Open Set Logo Detection and Retrieval

Andras Tüzkö[1], Christian Herrmann[1,2], Daniel Manger[1], Dieter Willersinn[1], Jürgen Beyerer[1,2]
[1] *Fraunhofer IOSB, Karlsruhe, Germany*
[2] *Karlsruhe Institute of Technology KIT, Vision and Fusion Lab, Karlsruhe, Germany*
{*andras.tuezkoe*|*christian.herrmann*|*daniel.manger*|*dieter.willersinn*}*@iosb.fraunhofer.de*

Keywords: Logo Detection, Logo Retrieval, Logo Dataset, Trademark Retrieval, Open Set Retrieval, Deep Learning.

Abstract: Searching for logos in image data allows several applications, with judging the effectiveness of advertisement in sports event broadcasts being one example. Given a query sample in shape of a logo image, the task is to find all further occurrences of this logo in a database of images or videos. Currently, common logo retrieval approaches are unsuitable for this task because of their closed world assumption. To address this issue, an open set logo retrieval method is proposed in this work which can search for previously unfamiliar logos only by one query sample. A two stage concept with an open set logo detection and comparison is proposed similar to other retrieval tasks. Both modules are based on task specific Convolutional Neural Networks (CNNs). To train the detector with appropriate in-the-wild data, an according novel Logos in the Wild dataset is collected and made publicly available. The proposed method extends the application field in comparison to closed set approaches and improvements over baseline methods derived from these state-of-the-art closed set approaches are shown.

1 INTRODUCTION

Automated search for logos is a desirable task in visual image analysis. A key application is the effectiveness measurement of advertisements. Being able to find all logos that match a query, for example, a logo of a specific company, in images allows to assess the visual frequency and prominence of logos in TV broadcasts. Typically, these broadcasts are sports events where sponsorship and advertisement is very common. This requires a flexible system where the query can easily be defined and switched according to the current task. Especially, also previously unseen logos should be found if one query sample is available. This requirement excludes basically all current logo retrieval approaches because they make a closed-world assumption where all searched logos are known beforehand. Instead, this paper focuses on open set logo retrieval where only one sample image of a logo is available.

Consequently, a novel processing strategy for logo retrieval based on a logo detector and a feature extractor is proposed as illustrated in figure 1. Similar strategies are known from other open set retrieval tasks, such as face or person retrieval (Bäuml et al., 2010; Herrmann and Beyerer, 2015). Both, the detector and the extractor are task specific CNNs. For detection, the Faster R-CNN framework (Ren et al., 2015) is employed and the extractor is derived from classification networks for the ImageNet challenge (Deng et al., 2009).

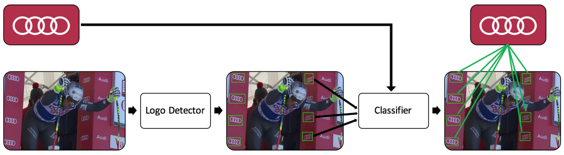

Figure 1: Proposed logo retrieval strategy.

The necessity for open set logo retrieval becomes obvious when having a look at the diversity and amount of existing logos and brands[1]. The METU trademark dataset (Tursun et al., 2017) contains, for example, over half a million different logos. Given this number, a closed set approach where all different logos are pretrained within the retrieval system is clearly inappropriate. This is why our proposed feature extractor generates a discriminative logo descriptor, which generalizes to unseen logos, instead of a mere classification between previously known brands. The well-known high discriminative capabilities of

[1] The term brand is used in this work as synonym for a single logo class. Thus, a brand might also refer to a product or company name if an according logo exists.

CNNs allow to construct such a feature extractor.

One challenge for training a general purpose logo detector lies in appropriate training data. Many logo or trademark dataset (Eakins et al., 1998; Tursun et al., 2017) only contain the original logo graphic but no in-the-wild occurrences of these logos which are required for the target application. The need for annotated logo bounding boxes in the images limits the number of suitable datasets. Existing logo datasets (Joly and Buisson, 2009; Kalantidis et al., 2011; Romberg et al., 2011; Letessier et al., 2012; Bianco et al., 2015; Su et al., 2016; Bianco et al., 2017) with available bounding boxes are often restricted to a very small number of brands and mostly high quality images. Especially, occlusions, blur and variations within a logo type are only partially covered. To address these shortcomings, a novel in-the-wild logo dataset is collected and made publicly available [2].

