# Logo Recognition Using CNN Features

Simone Bianco, Marco Buzzelli, Davide Mazzini(✉), and Raimondo Schettini

DISCo (Dipartimento di Informatica, Sistemistica E Comunicazione),
Universitàdegli Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy
{simone.bianco,davide.mazzini,schettini}@disco.unimib.it,
marco.buzzelli@unimib.it

**Abstract.** In this paper we propose a method for logo recognition based on Convolutional Neural Networks, instead of the commonly used keypoint-based approaches. The method involves the selection of candidate subwindows using an unsupervised segmentation algorithm, and the SVM-based classification of such candidate regions using features computed by a CNN. For training the neural network we augment the training set with artificial transformations, while for classification we exploit a query expansion strategy to increase the recall rate. Experiments were performed on a publicly-available dataset that was also corrupted in order to investigate the robustness of the proposed method with respect to blur, noise and lossy compression.

## 1 Introduction

Logo recognition in images and videos has been gaining considerable attention in the last decade, with such applications as copyright infringement detection, intelligent traffic-control systems, and automated computation of brand-related statistics. Such problems are typically addressed with keypoint-based detectors and descriptors like SIFT [1–3]. These methods are in fact best suited for well-defined shapes and affine transformations, like those found in the domain of logos. Since in many real applications the logo images could be highly degradated, in this paper we investigated Convolutional Neural Networks [4] as an alternative approach that is not based on keypoint detection. CNNs fall in the category of Deep Learning techniques, which have been employed in others fields such as speech recognition [5] and action recognition [6]. However, they are relatively new to logo recognition. The only relevant contribution is [7], where the authors trained a network to classify a custom dataset of vehicle logos and employed some very specific heuristics to localize the vehicle logo position. We instead make use of an object proposal algorithm [8,9] to produce a higher number of candidate windows, while drastically reducing the number of candidates obtainable from a multiscale sliding window approach. The underlying idea is similar to what was used in [10] for object recognition.

We report some experiments on the publicly available FlickrLogos-32 dataset [11] and show that our approach is able to achieve near-state-of-the-art performance on high-quality images, and still achieves good results on degradated

V. Murino and E. Puppo (Eds.): ICIAP 2015, Part II, LNCS 9280, pp. 438–448, 2015.
DOI: 10.1007/978-3-319-23234-8_41

images. More precisely, we tested our detection system after applying different types of image distortions to decrease the image quality in a controlled fashion. The proposed method can deal with JPEG compression of high intensity, and with noise and blur of medium intensity. Performance decrease with higher levels of blurriness, and further decrease with high-variance gaussian noise.

## 2 Proposed Method

In this section we outline the recognition pipeline used in this work. We follow the idea first investigated by Girshick et al. [10]. Given an input image, we extract regions which are more likely to contain an object. These regions are called object proposals. The algorithm used for the extraction of the objects proposals is class-agnostic, therefore it extracts regions of different aspect-ratios that can be used to recognize objects under different kinds of geometric transformation. These proposal are then warped to a common size (see Section 5.1) and processed for query expansion in order to increase recall. Finally we use a pre-trained CNN as feature extractor and a linear SVM for logo recognition and classification. Figure 1 shows the main steps of the recognition pipeline.

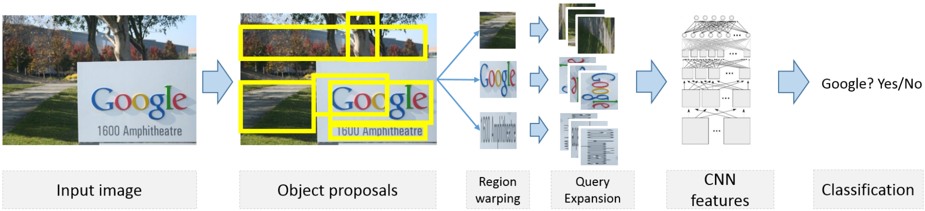

**Fig. 1.** Outline of the recognition pipeline: (1) Candidate objects regions are extracted from the image. (2) Regions are warped to a common size and multiplied through Query Expansion. (3) CNN features are computed over each region. (4) Classification is performed using linear SVMs.

