Context-Dependent Logo Matching and Recognition

Hichem Sahbi, Lamberto Ballan, *Member, IEEE*, Giuseppe Serra, and Alberto Del Bimbo, *Member, IEEE*

***Abstract*—We contribute through this work to the design of a novel variational framework able to match and recognize multiple instances of multiple reference logos in image archives. Reference logos as well as test images, are seen as constellations of local features (interest points, regions, etc.) and matched by minimizing an energy function mixing (i) a fidelity term that measures the quality of feature matching (ii) a neighborhood criterion which captures feature co-occurrence/geometry and (iii) a regularization term that controls the smoothness of the matching solution. We also introduce a detection/recognition procedure and we study its theoretical consistency. Finally, we show the validity of our method through extensive experiments on the challenging *MICC-Logos* dataset overtaking, by 20%, baseline as well as state-of-the-art matching/recognition procedures.**

***Index Terms*—Context-dependent kernel, logo detection, logo recognition.**

I. Introduction and Related Work

THE expanding and massive production of visual data from companies, institutions and individuals, and the increasing popularity of social systems like Flickr, YouTube and Facebook for diffusion and sharing of images and video, have more and more urged research in effective solutions for object detection and recognition to support automatic annotation of images and video and content-based retrieval of visual data [1], [2], [3]. Graphic logos are a special class of visual objects extremely important to assess the identity of something or someone. In industry and commerce, they have the essential role to recall in the customer the expectations associated with a particular product or service. This economical relevance has motivated the active involvement of companies in soliciting smart image analysis solutions to scan logo archives to find evidence of similar already existing logos, discover either improper or non-authorized use of their logo, unveil the malicious use of logos that have small variations with respect to the originals so to deceive customers, analyze videos to get statistics about how long time their logo has been displayed.

Logos are graphic productions that either recall some real world objects, or emphasize a name, or simply display some abstract signs that have strong perceptual appeal (see Fig. 1(a)). Color may have some relevance to assess the logo identity. But the distinctiveness of logos is more often given by a few details carefully studied by graphic designers, semiologists and experts of social communication. The graphic layout is equally important to attract the attention of the customer and convey the message appropriately and permanently. Different logos may have similar layout with slightly different spatial disposition of the graphic elements, localized differences in the orientation, size and shape, or – in the case of malicious tampering – differ by the presence/absence of one or few traits (see Fig. 1(b)).

(a)

(b)

(c)

Fig. 1. (a) Examples of popular logos depicting real world objects, text, graphic signs and complex layouts with graphic details; (b) Pairs of logos with malicious small changes in details or spatial arrangements; (c) Examples of logos displayed in real world images in bad light conditions, with partial occlusions and deformations.

Logos however often appear in images/videos of real world indoor or outdoor scenes superimposed on objects of any geometry, shirts of persons or jerseys of players, boards of shops or billboards and posters in sports playfields. In most of the cases they are subjected to perspective transformations and deformations, often corrupted by noise or lighting effects, or partially occluded. Such images – and logos thereafter – have often relatively low resolution and quality. Regions that include logos might be small and contain few information (see Fig. 1(c)). Logo detection and recognition in these scenarios has become important for a number of applications. Among them, several examples have been reported in the literature, such as the automatic identification of products on the web to improve commercial search-engines [4], the verification of

H. Sahbi is with the CNRS LTCI, Télécom ParisTech, 46 rue Barrault, 75013 Paris, France (e-mail: hichem.sahbi@telecom-paristech.fr).

L. Ballan, G. Serra and A. Del Bimbo are with Media Integration and Communication Center (MICC), Università degli Studi di Firenze, Viale Morgagni 65, 50134 - Firenze, Italy (e-mail: lamberto.ballan@unifi.it; serra@dsi.unifi.it; alberto.delbimbo@unifi.it).

Part of this work was conducted while L. Ballan and G. Serra were visiting scholars at Télécom ParisTech.

the visibility of advertising logos in sports events [5], [6], [7], the detection of near-duplicate logos and unauthorized uses [8], [9]. Special applications of social utility have also been reported such as the recognition of groceries in stores for assisting the blind [10].

A generic system for logo detection and recognition in images taken in real world environments must comply with contrasting requirements. On the one hand, invariance to a large range of geometric and photometric transformations is required to comply with all the possible conditions of image/video recording. Since in real world images logos are not captured in isolation, logo detection and recognition should also be robust to partial occlusions. At the same time, especially if we want to discover malicious tampering or retrieve logos with some local peculiarities, we must also require that the small differences in the local structures are captured in the local descriptor and are sufficiently distinguishing for recognition.

A. Related Work

Early work on logo detection and recognition was concerned with providing some automatic support to the logo registration process. The system must check whether other registered logos in archives of millions, exist that have similar appearance to the newcoming logo image, in order to ensure that it is sufficiently distinctive and avoid confusion [11], [12], [13], [8], [14]. Kato's system [15] was among the earliest ones. It mapped a normalized logo image to a 64 pixel grid, and calculated a global feature vector from the frequency distributions of edge pixels. More recently, Wei *et al.* [9] proposed a different solution, where logos were described by global Zernike moments, local curvature and distance to centroid. Other methods have used different global descriptors of the full logo image either accounting for logo contours or exploiting shape descriptors such as *shape context* [16], [17]. All these methods assume that a logo picture is fully visible in the image, is not corrupted by noise and is not subjected to transformations. According to this, they cannot be applied to real world images. Nevertheless, the use of global descriptors for logo detection in real world images has been proposed by several authors [18], [19], [20]. Phan *et al.* [19], [20] considered pairs of color pixels in the edge neighborhoods and accumulated differences between pixels at different spatial distances into a Color-Edge Co-occurence Histogram [18]. This global descriptor permits to perform fast approximate detection of logos, but is unsuited to deal with incomplete information or transformed versions of the original logo, nor to account for a precise representation of the locality of logo traits.

Interest points and local descriptors were used by many authors and appear much more appropriate to support detection and recognition of graphic logos in real world images. In fact, local visual descriptors like MSER [21], SIFT [22], SURF [23], have been proved to be able to capture sufficiently discriminative local elements with some invariant properties to geometric or photometric transformations and are robust to occlusions. In their seminal work, Sivic and Zisserman [24], [25], exploited the bag of visual words approach to represent affine covariant local regions from a codebook of SIFT descriptors; visual words were weighted with *tf-idf* for large-scale retrieval. They showed good capability to discriminate between objects, and gave also examples of logo matching in unconstrained environments. In their approach they did not account for relationships between near keypoints but simply defined a spatial proximity criterion, by checking the local context of the 15 nearest neighbors of each feature match. In [26], logos were described as bag of SIFT features for logo detection and recognition in sequences of sports video. Taking bag of SIFTs instead of bags of visual words has the advantage that only a few highly distinctive keypoints are searched for matching and the formation of the visual vocabulary is avoided. They accounted for spatial relationships between local features by performing iterative robust spatial clustering of the matched features, using M-estimation and outlier rejection. Although experiments showed that logos can be detected in very critical conditions and under partial occlusions with both systems, both methods only account for generic proximity and are therefore unable to capture the small differences in details or spatial layouts, and discover near duplicates.

Bag of SIFTs were also used by Joly and Buisson [27] and Costantinopoulos *et al.* [7]. To discard the outliers they performed geometric consistency checking, assuming the presence of affine geometric transformation between query and target images. Particularly, in [27] the authors applied the standard RANSAC algorithm to refine the initial set of feature matches. In this way they introduce a geometric verification according to a model (affine transformation) that could not be consistent in practice. Chum *et al.* [28], and Wu *et al.* [29] accounted for spatial proximity between visual words by performing spatial geometric hashing. In this way they were able to retrieve near duplicates in web images. However, while effective for searching in very large datasets, spatial geometric hashing does not permit a precise discrimination between local peculiarities.

Accounting of context geometry is crucial for recognition of individual objects in a scene and also for recognition of object with localized peculiarities, and appears therefore necessary to address the requirements of the problem at hand. Contextual information at the image level, such as in the spatial pyramid approach for whole-image categorization [30] is clearly not appropriate. The joint distribution of the geometry of object parts was considered by Fergus *et al.* [31] in constellation models. But such approach is impractical in most of the cases, since the complexity of the representation grows with the number of parts and the model becomes too difficult to learn when the number of parts is higher than a few units. Carneiro and Jepson [32] suggested to group local image features in flexible spatial models to improve matching accuracy between images. In their approach, matched features are refined by applying clustering and model verification based on semi-local spatial constraints. Chum and Matas [33] considered a special case where feature appearance is ignored and only spatial relations between pairs of features are used. Pantofaru et al. [34] introduced a method for object detection and localization which combines regions generated by image segmentation

with local patches. In particular, they defined the Region-based Context Feature as the histogram of the (quantized) local features near a segmented region, where the scale of the local patches is used to define spatial proximity. Similarly, Mortensen *et al.* [35] combined SIFT descriptors with a shape descriptor of the point neighborhood (the "Global Context") very similar to shape context. All of these methods do not appear appropriate to discriminate between slightly differing traits. Bronstein and Bronstein [36] have recently proposed to directly incorporate spatial information in the feature descriptor. They defined spatially sensitive bags of pairs of features, i.e. the distribution of near pairs of features. Particularly they showed that such pairs may have affine invariance if the feature transform and the canonical neighborhoods of the points are affine covariant. However in this approach, the representation is only affine covariant and has a very high dimensionality. Solutions for logo detection in unconstrained real world images, with explicit account of local contexts were recently presented by a few authors. Among them, Gao *et al.* [37] proposed a two-stage algorithm that accounts for local contexts of keypoints. They considered spatial-spectral saliency to avoid the impact of cluttered background and speed up the logo detection and localization. Unfortunately, their solution has revealed to be very sensitive to occlusions. Kleban *et al.* [38] employed a more complex approach that considers association rules between frequent spatial configuration of quantized SIFT features at multiple resolutions [39]. As reported also by the authors, a major limitation of this approach is image resolution since multiple local features are required to mine robust spatial configurations. This makes the method very weak in case of small or partially occluded logos.

