Karlsruhe Institute of Technology

Fraunhofer IOSB

Logo retrieval in mass data using deep learning

Master thesis

KIT - Karlsruhe Institute of Technology
Fraunhofer IOSB - Fraunhofer Institute of Optronics,
System Technologies and Image Exploitation

Andras Tüzkö

June 14, 2017

Main Advisor: Dipl.-Inform. Christian Herrmann
Co-Advisor: Dipl.-Inform. Daniel Manger

Interaktive Echtzeitsysteme

Statement of authorship

I hereby declare that I have produced this work by myself except the utilities known to the supervisor, that I have labeled all used utilities completely and detailed and that I have labeled all material that has been taken with or without modification from the work of others.

Karlsruhe, June 14, 2017 Andras Tüzkö Andras Tüzkö

Contents

1 Introduction

Advertising with static logos is one of the most important marketing methods. A very effective way to reach a lot of people with these static logos is, to sponsor sport teams or to buy advertising spaces in sport events broadcasted on the TV. However, the prices of these surfaces mean huge expenses for the advertiser. This is the reason why the need for logo appearance statistics of sport videos arises. In particular there is a desire for quantitative measurement of the proportional size of the logo to the screen, and of the time one particular logo is visible on the screen. This data is then used to judge the cost efficiency for the specific logo placement, i.e. to be able to decide on which sport event to advertise, with which size of logo, and where to place it.

In this work a system for logo retrieval with proposal based object detection and classification will be presented. The system consists of a logo detector, and a classifier used for feature extraction. The logo detector is a faster region based convolutional neural network [Ren15] trained to recognize logos on images. For feature extraction from the proposed region, there will be the performance of several classifier neural network tested. To decide the identity of the logo to be searched, and the proposed region, a similarity score will be calculated. To recognize logos in videos, a video is cut into frames, and then the system is run on every image.

The challenge of this task is manifold. The first problem is that the logos in these videos are far from being perfectly clear. They can be partially occluded, blurred - if the camera is moving fast, perspectively transformed, rotated and can have various coloring, suiting well to the design of the shirt or the arena. In addition, there is a problem with the ambient illumination variation just as for other computer vision tasks. Second challenge is the large variety of different brand's logos. This makes the detection of logos very challenging. In addition, a lot of company has a logo only with a name on it. To recognize complete words is a much harder task than only a letter or a simple shape. Furthermore, in classification tasks the objects to be classified have usually a 3d shape in the world, whereas logos have only a planar surface. This means, if the logo is photographed from any angle other then perpendicular to the plane of that, yields only less information. Unfortunately, there are only a few small publicly available datasets, with bounding box annotated logos. The majority of the images are adjusted to ensure a good visibility of the logos on them, not like on the frames of the sport videos.

In the decade before, hand-crafted feature extraction was prevalent in computer vision tasks. It needed an expert to create such a system, and it yielded often only mediocre results. Deep learning methods for computer vision problems are dominant since the success of convolutional neural networks in 2012 [Kri12]. The great improvement of deep learning methods is, compared to earlier systems (e.g. SIFT [Low04], HOG [Dal05] features), to learn how to extract features automatically. The development of deep nets is mainly powered by the annually organized ImageNet classification challenge [Rus15]. Since the aim of this contest is, to classify an object, which is filling out the majority of an image, the location of the particular object is irrelevant. To be able to classify and recognize objects which have a much smaller size relative to the size of the whole image, region based classification can be utilized.

The rest of this thesis is organized as follows. Section 2 reviews the related work within image retrieval, object detection and logo retrieval. In Section 3 the proposal based object detection with convolutional neural networks will be introduced. Section 4 describes the logo retrieval system. Afterwards, Section 5 includes evaluation and comparison of the system with another logo retrieval method. Finally, the last section concludes the work and gives prospects on future work.

