Homographic class template for logo localization and recognition

Raluca Boia *[1] and Corneliu Florea[1]

University Politehnica of Bucharest
rboia@imag.pub.ro; corneliu.florea@upb.ro

Abstract. We propose a method for localizing and recognizing brand logos in natural images. The task is extremely challenging, due to the various changes in the appearance of the logos. We construct class templates by matching features between examples of the same class to build homographies. An interconnections graph is developed for each class and the representative points are added to the class model. Finally, each class is depicted by the reunion of the suitable keypoints and descriptors, thus leading to a high precision of the proposed logo recognition system. Results show that we outperform the state of the art systems on the challenging Flickr-32 database.

Keywords: logo, localization, recognition, class model

1 Introduction

A logo is a graphic entity containing colors, shapes, textures, and perhaps text as well, organized in some spatial layout format. Logo localization and recognition is a subproblem of object detection and recognition and a challenging pattern recognition task. Being of interest for the marketing industry (e.g. to measure the impact of an advertising campaign), trademark registration or vehicle tracking, logo recognition has gained consistent attention in the last few years. Yet, the problem of integrated recognition (i.e. detection/localization + recognition) still remains unresolved. As the number of brands having personalized logos increases every day, such recognition systems require robust processing capabilities to support high numbers of classes.

The challenges of a logo detection system are due to perspective deformations, varying background, possible occlusions, scaling variability (from high resolutions of 1000×1000 to 20×20). Furthermore, although the objects are almost planar, there are situations when the pattern suffers from warping. Finally, the main difference to near-duplicate retrieval approaches is the high intra-class variability, as a certain brand logo can have variations in the used colors or even in shape. Examples that illustrate some of the mentioned issues are in figure 1.

* This work was supported by the Romanian Sectoral Operational Programme Human Resources Development 2007-2013 through the European Social Fund Financial Agreements POSDRU/159/1.5/S/132395 and POSDRU/159/1.5/S/134398.

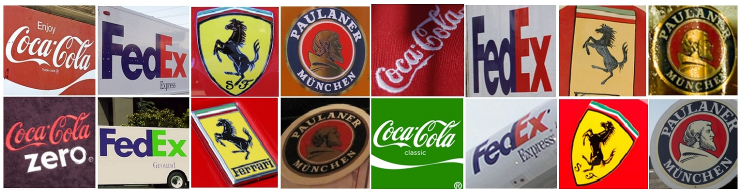


Fig. 1. Sample images from FlickrLogos-32 containing logos from the classes Coca Cola, FedEx, Ferrari and Paulaner. Note the variability in logo appearance or due to shadowing, color balance, warping, etc.

State of the art. The algorithms from the generic class of object recognition can be divided in two categories: generative [1], [2]. and discriminative [3], [4]. Discriminative techniques use the information concerning all the existing classes and train classifiers to distinguish between them. They are distressed by missing data and prior knowledge. Generative algorithms create object class models using, separately, the data of each class, being more suitable to high intra-class variation as is the case of logo recognition. The proposed method falls in the generative category.

To deal with extreme viewpoint changes, Schneiderman et al. [2] or Bernstein and Amit [5] used the aspect graphs for simulating the perspective point variation in mixture models, idea which is developed in the current work. Yet, they construct multiple models per class, while we use a single model.

Next, into the specific problem of the logo recognition, we note two main directions: general logo recognition and specific domain recognition such as vehicle logo. The first approaches [6], [7], [8] concerning generic logo recognition were limited in handling large image collections. Later methods [9] did recognize logos by performing frequent item-set mining to discover association rules in spatial pyramids of visual words. Revaud et al. [10] use a bag-of-words (BoW) based approach coupled with learned weights to penalize inter-class appearances, while Romberg et al. [11] enhanced the BoW system by embedding spatial knowledge into the cascaded index. Romberg and Lienhart, [12], extended the BoW by bundling on the min-hashing of SIFT-based visual words. However, most work done in this direction has the purpose of image retrieval, which is more permissive as compared to localization (aimed here), since the actual location is not reported. For vehicle logo detection, the problem of localization is handled [13], but on small databases with few classes.

