with Stochastic Gradient Descent with 0.9 of momentum factor, batch size of 32, and fixed learning rate of 0.0001. They were initialized with weights trained on the *ImageNet Dataset* [33] and fine-tuned with the train data of MPEG-7 CE-1 and MPEG-7 CE-2 databases, respectively. We performed data augmentation by randomly rotating, zooming, shifting, and flipping the testing database entries. The networks were trained in two phases. In the first phase, only the replaced layer of the networks was fine-tuned over 50 epochs. In the second phase, we performed fine-tuning of the whole networks over 20 epochs.

Evaluation. We used 10-fold cross-validation to assess the performance of the compared solutions on TIR tasks. We



(a) MPEG-7 CE-1, k vs. MAP



(b) MPEG-7 CE-1, h vs. MAP

Fig. 7: MAP values on the MPEG-7 CE-1 function of different sizes k of visual words codebook (a) and different bandwidth values h for visual hierarchies codebook learning (b).

[1]HoVW: https://github.com/Prograf-UFF/HoVW
[2]scikit-learn: https://scikit-learn.org library
[3]OpenCV: https://opencv.org
[4]AP-TED: http://tree-edit-distance.dbresearch.uni-salzburg.at

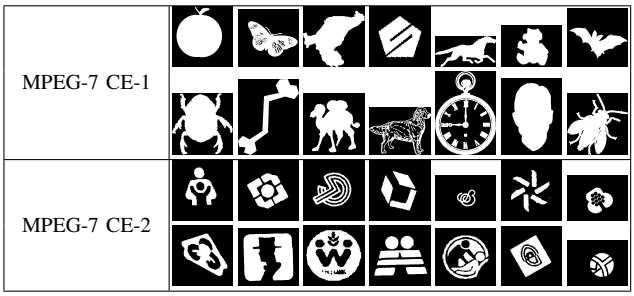


Fig. 6: Samples from the databases used in our experiments.

[5]TensorFlow: https://www.tensorflow.org/
[6]Keras: https://keras.io/

distributed the images randomly in the folds while keeping uniform distribution per class in each of them.

Precision values were computed as function of recall for the techniques proposed by ZM [27], Anuar *et al.* [5], Liu *et al.* [7], and our approach. For those techniques, we computed the F_1 score by taking precision at recall of 100%. The computation of F_1 score of the CNN-based approach is straightforward.

C. Precision-Recall Analysis

In experiments on MPEG-7 CE-1 (Fig. 8, top), HoVW yields a near-perfect result for the first 11 images retrieved in all categories, obtaining precision of 99.79% up to recall of 55%. For the last correlated image retrieved, HoVW's precision was 72%. The decay of the curve is easily explained by the way database entries are retrieved by HoVW and how we compute precision. As described in Section III, HoVW retrieves from the database all images having the same label as the query hierarchy and, gradually, chunks of entries with close labels. Conservatively, we include in the calculation of precision and recall all entries associated with the secondary sets of retrieved images. Therefore, it is expected that precision will decline with the increase of recall as the labels containing the remaining similar items become further from the original label. For instance, consider the example on Fig. 4 and a total of three similar images to retrieve, in which the green arrows point to correctly retrieved images, while the red arrow indicates the erroneous retrieved images. For the α cluster, two of the total similar images and one non-similar image were retrieved, accounting a mean precision and recall of 66%. When cluster β is reached, the precision steadies on 66%, while the recall reaches 100%. We believe that the adoption of specialized data structures to manage intra-label relationships would help mitigate this issue.

When compared to Anuar *et al.* and ZM, it can be seen in Fig. 8 (top) that the proposed approach has the best precision-recall curve in all ranks. HoVW outperforms Liu *et al.*'s approach up to recall of 80% and their performance are comparable from a recall of 85% to 100%. This result suggests that in practical applications of TIR specialists would have access to the expected top-ranked similar images while judging trademark infringement.

Fig. 8 (bottom) shows comparable results on the MPEG-7 CE-2 database for the competing techniques. In this database, HoVW presents slightly better performance than Liu *et al.*'s approach on 11 out 20 cases of the precision-recall curve and outperforms both Anuar *et al.* and ZM techniques.

D. F_1 Score Analysis

In this analysis, we used a state-of-the-art CNN-based approach as a baseline for comparison. Table I shows the F_1 score obtained by each compared technique on each database. On the MEPG-7 CE-1, the CNN-based approach accomplishes better F_1 score, reaching 88%, while HoVW reaches up to 85%. However, on MEPG-7 CE-2, HoVW outperforms the CNN-based approach by achieving a F_1 score of 89%.