The contributions of this work are threefold:

- A novel open set logo detector which can detect previously unseen logos.
- An open set logo retrieval system which needs only a single logo image as query.
- The introduction of a novel large-scale in-the-wild logo dataset.

2 RELATED WORK

Current logo retrieval strategies are generally solving a closed set detection and classification problem. Eggert et.al. (Eggert et al., 2015) utilized CNNs to extract features from logos and determined their brand by classification with a set of Support Vector Machines (SVMs). Fast R-CNN (Girshick, 2015) was used for the first time to retrieve logos from images by Iandola et al. (Iandola et al., 2015) and achieved superior results on the FlickrLogos-32 dataset (Romberg et al., 2011). Furthermore, R-CNN, Fast R-CNN and Faster R-CNN were used in (Bao et al., 2016), (Oliveira et al., 2016), (Qi et al., 2017). All these works use the same brands for training as for validation.

2.1 Open Set Retrieval

Retrieval scenarios in other domains are basically always considered open set, i.e.samples from the currently searched class have never been seen before. This is the case for general purpose image retrieval (Sivic and Zisserman, 2003), tattoo retrieval (Manger, 2012) or for person retrieval in image or video data where face or appearance-based methods are common (Bäuml et al., 2010; Weber et al., 2011; Herrmann and Beyerer, 2015). The reason is that these in-the-wild scenarios offer usually a too large and impossible to capture variety of object classes. In case of persons, a class would be a person identity with billions of persons existing. Consequently, methods have to be designed and trained on a limited set of classes and have to generalize to previously unseen classes. We argue that this approach is also required for logo retrieval because of the vast amount of existing brands and according logos which cannot be captured in advance. Typically, approaches targeting open set scenarios consist of an object detector and a feature extractor (**?**). The detector localizes the objects of interest and the feature extractor creates a discriminative descriptor regarding the target classes which can than be compared to query samples.

2.2 Object Detector Frameworks

Early detectors applied hand-crafted features, such as Haar-like features, combined with a classifier to detect objects in images (Viola and Jones, 2004). Nowadays, deep learning methods surpass the traditional methods by a significant margin. In addition, they allow a certain level of object classification within the detector which is mostly used to simultaneously detect different object categories (Sermanet et al., 2013). The YOLO detector (Redmon et al., 2015) introduces an end-to-end network for object detection and classification based on bounding box regressors for object localization. This concept is similarly applied by the Single Shot MultiBox Detector (SSD) (Liu et al., 2015). Faster Region-Based Convolutional Neural Network (R-CNN) (Ren et al., 2015) introduces a region proposal network (RPN) to detect object candidates in the feature maps and classifies the candidate regions by a fully connected network. Improvements of the Faster R-CNN are the Region-based Fully Convolutional Network (R-FCN) (Dai et al., 2016), which reduces inference time by an end-to-end fully convolutional network, and the Mask R-CNN (He et al., 2017), adding a classification mask for instance segmentation.

2.3 CNN-based Classification

AlexNet (Krizhevsky et al., 2012) was the first neural network after the conquest of SVMs, achieving impressive performance on image content classification and winning the ImageNet challenge (Deng

[2] `url://to.come`

et al., 2009). It consists of five convolutional layers, each followed by a max-pooling, which counted as a very deep network at the time. VGG (Simonyan and Zisserman, 2015) follows the general architecture of AlexNet with an increased number of convolutional layers achieving better performance. The inception architecture (**?**) proposed a multi-path network module for better multi-scale addressing, but was shortly after superseded by the Residual Networks (ResNet) (He et al., 2015; **?**). They increase network depth heavily up to 1000 layers in the most extreme configurations by additional skip connections which bypass two convolutional layers. The recent DenseNet (Huang et al., 2016) builds on a ResNet-like architecture and introduces “dense units”. The output of these units is connected with every subsequent dense unit’s input by concatenation. This results in a much denser network than a conventional feed-forward network.

3 LOGO DETECTION

The usual approach for scene retrieval is to create a global feature of the input image. This is achieved either by inferring from the complete image or by searching for key regions and then extracting features from the located regions, which are finally fused into a global feature. For logo retrieval, extraction of a global feature is counterproductive because it lacks discriminative power to retrieve small objects. Additionally, global features usually include no information about the size and location of the objects which is also an important factor for logo retrieval.