## 3 Selective Search

Selective Search has been introduced by van de Sande et al. [8,9] to enable the use of more expensive features and classifiers in object detection, eliminating the need of computing them for every possible sliding window.

The authors exploit a hierarchical grouping algorithm, in order to naturally generate locations at all scales, by continuing the grouping process until the whole image becomes a single region. They first use [12] to create initial regions, then, instead of using a single clustering technique, they use a variety of complementary grouping criteria to account for as many image conditions as possible. Such criteria include color similarity, texture similarity, and measures that encourage merging of small regions and overlapping regions. The final set of candidate locations is then obtained by combining the locations of these complementary partitionings.

# 4 Query Expansion

When working with object detection and classification, usually we have to deal with two kinds of variability. The *intrinsic* variability corresponds to the fact that two instances of the same object class can be visually different, even when viewed under similar conditions (e.g. different versions of a logo may differ for some details or colors). The *extrinsic* variability refers to those differences in appearance that are not specific to the object class (e.g. different viewpoints, lighting conditions, image compression artifacts).

The FlickrLogos-32 dataset exhibits high levels of extrinsic variability. To cope with this, we transform, at test time, each candidate location extracted with Selective Search, thus producing an expanded query. The candidate location is then assigned to the class with maximum confidence over all the expanded query.

# 5 Convolutional Neural Networks

Convolutional Neural Networks were first presented by Kunihiko Fukushima in [4] as a tool for visual pattern recognition. However, only in recent years they have become widespread in the scientific community, thanks to the development of high-performing architectures working on GPU [13].

A CNN takes an input image, which usually has undergone some minor preprocessing, processes it with a cascade of different transformation layers, to finally produce a prediction of the image class. Convolutional layers are the main type of layer used in CNNs. These are designed to perform a convolutional operation on the input data (which can be either the original image, or the result of hidden layers). The kernel involved in this operation, however, is not hand-encoded, but automatically learned through the backpropagation algorithm used to train the network. Non linearities are implemented in the CNN by activation functions and pooling layers. The most used activation function is ReLU (Rectified Linear Units), which keeps only the positive part of the input without applying any kind of upper bound to the signal. Pooling layers work as average or maximum filters, and as such are used to reduce the impact of small variations in the signal. Furthermore, they allow for a dimensionality reduction of the input data. Their effect depends on the filter size and stride. Dropout layers are used to reduce overfitting, which can occur when the number of training examples and number of CNN parameters are unbalanced, with the second one being greater than the first one. This is done, at every training iteration, by randomly dropping some neurons with probability $p$. At testing time instead, all the neurons are used, but their responses are weighted by $p$ itself.

## 5.1 CNN Features

Instead of learning an ad-hoc CNN for the logo recognition problem, we investigate how a pre-trained one works on this problem. It is in fact known that the features produced by CNNs in the last layers before the class assignment work effectively on other problems as well [14]. To this end, we employ a Caffe [13] implementation of the CNN described by Krizhevsky et al. [15] to extract a 4096-dimensional feature vector from each $227 \times 227$ RGB image. This is done by subtracting a previously computed mean RGB image, and forward-propagating the result through five convolutional layers and two fully connected layers. More details about the network architecture can be found in [13,15]. The CNN was originally trained on a large dataset (ILSVRC 2012) with image-level annotations to classify images into 1000 different classes. Features are obtained by extracting activation values of the last hidden layer. The extracted features for each candidate location are then used as input to a Support Vector Machine (SVM) [16] for classification as no-logo or as belonging to a specific logo class. We employed a multiclass one-vs-all linear SVM with regularization hyperparameter $C = 1$.

### 5.2 Transformation Pursuit

Since the Flickr-logo dataset has been collected to evaluate SIFT-like recognition algorithms [11] the training set contains only few examples for each logo (see Table 1). To handle the large extrinsic variability of the dataset and to prevent the learning algorithm to overfit, we significantly increase the training set following Transformation Pursuit [17]. For each region proposal extracted from images which overlaps with the groundtruth annotation we apply a set of predefined image transformations. In particular the applied tranfomations include: translation, scale, shear on the y axis and shear on the x axis. With this set we take into account also rotation transformation which is a combination of the two shear transform. By applying only the two extrema values of each of the complete set of geometric transformations we can increase the number of examples by a factor of $\sim 250$. In figure 2 is depicted a subset of the geometric transformations applied.