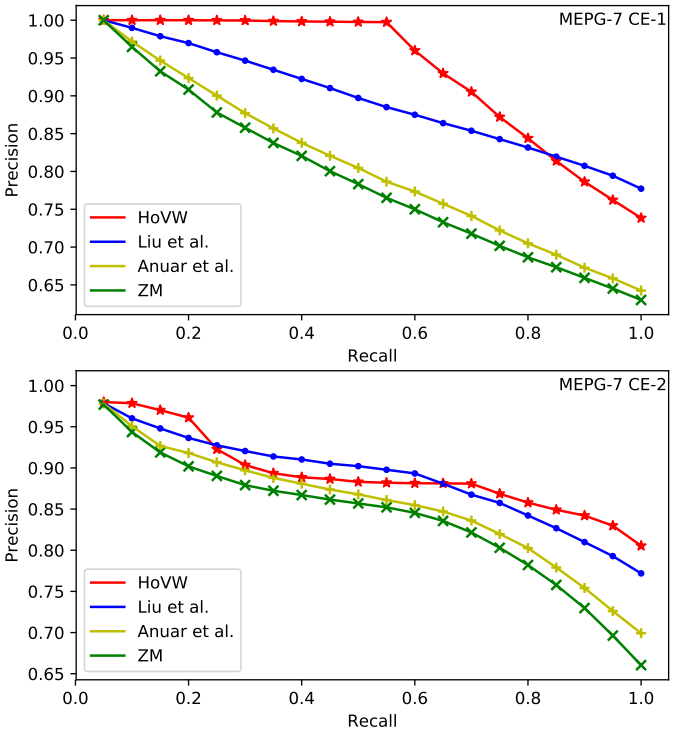


Fig. 8: The precision-recall curves of compared approaches.

TABLE I: The F_1 score obtained by each approach on the MPEG-7 CE-1 and MPEG-7 CE-2 databases.

Approach	MPEG-7 CE-1	MPEG-7 CE-2
HoVW	0.85	**0.89**
Liu *et al.* [7]	0.87	0.87
Anuar *et al.* [5]	0.78	0.82
ZM	0.77	0.79
CNN-based	**0.88**	0.81

E. Limitations

Recall that trademark images may include graphical or figurative shape patterns, text words, or both. The HoVW does not include the analysis of text in its conception. Also, the image decomposition does not consider Gestalt. As a result, some implicit organization of shape components may be disregarded while building the hierarchy of component shapes.

V. Conclusion and Future Works

We proposed a learning-based approach for TIR that uses two codebooks. The first codebook encodes basic shapes expected in the images using 29-dimension feature vectors combining region-based and contour-based descriptors. The second codebook encodes both local and global information of trademark images through hierarchical arrangements of their component shapes. The hierarchy is defined as a tree where

each node is related to a component shape while tree levels describe the topological relationship of the components. Tree dissimilarity is computed using an efficient tree edit distance algorithm proposed by Pawlik and Augsten [13]. The main contributions of our work are a new learning-based framework for the hierarchical representation of elements in binary images, and its application on trademark image description and retrieval from image databases. Experimental results on well-known image databases show that our approach outperforms state-of-the-art techniques.

As future work, we are exploring ways to incorporate optical character recognition and principles from Gestalt psychology while decomposing trademark images into basic shapes. The Gestalt properties could be useful to enhance the results obtained from the decomposition, providing normalized shapes grounded to a theory. Also, we are investigating whether state-of-the-art hierarchical image segmentation is suitable for the proposed technique.

ACKNOWLEDGMENTS

This work was partially supported by the Brazilian Council for Scientific and Technological Development (CNPq – Grant 311.037/2017-8) and the Rio de Janeiro State Foundation to Support Research (FAPERJ – Grant E-26/202.718/2018) agencies. Vítor N. Lourenço was sponsored by a CNPq fellowship.

REFERENCES

[1] T. Kato, K. Fujimura, and H. S. Nonmember, "TRADEMARK: multimedia image database system with intelligent human interface," *Systems and Comput. Japan*, vol. 21, pp. 33–46, 1990.

[2] J. K. Wu, C. P. Lam, B. M. Mehtre, Y. J. Gao, and A. D. Narasimhalu, "Content-based retrieval for trademark registration," *Multimed. Tools Appl.*, vol. 3, pp. 245–267, 1996.

[3] J. P. Eakins, M. E. Graham, and J. M. Boardman, "Evaluation of a trademark image retrieval system," in *Proc. Annual BCS-IRSG Conf. Inf. Retr. Research*, 1997.

[4] C. Hung Wei, Y. Li, W. Y. Chau, and C. T. Li, "Trademark image retrieval using synthetic features for describing global shape and interior structure," *Pattern Recognit.*, vol. 42, pp. 386–394, 2009.

