Context-dependent logo matching and retrieval

Recherche et localisation des logos

Hichem Sahbi
Lamberto Ballan
Giuseppe Serra
Alberto Del Bimbo

2010D009

mars 2010

Département Traitement du signal et des images
Groupe TII : Traitement et Interprétation des Images

Context-Dependent Logo Matching and Retrieval
Recherche et Localisation des Logos

Hichem Sahbi HICHEM.SAHBI@TELECOM-PARISTECH.FR
CNRS LTCI UMR 5141
Télécom ParisTech
Paris, France

Lamberto Ballan
Giuseppe Serra
Alberto Del Bimbo {BALLAN,SERRA,DELBIMBO}@DSI.UNIFI.IT
Media Integration and Communication Lab
University of Florence,
Florence, Italy

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 large scale images. 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 "Trademark-720" logo database overtaking, by 20%, baseline as well as standard matching/recognition procedures; furthermore, our method is able to process images of 1500×1500 pixels and checks for the existence of 13 reference logos in less than 1(s) using a standard 2 GHz PC.

Résumé

On introduit dans cet article, une nouvelle approche d'appariement et détection des logos basée sur une classe de fonctions de similarités dites dépendantes du contexte (CDS). Cette approche permet de construire ces fonctions de similarités, impliquant des points d'intérêts, en prenant en compte leurs propriétés intrinsèques visuelles ainsi que leurs contextes et configurations spatiales.

Les contributions de ce travail incluent : (i) une approche variationnelle permettant de construire CDS comme étant le point fixe d'une énergie comportant un terme "d'attache aux données", un critère de "contexte" et un terme de "régularisation" (ii) ainsi qu'une étude théorique de la consistance du processus d'appariement/détection des logos et ses propriétés d'invariance aux différentes transformations notamment la similitude et l'occultation. Enfin, la validité de la méthode est montrée à travers une évaluation sur des images réelles de logos.

Keywords: Context Dependent Similarity, Logo Detection and Localization, Image Retrieval.

1. Introduction

Automatic image and video annotation has received an increasing attention from the research and industrial community in the recent years (Datta et al., 2008). This is mainly due to the growing request for content based search and retrieval of interesting visual elements, resulting from the exponential growth of multimedia sharing systems such as Flickr and YouTube. In particular, a really challenging task is the detection and recognition of advertising trademarks/logos, which are of great interest for several real world applications. In fact, logos are key elements for companies[1] and play essential role in industry and commerce; they also recall the expectations associated with a particular product or service.

The early work on trademark detection and recognition addressed the problem of assisting the registration process. Since a trademark has to be formally registered, the idea of these approaches is to compare a newly designed trademark with archives of already registered ones, in order to ensure that it is sufficiently distinctive and avoid confusion (Kim and Kim, 1997; Schietse et al., 2007). Historically, the earliest approach was Kato's Trademark system (Kato, 1992). Its idea is to map normalized trademark images to an 8×8 pixel grid, and calculate a *GF-vector* for each image from frequency distributions of black and edge pixels appearing in each cell of the grid. Matching between logos was performed by comparing the GF-vectors. An other notable system was Artisan (Eakins et al., 1998) that achieves trademark retrieval using shape similarity. In this approach Gestalt principles were used in order to derive rules allowing individual image components to be grouped into perceptually significant parts. More recently, Wei et al. (2009) proposed a system that combines global Zernike moments and local curvature and distance to centroid features in order to describe logos. All these methods use synthetic images and rely on global logo descriptions, usually related to their contours or to particular shape descriptors (such as *shape context*) (Belongie et al., 2002; Rodriguez et al., 2008), so they require logos to be fully visible.

In the last years, other work on logo detection and recognition, in real world images/videos, has emerged and is targeted to automatically identify products (such as groceries in stores for assisting the blind) (Merler et al., 2007; Jing and Baluja, 2008) or to verify the visibility of advertising trademarks (e.g. billboards or banners) in sports events (Bagdanov et al., 2007; Watve and Sural, 2008). This problem is much harder, due to the relatively low resolution and quality of images (e.g. compression artifacts, color subsampling, motion blur, etc.) and also to the fact that trademarks are often small and may contain few information. Moreover their appearance is often characterized by occlusions, perspective transformations and deformations (see the examples in Fig. 1). Interest points and local descriptors have been successfully used in order to describe logos and obtain flexible matching techniques that are robust to partial occlusions as well as liner and non linear transformations. In Bagdanov et al. (2007), the authors provided a good evidence of

1. Companies are for instance interested in getting statistics about their logos in social media.

Figure 1: Realistic examples of trademark images characterized respectively by a bad light condition (Coca-Cola), occlusions (McDonald's), a deformation (Starbucks) and a small size (Ferrari).

trademark detection and localization in sport video; in their approach, each trademark is described as a bag of SIFT points (Lowe, 2004) which are classified and matched with the bags of SIFT features in video frames; localization is performed through robust clustering of matched SIFT features. Following the same approach, Joly and Buisson (2009) exploit SIFT point representation in order to detect logos in natural images. In order to refine their detection results, they also include geometric consistency constraints by estimating affine transformations between queries and retrieved images. Furthermore, they use a contrario adaptive thresholding in order to improve the accuracy of visual query expansion. From a more general point of view, Sivic and Zisserman (2003, 2008) proposed a text retrieval approach to object matching in videos, called "Video Google". They applied the traditional bag-of-words model in the visual domain by generating a codebook of affine covariant features, represented as SIFT descriptors, and using the *tf-idf* weighting scheme for indexing and retrieval. They also proposed a spacial consistency test by analyzing the 15 nearest neighbors of each match in order to improve scores of local features that share a similar neighborhood structure. The approach is proved to be effective in recognizing several kinds of objects (including logos) for very large scale retrieval tasks, but like other bag-of-visual-words approaches, it often suffers from poor recall. Improvements on this kind of approaches have been recently obtained by Chum et al. (2009), by introducing geometric hashing, and Wu et al. (2009) for the task of large-scale partial duplicate detection in web images.