B. Paper Contribution and Organization

In this paper, we present a novel solution for logo detection and recognition which is based on the definition of a "Context-Dependent Similarity" (CDS) kernel that directly incorporates the spatial context of local features [40], [41]. The proposed method is model-free, i.e. it is not restricted to any a priori alignment model. Context is considered with respect to each single SIFT keypoint and its definition recalls *shape context* with some important differences: given a set of SIFT interest points $\mathcal{X}$, the context of $x \in \mathcal{X}$ is defined as the set of points spatially close to x with particular geometrical constraints. Formally, the CDS function is defined as the fixed-point of three terms: (i) an energy function which balances a *fidelity* term; (ii) a *context* criterion; (iii) an *entropy* term. The fidelity term is inversely proportional to the expectation of the Euclidean distance between the most likely aligned interest points. The context criterion measures the spatial coherence of the alignments: given a pair of interest points (f_p, f_q) respectively in the query and target image with a high alignment score, the context criterion is proportional to the alignment scores of all the pairs close to (f_p, f_q) *but with a given spatial configuration*. The "entropy" term acts as a smoothing factor, assuming that with no a priori knowledge, the joint probability distribution of alignment scores is flat. It acts as a *regularizer* that controls the entropy of the conditional probability of matching, hence the uncertainty and decision thresholds so helping to find a direct analytic solution. Using the CDS kernel, the geometric layout of local regions can be compared across images which show contiguous and repeating local structures as often in the case of graphic logos. The solution is proved to be highly effective and responds to the requirements of logo detection and recognition in real world images.

The rest of the paper is organized as follows. In Section II, we report the definition of the "Context-Dependent Similarity" function. Hence, in Section III, we discuss the adaptation of this similarity function to the problem of logo detection in real world images, and apply this function to align interest points. We discuss the probability of point alignment in challenging conditions; invariance properties are also discussed. Results and comparative evaluations are presented in Section IV.

II. Context-Dependent Similarity

Let $\mathcal{S}_X = \{x_1, \ldots, x_n\}$, $\mathcal{S}_Y = \{y_1, \ldots, y_m\}$ be respectively the list of interest points taken from a reference logo and a test image (the value of n, m may vary with $\mathcal{S}_X$, $\mathcal{S}_Y$). We borrow the definition of context and similarity design from [40], [41], in order to introduce a new matching procedure applied to logo detection. The main differences with respect to [40], [41] reside in

- **The use of context for matching.** Context is used to find interest point correspondences between two images in order to tackle logo detection while in [40], context was used for kernel design in order to handle object classification using support vector machines.
- **The update of the design model.** Adjacency matrices are defined in order to model spatial and geometric relationships (context) between interest points belonging to two images (a reference logo and a test image). These adjacency matrices model interactions between interest points at different orientations and locations resulting into an anisotropic context, while in [40], context was isotropic.
- **The similarity diffusion process.** Resulting from the definition of context, similarity between interest points is recursively and anisotropically diffused.
- **The interpretation of the model.** Our designed similarity may be interpreted as a joint distribution (pdf) which models the probability that two interest points taken from $\mathcal{S}_X \times \mathcal{S}_Y$ match. In order to guarantee that this similarity is actually a pdf, a partition function is used as a normalization factor taken through all the interest points in $\mathcal{S}_X \times \mathcal{S}_Y$ (and not over all the objects in a training database as in [40]).

A. Context

The context is defined by the local spatial configuration of interest points in both $\mathcal{S}_X$ and $\mathcal{S}_Y$. Formally, in order to take into account spatial information, an interest point $x_i \in \mathcal{S}_X$ is defined as $x_i = (\psi_g(x_i), \psi_f(x_i), \psi_o(x_i), \psi_s(x_i), \omega(x_i))$ where the symbol $\psi_g(x_i) \in \mathbb{R}^2$ stands for the $2D$ coordinates of x_i while $\psi_f(x_i) \in \mathbb{R}^c$ corresponds to the feature of x_i (in

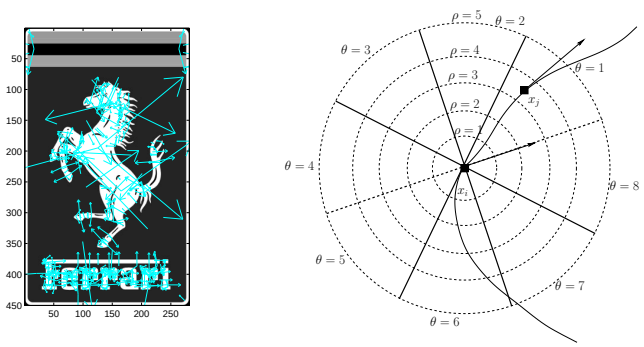

Fig. 2. This figure shows (on the left) a collection of SIFT points with their locations, orientations and scales, and (on the right) the definition and the partitioning of the context of an interest point x_i into different sectors (for orientations) and bands (for locations).

practice c is equal to 128, i.e. the coefficients of the SIFT descriptor [22]). We have also an extra information about the orientation of x_i (denoted $\psi_o(x_i) \in [-\pi, +\pi]$) which is provided by the SIFT gradient and about the scale of the SIFT descriptor (denoted $\psi_s(x_i)$). Finally, we use $\omega(x_i)$ to identify the image from which the interest point comes from, so that two interest points with the same location, feature and orientation are considered different when they are not in the same image; this is motivated by the fact that we want to take into account the context of the interest point in the image it belongs to. Let $d(x_i, y_j) = \|\psi_f(x_i) - \psi_f(y_j)\|_2$ measure the dissimilarity between two interest point features, where $\|\cdot\|_2$ is the "entrywise" L_2-norm (i.e. the sum of the square values of vector coefficients). The context of x_i is defined as in the following:

$$\mathcal{N}^{\theta,\rho}(x_i) = \{x_j : \omega(x_j) = \omega(x_i), x_j \neq x_i \text{ s.t. (i), (ii) hold}\},$$

with

$$\frac{\rho-1}{N_r}\,\epsilon_p \leq \|\psi_g(x_i) - \psi_g(x_j)\|_2 \leq \frac{\rho}{N_r}\,\epsilon_p, \qquad \text{(i)}$$

and

$$\frac{\theta-1}{N_a}\pi \leq \angle\big(\psi_o(x_i), \psi_g(x_j) - \psi_g(x_i)\big) \leq \frac{\theta}{N_a}\pi \qquad \text{(ii)}$$

where $\big(\psi_g(x_j) - \psi_g(x_i)\big)$ is the vector between the two point coordinates $\psi_g(x_j)$ and $\psi_g(x_i)$. The radius of a neighborhood disk surrounding x_i is denoted as ϵ_p and obtained by multiplying a constant value ϵ to the scale $\psi_s(x_i)$ of the interest point x_i. In the above definition, $\theta = 1, ..., N_a$, $\rho = 1, ..., N_r$ correspond to indices of different parts of that disk (see Fig. 2). In practice, as we will show in the experimental part of this paper (see Sect. IV), N_a and N_r correspond to 8 sectors and 8 bands. The definition of neighborhoods $\{\mathcal{N}^{\theta,\rho}(x_i)\}_{\theta,\rho}$ reflects the co-occurrence of different interest points with particular spatial geometric constraints. Fig. 3 shows an example taken from two different images containing the same logo ("Heineken"); the figure reports the context definition for two corresponding keypoints, showing a similar spatial configuration. All the definitions about interest points in $\mathcal{S}_Y$ and their context are similar to $\mathcal{S}_X$.

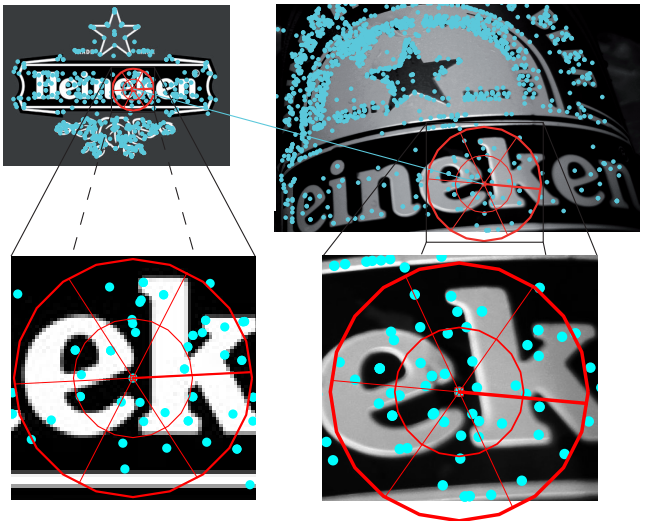

Fig. 3. This figure shows an example of real context definition. The two columns show the partitioning of the context of two corresponding interest points; which belong to two instances of "Heineken". In this example we consider a context definition including 6 sectors and 8 bands.