Therefore, we choose a two-stage approach consisting of logo detection and logo classification as illustrated in figure 1. First, the logos have to be detected in the input image. There are a lot of options to search for objects as explained in cha. Girshick et al. (Ren et al., 2015) proposed the Faster R-CNN, for end to end learning to detect and classify objects on an image. This network has a bounding box regressor for each trained class, thus it is capable to produce object type specific region proposals.

4 LOGO COMPARISON

5 LOGO DATASET

To train the proposed logo detector and feature extractor, a novel logo dataset is collected to supplement publicly available logo datasets. A comparison to other datasets is given in table 1. The goal is an in-the-wild logo dataset with pictures including logos instead of the pure original logo graphics. In addition, images where the logo does not represent the central dominant part of the image are preferred. See figure 2 for a few examples of the collected data. Following the general suggestions from (Bansal et al., 2017), we target for a dataset containing significantly more brands instead of collecting additional image samples for the already covered brands. This is the exact opposite strategy than performed by the Logos-32plus dataset. Starting with a list of well-known brands and companies, an image web search is performed. Because most other web collected logo datasets mainly rely on Flickr, we opt for Google image search to broaden the domain. Brand or company names are searched directly or in combination with a predefined set of search terms, e.g., ‘bmw advertisement’, ‘bmw building’, ‘bmw poster’ or ‘bmw store’.

For each search result, the first *N* images are downloaded, where *N* is determined by a quick manual inspection to avoid collecting too much garbage. After removing duplicates, this results in 4 to 608 images per searched brand. These images are then one-by-one annotated with logo bounding boxes or sorted out if unsuitable. Images are considered unsuitable if they contain no logos or fail the in-the-wild requirement, which is the case for the original raw logo graphics. Taken pictures of such logos and advertisement posters on the other hand are desired to be in the dataset. Annotations distinguish between textual and graphical logos as well as different logos from one company as exemplary indicated in figure 3. Altogether, the current version of the dataset, which is used in this paper, contains 631 brands with 17,738 annotated bounding boxes. 150 brands occur at least 10 times. An image may contain several logos with the maximum being 100 logos in one image. The complete distributions are shown in figures 4 and 5.

The collected Logos in the Wild dataset exceeds the size of all related logo datasets as shown in table 1. Even the union of all related logo datasets contains significantly less brands and RoIs which makes Logos in the Wild a valuable large-scale dataset. The annotation is still an ongoing process and further larger versions of the dataset are expected to be published in the future (??).

6 EXPERIMENTS

The proposed method is evaluated on the test set of the public FlickrLogos-32 dataset including the distractors. Additional experiments are performed on an

Table 1: Publicly available logo datasets in comparison with the novel dataset.

	dataset	brands	logo images	RoIs
public	BelgaLogos (Joly and Buisson, 2009; Letessier et al., 2012)	37	1,321	2,697
	FlickrBelgaLogos (Letessier et al., 2012)	37	2,697	2,697
	Flickr Logos 27 (Kalantidis et al., 2011)	27	810	1,261
	FlickrLogos-32 (Romberg et al., 2011)	32	2,240	3,404
	Logos-32plus (Bianco et al., 2015; Bianco et al., 2017)	32	7,830	12,300
	TopLogo10 (Su et al., 2016)	10	700	863
	total	80 (union)	15,598	23,222
new	Logos in the Wild	631	6,084	17,738

