function is used as a normalization factor taken through all the interest points in S_X , S_Y.

The context is defined by the local spatial configuration of interest points in both S_X and S_Y. Formally, in order to take into account spatial information, an interest point $x_i \in S_X$ is defined as $x_i = (\varphi_g(x_i), \varphi_f(x_i), \varphi_o(x_i), \varphi_s(x_i), w(x_i))$, where the symbol $(\varphi_g(x_i) \in R^2$ stands for the 2D coordinates of x_i while $\varphi_f(x_i) \in R^c$ corresponds to the feature of x_i (in practice c is equal to 128, i.e. the coefficients of the SIFT descriptor). We have also an extra information about the orientation of x_i (denoted , $\varphi_o(x_i) \in [-\pi, +\pi]$ which is provided by the SIFT gradient and about the scale of the SIFT descriptor (denoted $\varphi_s(x_i)$). Finally, we use $w(x_i)$ 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 $d(x_i, y_i) = \| \varphi_f(x_i) - \varphi_f(y_j) \|_2$ 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:

$$N^{\theta,\rho}(x_i) = \{x_j : w(x_j) = w(x_i), x_j \neq x_i\} \quad (1)$$

with

$$\frac{\rho-1}{N_r}\varepsilon_p \leq \|\varphi_g(x_i) - \varphi_g(x_j)\|_2 \leq \frac{\rho}{N_r}\varepsilon_p \quad (2)$$

and

$$\frac{\theta-1}{N_a}\pi \leq \angle(\varphi_o(x_i), \varphi_g(x_j) - \varphi_g(x_j)) \leq \frac{\theta}{N_a}\pi \quad (3)$$

Where $(\varphi_g(x_j) - \varphi_g(x_i))$ is the vector between the two point coordinates $\varphi_g(x_j)$ and $\varphi_g(x_i)$. The radius of a neighborhood disk surrounding x_i is denoted as p and obtained by multiplying a constant value to the scale $s(x_i)$ of the interest point x_i . In the above definition, $\theta = 1 \ldots \ldots, N_a, \rho = 1 \ldots \ldots \ldots, N_r$ correspond to indices of different parts of that disk. In practice Na and Nr correspond to 8 sectors and 8 bands. The definition of neighborhoods $\{N^{\theta,\rho}(x_i)\}_{\theta,\rho}$ reflects the co-occurrence of different interest points with particular spatial geometric constraints. Fig. 2 shows an example taken from two different images containing the same logo; the figure reports the context definition for two corresponding keypoints, showing a similar spatial configuration. All the definitions about interest points in S_Y and their context are similar to S_X.