B. Similarity Design

We define k as a function which, given two interest points $(x, y) \in \mathcal{S}_X \times \mathcal{S}_Y$, provides a similarity measure between them. For a finite collection of interest points, the sets $\mathcal{S}_X$, $\mathcal{S}_Y$ are finite. Provided that we put some (arbitrary) order on $\mathcal{S}_X$, $\mathcal{S}_Y$, we can view function k as a matrix $\mathbf{K}$, i.e. $\mathbf{K}_{x,y} = k(x, y)$, in which the "$(x, y)-$element" is the similarity between x and y. We also represent with $\mathbf{P}_{\theta,\rho}$, $\mathbf{Q}_{\theta,\rho}$ the intrinsic adjacency matrices that respectively collect the adjacency relationships between the sets of interest points $\mathcal{S}_X$ and $\mathcal{S}_Y$, for each context segment; these matrices are defined as $\mathbf{P}_{\theta,\rho,x,x'} = g_{\theta,\rho}(x, x')$, $\mathbf{Q}_{\theta,\rho,y,y'} = g_{\theta,\rho}(y, y')$ where g is a decreasing function of any (pseudo) distance involving (x, x'), *not necessarily symmetric*. In practice, we consider $g_{\theta,\rho}(x, x') = \mathbb{1}_{\{\omega(x)=\omega(x')\}} \times \mathbb{1}_{\{x' \in \mathcal{N}^{\theta,\rho}(x)\}}$, so the matrices $\mathbf{P}$, $\mathbf{Q}$ become sparse and binary. Finally, let $\mathbf{D}_{x,y} = d(x, y) = \|\psi_f(x) - \psi_f(y)\|_2$. Using this notation, the similarity $\mathbf{K}$ between the two objects $\mathcal{S}_X$, $\mathcal{S}_Y$ is obtained by solving the following minimization problem

$$\min_{\mathbf{K}} \quad \mathrm{Tr}\big(\mathbf{K}\,\mathbf{D}'\big) \;+\; \beta\,\mathrm{Tr}\big(\mathbf{K}\;\log\;\mathbf{K}'\big) - \alpha \sum_{\theta,\rho} \mathrm{Tr}\big(\mathbf{K}\,\mathbf{Q}_{\theta,\rho}\,\mathbf{K}'\,\mathbf{P}'_{\theta,\rho}\big) \qquad (1)$$

$$\text{s.t.} \quad \begin{cases} \mathbf{K} \geq 0 \\ \|\mathbf{K}\|_1 = 1 \end{cases}$$

Here $\alpha, \beta \geq 0$ and the operations log (natural), $\geq$ are applied individually to every entry of the matrix (for instance, $\log \mathbf{K}$ is the matrix with $(\log \mathbf{K})_{x,y} = \log k(x, y)$), $\|\cdot\|_1$ is the "entrywise" L_1-norm (i.e., the sum of the absolute values of the matrix coefficients) and Tr denotes matrix trace.

The first term, in the above constrained minimization problem, measures the quality of matching between two features $\psi_f(x)$, $\psi_f(y)$. In our case this is inversely proportional to the distance, $d(x, y)$, between the 128 SIFT coefficients of x

and y. A high value of $\mathbf{D}_{x,y}$ should result into a small value of $\mathbf{K}_{x,y}$ and vice-versa. The second term is a regularization criterion which considers that without any a priori knowledge about the aligned interest points, the probability distribution $\{\mathbf{K}_{x,y} : x \in \mathcal{S}_X, y \in \mathcal{S}_Y\}$ should be flat so the negative of the entropy is minimized. This term also helps defining a direct analytic solution of the constrained minimization problem (1). The third term is a neighborhood criterion which considers that a high value of $\mathbf{K}_{x,y}$ should imply high values in the neighborhoods $\mathcal{N}^{\theta,\rho}(x)$ and $\mathcal{N}^{\theta,\rho}(y)$. This criterion also makes it possible to consider the spatial configuration of the neighborhood of each interest point in the matching process. This minimization problem is formulated by adding an equality constraint and bounds which ensure a normalization of the similarity values and allow to see $\mathbf{K}$ as a probability distribution.

C. Solution

Let's consider the adjacency matrices $\{\mathbf{P}_{\theta,\rho}\}_{\theta,\rho}$, $\{\mathbf{Q}_{\theta,\rho}\}_{\theta,\rho}$ related to a reference logo $\mathcal{S}_X$ and a test image $\mathcal{S}_Y$ respectively, each of which collects the adjacency relationships between the image interest points for a specific context segment θ, ρ. It is possible to show that the optimization problem (1) admits a unique solution $\tilde{\mathbf{K}}$, under some constrains.

Proposition 1: Let $\mathbf{u}$ denote the matrix of ones and introduce

$$\zeta = \frac{\alpha}{\beta} \sum_{\theta,\rho} \|\mathbf{P}_{\theta,\rho}\mathbf{u}\mathbf{Q}'_{\theta,\rho} + \mathbf{P}'_{\theta,\rho}\mathbf{u}\mathbf{Q}_{\theta,\rho}\|_\infty,$$

where $\|\cdot\|_\infty$ is the "entrywise" L_∞-norm. Provided that the following two inequalities hold

$$\zeta \exp(\zeta) < 1 \tag{2}$$

$$\|\exp(-\mathbf{D}/\beta)\|_1 \geq 2 \tag{3}$$

the optimization problem (1) admits a unique solution $\tilde{\mathbf{K}}$, which is the limit of the *recursive form*

$$\mathbf{K}^{(t)} = \frac{G(\mathbf{K}^{(t-1)})}{\|G(\mathbf{K}^{(t-1)})\|_1}, \tag{4}$$

with

$$G(\mathbf{K}) = \exp\left\{ -\frac{\mathbf{D}}{\beta} + \frac{\alpha}{\beta} \sum_{\theta,\rho} \left(\mathbf{P}_{\theta,\rho}\mathbf{K}\mathbf{Q}'_{\theta,\rho} + \mathbf{P}'_{\theta,\rho}\mathbf{K}\mathbf{Q}_{\theta,\rho}\right) \right\}, \tag{5}$$

and

$$\mathbf{K}^{(0)} = \frac{\exp(-\mathbf{D}/\beta)}{\|\exp(-\mathbf{D}/\beta)\|_1}$$

Besides $\mathbf{K}^{(t)}$ satisfy the convergence property:

$$\|\mathbf{K}^{(t)} - \tilde{\mathbf{K}}\|_1 \leq L^t \|\mathbf{K}^{(0)} - \tilde{\mathbf{K}}\|_1 \tag{6}$$

with $L = \zeta \exp(\zeta)$.

Proof: This solution is a variant of the one found in [41]. The demonstration given in [41] still holds in this case. ■

Notice that at the convergence stage, we omit t in all $\mathbf{K}^{(t)}$ so the latter will simply be denoted as $\mathbf{K}$.

Algorithm 1: CDS Logo Detection and Recognition

Input: Reference logo image: I_X, Test image: I_Y, CDS parameters: $\epsilon, N_a, N_r, \alpha, \beta, \tau$.

Output: A boolean value determining whether the reference logo in I_X is detected in I_Y.

Extract SIFT from I_X, I_Y and let $\mathcal{S}_X := \{x_1, \ldots, x_n\}$, $\mathcal{S}_Y := \{y_1, \ldots, y_m\}$ be respectively the list of interest points taken from both images;

for $i \leftarrow 1$ **to** n **do**
  Compute the context of x_i, given ϵ, N_a, N_r;

for $j \leftarrow 1$ **to** m **do**
  Compute the context of y_j, given ϵ, N_a, N_r;

Set $t \leftarrow 1$, $max_t \leftarrow 30$;

repeat
  for $i \leftarrow 1$ **to** n **do**
    for $j \leftarrow 1$ **to** m **do**
      Compute CDS matrix entry $\mathbf{K}^{(t)}_{x_i,y_j}$, given α, β;
  Set $t \leftarrow t + 1$;

until *convergence (i.e.,* $\left\|\mathbf{K}^{(t)} - \mathbf{K}^{(t-1)}\right\|_2 \rightsquigarrow 0$*) OR* $t > max_t$;

$\mathbf{K} \leftarrow \mathbf{K}^{(t)}$;

for $i \leftarrow 1$ **to** n **do**
  for $j \leftarrow 1$ **to** m **do**
    Compute $\mathbf{K}_{y_j|x_i} \leftarrow \dfrac{\mathbf{K}_{x_i,y_j}}{\sum_{s=1}^{m} \mathbf{K}_{x_i,y_s}}$;
    A match between x_i and y_j is declared iff $\mathbf{K}_{y_j|x_i} \geqslant \sum_{s \neq j}^{m} \mathbf{K}_{y_s|x_i}$;

if *number of matches in* $\mathcal{S}_Y > \tau|\mathcal{S}_X|$ **then**
  return *true*; `// i.e. logo detection`
else
  return *false*;

III. Logo Detection and Recognition

Application of CDS to logo detection and recognition requires to establish a matching criterion and verify its probability of success.

Let $\mathcal{R} \subset \mathbb{R}^2 \times \mathbb{R}^{128} \times [-\pi, +\pi] \times \mathbb{R}^+$ denote the set of interest points extracted from all the possible reference logo images (see Section II-A) and X a random variable standing for interest points in $\mathcal{R}$. Similarly, we define $\mathcal{T} \subset \mathbb{R}^2 \times \mathbb{R}^{128} \times [-\pi, +\pi] \times \mathbb{R}^+$ as the set of interest points extracted from all the possible test images (either including logos or not) and Y a random variable standing for interest points in $\mathcal{T}$. X and Y are assumed drawn from existing (but unknown) probability distributions. Let's consider $\mathcal{S}_X = \{X_1, \ldots, X_n\}$, $\mathcal{S}_Y = \{Y_1, \ldots, Y_m\}$ as n and m realizations with the same distribution as X and Y respectively. To avoid false matches we have assumed that matching between Y_J and X is assessed iff

$$\mathbf{K}_{Y_J|X} \;\geq\; \sum_{j \neq J}^{m} \mathbf{K}_{Y_j|X}, \tag{7}$$

being $\mathbf{K}_{Y|X} = \mathbf{K}_{X,Y} / \big(\sum_{j=1}^{m} \mathbf{K}_{X,Y_j} \big)$.