[5] F. M. Anuar, R. Setchi, and Y. Lai, "Trademark image retrieval using an integrated shape descriptor," *Expert Syst. Appl.*, vol. 40, pp. 105–121, 2013.

[6] H. Qi, K. Li, Y. Shen, and W. Qu, "An effective solution for trademark image retrieval by combining shape description and feature matching," *Pattern Recognit.*, vol. 43, pp. 2017–2027, 2010.

[7] F. Liu, B. Wang, and F. Zeng, "Trademark image retrieval using hierarchical region feature description," in *Proc. IEEE Intl. Conf. Image Process.*, 2017, pp. 3620–3624.

[8] P. Sidiropoulos, S. Vrochidis, and I. Kompatsiaris, "Content-based binary image retrieval using the adaptive hierarchical density histogram," *Pattern Recognit.*, vol. 44, pp. 739–750, 2011.

[9] M. Yang, G. Qiu, J. Huang, and D. Elliman, "Near-duplicate image recognition and content-based image retrieval using adaptive hierarchical geometric centroids," in *Proc. Intl. Conf. Pattern Recognit.*, 2006, pp. 958–961.

[10] N. Alajlan, M. S. Kamel, and G. Freeman, "Multi-object image retrieval based on shape and topology," *Signal Process. Image Commun.*, vol. 21, pp. 904–918, 2006.

[11] N. Alajlan, M. S. Kamel, and G. H. Freeman, "Geometry-based image retrieval in binary image databases," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, pp. 1003–1013, 2008.

[12] I. Biederman, "Recognition-by-components: a theory of human image understanding," *Psychol. Rev.*, vol. 94, pp. 115–147, 1987.

[13] M. Pawlik and N. Augsten, "Tree edit distance: robust and memory-efficient," *Inf. Syst.*, vol. 56, pp. 157–173, 2016.

[14] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theor.*, vol. 28, pp. 129–137, 1982.

[15] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 603–619, 2002.

[16] M. Pelillo, K. Siddiqi, and S. W. Zucker, "Matching hierarchical structures using association graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, pp. 1105–1120, 1999.

[17] F. B. Silva, S. Goldenstein, S. Tabbone, and R. S. Torres, "Image classification based on bag of visual graphs," in *Proc. IEEE Intl. Conf. Image Process.*, 2013, pp. 4312–4316.

[18] F. B. Silva, R. O. Werneck, S. Goldenstein, S. Tabbone, and R. S. Torres, "Graph-based bag-of-words for classification," *Pattern Recognit.*, vol. 74, pp. 266–285, 2017.

[19] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Proc. European Conf. Comput. Vis. – Part I*, 2002, pp. 128–142.

[20] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. Intl. Conf. Comput. Vis.*, 1999, vol. 2, pp. 1150–1157.

[21] C. A. Perez, P. A. Estévez, F. J. Galdames, D. A. Schulz, J. P. Perez, D. Bastías, and D. R. Vilar, "Trademark image retrieval using a combination of deep convolutional neural networks," in *Proc. Intl. Joint Conf. on Neural Networks*, 2018, pp. 1–7.

[22] T. Lan, X. Feng, L. Li, and Z. Xia, "Similar trademark image retrieval based on convolutional neural network and constraint theory," in *Proc. Intl. Conf. on Image Process. Theor., Tools and Appl.*, 2018, pp. 1–6.

[23] T. Huang, G. Yang, and G. Tang, "A fast two-dimensional median filtering algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 13–18, 1979.

[24] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. Intl. Conf. Comput. Vis.*, 1998, pp. 839–846.

[25] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.*, vol. 9, pp. 62–66, 1979.

[26] S. Suzuki and K. Abe, "Topological structural analysis of digitized binary images by border following," *Comput. Vis. Graph. Image Process.*, vol. 30, pp. 32–46, 1985.

[27] A. Khotanzad and Y.H. Hong, "Invariant image recognition by Zernike moments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, pp. 489–497, 1990.

[28] Y. Mingqiang, K. Kidiyo, and R. Joseph, "A survey of shape feature extraction techniques," in *Pattern Recognition*, P. Yin, Ed., chapter 3. IntechOpen, 2008.

[29] I. T. Young, J. E. Walker, and J. E. Bowie, "An analysis technique for biological shape. i*," *Inf. Control*, vol. 25, pp. 357–370, 1974.

[30] L. Liu and M. T. Özsu, *Encyclopedia of Database Systems*, Springer US, 2009.

[31] W. Kim and Y Kim, "A region-based shape descriptor using Zernike moments," *Sig. Proc.: Image Comm.*, vol. 16, pp. 95–102, 2000.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.