2

Related Work

2.1 Image Retrieval

Many technics outside the scope of deep learning exists for images retrieval from videos. SIFT features [Low04] with bag-of-visual-words were used to efficiently get translation invariant descriptors around keypoints by Zisserman [Siv03]. HOG [Dal05] SIFT, HOG blockwise orientation histograms

2.2 Object Detection

Keypoint detection: translation, rotation invariant Viola Jones

2.3 Logo Retrieval

3

Proposal Based Object Detection and Classification

In this section the theoretical overview of the logo retrieval system will be presented. First of all the fully convolutional networks will be introduced in section 3.1. Section 3.2 explains region proposal systems for generating candidate object locations on an image. Afterwards, the section 3.3 describes region based convolutional neural networks for object detection and classification. The improvement of this method, the fast region based convolutional neural networks will be detailed in the section 3.3.2. Folowing this, the further development, the faster region based convolutional neural networks will be reviewed in the section 3.3.3.

3.1 Fully Convolutional Neural Networks

A neural network is fully convolutional, if it does not contain any fully connected layers. Firstly Matan et.al. used FCNs for recognizing strings of digits. Long et. al. proposed [Lon14] how to transform a deep neural network with fully connected classifier layers at the end, to a fully convolutional network. For this purpose the fully connected layers at the end of the network are to be converted to convolutional layers.

Since the number of weights of a neuron in a fully connected layer is defined by the shape of the data of the layer, the trained network can process only a fix-sized input. As a fully convolutional network does not have fully connected layer anymore, it has the advantage of being able to train and test with images of arbitrary sizes.

The outputs of such a network are two dimensional feature maps, which can be used as heatmaps per class. These convolutional maps can also be used directly for semantic segmentation, where each pixel of an image should be classified. Nowadays fully convolutional networks are essential part of state-of-the-art object detectors, yielding better performance, image size agnosticism, as well as shorter training and inference times.

3.2 Region proposal systems

To recognize different objects on an image, like logos, small regions should be considered. The easiest way to search for these locations is the exhaustive sliding window search, applied on multiple scales. Although, as section 2.2 presents, this induces a lot of computational costs. In order to reduce this computational burden, region proposal systems can be utilized. Region proposals are possible object locations on an image.

Earlier computer vision solutions used external proposal systems. This means that the proposals of every image should be pre-calculated before training or inference. One of the most popular region proposal methods is selective search [Uij13]. It merges neighbor regions according to a similarity score in a bottom-up fashion. It processes an image under 2s on the CPU, which precludes the possibility of real time applications. Edge Boxes [Zit14] are efficiently calculating the number of contours in a box, and ranking them according to that almost real time. Today, as section 3.3.3 introduces, the proposal system is already part of the neural network.

3.3 Region Based Convolutional Neural Networks

This section gives a brief overview about how faster region based convolutional neural networks evolved.

3.3.1 Regions with Convolutional Neural Network Features

Although this network is already historical, it is worth to mention it, because it helps to understand the improvements of the later systems. Region based convolutional neural networks [Gir13] consist of four separate systems. Firstly, region proposals are generated external with selective search. There will be altogether 2000 object positions considered. Secondly, each region of the possible object locations is warped to a size of 227x227, and then the feature vector of every single region is extracted with a CNN. The network is pretrained on the ImageNet dataset [Kri12], and then fine-tuned on the final classes. The network is run on every region proposal bounding boxes, to extract vectors with a fixed-size. These vectors will to be written to the disk. Thirdly, a set of class-specific linear SVM is used to classify the specific region. At last bounding box regressions is run, to reduce the mislocalization of the object. This happens outside of the network.

