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*—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]–[8] 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]–[3] or by partitioning the image [4]–[6] 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 presents 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. Also, image structure and the layout of individual image elements are essential when judging similarity [9]. 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 [10]. Clustering in (e) and (g) are performed using, respectively, point-based k-means [11] and mean shift for distance matrices [12].

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.

II. Related Work

Some recent solutions on TIR systems follow the strategy of integrating global features to capture the gross essence of the patterns and local features to describe the interior details of trademark images. For instance, Wei *et al.* [1], Anuar *et al.* [2], and Qi *et al.* [3] build feature vectors by concatenating global descriptors like ZM with local measures like curvature and distance to centroid, among others. In contrast to those approaches, the descriptors produced by HoVW do not encode global and local information separately.

Liu *et al.* [4], Sidiropoulos *et al.* [5], and Yang *et al.* [6] propose hierarchical region feature descriptors where the binary trademark image is iteratively partitioned into progressively smaller ones along various directions. 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.

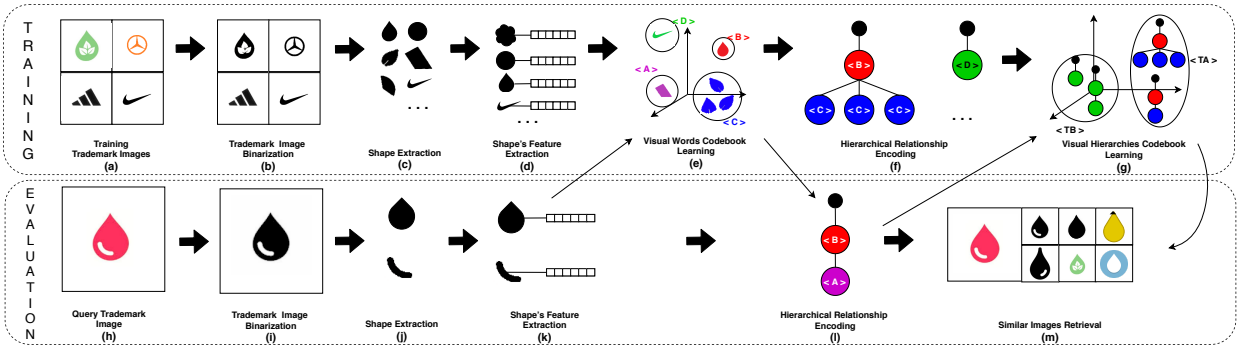

Fig. 1: Overview of the training and evaluation stages of the Hierarchy-of-Visual-Words approach.

The image retrieval system proposed by Alajlan *et al.* [7], [8] uses a structured representation called Curvature Tree to encode both shape and topology of objects and holes (Definiton 1) 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 [10] instead of maximum similarity subtree isomorphism [13].

The use of codebooks and structured representation of images was also considered by Silva *et al.* [14], [15]. However, in this case, local information of grayscale images is extracted by Hessian Affine and SIFT 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 instead of the tree structure applied by our approach to organizing the component shapes of the trademark image.

III. The Proposed Framework

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. Than, we apply a median filter [16] to reduce impulsive noise and a bilateral filter [17] to remove texture without losing overall shapes since sharp edges are preserved. The final binary image is obtained by applying Otsu's method [18] 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 connect 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 rectangle and the square).

In this work, we extract shapes using the border-following method proposed by Suzuki and Abe [19]. 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.

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 of order 0 to 8. ZM are rotation invariant by construction. We use Khotanzad and Hong's approach [20] to obtain translation and scale invariance too. We have used four measures to describe shape's contour [21]:

- *Circularity* defines the relation between the perimeter of the shape and its area;
- *Average bending energy* defines the mean sum of the shape's curvature;
- *Eccentricity* characterizes the statistical distribution of contour points around the principal axes of shape's contour;
- *Convexity* defines the relation between the perimeter of the convex hull and the perimeter of the shape.