The intuition behind the strong criterion above comes from the fact that when $\mathbf{K}_{Y_J|X} \gg \sum_{j \neq J}^{m} \mathbf{K}_{Y_j|X}$, the entropy of the conditional probability distribution $\mathbf{K}_{.|X}$ will be close to 0, so the uncertainty about the possible matches of X will be reduced. The reference logo $\mathcal{S}_X$ is declared as present into the test image if, after that the match in $\mathcal{S}_Y$ has been found for each interest point of $\mathcal{S}_X$, the number of matches is sufficiently large (at least $\tau |\mathcal{S}_X|$ for a fixed $\tau \in [0,1]$, being $1-\tau$ the occlusion factor tolerated). We summarize the full procedure for logo detection and recognition in Algorithm 1.

A. Theoretical Foundation of Our Matching Algorithm

A theoretical lower bound to the probability of finding correct matches using criterion (7) can be obtained from Eq. 5, under the hypothesis of correct matches in $\mathcal{S}_X \times \mathcal{S}_Y$ (i.e. the reference logo exists in the image). This hypothesis is referred to as $\mathbf{H}_1$. Similarly $\mathbf{H}_0$ (the null hypothesis) stands for the incorrect matches in $\mathcal{S}_X \times \mathcal{S}_Y$.

Assuming without a loss of generality, that all the entries of the left-hand side term of Eq. 5 (i.e, $\exp(-\mathbf{D}/\beta)$) are identical, for a fixed $\tau \in [0,1]$, it appears clearly that the context term (the right-hand side term inside the exponential) is highly influential and that the probability of finding correct matches is dependent on setting of the parameters α/β and $q = N_a N_r$ (i.e. the fixed number of cells in the context) and also n (i.e. the number of SIFT points in the query image).

Proposition 2: Let $(.)_+$ denote the positive part of any real valued function. For a fixed $\tau \in]0,1]$, one may show that

$$P\left(\mathbf{K}_{Y_J|X} \geq \sum_{j \neq J}^{m} \mathbf{K}_{Y_j|X} \right) \geq \left(\frac{1-\nu}{1+\nu} \right)_+, \qquad (8)$$

here $\nu = (m-1) \left(\dfrac{q^2 - 1 + \exp(2\alpha/\beta)}{q^2 - \tau q + \tau q \exp(2\alpha/\beta)} \right)^{qn}$ and the probability is w.r.t. $X, Y_1, \ldots, Y_m$, with $(X, Y_J) \in \mathbf{H}_1$, $(X, Y_j) \in \mathbf{H}_0$.
Provided that $\tau \gg 1/q$,

$$\nu \overset{n \to +\infty}{\longrightarrow} 0 \quad \text{and} \quad P\left(\mathbf{K}_{Y_J|X} \geq \sum_{j \neq J}^{m} \mathbf{K}_{Y_j|X} \right) \overset{n \to +\infty}{\longrightarrow} 1.$$

Proof: The proof of the proposition above is given in Appendix A. ∎

Fig. 4 compares theoretical expectations with measured performance, as a function of α/β, q, n and shows that with appropriate settings of these parameters, criterion (7) is able to find (almost all) the correct matches while discarding the incorrect ones. Empirical matchings are obtained on a validation set including a subset of "matches" and a subset of "non matches". The two sets were automatically generated (i) by embedding reference logos into test images at random locations so the ground-truth of "matches" and "non-matches" can be *automatically recovered* (these reference logos and test images belong to the MICC-Logos dataset), and (ii) by adding a uniformly distributed noise to the test images. Since logos can be partially occluded, it has been assumed that the reference logo is still detectable even though half-occluded in the test image, so setting $\tau = 0.5$ in the first three curves reported in the figure.

Fig. 4 shows also the evolution of the lower bound in (8) and empirical matching results with respect to the occlusion factor $1-\tau$ (in the fourth curve reported in the figure). For each value of τ, we automatically generate a validation set as described earlier but the logos of test images are now partially and randomly occluded with a factor $1-\tau$. According to Fig. 4, as the amount of occlusion decreases (i.e., τ increases), the probability of finding correct matches increases, when using criterion (7), and reaches a very high value just when $\tau = 0.5$, i.e. even though logos are *half*-occluded. Can be noticed that though the method is tolerant with respect to $\tau < 1$, it remains highly selective, so it can be used effectively also to detect near-duplicates.

B. Properties and Considerations

The adjacency matrices $\mathbf{P}_{\theta,\rho}$, $\mathbf{Q}_{\theta,\rho}$ in $\mathbf{K}$ (see Eq. 4 and 5), provide the context and the intrinsic geometry of the reference and the test logos $\mathcal{S}_X$, $\mathcal{S}_Y$. It is easy to see that $\mathbf{P}_{\theta,\rho}$, $\mathbf{Q}_{\theta,\rho}$ are translation and rotation invariant and can also be made scale invariant when the support (disk) of the context (i.e. its radius ϵ_p) is adapted to the scales of $\psi_g(\mathcal{S}_X)$ and $\psi_g(\mathcal{S}_Y)$ respectively. It follows that the right-hand side of our similarity $\mathbf{K}$ is invariant to any 2D similarity transformation. Notice, also, that the left-hand side of $\mathbf{K}$ may involve similarity invariant features $\psi_f(.)$ (actually SIFT features), therefore $\mathbf{K}$ – and also our matching criterion (7) – is similarity invariant. The context can also be defined on other supports (rectangles, etc.) and can be made invariant to other transformations including affine and non-linear.

By taking β "not too large", one can ensure that (3) holds. Then by taking "small enough" α, inequality (2) can also be satisfied. Note that $\alpha = 0$ corresponds to a similarity which is not context-dependent (i.e. *context-free*, following our nomenclature). So, in this case, the similarities between neighbors are not taken into account to assess the similarity between two interest points. Besides our choice of $\mathbf{K}^{(0)}$ is exactly the optimum (and fixed point) for $\alpha = 0$.

One important aspect of the method that has influence on the performance and suits to logo detection/recognition is that the local context is recursively defined. In particular, we assess that *two interest points match if their local neighbors match, and if the neighbors of their local neighbors match too, etc.* The recursive form of our solution allows us to iteratively diffuse the similarity using larger and more precise context so providing increased precision of matching (see Fig. 5). Another interesting aspect is that the energy function in (1) is model-free, so no a-priori alignment model is used in order to design the similarity and to find the set of matches in $\mathcal{S}_X \times \mathcal{S}_Y$. This avoids to assume a-priori hypothesis that could not fit with the observations.

To have partitioned the neighborhood into several cells corresponding to different degrees of proximity has lead to significant improvements of our experimental results. On the one hand, the constraint (2) becomes easier to satisfy, for

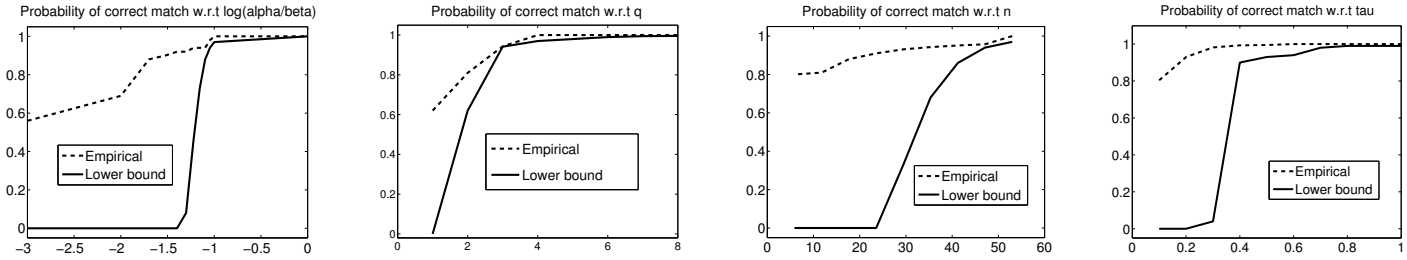

Fig. 4. These figures show the evolution of the probability of finding correct matches using criterion (7). Dashed curves correspond to the empirical measures found experimentally, while solid curves correspond to the lower bounds in (8). The evolution of these curves is shown with respect to $\log(\alpha/\beta)$, q, n and τ respectively. Settings used are $\alpha/\beta = 1$, $q = 4$ and $\tau = 0.5$; n and m vary with respect to reference and test images respectively. Note that $q = 1$ corresponds to isotropic context and $\alpha/\beta \to 0$ corresponds to context-free setting.

Fig. 5. This figure shows the reduction of false matches with respect to the number of iterations in CDS evaluation. At $t = 0$, CDS does not take into account the context and this results into the many wrong matches. As t increases, matching results become precise as the diffusion of the similarity takes into account larger and more precise context (dashed in figures). For ease of visualization only a subset of interest points and their matches are shown.

larger α with partitioned neighborhood, compared to [40]. On the other hand, when the context is split into different parts, we end up with a context term, in the right-hand side of the exponential (5), which grows slowly compared to the one presented in our previous work [40] and grows only *if similar spatial configurations* of interest points have high similarity values. Therefore, numerically, the evaluation of that term is still tractable for large values of α which apparently produces a more positively influencing (and precise) context-dependent term in (1). Fig. 6 shows an example of our context dependent matching and detection results (figures on the right) with respect to context-free ones (figures on the left). Bottom histograms show the conditional probability distribution $\mathbf{K}_{.|X}$ for a particular interest point X in the reference logo. This distribution is peaked when using context dependent similarity so the underlying entropy is close to 0 and the uncertainty about possible matches is dramatically reduced.

From criterion (7) and its theoretical bound (8), several considerations follow. Under the $\mathbf{H}_1$ hypothesis, i.e. the hypothesis that the reference logo exists in the image, the lower bound in (8) increases with respect to n, q, while it decreases with respect to m. Notice that typically $n \ll m$ and also that this bound is useless when $q = 1$ (i.e., when the context is isotropic) and when $q \to \infty$ (i.e., when the number of cells in the context is extremely large leading to overfitting).