Furthermore, few other interesting work includes spatial informations into object or logo representations in order to improve the detection performances. First, Carneiro and Jepson (2004) introduced the idea of grouping local image features in flexible spatial models to improve matching accuracy between images. Pantofaru et al. (2006) defined region-based context features (RCF) by combining image regions - obtained through image segmentation - with local patches such as SIFT descriptors. Similarly, Mortensen et al. (2005) modelled a global context by integrating Shape Context with SIFT local descriptors. Quack et al. (2006, 2007) then introduced the idea of employing association rules that capture frequent spatial configuration of quantized SIFT features at multiple resolutions, for object categorization and retrieval. These configurations are indexed in order to retrieve representative training templates for matching, nevertheless image resolution is a major limitation. This idea was followed later also by Kleban et al. (2008) for the specific case of logo detection in natural images. Gao et al. (2009) presented a two-stage logo detection algorithm which also

achieves localization by adapting a spatial-spectral saliency in order to improve the matching precision. They proposed a spatial context descriptor in order to estimate the spatial distribution of the set of matching points. In particular, they find minimum boundary round of matched points and partition it into nine areas. Finally, they describe the distribution of these points using a nine-dimensional histogram. However, this global logo representation is sensitive to occlusion.

Following the above discussed work, logo detection algorithms, based on interest points, are known to be very effective and also flexible in order to handle invariance (including occlusion and affine transformations). Nevertheless, their success strongly depends on the quality of matching (also referred to as alignment) mainly when images contain repeatable and redundant structures. On the one hand, a *naive* matching strategy, which given two images (a reference logo and a test image), looks for all pairs of interest point matches, using a context[2]-free similarity, such as the laplacian or the Gaussian, might result into many false matches. Figure (3, Top, Middle left) illustrates the deficiency of such naive approach when used between two groups of interest points; any slight perturbation of the values of the underlying features will result into unstable matching results *if no context* is taken into account. On the other hand, putting strong model assumptions about possible transformations (homography, affine, etc.) between reference logos and test images, might not capture the actual inter-logo transformations; for instance when logos deform.

In this paper, we introduce an alternative matching framework, for logo detection, based on a new class of similarity functions, called "context-dependent similarities" ("CDS") and defined as the fixed-point of an energy function which balances a "fidelity" term, a "context" criterion and an "entropy" term. The fidelity term is inversely proportional to the expectation of the Euclidean distance between the most likely aligned interest points while the context criterion measures the spatial coherence of the alignments, i.e., how good two interest points, with close geometric context, match. Given a pair of interest points (f_p, f_q) with a high alignment score (defined by our "CDS" values), 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, as a key smoothing factor, considers that without any a priori knowledge about the alignment scores between pairs of interest points, the joint probability distribution related to these scores should be as flat as possible so this term acts as a *regularizer* that controls the entropy of the conditional probability of matching, hence the uncertainty and decision thresholds; furthermore this term helps finding a direct analytic solution, otherwise, the variational problem will be difficult to solve. In a second major part of this work, we introduce a matching, detection and recognition procedure based on our similarity measure and we show, under the hypothesis of the existence of reference logos into test images, that the probability of success of this procedure is high, which is also corroborated through experiments. Moreover, we will show through this theoretical analysis, that our context dependent logo detection has more easy to set decision thresholds, than context free approaches. Finally, note also that the proposed alignment and logo detection method is model-free, i.e., it is not based on any a priori alignment model such as homography which might not capture the actual inter-logo transformations.

2. Given a set of interest points $\mathcal{X}$, the context of $x \in \mathcal{X}$ is defined as the set of points spatially close to x and with some particular geometrical constraints (see section 2.1 for a detailed and a formal definition of the context.)

The paper is organized as follows: we first discuss in Section 2, our energy function which makes it possible to design our context-dependent similarity, then we show in Section 3, the application of this similarity in order to align interest points and perform logo detection. We will also show some theoretical properties about our alignment procedure mainly its probability of success even in challenging conditions such as presence of partial occlusion and its rotation, scale and translation invariance. In Section 4, we show logo detection results and comparison on challenging logo images, and we conclude in Section 5 while providing possible extensions for a future work.

2. Context-Dependent Similarity

Let $\mathcal{S}_p = \{x_1^p, \ldots, x_n^p\}$, $\mathcal{S}_q = \{x_1^q, \ldots, x_m^q\}$ be the list of interest points taken respectively from a reference logo and a test image ($n \ll m$ and the value of n, m may vary with the objects p, q).

2.1 Context

In order to take into account contextual information, an interest point x is formally defined as $x = (\psi_g(x), \psi_f(x), \psi_o(x), \omega(x))$ where the symbol $\psi_g(x) \in \mathbb{R}^2$ stands for the $2D$ coordinates of x while $\psi_f(x) \in \mathbb{R}^s$ corresponds to the feature of x (in practice the 128 coefficients of the SIFT; (Lowe, 2004)). We have an extra information about the orientation of x (denoted $\psi_o(x) \in [-\pi, +\pi]$) which is provided by the SIFT gradient. Finally, we use $\omega(x)$ to denote the object 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 not surprising since we want to take into account the context of the interest point in the image it belongs to).