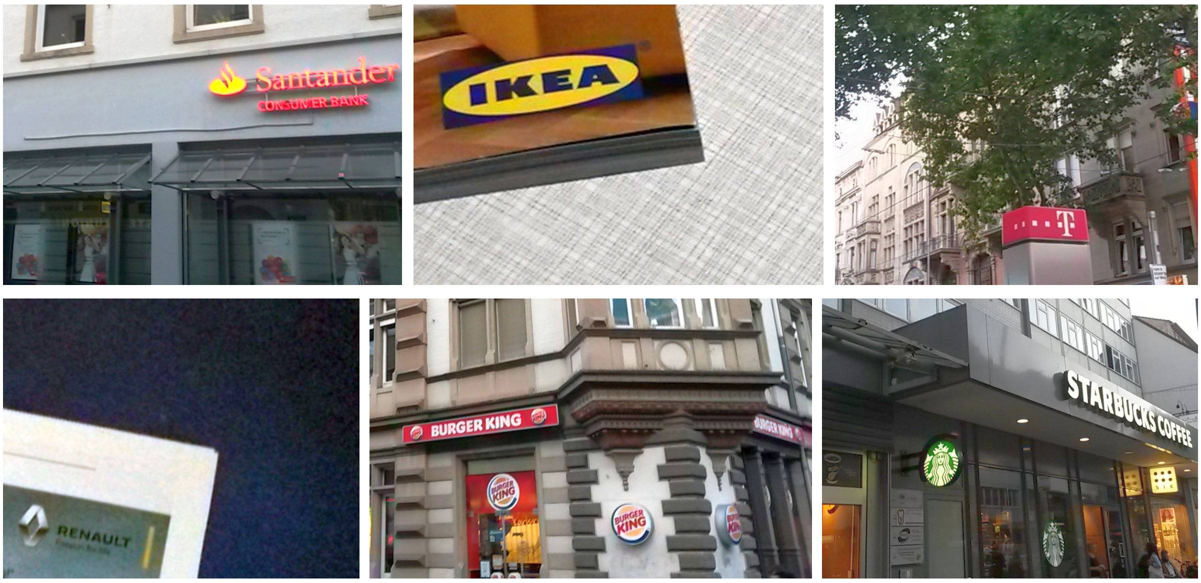

Figure 2: Examples from the collected Logos in the Wild dataset.

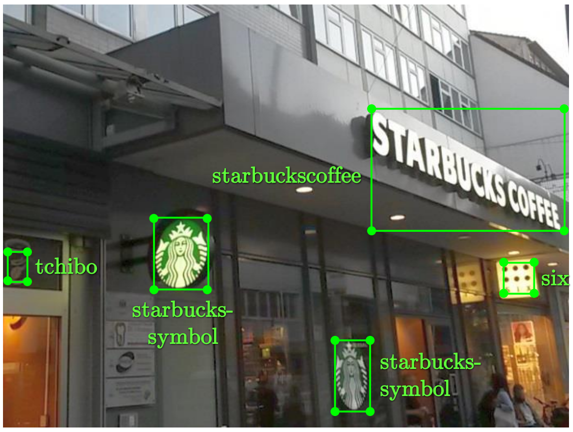

Figure 3: Annotations differentiate between textual and graphical logos.

internal dataset of sports event TV broadcasts. For a proper separation of train and test data, all brands which are present in the FlickrLogos-32 test set are

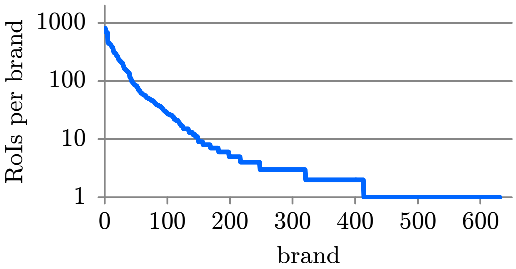

Figure 4: Distribution of number of RoIs per brand.

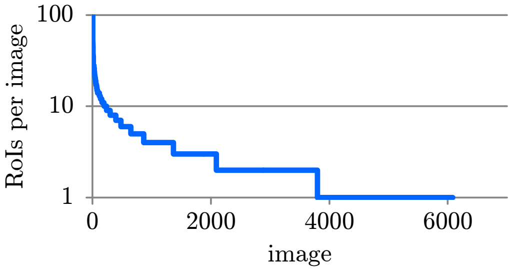

Figure 5: Distribution of number of RoIs per image.

Table 2: Train and test set statistics.

?? | ??

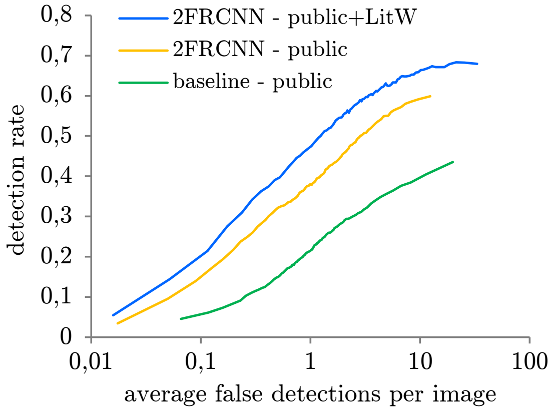

Figure 6: Detection FROC curve for FlickrLogos-32 test set.

removed from the training data. The training set consists of two parts. The union of all public logo datasets and the novel Logos in the Wild dataset. The respective training and test set sizes are listed in table **??**.