Database For a realistic evaluation of the proposed method, we chose the Flickr Logos-32 database [11], which was formed by careful selecting images from collections of photos in a real word environment, depicting brand logos. The testing/ training scheme is the same as in the case of Romberg et al. [11]: 30 images per class for training and 30 images per class for testing phase for a total of 32 classes. For the training phase, we used only the crops of the logos in the images, while for the test part, we scan the entire images.

We chose FlickrLogos-32 over the BelgaLogos dataset [7], as the latter was originally used for logo retrieval rather than for classification and it only defines a small number of images per class with limited variability. Taking into account the average object size, when compared to other databases for object detection, FlickrLogos-32 can be considered a small-object dataset.

2 Class description by class model

The proposed method builds a class model by starting with SIFT features extraction from the training logo crops. The features are then matched and, using random sample consensus, a homography transform is found to stitch each 2 images of the same class in the training set. This pairing in fact builds a graph of the interconnections of images. The image with most links will represent its class and the entire model will be built on top of it. Using the graph and the homopraphies found, all the images are projected on the plane of the central one. Using a quality map for each matching, the suitable keypoints and features are chosen to be part of the model that is further used in detection and recognition.

Feature extraction To learn the logo classes, the most relevant features are extracted using the Scale-invariant feature transform (SIFT) [14] algorithm: the Difference-of-Gausssians (DoG) locator of keypoints and the description of the keypoints' vicinity by the SIFT local features. We used the following adaptation: the edge threshold that eliminates peaks of the DoG scale space was increased (from 10 to 100 - value empirically found) to enforce a high number of features from the logo area.

To thwart the very small size of some logos, we increased the number of features extracted, by upscaling the small images at 200 pixels while keeping the aspect ratio. A similar idea is in [10], but we differ by the fact we did not enlarge all the images, to keep the running time low.

Image matching In this stage we develop the process of image stitching for finding correspondences between the features of each two training images from every class. The basic image stitching algorithm uses the VLFeat open source code. Given the features from two input images, we match them with the algorithm from [15], which rejects the correspondences that are too ambiguous.

Once the features are matched, the correspondences of their locations should indicate the transform that projects the second image onto the plane of the first one. This transform, called the *homography* transform and denoted by H, has the role of moving a point (a, b) from the plane of the first image to the coordinates (x, y) on the plane of the second image:

$$\begin{bmatrix} a \\ b \\ 1 \end{bmatrix} H = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, where \;\; H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{1}$$

Since each point correspondence provides 2 equations, 4 correspondences suffice in solving the 8 degrees of freedom of H. Often, more than 4 correspondences are available for a more robust solution.

To address the problem of outliers, the RANSAC (random sample consensus) algorithm is employed to estimate H [15]. For each 4 feature correspondences, the homography H between them is found with the direct linear transformation (DLT) [16]. This is repeated n times and the solution with most inliers is selected: the winner is the case when the projections are consistent with H within a tolerance of ϵ pixels. Our experiments proved that at least 20 pairs of points should be matched in order to obtain a correct homography.

The algorithm should iterate enough to maximize the chance to find the best match. Given the probability p_i that a feature match is correct between a pair of matching images (the inlier probability), the probability of finding the correct transformation $p(H^{Correct})$ after n trials is:

$$p(H^{Correct}) = 1 - (1 - (p_i)^r)^n \tag{2}$$

We modify the algorithm by significantly augmenting the number of trials to 200,000 iterations, compared to 500 used in [15] as logos are smaller, possibly occluded and with fewer keypoints than panoramas. If the number of inliers is high, then the homography is quickly found and, to limit the calculus, we introduced a stopping criteria based on obtaining a score above a threshold for the homography.

In the test phase, this same algorithm is used to match the test images to the class models and often no matches are found. Here, also to limit the time, if the initial number of matching pairs is below 20, then the algorithm decides that there is no chance of finding a suitable homography and exits.

The interconnections graph To sum up, the training process consists in estimating the homography between each 2 training crops of logos of the same class with RANSAC. Thus for n crops images per class, $n(n-1)/2$ image pairs are matched. Due to occlusions, inverted colors, or large variations or distortions in shape, not all the pairs of images have enough matching points to output an appropriate homography. In the end, the output of the matching procedure is a graph if only some nodes (images) are connected, similar to the idea in [11].