Let $d(x, x') = \|\psi_f(x) - \psi_f(x')\|_2$ measure the dissimilarity between two interest point features, $\|\cdot\|_2$ is the "entrywise" L_2-norm (i.e., the sum of the square values of vector coefficients). Introduce the context of x

$$\mathcal{N}^{\theta,\rho}(x) = \{x' : \omega(x') = \omega(x), x' \neq x \text{ s.t. (i) and (ii) hold}\},$$

with

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

and

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

Here ϵ_p is the radius of a neighborhood disk surrounding x and $\theta = 1, ..., N_a$, $\rho = 1, ..., N_r$ correspond to indices of different parts of that disk (see Fig. 2). In practice, N_a and N_r correspond to 8 sectors and 8 bands. The definition of neighborhoods $\{\mathcal{N}^{\theta,\rho}(x)\}_{\theta,\rho}$ reflects the co-occurrence of different interest points with particular spatial geometric constraints (see again Fig. 2).

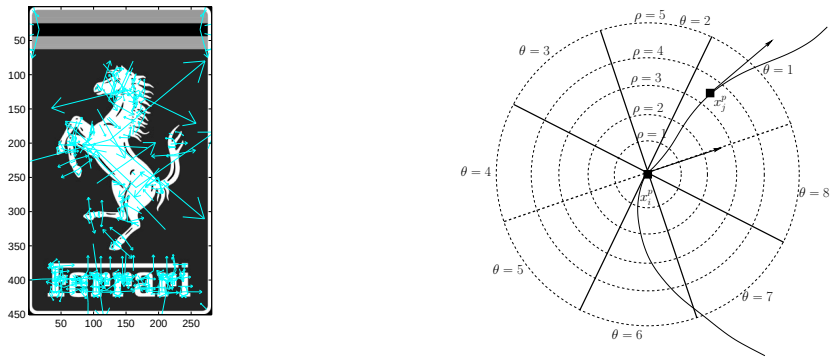

Figure 2: This figure shows a collection of SIFT interest points (with their locations, orientations and scales) (left) and the partitioning of the context (also referred to as neighborhood) of an interest point into different sectors for orientations and bands for locations (right).

2.2 Similarity Design

The set $\mathcal{X}$ of all possible interest points is the union over all possible objects $\mathcal{S}_p$, $\mathcal{S}_q$:

$$\mathcal{X} = \Big\{ \cup_p \mathcal{S}_p \Big\} \cup \Big\{ \cup_q \mathcal{S}_q \Big\}.$$

We consider $k : \mathcal{S}_p \times \mathcal{S}_q \rightarrow \mathbb{R}$ as a function which, given two interest points (x_i^p, x_j^q), provides a similarity measure between them.

For a finite collection of interest points, the sets $\mathcal{S}_p$, $\mathcal{S}_q$ are finite. Provided that we put some (arbitrary) order on $\mathcal{S}_p$, $\mathcal{S}_q$, we can view function k as a matrix $\mathbf{K}$ in which the "$(x, x')-$element" is the similarity between x and x': $\mathbf{K}_{x,x'} = k(x, x')$. Let $\mathbf{P}_{\theta,\rho}$ be the intrinsic adjacency matrices respectively defined as $\mathbf{P}_{\theta,\rho,x,x'} = g_{\theta,\rho}(x, x')$, where g is a decreasing function of any (pseudo) distance involving (x, x'), *not necessarily symmetric*. In practice, we consider $g_{\theta,\rho}(x, x') = \mathbb{1}_{\{\omega(x)=\omega(x')\}} \times \mathbb{1}_{\{x' \in \mathcal{N}^{\theta,\rho}(x)\}}$. Let $\mathbf{D}_{x,x'} = d(x, x')$. We propose to use the function on $\mathcal{S}_p \times \mathcal{S}_q$ defined by solving

$$\min_{\substack{\mathbf{K} \geq 0, \\ \|\mathbf{K}\|_1 = 1}} \quad \mathrm{Tr}\big(\mathbf{K}\ \mathbf{D}^{'}\big) \ + \ \beta\ \mathrm{Tr}\big(\mathbf{K}\ \log\ \mathbf{K}^{'}\big) \ - \ \alpha \sum_{\theta,\rho} \mathrm{Tr}\big(\mathbf{K}\ \mathbf{P}_{\theta,\rho}\ \mathbf{K}'\ \mathbf{P}'_{\theta,\rho}\big) \qquad (1)$$

Here $\alpha, \beta \geq 0$ and the operations log and $\geq$ are applied individually to every entry of the matrix (for instance, $\log \mathbf{K}$ is the matrix with $(\log \mathbf{K})_{x,x'} = \log k(x, x')$), $\| \cdot \|_1$ 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 two features $\psi_f(x)$, $\psi_f(x')$. In the case of SIFT, this is considered as the distance, $d(x, x')$, between the 128 SIFT coefficients of x and x'. A high value of $d(x, x')$ should result into a small value of $k(x, x')$ 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 $\{k(x,x')\}$ 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 $k(x,x')$ should imply high values in the neighborhoods $\mathcal{N}^{\theta,\rho}(x)$ and $\mathcal{N}^{\theta,\rho}(x')$. This criterion also makes it possible to consider the spatial configuration of the neighborhood of each interest point in the matching process. We formulate the minimization problem by adding an equality constraint and bounds which ensure a normalization of the similarity values and allow to see $\{k(x,x')\}$ as a probability distribution on $\mathcal{S}_p \times \mathcal{S}_q$.