In the first step, the detector stage alone is assessed. Then, the combination of detection and comparison for logo retrieval is evaluated. Detection and matching performance is measured by the Free-Response Receiver Operating Characteristic (FROC) curve (Miller, 1969) which denotes the detection or detection and identification rate versus the number of false detections.

In all cases the CNNs are trained until convergence which requires ?? to ??k iterations with a batchsize of ??. Training duration depends on the architecture as well as the amount of training data.

6.1 Detection

As baseline, the state-of-the-art closed set logo retrieval method from (**?**) is employed and trained on the public portion of the training data. It is adapted to open set detection by using the RPN scores as detections. This skips the closed set classification part of the network which is pre-trained on different logos than should be detected on the test set. The proposed logo detector is first trained on the same public data for comparison. The results in figure 6 indicate that this strategy is superior by a significant margin.

Further improvement is achieved by combining the public training data with the novel data. Adding the Logos in the Wild dataset as additional training data improves the detection results with its large variety of additional training brands. This confirms findings from other domains, such as face analysis, where wider training datasets are preferred over deeper ones (Bansal et al., 2017). This means it is better to train on additional different brands than on additional samples per brand. As direction for future dataset collection, this suggests to focus on additional brands.

6.2 Retrieval

For the retrieval experiments, the state-of-the-art closed set logo retrieval method from the previous section is again used as baseline. The class probabilities are interpreted as feature vector which is then used to match previously unseen logos. For the proposed open set strategy, the best logo detection network from the previous section is used in all cases. Detected logos are described by the classification network's output feature. Descriptor matching is performed in all cases with cosine similarity. Three different state-of-the-art classification architectures, namely VGG16 (Simonyan and Zisserman, 2015), ResNet101 (He et al., 2015) and DenseNet?? (Huang et al., 2016), serve as base for the logo classification stage. All networks are pretrained on ImageNet and afterwards fine-tuned either on the public logo train set or the combination of the public and the Logos in the Wild train data.

In 10 iterations, each of the 10 FlickrLogos-32 train samples for each brand serves as query sample. This allows to assess the statistical significance of results similar to a 10-fold-cross-validation strategy. Figure 7 shows the FROC results for the trained networks including indicators for the standard deviation of the measurements. The detection identification rate denotes the amount of ground truth logos which are correctly detected and are assigned the correct brand. While the baseline is only able to find a minor amount of the logos, our best performing approach is able to correctly retrieve 25 percent of the logos if tolerating only one false alarm every 100 images. As expected, the more recent network architectures provide better results. Also, including the Logos in the Wild data in the training yields a significant boost in performance. Specifically, the larger training dataset has a larger impact on the performance than a better network architecture. Table 3 compares our open set results with closed set results from the literature in terms of the mean average precision (map). We achieve ?? percent of the closed set performance with only one sample for a brand at test time instead of dozens or hundreds of brand samples at training time.

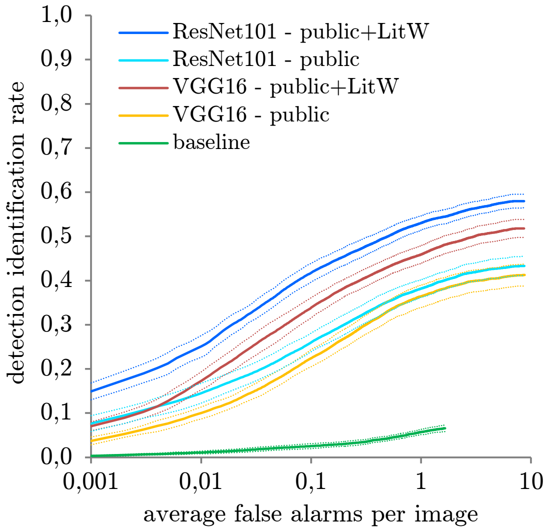

Figure 7: Detection+Classification FROC curve for FlickrLogos-32 test set. Including dashed indicators for one standard deviation.