## 6 Experimental Setup and Results

Experiments were performed on the publicly-available FlickrLogos-32 dataset [11]. This is a collection of photos showing 32 different logo brands, and is meant for the evaluation of logo retrieval and multi-class logo detection/recognition systems on real-world images. All logos have an approximately planar or cylindrical surface. The whole dataset is split into three disjoint subsets $P_1$, $P_2$, and $P_3$ as reported in Table 1, each containing images of all 32 classes.

### 6.1 Selective Search Evaluation

The Selective Search algorithm [9] can extract the candidate object regions upon different color spaces. In this section we report an evaluation of Selective Search object proposals quality using different color spaces on the FlickrLogos-32 dataset.

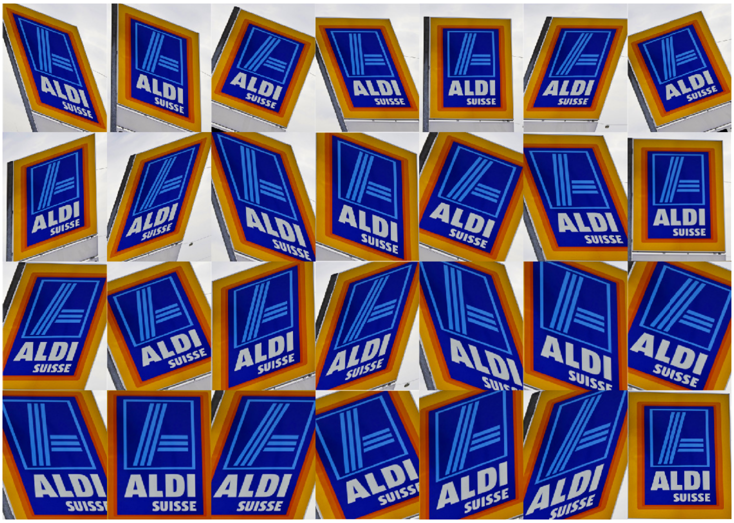

**Fig. 2.** Representative subset of geometric tranformations applied to an extracted region proposal. The original image is in the lower-right corner.

**Table 1.** FlickrLogos-32 dataset partitions

| Partition | Description | Images | Total |
|---|---|---|---|
| $P_1$ (training set) | Single logo images, clean background | 10 per class | 320 images |
| $P_2$ (validation set) | Images showing at least a single logo<br>Non-logo images | 30 per class<br>3000 | 3960 images |
| $P_3$ (test set) | Images showing at least a single logo<br>Non-logo images | 30 per class<br>3000 | 3960 images |

Hosang et al. [18] introduced a class agnostic metric to evaluate the effectiveness of an object proposal algorithm: the Recall versus IoU (Intersection over Union). It is computed by varying the IoU rejection threshold, then for each threshold value, the number of overlapping bounding-boxes is counted.

In Figure 3 we report the curves for five different color spaces: HSV, Lab, rg plus the Intensity channel, the Hue channel and the Intensity channel only.

We also report in Table 2 the list of the mean number of object proposals extracted and the Average Recall for each colorspace tested. The Average Recall was computed for levels of IoU from 0.5 to 1. Lower levels haven't been considered because we are only interested in high levels of overlap.

Finally we chose to use the Selective Search based on the HSV colorspace as building-block for our pipeline because it shows the higher recall among all the others colorspaces.