2.3 Solution

The above stated minimization problem admits one solution under some constraints

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

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

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*

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

with

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

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 the proof is omitted and may be found in Sahbi et al. (2009). ■

By taking not too large β, 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: 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 $\alpha = 0$.

To have partitioned the neighborhood into several cells corresponding to different degrees of proximity (as shown in Fig. 2) has lead to significant improvements of our experimental results. On the one hand, the constraint (2) becomes easier to satisfy, for larger α with partitioned neighborhood, compared to Sahbi et al. (2008). 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 (Sahbi et al., 2008) 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, i.e., last term in (1) (see also equation (9) and discussion in Section 3.1). Finally, notice that at the convergence stage, we omit t in all $\mathbf{K}^{(t)}$ so the latter will simply be denoted as $\mathbf{K}$.

3. Logo Detection and Consistency

Let X, Y be two random variables standing respectively for interest points in $\mathcal{S}_p$, $\mathcal{S}_q$, and $\{X_1, \ldots, X_n\}$ (resp. $\{Y_1, \ldots, Y_m\}$) as n (resp. m) realizations with the same distribution as X (resp. Y). Define also H_1 (resp. H_0) as the set of all possible matching points (resp. non matching points) taken from $\{\mathcal{S}_p\} \times \{\mathcal{S}_q\}$ according to a well defined ground truth.

3.1 Matching

Given X, a good matching strategy consists in declaring Y_J as a match iff the conditional probability on (X, Y_J) is larger than the sum of the conditional probabilities on $\{(X, Y_j),\ j \neq J\}$; leading to

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

being $\mathbf{K}_{Y|X} = \mathbf{K}_{X,Y} / \big(\sum_{j=1}^{m} \mathbf{K}_{X,Y_j} \big)$; the intuition behind the above criterion 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 given X, the uncertainty about its possible matches will be reduced.
Considering (7), we define its probability of success

$$p_s = P\bigg(\mathbf{K}_{Y_J|X} \ > \ \sum_{j \neq J}^{m} \mathbf{K}_{Y_j|X} \bigg), \tag{8}$$

this probability is with respect to $\{X, X_1, \ldots, X_n\}$, $\{Y_1, \ldots, Y_m\}$. In the remainder of this section, we will discuss the consistency of the matching criterion (7), mainly its probability of success under H_1 and H_0.

Proposition 2 *fix $Q = N_r N_a$ (see (i),(ii)) and consider X. Under the hypothesis of existence of a reference logo into a test image (i.e. $\exists Y_J : (X, Y_J) \in H_1$), the probability of*

success p_s (in 8 also denoted P_1 under H_1) is at least

$$\frac{\exp\big(n(1-1/Q)\big)}{\exp\big(n(1-1/Q)\big)+(m-1)}, \tag{9}$$

and if $Q>1$, p_s will be exponentially-convergent to 1 with respect to n and decreasing with respect to m. Whereas, under the hypothesis of non existence of a reference logo into a test image (i.e. $\nexists Y_J : (X,Y_J)\in H_1$), the probability of success p_s (also denoted P_0 under H_0) is $1/m \underset{m\to+\infty}{\longrightarrow} 0$.

Proof see appendix. ■

It results from the above proposition, that under H_1 (in contrast to H_0), p_s is an increasing function of n, Q and a decreasing function of m. For instance, if $m = 10.000$, $Q = 64$, then p_s reaches 1 with only $n \geq 20$ sample points in the reference logo. Clearly, this shows that the procedure is able to correctly match very few interest points (in $\mathcal{S}_p$) into a very large collection (in $\mathcal{S}_q$), with small uncertainty (i.e., $P_1 \gg P_0$), so the underlying detection thresholds are easy to set as also corroborated through the next section and detection results.

3.2 Logo Detection

Given a test image $\mathcal{S}_q$ and a reference logo $\mathcal{S}_p$, the latter is declared as present into $\mathcal{S}_q$ *if the number of times the inequality (7) is satisfied is larger than τn ($\tau \in]0,1]$)*; here $(1-\tau)$ is the amount of occlusion that $\mathcal{S}_p$ might have in $\mathcal{S}_q$ while still can be detected[3]. Let $\mathcal{X}_s$ ($\mathcal{X}_s \to \mathcal{B}(n,p_s)$) be a binomial random variable standing for the number of times good matches are found in $\mathcal{S}_q$, for the n points in $\mathcal{S}_p$, using (7). In this section, we are interested in lower bounding

$$P\big(\mathcal{X}_s \geq \tau n\big), \quad \tau \in]0,1], \tag{10}$$

here P is the probability distribution of $\mathcal{X}_s$. Now, we provide our main result which allows us under some conditions to lower bound (10)

Proposition 3 *fix τ and consider $\mathcal{X}_s$ as a binomial random variable with parameter p_s. If p_s ($\in [0,1]$) is at least $\sqrt{-\frac{\ln(\delta/2)}{2n}}+\tau$, then*

$$P\big(\mathcal{X}_s \geq \tau n\big) \geq 1-\delta \tag{11}$$

here $\delta \ll 1$ is a fixed error rate.