Table 3: FlickrLogos-32 test set retrieval results.

setting	method	map
open set	baseline (?)	0,036
	VGG16 - public	0,286
	ResNet101 - public	0,327
	VGG16 - public+LitW	0,382
	ResNet101 - public+LitW	0,464
closed set	BD-FRCN-M (Oliveira et al., 2016)	0,735
	DeepLogo (Iandola et al., 2015)	0,744
	Faster-RCNN (Su et al., 2016)	0,811
	Fast-M (Bao et al., 2016)	0,842

In addition to public data, target domain specific experiments are performed on TV broadcasts of sports events. In total, 1,218 annotated frames with more than 10,000 logos from four different events are available in our SportLogo dataset where 3 events are used for training and one as test set. Refer to table 4 for details. In comparison to public logo datasets, the logos are usually significantly smaller in these cases and cover only a tiny fraction of the image area as illustrated in figure 8.

7 CONCLUSIONS

- significant improvement over baseline - enables novel applications - novel large scale in-the-wild logo dataset

Table 4: SportLogo dataset statistics.

	phase	brands	logo images	RoIs
football-1	train	104	331	3,329
ski		27	179	701
ice hockey		19	410	3,920
football-2	test	40	298	2,348

REFERENCES

Bansal, A., Castillo, C., Ranjan, R., and Chellappa, R. (2017). The Do's and Don'ts for CNN-based Face Verification. *arXiv preprint arXiv:1705.07426*.

Bao, Y., Li, H., Fan, X., Liu, R., and Jia, Q. (2016). Region-based CNN for Logo Detection. In *International Conference on Internet Multimedia Computing and Service*, ICIMCS'16, pages 319–322, New York, NY, USA. ACM.

Bäuml, M., Bernardin, K., Fischer, M., Ekenel, H., and Stiefelhagen, R. (2010). Multi-pose face recognition for person retrieval in camera networks. In *International Conference on Advanced Video and Signal-Based Surveillance*. IEEE.

Bianco, S., Buzzelli, M., Mazzini, D., and Schettini, R. (2015). Logo recognition using cnn features. In *International Conference on Image Analysis and Processing*, pages 438–448. Springer.

Bianco, S., Buzzelli, M., Mazzini, D., and Schettini, R. (2017). Deep learning for logo recognition. *Neurocomputing*, 245:23–30.

Dai, J., Li, Y., He, K., and Sun, J. (2016). R-FCN: Object Detection via Region-based Fully Convolutional Networks. *arXiv preprint arXiv:1605.06409*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE.

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

Eggert, C., Winschel, A., and Lienhart, R. (2015). On the Benefit of Synthetic Data for Company Logo Detection. In *ACM Multimedia Conference*, MM '15, pages 1283–1286, New York, NY, USA. ACM.

Girshick, R. (2015). Fast R-CNN. In *International Conference on Computer Vision*.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask R-CNN. *arXiv preprint arXiv:1703.06870*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385*.

Herrmann, C. and Beyerer, J. (2015). Face Retrieval on Large-Scale Video Data. In *Canadian Conference on Computer and Robot Vision*, pages 192–199. IEEE.