We have chosen the above mentioned descriptors 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 (Fig. 1 (e)). In this step, we apply k-means clustering [11] 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.

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.

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 [12]. 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 of elements. Also, it must be computationally and memory-efficient. We satisfy those requirements by using the AP-TED algorithm developed by Pawlik and Augsten [10], 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:

$$\delta_r(n_a, n_b) = \mathrm{dist}_E(\lambda_a, \lambda_b), \tag{1}$$

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 in function 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 or the number of siblings of n at same level $\mathcal{L}$ of the tree:

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

where $\alpha = \min\{\log_2^{-1} \mathcal{L}, \log_2^{-1} \mathcal{D}\}$. In (2), 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.

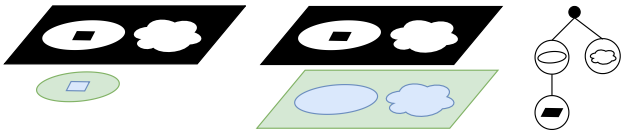

Fig. 2: Examples of inclusion and exclusion relationships.

Similar images retrieval step receives the hierarchical representation of the query image as input (Fig. 1 (m)). 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 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 presented in ascending order of dissimilarity.

IV. Experiments and Results

We have implemented our algorithms using Python. Our code[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` library. Image filtering and shape extraction are performed using `OpenCV`. The dissimilarity of hierarchies is computed using the AP-TED's implementation provided by the authors.

We have used two image databases in our experiments: *MPEG-7 Core Experiment CE-Shape-1* and *MPEG-7 Region Shape Dataset CE-2* [22]. The MPEG-7 CE-1 database is comprised of $1{,}400$ images organized into 70 classes having 20 similar images each. The MPEG-7 CE-2 database includes 871 images organized into 51 categories having from 11 to 21 images each, and 2750 images that do not belong to any category. We have used only the categorized images in our experiments.

In our experiments, we have performed median filtering using a 5×5 window. Here, bilateral filtering has no effect 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 [23] computed as function of the number of clusters in k-means, for $k \in \{100, 200, \cdots, 1200\}$.

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 increased number of labels, which makes it more detailed in terms of nuances between visual hierarchies. On the other hand, a large h value produces a codebook with fewer labels, which makes it more resilient to small changes. The parameter h was set to 0.7 and 1.7 for, respectively, MPEG-7 CE-1 and

[1] HoVW: https://github.com/Prograf-UFF/HoVW

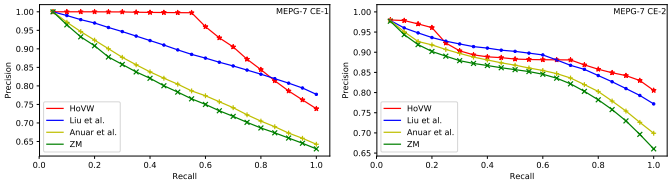

Fig. 3: The precision-recall curves of compared approaches.

MPEG-7 CE-2 databases. Both values maximized the MAP value and produced codebooks having 25 and 21 labels, respectively. According to our experience, the advantage on using mean shift 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 trademark images than for visual words representing simpler shapes.

We have evaluated our approach by comparing its precision-recall curves [23] to results presented by Liu *et al.* [4], Anuar *et al.* [2], and the traditional use of ZM as baseline [20]. In experiments on MPEG-7 CE-1 (Fig. 3, left), HoVM 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, HoVM'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. Recall from Section III that 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. 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. 3 (left) 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 recall of 85% to 100%. This result suggests that in practical applications of TIR specialists would have access to the expected top-ranked similar trademark images while judging copyright infringement.

Fig. 3 (right) 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.

V. Conclusion and Future Works

We presented a learning-based approach for TIR that uses two codebooks. The first codebook encodes basic shapes expected in the images. 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. 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 principles from Gestalt psychology while decomposing trademark images into basic shapes.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

[19] 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.

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

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

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

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