Proof the left-hand side of the above inequality is equal to

$$\begin{aligned} & P\left(\sum_i^n Z_i \geq \tau n\right), \text{ here } \mathcal{X}_s = \sum_i^n Z_i,\ Z_i \to \mathcal{B}(1,p_s) \\ = \ & 1 - P\left(p_s - \frac{1}{n}\sum_i^n Z_i \geq p_s - \tau\right) \\ \geq \ & 1 - 2\exp\Big(-2n(p_s-\tau)^2\Big), \quad \text{(by Hoeffding's inequality)} \end{aligned} \tag{12}$$

3. *It reasonable to set $\tau = 0.5$, which means that a reference logo is still detectable event-though half-occluded in a test image.*

the sufficient condition is

$$2\exp\Big(-2n(p_s-\tau)^2\Big) \leq \delta \Rightarrow p_s \geq \sqrt{-\frac{\ln(\delta/2)}{2n}} + \tau, \tag{13}$$

when $n \to +\infty$, and if p_s is at least equal to τ, then $P\big(\mathcal{X}_s \geq \tau n\big) \underset{n\to+\infty}{\longrightarrow} 1$. ■

Now combining (9) and (13), the sufficient condition which guarantees (11) becomes under H_1

$$\frac{\exp\big(n(1-1/Q)\big)}{\exp\big(n(1-1/Q)\big)+(m-1)} \;\geq\; \sqrt{-\frac{\ln(\delta/2)}{2n}} + \tau, \tag{14}$$

which holds true mainly for larger n, Q, but $\tau < 1$, and even large m. For instance if $n = 20$, $m = 10.000$, $Q = 64$, the left hand side is very close to 1 and hence the inequality (14) will be satisfied even when $\tau \to 1$ (low occlusion factor) and $\delta \to 0$ (high lower bound).

3.3 Invariance Properties

The adjacency matrices $\mathbf{P}_{\theta,\rho}$, in $\mathbf{K}$, provide the intrinsic properties and also characterize the geometry of logos $\{\mathcal{S}_p\}$ in $\mathcal{X}$. It is easy to see that $\mathbf{P}_{\theta,\rho}$ is translation and rotation invariant and can also be made scale invariant when ϵ_p (see (i)) is adapted to the scales of $\psi_g(\mathcal{S}_p)$. 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 $\mathbf{K}^{(t)}$ may involve similarity invariant features $\psi_f(.)$ (actually SIFT features), so $\mathbf{K}^{(t)}$ (and also the matching process) is similarity invariant.

4. Benchmarking

4.1 Test Data and Settings

In order to show the extra-value of our context dependent matching strategy (i.e., based on "CDS") with respect to context free one and other approaches, we evaluate the performances of multiple-logo detection on a novel challenging dataset called TradeMark-720, containing 13 trademark classes each one represented with $14-87$ real world pictures downloaded from the web, resulting into a collection of 720 images. 13 reference logos are used and correspond to trademarks: 1: "Agip", 2: "Apple", 3: "Barilla", 4: "Birra_Moretti", 5: "Cinzano", 6: "CocaCola", 7: "Esso", 8: "Ferrari", 9: "Heineken", 10: "Marlboro", 11: "McDonald", 12: "Pepsi", 13: "Starbucks". Note that each reference logo is synthetically transformed in order to generate 4 affine transformations. Interest points are extracted from test images as well as reference logos and encoded using the usual SIFT features.

Each test image $\mathcal{S}_q$ is processed in order to evaluate the similarity function $\mathbf{K}$ (shown in 4) with respect to each reference logo $\mathcal{S}_p$, using Gaussian power assist setting, i.e., $\mathbf{K}^{(0)}_{x,x'} = \exp(-d(x,x')/\beta)$. Our goal is to show the improvement brought when using $\mathbf{K}^{(t)},\ t \in \mathbb{N}^+$, so we compared it first against the standard context-free similarity (i.e., $\mathbf{K}^{(t)},\ t = 0$), then with respect to standard matching techniques including RANSAC. First, the setting of β is performed by maximizing the performance of the Gaussian similarity as the latter

Thresholds (τ)	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
Errors	FRR/FAR	FRR/FAR	FRR/FAR	FRR/FAR	FRR/FAR	FRR/FAR	FRR/FAR	FRR/FAR	FRR/FAR	FRR/FAR
Sift Matching	.29/.31	.50/.14	.63/.09	.73/.05	.77/.03	.82/.02	.85/.02	.89/.01	.92/.01	.95/.01
Ransac Matching	.39/.08	.52/.03	.62/.01	.70/.01	.75/.01	.80/0.01	.82/.01	.84/.01	.86/.01	.87/.01
"CDS" Matching	**.09/.27**	**.10/.22**	**.11/.20**	**.12/.18**	**.12/.18**	**.12/.17**	**.13/.17**	**.13/.17**	**.13/.16**	**.13/.16**

Table 1: This table shows a comparison of our method, with respect to Sift and Ransac matching; in all these experiments we clearly see that the global error rates (defined as $\frac{1}{2}(FAR + ERR)$) of our method are better than those reported for standard matching techniques. Notice also that FAR is an increasing function of the occlusion factor $(1 - \tau)$ while FRR is a decreasing function.

corresponds to the left-hand side (and the baseline form) of $\mathbf{K}^{(t)}$, i.e., when $\alpha = 0$.[4] For our database, we found that the best performances are achieved for $\beta = 0.1$ and this also guarantees condition (3) in practice. 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 (2) is satisfied. Intuitively, the weight parameter α should then be relatively high while also satisfying condition (2). We found that the best α is 0.1.