The τ element provides a measure of the fraction of

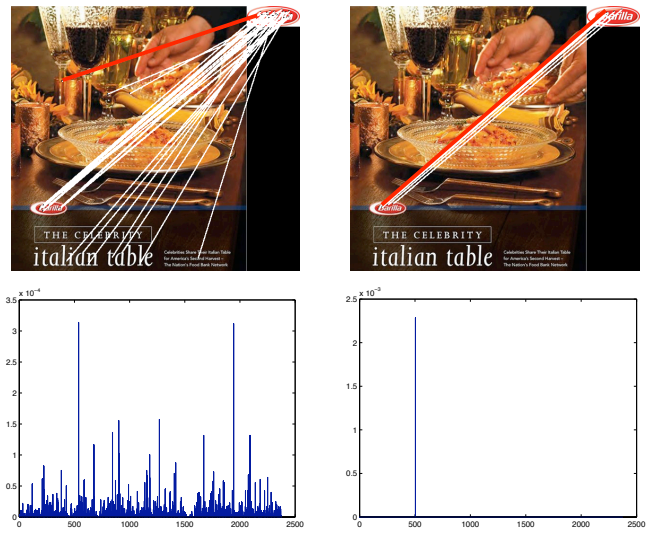

Fig. 6. This figure shows a comparison of the matching results when using a context-free strategy and our context dependent matching. Bottom figures show the conditional probability distribution $K_{\cdot|X}$ for a particular interest point X in the reference logo. This distribution is peaked when using context dependent similarity so the underlying entropy is close to 0 and the uncertainty about possible matches is dramatically reduced. Top figures show the matching results between the reference logo and the test image which are correct using the context dependent matching framework.

interest points that are considered sufficient to assess the presence/absence of a reference logo in a test image. Typically we cannot know a priori what is the amount of occlusion that we may have in test images. Setting τ to a very small value makes the false acceptance rate high, while setting it to a high value makes the false rejection rate high; therefore, setting τ to 0.5 is a kind of compromise that works satisfactorily when no knowledge is available. If we want to detect a portion of a logo even though manipulated or even it has many variants, then we should have tolerance to occlusion. As an example of this aspect, Fig. 7 shows logo detections with different values of τ. Bound (8) shows that performance does not degrade too much when logo structure is different, i.e. some points in reference logo do not have matches in test images. Context remains stable and discriminative. If we want to detect only "exact copies" of logos with only some noise and geometric (similarity) transformations, then we should set τ close to 1 (Fig. 4 also corroborates this aspect showing that the method is very selective without the need of rising the threshold too much). Under $\mathbf{H}_0$, criterion (7) is very strong and difficult to satisfy (i.e. its probability of success is $O(1/m) \rightarrow 0$) and this prevents from creating wrong matches.

IV. Benchmarking

In order to show the effectiveness of our context dependent matching strategy (i.e., based on CDS) with respect to other approaches, we evaluate the performances of multiple-logo detection on a novel challenging dataset called MICC-Logos, containing 13 logo classes each one represented with $15-87$ real world pictures downloaded from the web, resulting into a

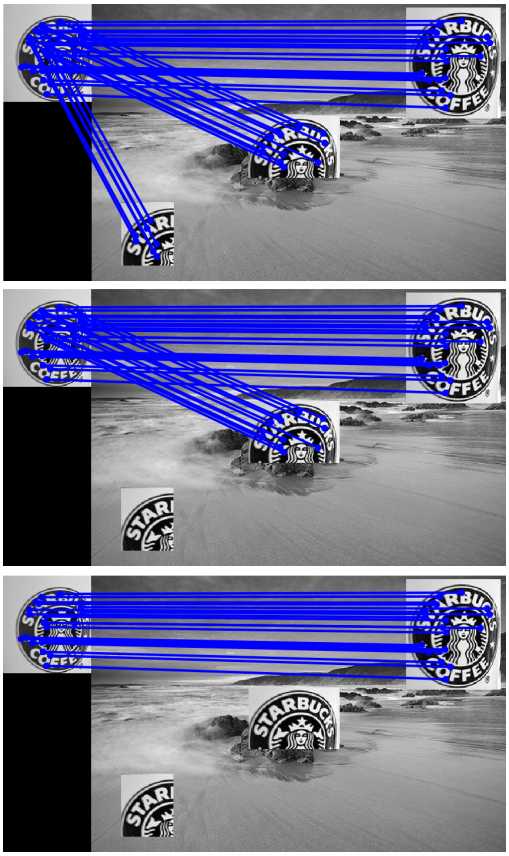

Fig. 7. Examples of logo detections with different parameters of τ (0.25, 0.5 and 0.8, respectively). As τ increases logo detection is more sensitive to occlusion. In this experiment, $\alpha = \beta = 0.1$ and $N_a = N_r = 8$.

collection of 720 images (see Fig. 8) [1]. The image resolution varies from 480×360 to 1024×768 pixels. Interest points are extracted from test images as well as reference logos, and encoded using SIFT features. Each test image $\mathcal{S}_Y$ is processed in order to evaluate the similarity function $\mathbf{K}$ (shown previously in Eq. 4) with respect to each reference logo $\mathcal{S}_X$, using Gaussian power assist setting: $\mathbf{K}^{(0)}_{x,y} = \exp(-d(x,y)/\beta)$.

A. Setting

The setting of β is related to the Gaussian similarity (i.e., $\exp(-\mathbf{D}/\beta)$) as the latter corresponds to the left-hand side (and the baseline form) of $\mathbf{K}^{(t)}$, i.e. when $\alpha = 0$. Since the 128 dimensional SIFT features, used to compute $\mathbf{D}$, have a unit L_2 norm and hence belong to a hypersphere of radius r ($r = 1$), a reasonable setting of β is $0.1r$ which also satisfies condition (3) in our experiments. The influence (and the performance) of the right-hand side of $\mathbf{K}^{(t)}, \alpha \neq 0$ (context term) increases as α increases nevertheless and as shown earlier, the convergence of $\mathbf{K}^{(t)}$ to a fixed point is guaranteed only if Eq. 2 is satisfied. Intuitively, the weight parameter α should then be relatively high while also satisfying condition (2). Following the lower bounds and the empirical measures shown in Fig. 4, it is easy to see that the best matching performance is achieved when $\alpha/\beta = 1$ (in our experiments we set $\alpha = \beta = 0.1$ and $N_r =$

[1] The MICC-Logos dataset is available on request at the following webpage: http://www.micc.unifi.it/vim/datasets/micc-logos

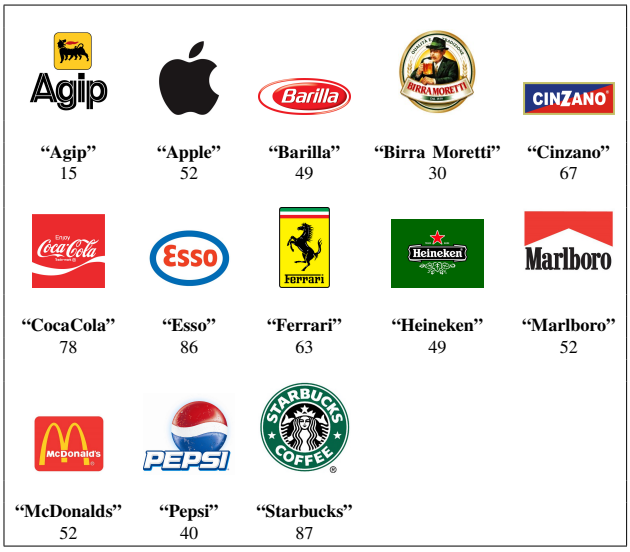

Fig. 8. **MICC-Logos** dataset. Logo classes: the number of test images is reported for each class.

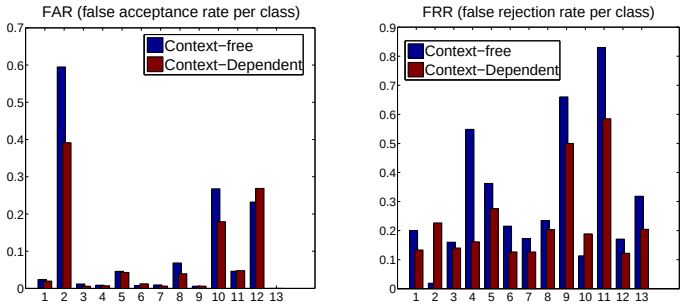

Fig. 9. This figure shows a comparison of logo detection using our (i) context-dependent similarity and (ii) context-free one (actually Gaussian). FAR and FRR rates are shown for each class. In these experiments, $\beta = \alpha = 0.1$ and $\tau = 0.5$ while n and m vary of course with reference logos and test images. Excepting the logos "Apple" and "Mc Donald's" (which contain very few interest points $n < 12$), the FRR errors are almost always significantly reduced while FAR is globally reduced.

$N_a = 8$) and this setting also guarantees conditions in Eqs. 2, 3 and therefore the convergence of CDS to a fixed point. In practice, we observe that convergence usually happens in less than 3 iterations. However, the other interest from convergence is to save time, as one may stop the iterative process before reaching the upper bound on the number of iterations (we set the max number of iterations to 30).