Each class will finally have a graph expressing the doable connections between its training images and most likely a core image (i.e. the one most connected to the others). The model of the class will be built on it, since it is clearly the most representative image in the class. We illustrate this in figure 2 by showing an example with a high number of image connections.

The class model Each link between two images indicates different keypoints that are being used in the matching, since each image in particular has its own representative features. For example figure 3 proves that in the first case some keypoints are selected, while in the second, others are highlighted in the matching process. The consequence is that if only one of the images in the training set is used to represent the class, many important features can be lost.

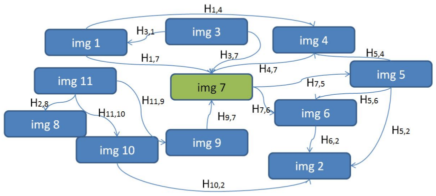


Fig. 2. Small part of a class graph. H_{ij} projects image j onto image i while the inverse of H_{ij} projects i onto j. The image most connected to the others is highlighted, and is the *central image*.

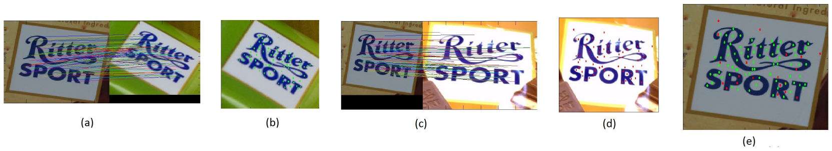


Fig. 3. The merging onto the central image of the representative keypoints coming from 2 images. (a) the matching pairs of descriptors between first image and central one, (b) the important keypoints of the first image, (c) the matching pairs of descriptors between second image and central one, (d) the important keypoints of the second image, (e) the reunion of the important keypoints on the central image

The main idea of the training stage is to conglomerate all the representative keypoints and their corresponding descriptors. We choose the central image to be the one on which this aggregation takes place, since it is obviously the best to represent the class. Using the homographies found, all the images are projected on the plane of the central image. The projected locations of the important keypoints from these images are computed and, in the end, the model of the class will be the central image described by the reunion of all the suitable points and descriptors in the class. Figure 3 shows the result of the aggregation of the keypoints from the two images, proving that each matching process reveals different pairs of keypoints that must be merged in order to obtain the best representation of the logo.

The merging is an easy task if the images are directly connected to the core image. This part of the training stage resembles [13], with the difference that there, the most representative image is manually selected. Moreover, they consider a smaller database where all the images connect to the chosen one. Contrary, we take into account also the case of the images that have no direct link to the central image by considering the path from that image to the central one. For example, in figure 2 images 1 and n are connected through images $2, 3, \ldots n-1$. The homography between image 1 and n is the composition of the homographies of the images connecting them:

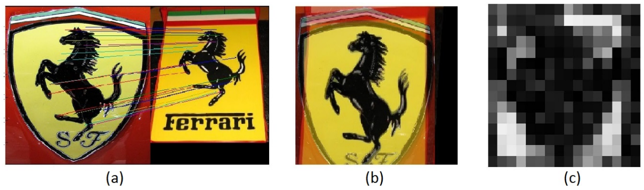


Fig. 4. Building the quality map (a) The matching pairs of points (b) The mosaic of images after applying the found homography. (c) The quality map. The darker areas show good quality of matching.

$$H_{1,n} = H_{1,2} \circ H_{2,3} \circ \cdots \circ H_{n-1,n} \tag{3}$$

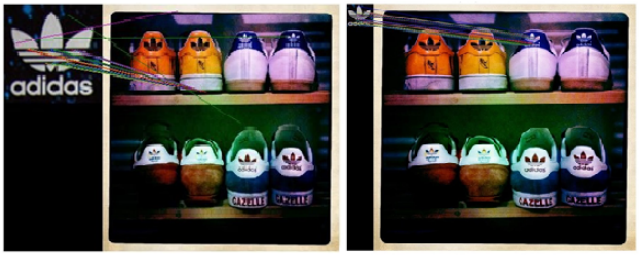
To select the shortest path (as it introduces fewest errors), from the many possible ones existing between two images, we use the Djikstra algorithm on the image connections graph. Given the corresponding coordinates of the points between any training image to the core image, we select the most representative keypoints for further use by computing the quality of match.