4.2 Performance, Comparison and Discussion

We used criteria (7), (10) in order to decide whether a given reference logo $\mathcal{S}_p$ exists into a test image $\mathcal{S}_q$. Different values were experimented for the tolerance factor τ and performances are measured using False Acceptance (FAR) and False Rejection Rates (FRR) defined as

$$\text{FAR} = \mathbb{E}\left(\frac{\text{false positive}}{\text{false positive+ true negative}}\right),\ \text{FRR} = \mathbb{E}\left(\frac{\text{false negative}}{\text{false negative + true positive}}\right),$$

here the expectation is with respect to all possible test images. Diagrams in (3, Bottom), show the FAR, FRR errors for different classes (trademarks) of our test set; we clearly see the out-performance and the improvement of the our context dependent similarity function (i.e., $\mathbf{K}^{(t)}, t \in \mathbb{N}^+$), in logo detection, with respect to the baseline, i.e., context-free similarity ($\mathbf{K}^{(0)}$). For almost all the classes of the test set, the improvement brought by the "CDS" similarity is clear and consistent; except the classes "Apple" and "McDonald" as their reference logos contain very few interest points ($n < 12$), and this makes (*consistently with our theoretical analysis*) inequality (14) difficult to satisfy mainly for high expectations about the lower bound in (11) (i.e., low δ) and when τ is relatively high.

Table 1 shows a comparison of our context dependent similarity for logo matching and detection with respect to other techniques including SIFT matching and also with respect to (iterative) Ransac matching using the inliers/outliers, obtained by estimating the affine transformation between SIFT correspondences. In both cases, a match is declared if Lowe's second nearest neighbor test is satisfied. In particular, the SIFT matching technique follows the approach presented in Bagdanov et al. (2007) where a logo is detected if the overall

4. Notice that selecting β independently from α is obviously *"not sub-optimal"* for the context dependent similarity but *"sub-optimal"* for the Gaussian similarity.

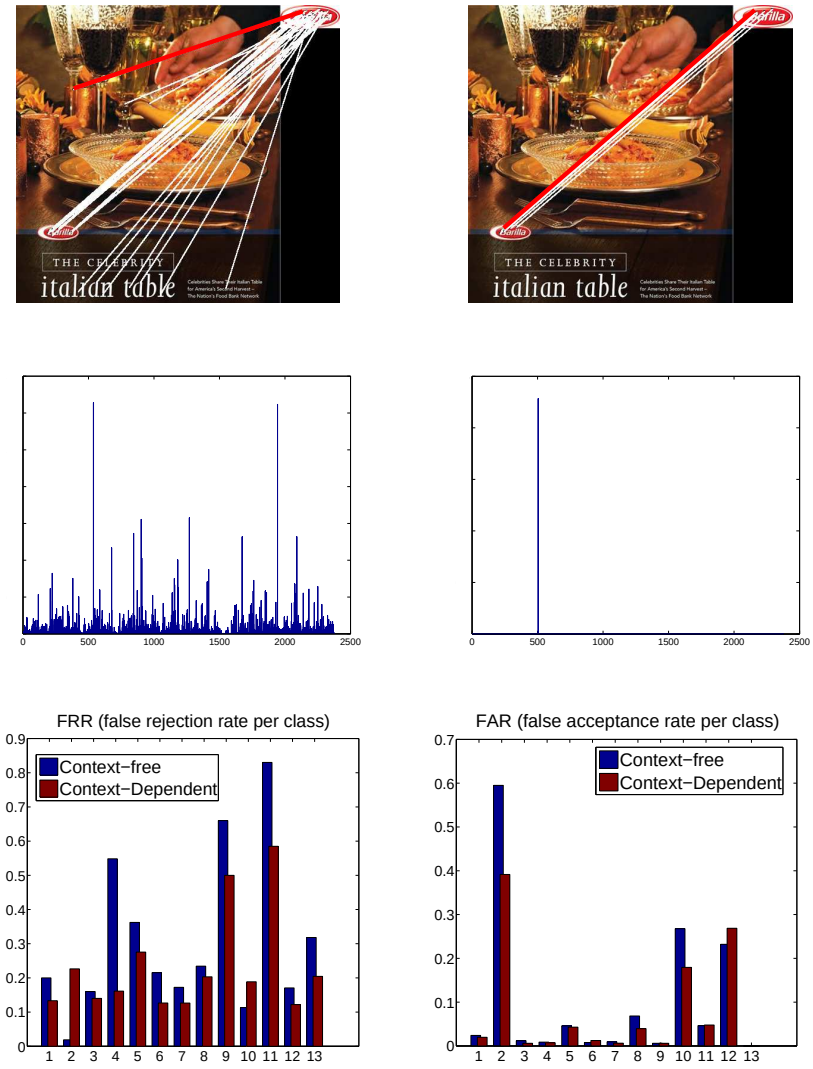

Figure 3: (Top) This figure shows a comparison of the matching results when using a naive matching strategy without context and our context dependent matching (i.e., based on "CDS"). (Middle) Figures show the conditional probability distribution $K_{.|X}$ for a particular interest point X in the reference logo. This distribution is peaked when using context dependent similarity (Middle, right) so the underlying entropy is close to 0 and the uncertainty about possible matches is dramatically reduced. (Bottom) 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.

Figure 4: These pictures show logo detection results; for every test images, all the 13 reference logos were checked using criterion (10). Match points are also displayed.

number of SIFT matches is above a fixed (trained) threshold. The Ransac approach follows the same idea but it introduce a model-based (not necessarily always consistent) geometric consistency test by selecting only matches that satisfy the affine transformation between the query and the retrieved images (see Joly and Buisson (2009)). Even though, the FAR and FRR results are variable depending on the setting of τ, in all these cases, the average error rates, defined as (FAR+FRR)/2, of our method are lower than those reported for SIFT matching and Ransac. Finally, our method is able to process images of 1500×1500 pixels and checks for the existence of 13 reference logos in less than 1(s) using a standard 2 GHz PC (see results in Figs. 4, 5, 6).

5. 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 information about the spatial configuration in similarity design as well as visual features, (ii) the ability to control the regularization of the solution via our energy function, (iii) the invariance to many transformations including

Figure 5: Other logo detection results.