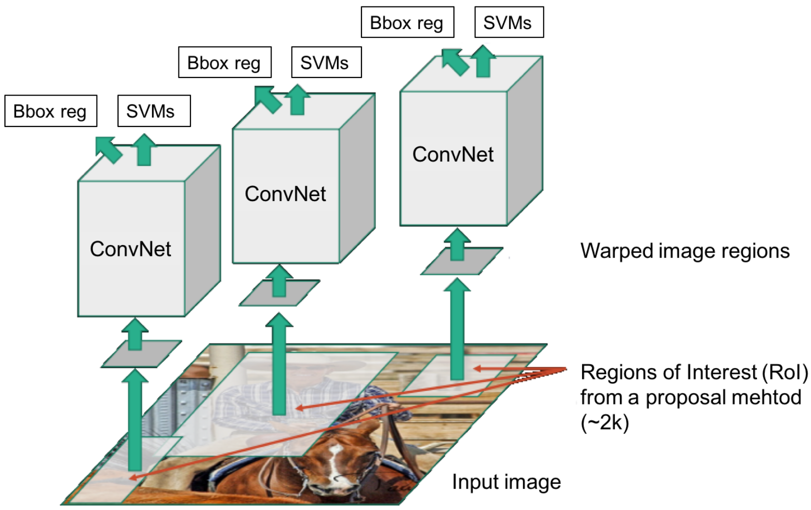

Figure 3.1: R-CNN takes external region proposals, warps the region to a uniform shape, and extracts the features separately. The extracted features are then used to classify the region, and calculate bounding box regression externally.

3.3.2 Fast Region Based Convolutional Neural Network

Fast Region based CNNs [Gir16] are aimed to improve the classification accuracy and feature vector extraction speed of the interest regions, generated also with selective search. An intermediate convolutional feature map is extracted from the whole input image with a fully convolutional neural network, also called as base network in [Ser13]. The output is a downscaled feature map, which is fed to the so called RoI (region of interest) pooling layer. This layer crops regions from the map according to the appropriate downscaled region proposals, and executes a modified version of max pooling on each regions, which results in a convolutional map with a fixed-shape, regardless the size of the region.

After the pooling, fully connected layers are used to calculate the final class probabilities and bounding box regressions for each region. The output of the bounding box regression are class specific small position and size adjustments, needed to refine the rough object locations.

The improvements of this method compared to the previous region based CNN introduced in section 3.3 are as follows:

- **Joint feature extraction:** much shorter training and inference time is achieved by the lower computational redundancy of running convolutional layers on the whole image only once, rather then for every proposed regions.
- **One network:** the feature extraction and the classification happens in the same network. This has more advantages:
  - This results again in faster test and training times, due to the unnecessity of writing the extracted feature vectors to disk, which incidentally could require hundreds of gigabytes of storage [Gir16] for the VOC07 trainval set [Eve].
  - As the backpropagation is implemented through the RoI pooling layer, the whole network, together with the convolutional layers, can be trained jointly, against earlier implementations, like R-CNN [Gir13] or spatial pyramid pooling networks (SPPnet) [He14].
- **Minibatch from a few images:** Faster training speed is achieved by collecting a minibatch only from two images, rather than every region from different images. This method is proved to converging within similar times, despite the high correlated regions.

The network is trained with a multi-task loss function for classification and bounding box regression, defined as:

$$L(p,u,t^u,v) = L_{cls}(p,u) + \lambda[u \geq 1] L_{loc}(t^u,v) \tag{3.1}$$

where p is the computed class probabilities, u is the groundtruth class, t^u is the predicted bounding box offsets for every classes, and v is the groundtruth bounding box position and size. Since the probabilities are calculated with softmax as usual, the log loss is used for the classification error: $L_{cls}(p,u) = -log p_u$. For bounding box regression loss, a smooth version of L1 loss is used, which is defined as follows:

$$L_{loc}(t^u,v) = \sum_{i \in (x,y,w,h)} smooth_{L_1}(t_i^u, v_i) \tag{3.2}$$

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{3.3}$$

is As the background class has the

The network is trained for K+1 classes, where K is the number of object classes, and the background is also modelled as a separate class. During training, the positive examples are chosen, regarding the intersection over union (IoU) value to the groundtruth. This value is widely used for measuring the overlapping between regions, regardless the actual size of the regions. The calculation between two regions, R_1 and R_2 is as follows:

$$\frac{area(R_1 \cap R_2)}{area(R_1 \cup R_2)} \tag{3.4}$$

For positive training examples there are thoose regions applied, which have an IoU with the groundtruth at least 0.5. For the background class are the examples with IoU [0.1,0.5) used.