B. Logo Detection Performance

Logo detection is achieved by finding for each interest point in a given reference logo $\mathcal{S}_X$ its best match in a test image $\mathcal{S}_Y$; if the number of matches is larger than $\tau|\mathcal{S}_X|$ (for a fixed $\tau \in]0, 1]$), then the reference logo will be declared as present into the test image. Different values of τ were experimented and performances are measured using False Acceptance and False Rejection Rates (denoted as FAR and FRR, respectively):

$$\text{FAR} = \frac{\text{\# of incorrect logo detection}}{\text{\# of logo detections}};$$

$$\text{FRR} = \frac{\text{\# of unrecognized logo appearance}}{\text{\# of logo appearances}}.$$

Table I reports these FAR and FRR results; setting τ to 0.5 guarantees a high detection rate at the detriment of a small increase of false alarms. Diagrams in Fig. 9 show FAR and FRR for the different classes in the MICC-Logos dataset. We clearly see the out-performance of our context dependent similarity (i.e., $\mathbf{K}^{(t)}$, $t \in \mathbb{N}^+$) with respect to the baseline context-free similarity (i.e., $\mathbf{K}^{(0)}$). For almost all the classes, the improvement brought by CDS is clear and consistent. Figure 11 shows some examples of logo detection results, obtained using the parameters reported in the previous subsection.

TABLE I
PERFORMANCE OBTAINED USING CDS AND DIFFERENT VALUES OF τ. NOTICE THAT FAR IS A DECREASING FUNCTION OF τ WHILE FRR IS AN INCREASING FUNCTION.

τ	0.1	0.2	0.3	0.4	**0.5**	0.6	0.7	0.8	0.9
FAR	0.28	0.22	0.2	0.19	**0.18**	0.18	0.17	0.17	0.17
FRR	0.1	0.11	0.11	0.12	**0.12**	0.13	0.13	0.14	0.14

C. Comparison and Discussion

Firstly, we compare our proposed CDS matching and detection procedure against nearest-neighbor SIFT matching and nearest-neighbor matching with RANSAC verification.

SIFT based logo detection follows the idea in [26] where a reference logo is detected, in a test image, if the overall number of SIFT matches is above a fixed threshold. SIFT matches are obtained by computing for each interest point in $\mathcal{S}_X$ its Euclidean distance to all interest points in $\mathcal{S}_Y$, and keeping only the nearest-neighbors. RANSAC based logo detection follows the same idea but it introduces a model (transformation) based criterion not necessarily consistent in practice. This criterion selects only the matches that satisfy an affine transformation between reference logos and test images. The (iterative) RANSAC matching process, is applied as a "refinement" of SIFT matching (a similar approach is used in [27]). In both cases a match is declared as present iff Lowe's second nearest neighbor test is satisfied [22].

Secondly, we also compare our CDS logo detection algorithm to two relevant methods that use context in their matching procedure [25], [37]. The Video Google approach [25] is closely related to our method as it introduces a spatial consistency criterion, according to which only the matches which share similar spatial layouts are selected. The spatial layout (context) of a given interest point includes 15 nearest neighbors that are spatially close to it. Given $X \in \mathcal{S}_X$, $Y \in \mathcal{S}_Y$, points in the layouts of X and Y which also match casts a vote for the final matching score between X and Y. The basic idea is therefore similar to ours, but the main difference resides in the definition of context in Video Google which is strictly local [2]. In our method the context is also local but recursive; *two interest points match if their local neighbors match, and if the neighbors of their local neighbors match too, etc*, resulting into a recursive diffusion of the similarity through the context (see Fig. 5).

[2]The context in Video Google is empirically set to 15 neighbors and it does not take into account the scale of SIFT points. The number of accounted neighbors may span random image areas depending on the content.

Partial Spatial Context (PSC) logo matching [37] relies on a similar context definition. Given a set of matching interest points, it formulates the spatial distribution for this set (i) by selecting a circular region that contains all these points, (ii) by computing the scale and orientation of the set as the average value of, respectively, all the scales and orientations of the points, (iii) by partitioning the distribution of these points in 9 cells. Starting from this context definition, PSC histograms are computed for both reference logos and test images. A PSC histogram is defined as the number of matches lying in each cell, and logo matching is performed by computing the similarity between two PSC histograms. This schema is efficient and quick to be computed, but its spatial (context) definition is rough and is very sensible to outliers.

Table II and Fig. 10 show a comparison of the results obtained by the five methods. Table II illustrates the FRR performance for fixed FAR values and clearly shows that our CDS method produces the lowest error rates compared to the other methods. Fig. 10 shows the FAR and FRR errors class-by-class on the MICC-Logos dataset.

D. Experiments on FlickrLogos-27

We report also results on another public dataset, the FlickrLogos-27 image collection, to demonstrate the generality of our method. It is a very recent dataset, obtained from Flickr as our dataset, and the authors provide ground-truth for 27 logo classes and annotations for 4536 logo appearances. They proposed a scalable logo recognition approach that extends the common bag-of-words model and incorporates local geometry in the indexing process. In their paper [42] are reported results obtained using a common bag-of-words (bow) model vs their multi-scale Delaunay Triangulation approach (msDT). Both these methods use a codebook of quantized SIFT features. Performances are reported in terms of accuracy by varying the number of training images per class (within the interval $[5, 30]$).

We performed experiments on the FlickrLogos-27 dataset using our CDS method and following the same experimental protocol proposed by the authors (please refer to [42] for more details). Since our method does not provide a learning phase, we followed the same procedure presented in the previous sections using the "training images" as reference logos. Therefore, if we have k training images, we iterate k times our logo detection procedure (as reported in Algorithm 1) and finally we assign to each test image the label corresponding to the reference image that maximizes criterion (7). We report the results in Fig. 12 compared to bow and msDT. As demonstrated by this figure, our method guarantees very good performance also using a single reference logo (i.e. 0.57 in accuracy, that is close to the best performance obtained by the other two methods) and substantially outperforms both methods with more reference images.

E. Computational Cost

The computational cost of our logo detection procedure is mainly dominated by CDS evaluation. In particular, the key part of the algorithm is the computation of the context term.

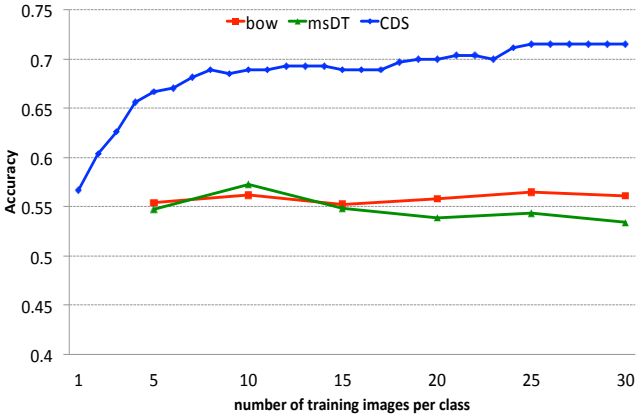

Fig. 12. Performance of our approach vs bow and msDT [42] methods on the FlickrLogos-27 dataset (*query set*). CDS is computed using $\alpha = \beta = 0.1$, $N_r = N_a = 8$ and $\tau = 0.5$.

Assuming $\mathbf{K}^{(t-1)}$ known for a given pair of points (x, y), the complexity is $O(\max(N^2, s))$; here s is the dimension of $\psi_f(x)$ (i.e. 128 since we use SIFT features) and N is given by the $\max_{x,\theta,\rho} \#\{\mathcal{N}^{\theta,\rho}(x)\}$ (i.e. the max number of points in all the neighborhoods). When $N < \sqrt{s}$, evaluating our CDS is equivalent to efficient kernels such as linear or intersection. In worst cases $N \gg \sqrt{s}$ and the evaluation of CDS should be prohibitive. In practice it may only happen when the context is too large (see Fig. 13). Anyway, using the same setting for CDS used in the previous experiments, our method is able to process images and checks for the existence of a reference logo in less than $1s$. This running time is achieved, on average on our MICC-Logos dataset, on a standard 2.6GHZ PC with 2GB memory.

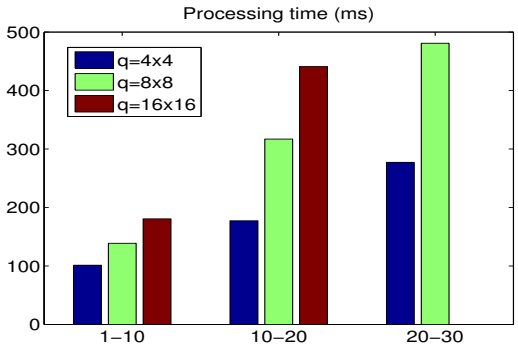

Fig. 13. This figure shows the processing time in order to detect a reference logo in a test image. Results are reported w.r.t. N the maximum number of interest points in the cells of the context; here N varies w.r.t. reference images (and of course increases as ϵ and n increase) and its values are quantized into intervals $[1, 10]$, $]10, 20]$, $]20, 30]$. Note that all these performances were obtained on a test image with 2568 interest points, $\alpha = \beta = 0.1$ and $\tau = 0.5$. Results are not available when $q = 16 \times 16$ and $N \in]20, 30]$ as the context is split into a *large* number of *small* cells so no cell in the context includes more than 20 interest points.

V. Conclusion

We introduced in this work a novel logo detection and localization approach based on a new class of similarities referred to as context dependent. The strength of the proposed method resides in several aspects: (i) the inclusion of the

TABLE II

THIS TABLE SHOWS A COMPARISON OF OUR CDS METHOD WITH RESPECT TO SIFT, RANSAC, VIDEO GOOGLE AND PARTIAL SPATIAL CONTEXT (PSC) MATCHING. THE FIRST ROW REPORTS FAR VALUES, WHILE EACH OTHER ROW THE CORRESPONDING FRR VALUE OBTAINED WITH EACH METHOD. IN THESE EXPERIMENTS, CDS IS COMPUTED BY SETTING $\alpha = \beta = 0.1$, $N_r = N_a = 8$ WHILE τ VARIES IN ORDER TO HAVE FRR FOR DIFFERENT FAR.