Figure 6: Further logo detection results.

translation, scale, rotation and also partial occlusion, 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 while very low under background.
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. Other extensions include searching duplicate objects/images in the web and extensions to other objects including driving signs.

Appendix

Proof [Proposition 2] Let $\mathcal{N}^{\theta,\rho}(X)$, $\mathcal{N}^{\theta,\rho}(Y)$ be two random variables standing for the number of interest points falling inside the context cell (θ,ρ) of respectively a reference logo and a test image. Here $\mathcal{N}^{\theta,\rho}(X) \rightarrow \mathcal{B}(n, 1/Q)$, $\mathcal{N}^{\theta,\rho}(Y) \rightarrow \mathcal{B}(m, 1/Q)$ and Q is the number of cells in the context ($Q = N_a \times N_r$, in practice $Q = 64$). Following the definition of our fixed point $\mathbf{K}_{X,Y}$ in (4), we have

$$\mathbf{K}_{Y|X} \quad \propto \quad \frac{1}{C} \exp\big(\mathcal{N}(X,Y)\big), \tag{15}$$

where $\mathcal{N}(X,Y)$, stands for the number of matching points in the context of X, Y

$$\mathcal{N}(X,Y) \;=\; \sum_{\theta,\rho}^{Q} \mathcal{N}^{\theta,\rho}(X)\, \mathcal{N}^{\theta,\rho}(Y). \tag{16}$$

Under $H_1 \longrightarrow \exists Y_J$ s.t. $(X, Y_J) \in H_1$

Since $\mathbf{K}_{Y_J|X} + \sum_{j\neq J}^{m} \mathbf{K}_{Y_j|X} = 1$, using (8), p_s, $q_s = 1 - p_s$ are respectively

$$\begin{aligned} & \mathbb{E}\big(\mathbf{K}_{Y_J|X} \big| (X, Y_J) \in H_1\big), \\ \text{and} \quad & \\ & \sum_{j\neq J}^{m} \mathbb{E}\big(\mathbf{K}_{Y_j|X} \big| (X, Y_j) \in H_0\big), \end{aligned} \tag{17}$$

here the expectation $\mathbb{E}$ is with respect to $\{X, X_1, \ldots, X_n\}$, $\{Y_1, \ldots, Y_m\}$. Now, combining (15), (17), p_s and q_s are at least

$$\begin{aligned} & \frac{1}{C} \exp\Big(\mathbb{E}_{H_1}\big(\mathcal{N}(X,Y)\big)\Big) \\ \text{and} \quad & \\ & \frac{1}{C}(m-1) \exp\Big(\mathbb{E}_{H_0}\big(\mathcal{N}(X,Y)\big)\Big), \quad \text{(by Jensen's inequality)} \end{aligned} \tag{18}$$

$\mathbb{E}_{\mathbb{H}_1}$ (resp. $\mathbb{E}_{\mathbb{H}}$) denotes the expectation w.r.t data in H_1 (resp. H_0) equal to

$$\begin{aligned}
\mathbb{E}_{H_0}\big(\mathcal{N}(X,Y)\big) &= \mathbb{E}_{H_0}\left(\sum_{\theta,\rho}\mathcal{N}^{\theta,\rho}(X)\mathcal{N}^{\theta,\rho}(Y)\right)\\
&= \sum_{\theta,\rho}\mathbb{E}_{H_0}\Big(\mathcal{N}^{\theta,\rho}(X)\mathcal{N}^{\theta,\rho}(Y)\Big)\\
&= \sum_{\theta,\rho}\mathbb{E}_{H_0}\left(\mathcal{N}^{\theta,\rho}(X)\right)\mathbb{E}_{H_0}\left(\mathcal{N}^{\theta,\rho}(Y)\right), \qquad (19)\\
&\quad \mathcal{N}^{\theta,\rho}(X),\mathcal{N}^{\theta,\rho}(Y) \overset{i.i.d}{\rightarrow} \mathcal{B}(n,1/Q)\\
&= n^2\ (1/Q)^2 Q\\
&= n^2/Q.
\end{aligned}$$

$$\mathbb{E}_{H_1}\big(\mathcal{N}(X,Y)\big) = \mathbb{E}_{H_1}\left(\sum_{\theta,\rho}\mathcal{N}^{\theta,\rho}(X)\mathcal{N}^{\theta,\rho}(Y)\right). \qquad (20)$$

Under H_1, $\mathcal{N}^{\theta,\rho}(X) \simeq \mathcal{N}^{\theta,\rho}(Y)$ and

$$\begin{aligned}
\mathbb{E}_{H_1}\big(\mathcal{N}(X,Y)\big) &\simeq \mathbb{E}_{H_1}\left(\sum_{\theta,\rho}\mathcal{N}^{\theta,\rho}(X)^2\right)\\
&= \sum_{\theta,\rho}\mathbb{E}_{H_1}\Big(\mathcal{N}^{\theta,\rho}(X)^2\Big)\\
&= \sum_{\theta,\rho}\mathbb{E}_{H_1}\left((\sum_{i}^{n} Z_{\theta,\rho,i})^2\right),\\
&\quad Z_{\theta,\rho,i} \rightarrow \mathcal{B}(1,1/Q) \qquad (21)\\
&= \sum_{\theta,\rho}\mathbb{E}_{H_1}\left(\sum_{i,j}^{n} Z_{\theta,\rho,i}Z_{\theta,\rho,j}\right)\\
&= \sum_{\theta,\rho}\mathbb{E}_{H_1}\left(\sum_{i}^{n} Z^2_{\theta,\rho,i}\right)\\
&\quad + \quad \mathbb{E}_{H_1}\left(\sum_{i,j\neq i}^{n} Z_{\theta,\rho,i}Z_{\theta,\rho,j}\right).
\end{aligned}$$