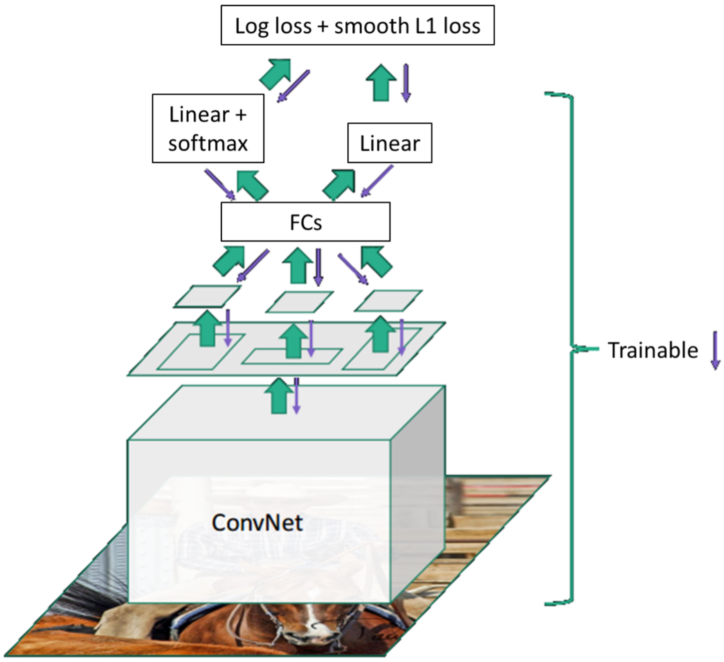

Figure 3.2: Fast R-CNN uses external proposals, then infers the complete image with a fully convolutional network. The proposals is then used to crop regions from the feature map with RoI pooling. The cropped region is then classified and the region coordinates are adjusted with a fully connected network.

3.3.3 Faster Region Based Convolutional Neural Network

A great disadvantage of the Fast R-CNN method is, that the region proposals are generated externally. Girshick et.at. introduces Faster R-CNN [Ren15], which generates the interest regions within the neural network nearly cost-free (10ms pro image). This system consists of a region proposal system and a Fast R-CNN object detector.

Region Proposal Network

The convolutional feature map, extracted by the base network, is processed by the RoI pooling layer. Additionally a thin fully convolutional network, the region proposal network (RPN), is also utilized on the convolutional maps, to generate the proposals. Reference boxes, called "anchors" are generated at every position of the conv feature map, in different scales and different aspect ratios. This ensures the scale invariance of the objects. A convolutional layer with 3x3 kernel extracts a fixed-size vector from every window of the conv map, where in each window 9 anchors are considered. As the fully convolutional network iterates through the conv map in a sliding window fashion, translation invariance is granted. An objectness score and bounding box offset is then calculated for every anchor, by different classifiers and regressors, specialized in a specific scale and aspect ratio.