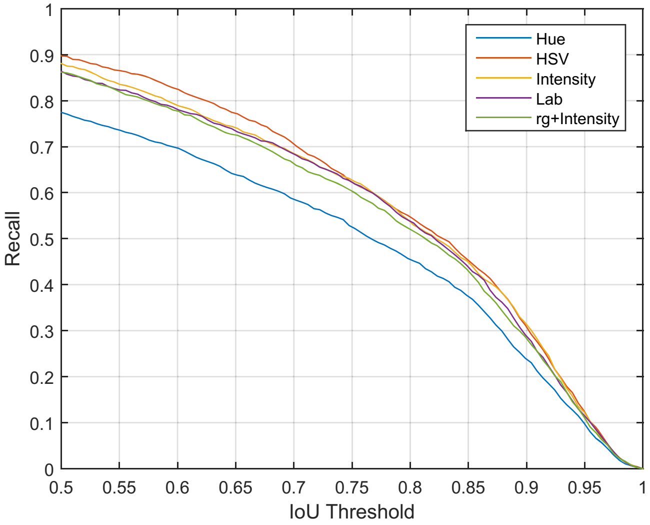

**Fig. 3.** Recall versus IoU threshold for different color spaces on Flickr-Logos dataset.

**Table 2.** Mean Number of Object Proposals per image and Average Recall value for each color space

| Color space | #Proposals/image | Average Recall |
|---|---|---|
| Hue | 486 | 0.472 |
| HSV | 642 | 0.566 |
| Intensity | 412 | 0.552 |
| Lab | 292 | 0.545 |
| rg + Intensity | 352 | 0.535 |

### 6.2 Results

The system was trained on FlickrLogos-32 $P_1$ set, and validated on $P_2$ for hyperparameters selection with a target precision of 98% (for better comparison with the state of the art [1,11,19]). Finally, it was tested on $P_3$ with the selected hyperparameters.

Figure 4 shows the performance level obtained by the proposed method in terms of precision and recall, on validation set and test set. For the test set we report both performance with and without Query Expansion, showing the significant gain obtained with this step.

Table 3 reports a comparison with other state of the art approaches. [11] and [1] are based on the bundling of neighboring SIFT-based visual words into

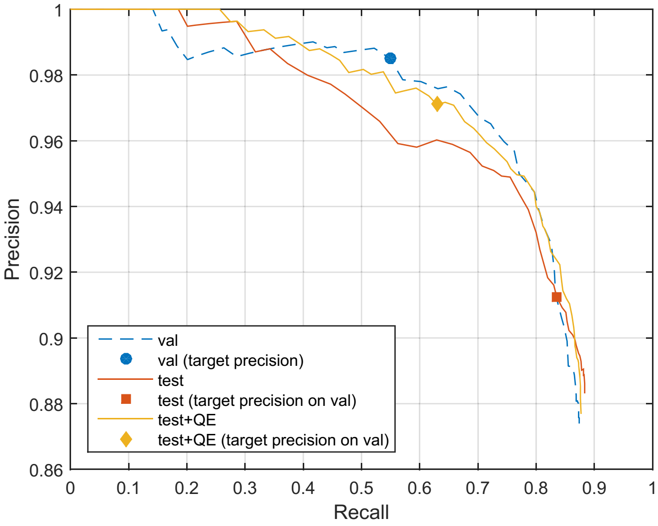

**Fig. 4.** Precision-Recall curve on validation set, test set, and test set with Query Expansion. Selected points are obtained by setting target precision at 98% in validation set.

unique signatures. [19] uses a statistical model for identifying incorrect detections output by keypoint-based matching algorithms.

Results show how even though the underlying CNN was trained for recognition in a different domain, it is still able to achieve near-state-of-the-art performance.

**Table 3.** Performance comparison with other approaches

| Method | Precision | Recall |
|---|---|---|
| Romberg et. al [11] | 0.98 | 0.61 |
| Revaud et. al [19] | $\geq 0.98$ | 0.73 |
| Romberg et. al [1] | 0.999 | 0.832 |
| Proposed method | 0.91 | 0.84 |
| Proposed method + QE | 0.97 | 0.63 |

## 7 Results under Image Distortions

We want to test the robustness of the proposed method with respect to three kinds of image distortion as reported in Table 4 and Figure 5.