FRR \ FAR	0.299	0.181	0.125	0.094	0.075	0.06	0.051	0.043	0.037
CDS	**0.093**	**0.151**	**0.187**	**0.216**	**0.249**	**0.279**	**0.292**	**0.309**	**0.325**
SIFT	0.264	0.348	0.394	0.452	0.503	0.544	0.571	0.589	0.622
RANSAC	0.253	0.340	0.381	0.407	0.423	0.434	0.444	0.457	0.477
Video Google [25]	0.237	0.304	0.350	0.395	0.427	0.448	0.469	0.508	0.538
PSC matching [37]	0.248	0.330	0.371	0.403	0.433	0.467	0.493	0.524	0.551

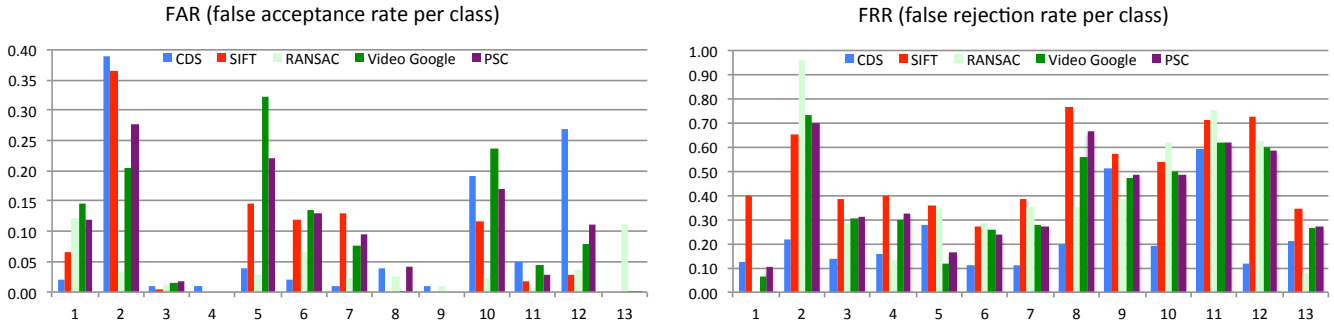

Fig. 10. This figure shows a comparison of logo detection using our (i) context-dependent similarity, (ii) SIFT, (iii) RANSAC and (iv) Video Google. FAR and FRR rates are shown for each class. In these experiments, $\alpha = \beta = 0.1$, $N_r = N_a = 8$ and $\tau = 0.5$.

(a) (b)

Fig. 11. These pictures show some examples of logo detection results. In particular (a) shows examples of matching in case of partial appearance, perspective transformations and low resolution, while (b) shows examples of matching in case of deformations. The default parameters used in these experiments correspond to $\alpha = \beta = 0.1$, $N_r = N_a = 8$ and $\tau = 0.5$.

information about the spatial configuration in similarity design as well as visual features, (ii) the ability to control the influence of the context and the regularization of the solution via our energy function, (iii) the tolerance to different aspects including partial occlusion, makes it suitable to detect both near-duplicate logos as well as logos with some variability in their appearance, and (iv) the theoretical groundedness of the matching framework which shows that under the hypothesis of existence of a reference logo into a test image, the probability of success of matching and detection is high.

Further extensions of this work include the application of the method to logo retrieval in videos and also the refinement of the definition of context in order to handle other rigid and non-rigid logo transformations.

APPENDIX
PROOF OF PROPOSITION 2

Proof: Let $\mathcal{N}_{X,Y}^{\theta,\rho}$ be a random variable standing for the number of matches falling in the context cell (θ, ρ) of X, Y (here X, Y belong respectively to a reference logo and a test image). It is easy to see that under $\mathbf{H_1}$, $\mathcal{N}_{X,Y}^{\theta,\rho} \rightarrow \mathcal{B}(n, \tau/q)$ while under $\mathbf{H_0}$, $\mathcal{N}_{X,Y}^{\theta,\rho} \rightarrow \mathcal{B}(n, 1/q^2)$, q is the fixed number of cells in the context. Again, assuming the left-hand side term

in (4) constant, $\mathbf{K}_{Y_J|X}$ may be written

$$\mathbf{K}_{Y_J|X} \;\; := \;\; \frac{1}{\mathbf{Z}_X} \exp\big(\gamma\, \mathcal{N}_{X,Y_J}\big), \tag{9}$$

here $\gamma = \frac{2\alpha}{\beta}$ and $\mathcal{N}_{X,Y}$ denotes the number of matching pairs in the context of X, Y

$$\mathcal{N}_{X,Y} \;\; = \;\; \sum_{\theta,\rho}^{q} \mathcal{N}_{X,Y}^{\theta,\rho}, \tag{10}$$

and $\mathbf{Z}_X$ is the partition function of $\mathbf{K}_{.|X}$ given by

$$\mathbf{Z}_X = \exp\big(\gamma\, \mathcal{N}_{X,Y_J}\big) + \sum_{j\neq J} \exp\big(\gamma\, \mathcal{N}_{X,Y_j}\big). \tag{11}$$

Under the hypothesis that $(X, Y_J) \in \mathbf{H}_1$ and $(X, Y_j) \in \mathbf{H}_0$, $j \neq J$, our goal is to lower bound p_s

$$p_s = P\bigg(\mathbf{K}_{Y_J|X} \;\geq\; \sum_{j\neq J} \mathbf{K}_{Y_j|X}\bigg). \tag{12}$$

Since $\mathbf{K}_{Y_J|X} + \sum_{j\neq J} \mathbf{K}_{Y_j|X} = 1$, the above probability is

$$\begin{aligned} p_s &= P\bigg(\mathbf{K}_{Y_J|X} \;\geq\; 1/2\bigg) \\ &= P\bigg(\sum_{j\neq J} \mathbf{K}_{Y_j|X} \;<\; 1/2\bigg) \\ &= 1 - P\bigg(\sum_{j\neq J} \mathbf{K}_{Y_j|X} \;\geq\; 1/2\bigg). \end{aligned} \tag{13}$$

By Markov Inequality,

$$\begin{aligned} p_s &\geq 1 - 2\;\mathbb{E}\bigg(\sum_{j\neq J} \mathbf{K}_{Y_j|X}\bigg) \\ &= 1 - 2\;\mathbb{E}\bigg(1 - \mathbf{K}_{Y_J|X}\bigg) \\ &= 2\;\mathbb{E}\bigg(\mathbf{K}_{Y_J|X}\bigg) - 1 \\ &= 2\;\mathbb{E}\bigg(\frac{1}{1+\mathbf{C}}\bigg) - 1, \end{aligned} \tag{14}$$

here the expectation is with respect to $\{X, X_1, \ldots, X_n\}$, $\{Y_1, \ldots, Y_m\}$ and

$$\mathbf{C} = \frac{1}{\exp\big(\gamma\, \mathcal{N}_{X,Y_J}\big)} \sum_{j\neq J} \exp\big(\gamma\, \mathcal{N}_{X,Y_j}\big). \tag{15}$$

As $1/(1+\mathbf{C})$ is convex w.r.t. $\mathbf{C}$, Jensen's Inequality leads to

$$\begin{aligned} p_s &\geq 2\bigg(\frac{1}{1+\mathbb{E}(\mathbf{C})}\bigg) - 1 \\ &= \frac{2\;\mathbb{E}\exp(\gamma\, \mathcal{N}_{X,Y_J})}{\mathbb{E}\exp(\gamma \mathcal{N}_{X,Y_J}) + \sum_{j\neq J} \mathbb{E}\exp(\gamma\, \mathcal{N}_{X,Y_j})} - 1, \end{aligned} \tag{16}$$

which also results from the independence of terms involved in the sum and the product in (15).

Let $\Phi_t(Z) = \mathbb{E}\exp(tZ),\ t \in \mathbb{R}$ be the moment generating function of a given random variable Z. For a collection of i.i.d random variables $Z_1, \ldots, Z_n$ with the same distribution as Z, one may show that

$$\Phi_t(Z_1 + \cdots + Z_n) = \big(\Phi_t(Z)\big)^n \tag{17}$$

As (10) is the sum of mutually independent binomials, Equations (16) and (17) imply

$$p_s \;\; \geq \;\; \frac{2\;\Phi_\gamma(\mathcal{N}_{X,Y_J}^{\theta,\rho})^q}{\Phi_\gamma(\mathcal{N}_{X,Y_J}^{\theta,\rho})^q + (m-1)\Phi_\gamma(\mathcal{N}_{X,Y_j}^{\theta,\rho})^q} - 1, \tag{18}$$

with

$$\begin{aligned} \Phi_\gamma(\mathcal{N}_{X,Y_J}^{\theta,\rho}) &= \big(1 - \tau/q + \tau\exp(2\alpha/\beta)/q\big)^n \\ \Phi_\gamma(\mathcal{N}_{X,Y_j}^{\theta,\rho}) &= \big(1 - 1/q^2 + \exp(2\alpha/\beta)/q^2\big)^n \end{aligned} \tag{19}$$

If we replace (19) into (18), and provided that $\tau \gg 1/q$, we obtain our main result

$$p_s \;\; \geq \;\; \bigg(\frac{1-\nu}{1+\nu}\bigg)_+ \overset{n\to+\infty}{\longrightarrow} 1, \tag{20}$$

with $\nu = (m-1)\bigg(\dfrac{q^2 - 1 + \exp(2\alpha/\beta)}{q^2 - \tau q + \tau q \exp(2\alpha/\beta)}\bigg)^{qn} \overset{n\to+\infty}{\longrightarrow} 0$ ■

REFERENCES

[1] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 1–60, 2008.

[3] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, "Event Detection and Recognition for Semantic Annotation of Video," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 279–302, 2011.

[4] Y. Jing and S. Baluja, "Pagerank for product image search," in *Proc. of WWW*, Beijing, China, 2008, pp. 307–316.

[5] L. Ballan, M. Bertini, and A. Jain, "A system for automatic detection and recognition of advertising trademarks in sports videos," in *Proc. of ACM Multimedia*, Vancouver, BC, Canada, 2008, pp. 991–992.