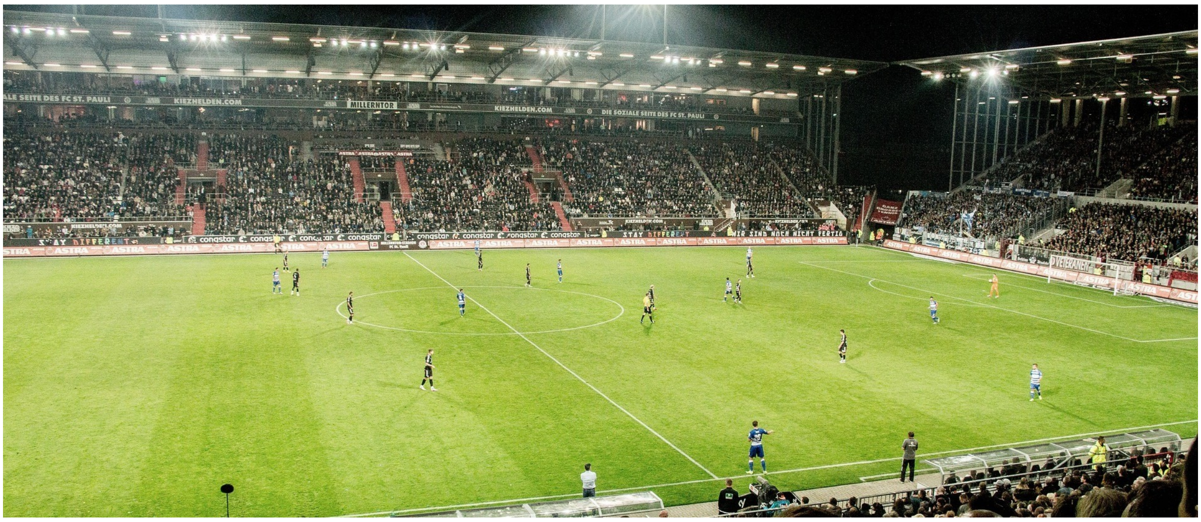

Figure 8: Example football scene with small logos in the perimeter advertising.

Huang, G., Liu, Z., and Weinberger, K. Q. (2016). Densely Connected Convolutional Networks. *CoRR*, abs/1608.06993.

Iandola, F. N., Shen, A., Gao, P., and Keutzer, K. (2015). DeepLogo: Hitting Logo Recognition with the Deep Neural Network Hammer. *CoRR*, abs/1510.02131.

Joly, A. and Buisson, O. (2009). Logo retrieval with a contrario visual query expansion. In *ACM Multimedia Conference*, pages 581–584.

Kalantidis, Y., Pueyo, L., Trevisiol, M., van Zwol, R., and Avrithis, Y. (2011). Scalable Triangulation-based Logo Recognition. In *ACM International Conference on Multimedia Retrieval*, Trento, Italy.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, pages 1097–1105. Curran Associates, Inc.

Letessier, P., Buisson, O., and Joly, A. (2012). Scalable mining of small visual objects. In *ACM Multimedia Conference*, pages 599–608. ACM.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). SSD: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., and Reed, S. E. (2015). SSD: Single Shot MultiBox Detector. *CoRR*, abs/1512.02325.

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

Manger, D. (2012). Large-scale tattoo image retrieval. In *Computer and Robot Vision (CRV), 2012 Ninth Conference on*, pages 454–459. IEEE.

Miller, H. (1969). The FROC Curve: a Representation of the Observer's Performance for the Method of Free Response. *The Journal of the Acoustical Society of America*, 46(6(2)):1473–1476.

Oliveira, G., Frazão, X., Pimentel, A., and Ribeiro, B. (2016). Automatic Graphic Logo Detection via Fast Region-based Convolutional Networks. *CoRR*, abs/1604.06083.

Qi, C., Shi, C., Wang, C., and Xiao, B. (2017). Logo Retrieval Using Logo Proposals and Adaptive Weighted Pooling. *IEEE Signal Processing Letters*, 24(4):442–445.

Redmon, J., Divvala, S. K., Girshick, R. B., and Farhadi, A. (2015). You Only Look Once: Unified, Real-Time Object Detection. *CoRR*, abs/1506.02640.

Redmon, J. and Farhadi, A. (2016). YOLO9000: better, faster, stronger. *arXiv preprint arXiv:1612.08242*.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.

Romberg, S., Pueyo, L. G., Lienhart, R., and van Zwol, R. (2011). Scalable Logo Recognition in Real-world Images. In *ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 25:1–25:8, New York, NY, USA. ACM.

Schapire, R. E. (1999). A Brief Introduction to Boosting. In *International Joint Conference on Artificial Intelligence*, IJCAI'99, pages 1401–1406, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *CoRR*, abs/1312.6229.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Sivic, J. and Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, pages 1470–1477. IEEE.

Su, H., Zhu, X., and Gong, S. (2016). Deep Learning Logo Detection with Data Expansion by Synthesising Context. *CoRR*, abs/1612.09322.

Tursun, O., Aker, C., and Kalkan, S. (2017). A Large-scale Dataset and Benchmark for Similar Trademark Retrieval. *arXiv preprint arXiv:1701.05766*.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Weber, M., Bäuml, M., and Stiefelhagen, R. (2011). Part-based clothing segmentation for person retrieval. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, pages 361–366. IEEE.