Since $Z_{\theta,\rho,i}, Z_{\theta,\rho,j} \overset{i.i.d}{\rightarrow} \mathcal{B}(1,1/Q)$

$$\begin{aligned}
\mathbb{E}_{H_1}\big(\mathcal{N}(X,Y)\big) &\simeq \sum_{\theta,\rho}\sum_{i}^{n}\mathbb{E}_{H_1}\left(Z^2_{\theta,\rho,i}\right)\\
&\quad + \sum_{i,j\neq i}^{n}\mathbb{E}_{H_1}\left(Z_{\theta,\rho,i}\right)\mathbb{E}_{H_1}\left(Z_{\theta,\rho,j}\right) \qquad (22)\\
&= Q\big(n/Q + n(n-1)(1/Q)^2\big)\\
&= n^2/Q + n(1-1/Q),
\end{aligned}$$

therefore,

$$\begin{aligned} p_s &\geq \frac{1}{C}\exp\big(n^2/Q + n(1-1/Q)\big), \\ q_s &\geq \frac{1}{C}(m-1)\exp\big(n^2/Q\big). \end{aligned} \tag{23}$$

Accordingly p_s is at least

$$\frac{\exp\big(n(1-1/Q)\big)}{\exp\big(n(1-1/Q)\big) + (m-1)} \tag{24}$$

Under $H_0 \longrightarrow \nexists\ Y_J$ s.t. $(X, Y_J) \in H_1$

Equations (17) are updated as

$$\begin{aligned} &\mathbb{E}\big(\mathbf{K}_{Y_J|X}\big|(X,Y_J) \in H_0\big), \\ \text{and} \\ &\sum_{j\neq J}^{m} \mathbb{E}\big(\mathbf{K}_{Y_j|X}\big|(X,Y_j) \in H_0\big) \end{aligned} \tag{25}$$

p_s is then $1/m$. ■

Acknowledgement

The research of the first author was supported by the French National Research Agency (ANR) under the AVEIR project.

References

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.

S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on PAMI*, 24(4):509–522, 2002.

G. Carneiro and A. Jepson. Flexible spatial models for grouping local image features. In *Proc. of IEEE CVPR*, Washington, DC, USA, 2004.

O. Chum, M. Perdoch, and J. Matas. Geometric min-hashing: Finding a (thick) needle in a haystack. In *Proc. of IEEE CVPR*, Miami, USA, 2009.

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008.

J. P. Eakins, J. M. Boardman, and M. E. Graham. Similarity retrieval of trademark images. *IEEE Multimedia*, 5(2):53–63, 1998.

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, USA, 2009.

Y. Jing and S. Baluja. Pagerank for product image search. In *Proc. of WWW*, Beijing, 2008.

A. Joly and O. Buisson. Logo retrieval with a contrario visual query expansion. In *Proc. of ACM Multimedia*, Beijing, China, 2009.

T. Kato. Database architecture for content-based image retrieval. *Proc. of SPIE Image Storage and Retrieval Systems,*, 1662:112–123, 1992.

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.

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.

D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

Michele Merler, Carolina Galleguillos, and Serge Belongie. Recognizing groceries in situ using in vitro training data. In *Proc. of IEEE CVPR SLAM-Workshop*, Minneapolis, MN, USA, June 2007.

E. Mortensen, H. Deng, and L. Shapiro. A SIFT descriptor with global context. In *Proc. of IEEE CVPR*, San Diego, CA, USA, 2005.

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, 2006.

Till Quack, Vittorio Ferrari, and Luc Van Gool. Video mining with frequent itemset configurations. In *Proc. of ACM CIVR*, Tempe, AZ, USA, 2006.

Till Quack, Vittorio Ferrari, Bastian Leibe, and Luc Van Gool. Efficient mining of frequent and distinctive feature configurations. In *Proc. of ICCV*, Rio de Janeiro, Brazil, 2007.

J.J. Rodriguez, P.M.Q. Aguiar, and J.M.F. Xavier. ANSIG - an analytic signature for permutation-invariant two-dimensional shape representation. In *Proc. of IEEE CVPR*, Nice, France, 2008.

H. Sahbi, J.Y. Audibert, J. Rabarisoa, and R. Keriven. Context dependent kernel design for object matching and recognition. In *Proc. of IEEE CVPR*, Anchorage, USA, 2008.

H. Sahbi, J-Y. Audibert, and R. Keriven. Incorporating context and geometry in kernel design. *technical report N 2009D002*, 2009.

J. Schietse, J. P. Eakins, and R. C. Veltkamp. Practice and challenges in trademark image retrieval. In *Proc. of ACM CIVR*, Amsterdam, NL, 2007.

J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. of ICCV*, Nice, France, 2003.

J. Sivic and A. Zisserman. Efficient visual search for objects in videos. *Proceedings of the IEEE*, 96(4):548–566, 2008.

Alok Watve and Shamik Sural. Soccer video processing for the detection of advertisement billboards. *Pattern Recognition Letters*, 29(7), 2008.

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*, 2009.

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.

Dépôt légal : 2010 – 1er trimestre
Imprimé à Télécom ParisTech – Paris
ISSN 0751-1345 ENST D (Paris) (France 1983-9999)

Télécom ParisTech

Institut TELECOM - membre de ParisTech

46, rue Barrault - 75634 Paris Cedex 13 - Tél. + 33 (0)1 45 81 77 77 - www.telecom-paristech.frfr

Département TSI