**Table 4.** Types of distortions applied to the images of the FlickrLogos-32 dataset.

| Type | Amount |
|---|---|
| Gaussian Blur | Filter Size 10px |
| Gaussian Blur | Filter Size 20px |
| JPEG Compression | Quality 20% |
| JPEG Compression | Quality 10% |
| Gaussian Noise | $\sigma^2 = 0.005$ |
| Gaussian Noise | $\sigma^2 = 0.02$ |

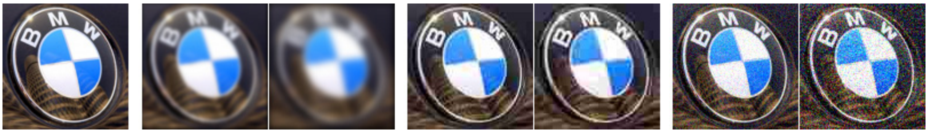

**Fig. 5.** Types of distortions applied to the images of the FlickrLogos-32 dataset.

### 7.1 Selective Search Evaluation

Figure 6 shows values of the Recall versus IoU threshold for the original and distorted images.

Image distortion has a low impact on the overall quality of the extracted Object Proposals. Blur has the biggest impact especially for low levels of IoU but in the worst case performance dropped by 10% only. On the other hand, jpeg compression seems to have a very low impact on the Object Proposals quality even at high levels of compression. Our tests confirm the results obtained by Hosang et al. in [18] which found the Selective Search to be one of the most robust Object Proposals algorithms.

### 7.2 Results

In order to test the complete recognition pipeline on the distorted images, we augmented the training set by a factor of six. For each original image we add six deformed images, each with a single distortion applied. The magnitude of each distortion, shown in table 4, is the same for train and test. We run the recognition pipeline on images of the Flickr-logo test set modified with a single distortion at a time. This controlled environment makes it possible to check the impact of every single distortion without masking effects.

Figure 7 shows the results of the performed tests. Gaussian noise and blur with high magnitude ($\sigma^2 = 0.02$ and $size = 20px$) have the highest impact on the overall results. The results on the complete recognition pipeline reflect those on the Selective Search part: the same types of distortions having the highest impact on the overall performance affect also the Recall measure of the Selective Search evaluation. This clue leads us to consider the quality of the Object Proposals as one of the most important aspects to care about in our recognition pipeline.

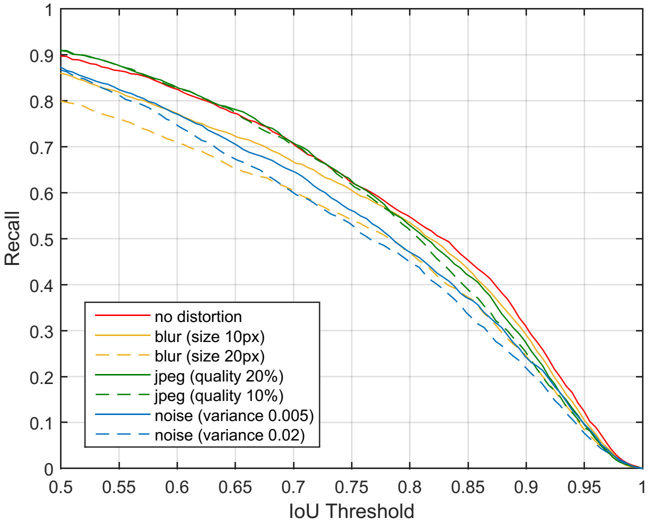

**Fig. 6.** Recall versus IoU threshold for 6 different types of image distortion on the Flickr-Logos dataset. Only the best overlapping bounding-box for each groundtruth annotation is considered.

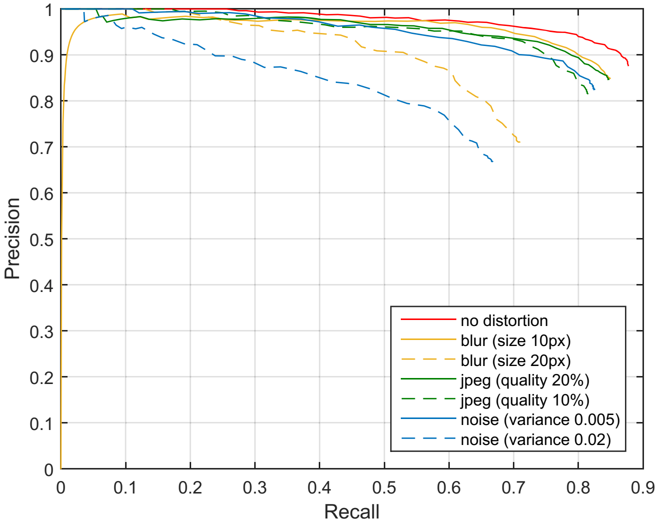

**Fig. 7.** Precision-Recall curves on the Flickr-logo test dataset. Different image distortions have been applied to obtain different curves.