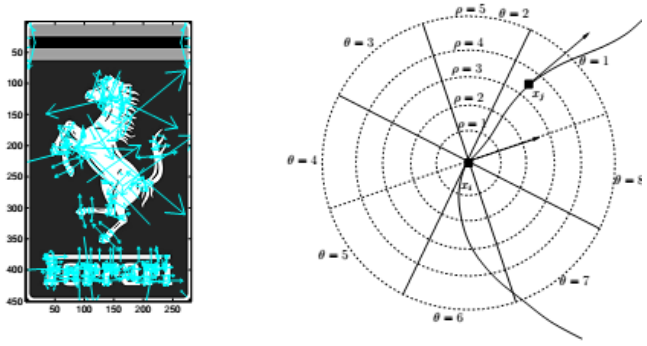

Figure 1: Collection of SIFT points

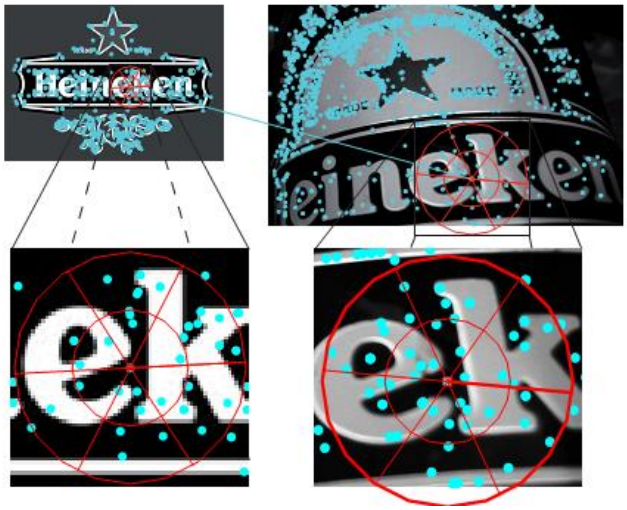

Figure 2: Example of real context

B. Similarity Design

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

$$\min_K Tr(K\,D') + \beta\, Tr(K \log K') - \alpha \sum_{\theta,\rho} Tr(K\, Q_{\theta,\rho}\, K'\, P'_{\theta,\rho}) \quad (4)$$

$$\text{s.t} \quad \begin{Bmatrix} K \geq 0 \\ \|K\|_1 = 1 \end{Bmatrix} \quad (5)$$

Here α,β ≥ 0 and the operations log (natural), ≥ are applied individually to every entry of the matrix (for instance, log K is the matrix with $(log\ K)_{x,y} = \log k(x,y)$, $\| \ . \|_1$ is the "entrywise" L1-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 f (x), 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 $D_{x,y}$ should result into a small value of $K_{x,y}$ 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 $\{K_{x,y} : x \in S_x, : y \in S_y\}$ 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. The third term is a neighborhood criterion which considers that a high value of $K_{x,y}$ should imply high values in the neighborhoods $N^{\theta,\rho}(x)$, $N^{\theta,\rho}(y)$ .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 K as a probability distribution.

4. Logo Detection

There are two inputs one is the reference logo which is the one to detect from the real world video. The second input is a real world test video which contains lot of images logos etc. The test input is a real world video that means taken without any isolation. A real world video or image consists of many limitations like background cluttering, change in illumination, less quality, low resolution, lighting effects, partial occlusion etc. due to this the logo detection become a really challenging task.

The two inputs are accepted by the program. The next step is to convert the video into frames. The real world video is converted to frames as per the given number of frames. This is for the ease of calculation. This converted frames as well as the reference image are taken to the next step called feature extraction and context dependent algorithm. With the help of this we get the probability of matching. In feature extraction the local features of both images are extracted. In the case of test video each frame is considered as one image and the features of each image is taken. With the help of feature extraction the keypoints and descriptors are extracted from each figure.
The extracted features are then used to calculate the context and the adjacency matrix is created with the help of context dependent similarity algorithm. According to this a probability of keypoint matching is calculated. If this value exceeds the threshold value then the logo is detected in that particular frame. This is repeated for each frame and calculated the number of frames containing the logo images.

5. Results

As we had described two inputs are there, one is an image and the second one is real world video.

Figure 3: input logo image
The above figure represents the input logo image. This image is matched with the each and every frame of the real world video that given as the second input.

Figure 4: matching with the frames that does not contain logo

Figure 5: matching with the frames that contains logo

The real world video frames may or may not contain logo image. There are frames that contain logo and does not contain logo. Each and every frame matched with the input and some of example output for both the frames contain and does not contain logo are shown above.

6. Conclusion

This context based similarity detection introduces a novel logo detection and localization approach based on a new class of similarities referred to as context dependent similarity. The companies use logos for conveying their messages about the product or service. Searching a particular logo from a group of images present in the real world video is extracted using a context dependent algorithm. The strength of this method resides in several aspects: the inclusion of the information about the spatial configuration in similarity design as well as visual features, the ability to control the influence of the

context and the regularization of the solution, 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.

References

[1] Hichem Sahbi, Lamberto Ballan, Giuseppe Serra, and Alberto Del Bimbo, ‘‘Context-Dependent Logo Matching and Recognition’’, 2013, pp. 1-13.
[2] 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.
[3] 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.
[4] 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
[5] R. Phan, J. Chia, and D. Androutsos, “Unconstrained logo and trademark retrieval in general color image database using color edge gradient cooccurence histograms,” in Proc. of IEEE ICASSP, Las Vegas, NV, USA 2008, pp. 1221–1224.
[6] 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.
[7] 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.
[8] 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.
[9] D. Lowe, “Distinctive image features from scale-invariant keypoints,” International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004
[10] 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.
[11] 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.