The quality map The quality of matching is retrieved by means of quality map, which is built for each pair of images stitched. The map values are directly related to the correctness of the matching in that area. This procedure is similar to shape matching score from [17]: given a training set of shapes the joint distribution is computed; given an actual pair, the score is retrieved by back projecting the joint distribution.

A pixel having a good quality value is a point that represents a suitable connection between the images and is not an occlusion or distortion of the shape. Figure 4 describes the matching process and the overlaid images after applying the homography. Figure 4(c) shows the quality map created for this matching, where the darker regions in the map show the areas where the matching is correct. The areas of occlusion or difference in the shape of the logo are indicated by the lighter values in the map showing a poor quality of the matching.

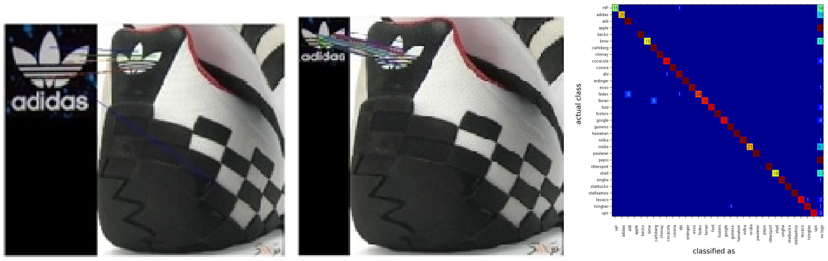
Quantization of descriptors To keep the descriptor invariant to perspective transform, the original SIFT descriptor is stored. The final descriptor is formed by the merged keypoint vector and the merged descriptor vector of the core image, as it is more robust to perspective and more comprehensive than the central image as a model. This fact is illustrated by figure 5.

Some of the positions found originate from the same interest points of the logo and, thus, become adjacent on the model image. Evidently, their descriptors are extremely similar. This requires a quantization step that keeps the unique keypoints and features describing the image. The quantization has the purpose of reducing the testing computation time.



(a) (b)

Fig. 5. (a) Matching fails when using only the keypoints and descriptors of the central image. (b) Successful detection when using the model of the class.



(a) (b) (c)

Fig. 6. (a)The failed detection of the very small logo. (b) The successful detection after resizing the test image. (c) The confusion matrix of the proposed method.

3 Implementation and results

3.1 Testing

The purpose of testing is to locate logos and classify them. Given a model for each class, the testing phase tries to match the current image to be tested against all the class models. The matching is done just as in the training phase, using SIFT feature matching and RANSAC search for the correct homography. Since now the logos are part of natural images, with large areas of non-uniform background yielding a considerable number of keypoints, the ratio of outliers versus inliers is higher than in training phase, where we used only the logo crops. This motivates the use of a high number of iterations in the RANSAC stage.

The same type of quality map is built for each matching result and its average is used as an indicator of the quality of the image stitching. If the score is high enough, the decision is taken that the logo is present. The system will indicate its position and the corresponding homography that stitches the model of the logo class to the test image. An example of detection after matching a high number

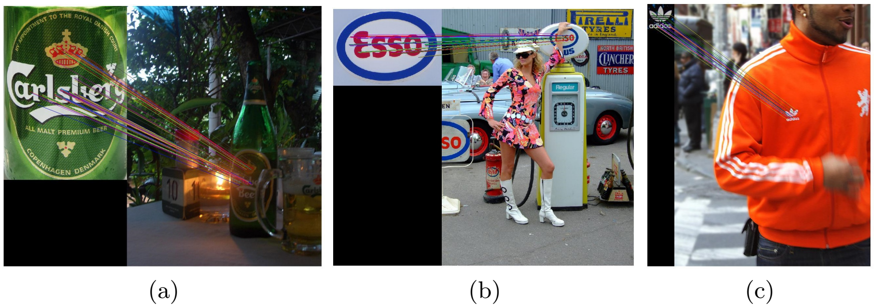


Fig. 7. Examples of detections: (a) with blurry and shadowed logo, (b) with occluded over the logo or (c) for a very small logo (30×30).

Table 1. Classification results for the compared methods for 5 example classes and respectively entire set.