## 8 Conclusions

In this work we treated the problem of logo recognition. This is usually addressed with keypoint-based methods on high-quality images. We instead used Convolutional Neural Networks as a robust alternative for low-quality images. The proposed pipeline involved selecting candidate subwindows using Selective Search, augmenting the training set using Transformation Pursuit, and performing Query Expansion for increasing recall. The method proved to be effective even with CNN features that were trained for a different task, producing results close to the state of the art. The robustness of the method has been investigated with respect to three different kinds of distortion: blur, noise and lossy compression. Results showed that noise was the most affecting one, while lossy compression produced little to no performance loss.

As future developments, we will investigate the use of a Viola-Jones-like AdaBoost detector [20] for object proposal in place of Selective Search. In order to further improve the recognition performance, we will also investigate the application of a pre-processing step to the image aimed to obtain a faithful color description [21].

## References

1. Romberg, S., Lienhart, R.: Bundle min-hashing for logo recognition. In: Proceedings of the 3rd ACM conference on International conference on multimedia retrieval, pp. 113–120. ACM (2013)
2. Bianco, S., Schettini, R., Mazzini, D., Pau, D.P.: Quantitative review of local descriptors for visual search. In: IEEE Third International Conference on Consumer Electronics Berlin (ICCE-Berlin), ICCEBerlin 2013, pp. 98–102. IEEE (2013)
3. Bianco, S., Schettini, R., Mazzini, D., Pau, D.P.: Local detectors and compact descriptors: a quantitative comparison. Digital Signal Processing (2015) (accepted)
4. Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics **36**(4), 193–202 (1980)
5. Abdel-Hamid, O., Mohamed, A.-R., Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) **22**(10), 1533–1545 (2014)
6. Foggia, P., Saggese, A., Strisciuglio, N., Vento, M.: Exploiting the deep learning paradigm for recognizing human actions. In: 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 93–98. IEEE (2014)
7. Pan, C., Yan, Z., Xu, X., Sun, M., Shao, J., Wu, D.: Vehicle logo recognition based on deep learning architecture in video surveillance for intelligent traffic system (2013)
8. van de Sande, K.E.A., Uijlings, J.R.R., Gevers, T., Smeulders, A.W.M.: Segmentation as selective search for object recognition. In: IEEE International Conference on Computer Vision (2011)
9. Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. International Journal of Computer Vision **104**(2), 154–171 (2013)

10. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580–587. IEEE (2014)
11. Romberg, S., Pueyo, L.G., Lienhart, R., Van Zwol, R.: Scalable logo recognition in real-world images. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, p. 25. ACM (2011)
12. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision **59**(2), 167–181 (2004)
13. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
14. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: An astounding baseline for recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2014
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
16. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995)
17. Paulin, M., Revaud, J., Harchaoui, Z., Perronnin, F., Schmid, C.: Transformation pursuit for image classification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3646–3653. IEEE (2014)
18. Hosang, J., Benenson, R., Dollár, P., Schiele, B.: What makes for effective detection proposals? (2015). arXiv preprint arXiv:1502.05082
19. Revaud, J., Douze, M., Schmid, C.: Correlation-based burstiness for logo retrieval. In: Proceedings of the 20th ACM International Conference on Multimedia, pp. 965–968. ACM (2012)
20. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, pp. I-511. IEEE (2001)
21. Bianco, S., Bruna, A., Naccari, F., Schettini, R.: Color space transformations for digital photography exploiting information about the illuminant estimation process. Journal of the Optical Society of America A **29**(3), 374–384 (2012)