[6] A. Watve and S. Sural, "Soccer video processing for the detection of advertisement billboards," *Pattern Recognition Letters*, vol. 29, no. 7, pp. 994–1006, 2008.

[7] C. Constantinopoulos, E. Meinhardt-Llopis, Y. Liu, and V. Caselles, "A robust pipeline for logo detection," in *Proc. of IEEE ICME*, Barcelona, Spain, 2011, pp. 1–6.

[8] J.-L. Shih and L.-H. Chen, "A new system for trademark segmentation and retrieval," *Image and Vision Computing*, vol. 19, no. 13, pp. 1011–1018, 2001.

[9] C.-H. Wei, Y. Li, W.-Y. Chau, and C.-T. Li, "Trademark image retrieval using synthetic features for describing global shape and interior structure," *Pattern Recognition*, vol. 42, no. 3, pp. 386–394, 2009.

[10] M. Merler, C. Galleguillos, and S. Belongie, "Recognizing groceries in situ using in vitro training data," in *Proc. of IEEE CVPR SLAM Workshop*, Minneapolis, MN, USA, 2007, pp. 1–8.

[11] Y. S. Kim and W. Y. Kim, "Content-based trademark retrieval system using visually salient feature," in *Proc. of IEEE CVPR*, San Juan, Puerto Rico, 1997, pp. 307–312.

[12] A. Jain and A. Vailaya, "Shape-based retrieval: a case study with trademark image databases," *Pattern Recognition*, vol. 31, no. 9, pp. 1369–1390, 1998.

[13] J. P. Eakins, J. M. Boardman, and M. E. Graham, "Similarity retrieval of trademark images," *IEEE Multimedia*, vol. 5, no. 2, pp. 53–63, 1998.

[14] J. Schietse, J. P. Eakins, and R. C. Veltkamp, "Practice and challenges in trademark image retrieval," in *Proc. of ACM CIVR*, Amsterdam, NL, 2007, pp. 518–524.

[15] T. Kato, "Database architecture for content-based image retrieval," *Proc. of SPIE Image Storage and Retrieval Systems,*, vol. 1662, no. 1, pp. 112–123, 1992.

[16] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.
[17] J. Rodriguez, P. Aguiar, and J. Xavier, "ANSIG - An analytic signature for permutation-invariant two-dimensional shape representation," in *Proc. of IEEE CVPR*, Anchorage, AK, USA, 2008, pp. 1–8.
[18] J. Luo and D. Crandall, "Color object detection using spatial-color joint probability functions," *IEEE Transactions on Image Processing*, vol. 15, no. 6, pp. 1443–1453, 2006.
[19] R. Phan, J. Chia, and D. Androutsos, "Unconstrained logo and trademark retrieval in general color image database using color edge gradient co-occurence histograms," in *Proc. of IEEE ICASSP*, Las Vegas, NV, USA 2008, pp. 1221–1224.
[20] R. Phan and D. Androutsos, "Content-based retrieval of logo and trademarks in unconstrained color image databases using color edge gradient co-occurrence histograms," *Computer Vision and Image Understanding*, vol. 114, no. 1, pp. 66–84, 2010.
[21] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.
[22] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[23] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.
[24] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. of ICCV*, vol. 2, Nice, France, 2003, pp. 1470–1477.
[25] ——, "Efficient visual search of videos cast as text retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 591–606, 2009.
[26] A. D. Bagdanov, L. Ballan, M. Bertini, and A. Del Bimbo, "Trademark matching and retrieval in sports video databases," in *Proc. of ACM MIR*, Augsburg, Germany, 2007, pp. 79–86.
[27] A. Joly and O. Buisson, "Logo retrieval with a contrario visual query expansion," in *Proc. of ACM Multimedia*, Beijing, China, 2009, pp. 581–584.
[28] O. Chum, M. Perdoch, and J. Matas, "Geometric min-hashing: Finding a (thick) needle in a haystack," in *Proc. of IEEE CVPR*, Miami, FL, USA, 2009, pp. 17–24.
[29] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Proc. of IEEE CVPR*, Miami, USA, 2009, pp. 25–32.
[30] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. of CVPR*, vol. 2, New York, NY, USA, 2006, pp. 2169–2178.
[31] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. of CVPR*, vol. 2, Madison, WI, USA, 2003, pp. 264–271.
[32] G. Carneiro and A. Jepson, "Flexible spatial models for grouping local image features," in *Proc. of IEEE CVPR*, vol. 2, Washington, DC, USA, 2004, pp. 747–754.
[33] O. Chum and J. Matas, "Unsupervised discovery of co-occurrence in sparse high dimensional data," in *Proc. of CVPR*, San Francisco, CA, USA, 2010, pp. 3416–3423.
[34] C. Pantofaru, G. Dorko, C. Schmid, and M. Hebert, "Combining regions and patches for object class localization," in *Proc. of IEEE CVPR Beyond Patches Workshop*, New York, NY, USA, 2006, pp. 1–8.
[35] E. Mortensen, H. Deng, and L. Shapiro, "A SIFT descriptor with global context," in *Proc. of IEEE CVPR*, San Diego, CA, USA, 2005, pp. 184–190.
[36] A. M. Bronstein and M. M. Bronstein, "Spatially-sensitive affine-invariant image descriptors," in *Proc. of ECCV*, vol. 2, Crete, Greece, 2010, pp. 197–208.
[37] K. Gao, S. Lin, Y. Zhang, S. Tang, and D. Zhang, "Logo detection based on spatial-spectral saliency and partial spatial context," in *Proc. of IEEE ICME*, New York, NY, USA, 2009, pp. 322–329.
[38] J. Kleban, X. Xie, and W.-Y. Ma, "Spatial pyramid mining for logo detection in natural scenes," in *Proc. of IEEE ICME*, Hannover, Germany, 2008, pp. 1077–1080.
[39] T. Quack, V. Ferrari, B. Leibe, and L. Van Gool, "Efficient mining of frequent and distinctive feature configurations," in *Proc. of ICCV*, Rio de Janeiro, Brazil, 2007, pp. 1–8.
[40] H. Sahbi, J.-Y. Audibert, J. Rabarisoa, and R. Kerivan, "Context-dependent kernel design for object matching and recognition," in *Proc. of IEEE CVPR*, Anchorage, AK, USA, 2008, pp. 1–8.
[41] H. Sahbi, J.-Y. Audibert, and R. Kerivan, "Context-dependent kernels for object classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 699–708, 2011.
[42] Y. Kalantidis, L. G. Pueyo, M. Trevisiol, R. van Zwol, and Y. Avrithis, "Scalable Triangulation-based Logo Recognition," in *Proc. of ACM ICMR*, Trento, Italy, 2011, pp. 1–7.

Hichem Sahbi received his MSc degree in theoretical computer science from the University of Paris Sud, Orsay, France, and his PhD in computer vision and machine learning from INRIA/Versailles University, France, in 1999 and 2003, respectively. From 2003 to 2006 he was a research associate first at the Fraunhofer Institute in Darmstadt, Germany, and then at the Machine Intelligence Laboratory at Cambridge University, UK. From 2006 to 2007, he was a senior research associate at the École des Ponts ParisTech, Paris, France. Since 2007, he has been a CNRS CR1 associate professor at Télécom ParisTech/ENST, Paris. His research interests include statistical machine learning, kernel and graph based inference, computer vision, and image retrieval.

Lamberto Ballan received the Laurea degree in computer engineering and the Ph.D. degree in computer engineering, multimedia and telecommunication from the University of Florence, Florence, Italy, in 2006 and 2011, respectively. He is currently a postdoctoral researcher in the Media Integration and Communication Center at the University of Florence. He was a visiting scholar at Télécom ParisTech/ENST, Paris, France, in 2010. His main research interests focus on multimedia information retrieval, image and video analysis, pattern recognition, and computer vision. His work was conducted in the context of several EU and national projects, and his results have led to around 25 publications in international journals and conferences, mainly in multimedia and image analysis. Dr. Ballan received the best paper award by the ACM-SIGMM Workshop on Social Media in 2010. He co-organized the 1st Int'l Workshop on Web-scale Vision and Social Media in conjunction with ECCV 2012.

Giuseppe Serra is a postdoctoral researcher at the Media Integration and Communication Center, University of Florence, Italy. He received the Laurea degree in computer engineering in 2006 and the Ph.D. degree in computer engineering, multimedia and telecommunication in 2010, both from the University of Florence. He was a visiting scholar at Carnegie Mellon University, Pittsburgh, PA, and at Télécom ParisTech/ENST, Paris, in 2006 and 2010 respectively. His research interests include image and video analysis, multimedia ontologies, image forensics, and multiple-view geometry. He has published more than 25 publications in scientific journals and international conferences. He has been awarded the best paper award by the ACM-SIGMM Workshop on Social Media in 2010.

Alberto Del Bimbo is a full professor of computer engineering at the University of Florence, Italy, where he is also the director of the Master in Multimedia, and the director of the Media Integration and Communication Center. His research interests include pattern recognition, multimedia information retrieval, computer vision, and human-computer interaction. He has published more than 300 publications in some of the most distinguished scientific journals and international conferences, and is the author of the monograph Visual Information Retrieval. Prof. Del Bimbo is an IAPR fellow and Associate Editor of Multimedia Tools and Applications, Pattern Analysis and Applications, Journal of Visual Languages and Computing, International Journal of Image and Video Processing, and International Journal of Multimedia Information Retrieval and was an Associate Editor of Pattern Recognition, IEEE Transactions on Multimedia, and IEEE Transactions on Pattern Analysis and Machine Intelligence. He was general co-chair of ACM Multimedia in 2010 and of the 12th European Conference on Computer Vision in 2012.