Method	**Detection Rate** [%]					
Classes	**Aldi**	**Coca cola**	**DHL**	**Esso**	**Paulaner**	**All classes**
Romberg et al. [11]	56.66	60	16.6	76.6	60	61,14
Central image model [13]	76.66	66.66	70	63.3	90	60.1
Proposed method	*100*	*86.6*	*96.6*	*96.6*	*100*	*84,06*

of keypoints is in figure 7 (a). If after being confronted to all the class models, no score is large enough, then the test image will be classified as ”no-logo”.

The training phase has taught us that small sized logos do not present enough features to be correctly represented and then classified. Thus, the test images might contain also very small logos. Since there is no information about their sizes or locations, we doubled the size of test images before trying the matching. Figure 6 shows a case when the detection fails as the logo in the test image is extremely small. (b) presents the solution of the problem by enlarging the test image.

3.2 Results. Comparison with state of the art

We have obtained 100% classification rate for 13 classes and over 90% for 20 classes and respectively 84,06% for the entire dataset. A true detection is if the found logo is present in that image. The localization is correct if the intersection-over-union, (i.e. Jaccard index), is above 50% [18]. The results are better described by the confusion matrix presented in figure 6 (c). Again we have used the same scenario as Romberg et al. [11]. Comparative results may be seen in table 1. To show the benefits of the proposed homography based construction, we considered the central image as class model as discussed in [13] for vehicle logo recognition. Examples of the method detecting logos in extreme situations, such as small sizes, highly occluded or very blurred are in figure 7.

4 Discussion and continuations

The proposed method falls short for symmetric and circular logos with too few keypoints, and which do not represent well the area, leading to an inability to compute homographies. While normally we find over 300 pairs of images that match, for "Pepsi" and "Apple" only ≈ 5 connections are in the class describing graph, thus, leading to wrong class model and low classification performance.

Yet, overall, the method is effective in detecting the majority of classes, surpassing many challenges of logo detection in natural images. Continuation envisages the cases of failure by changing the matching process so to take into account the vicinity of the points, thus improving the homography building.

References

1. Lowe, D.: Object recognition from local scale-invariant features. In: ICCV. (1999) 1150–1157
2. Schneiderman, H., Kanade, T.: A statistical method for 3d ob-ject detection applied to faces and cars. In: CVPR. (2003) 746–751
3. Torralba, A., Murphy, K., Freeman, W.: Sharing visual fea-tures for multiclass and multiview object detection. In: CVPR. (2004) 762–769
4. Opelt, A., Fusseneger, M., Pinz, A., Auer, P.: Generic object recognition with boosting. IEEE T. PAMI **28(3)** (2006) 416 – 431
5. Bernstein, E., Amit, Y.: Part-based statistical models for object classification and detection. In: CVPR. (2005) 734–740
6. Bagdanov, A., Ballan, L., Bertini, M., Del Bimbo, A.: Trademark matching and retrieval in sports video databases. In: ACM MIR. (2007) 79–86
7. Joly, A., Buisson, O.: Logo retrieval with a contrario visual query expansion. In: ACM MM. (2009) 581–584
8. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: ICCV. (2003) 1470 – 1477
9. Kleban, J., Xie, X., Ma, W.Y.: Spatial pyramid mining for logo detection in natural scenes. In: IEEE ICME. (2008) 1470 – 1477
10. Revaud, J., Douze, M., Schmid, C.: Correlation-based burstiness for logo retrieval. In: ACM MM. (2012) 965–968
11. Romberg, S., Garcia Pueyo, L., Lienhart, R., van Zwol, R.: Scalable logo recognition in real-world images. In: ACM ICMR. (2011) 965–968
12. Romberg, S., Lienhart, R.: Bundle min-hashing for logo recognition. In: ACM ICMR. (2013)
13. Psyllos, A.P., Anagnostopoulos, C.N.E., Kayafas, E.: Vehicle logo recognition using a sift-based enhanced matching scheme. IEEE T. TITS **11(2)** (2010) 322 – 328
14. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **62(2)** (2004) 91 – 110
15. Brown, M., Lowe, D.: Automatic panoramic image stitching using invariant features. IJCV **74(1)** (2006) 59–73
16. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2004)
17. Florea, L., Florea, C., Vranceanu, R., Vertan, C.: Can your eyes tell me how you think? a gaze directed estimation of the mental activity. In: BMVC. (2013)
18. Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. In: IJCV. Volume 1. (2010) 303–338