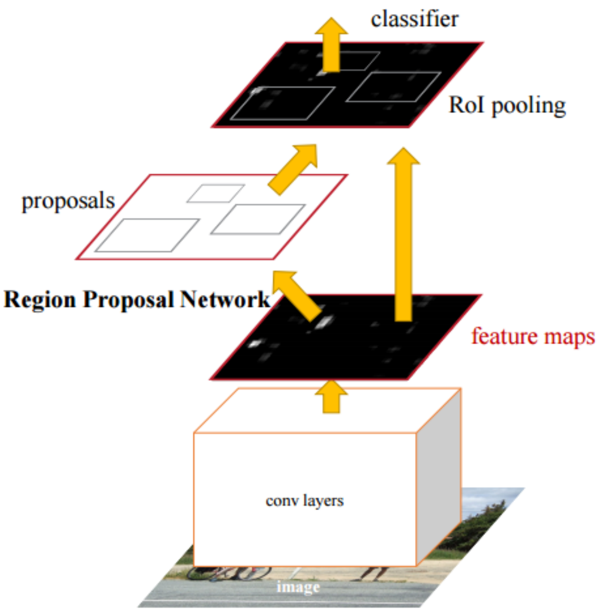

Figure 3.3: Faster R-CNN consists of a Fast R-CNN and a region proposal network (RPN). Fast R-CNN is responsible for the feature map extraction from the whole image, and classify regions from that. The RPN is an in-network implemented proposal system, for generating candidate object locations in a fast way.

Training

As the RPN is a class agnostic object detector, it should detect all the types of objects, which the network is trained on. For this purpose, the class informations, during training of RPN, can be discarded. Positive examples are collected from the proposals with an IoU higher than 0.7 with any groundtruth. Negative label is assigned for regions, which have an IoU, lower than 0.3.

Since the objective of the RPN is the same as for Fast R-CNN, namely to classify regions and regress bounding box coordinates, the loss functions for training the RPN can the same multi-task loss, which are used for training a Fast R-CNN, detailed in section 3.3.2.

Earlier versions of the system suggested to train the complete system in an alternating way. First the RPN will be trained with a CNN, pretrained on e.g. the ImageNet dataset. The trained RPN will then be used to propose regions, and train the Fast R-CNN part with them. The CNN of the fine-tuned Fast R-CNN is then used to train the RPN again, and this process is repeated iteratively. Later, it has been proven, that training the whole network of the faster R-CNN can be performed in an end to end manner, with marginal performance drop. This means, that all the parts of the system can be trained jointly.

4

Logo Retrieval System

The difficultness of the problem logo retrieval was highlighted in the introduction, in chapter 1. Nowadays, deep neural networks are utilized to solve such difficult computer vision problems. In this work, end to end neural network solutions will be used, to retrieve logos from videos. The set of logos to be searched is called the query set, the frames of videos, where the logos should be retrieved from, is called the search set.

supervised pre-training on a large auxiliary dataset (ILSVRC), followed by domain- specific fine-tuning on a small dataset (PASCAL), is an effective paradigm for learning high-capacity CNNs when data is scarce. RCNN paper

4.1 Logo Datasets

The hunger of deep learning methods for training data is well-known. As the publicly available logo datasets are relatively small, a better training result can be achieved if the datasets are merged. The different logo datasets with the number of brands, images and bounding box RoIs can be seen in table 4.1. The total number of brands means the number of different brands altogether.

This training data is insufficient to train a network from scratch (with randomly initialized weights). Girshick et. al. showed [Gir13], that initializing the weights of the network from a CNN, which was trained on a not related dataset (e.g. ImageNet classification), and fine-tuning that on the target dataset, can boost on performance significantly. It is because of the hierarchical learning of shapes by the layers of the convolutional neural network. As a result the learning of the first several convolutional layers can be switched off during fine-tuning.

There are also trademark datasets available having a much greater cardinality [Tur17]. The images of this dataset contain however only the logo of a company, without any context. This dataset turned out to have no use for region based deep learning methods, since this approach needs to learn to distinguish between objects to be learned and the background. The network was trained with the fusion of FlickrLogos-32 and the trademark dataset, and tested with the evaluation method of FlickrLogos-32.

Table 4.1: Publicly available logo datasets with with bounding box annotations

	Number of brands	**Number of logo images**	**Number of RoIs**
BelgaLogos [Jol09], [Let12]	37	1321	2697
FlickrBelgaLogos [Let12]	37	2697	2697
Flickr Logos 27 [Kal11]	27	810	1261
FlickrLogos-32 [Rom11]	32	$70 \cdot 32 = 2240$	3404
Logos-32plus [Bia17], [Bia15]	32	7830	12300
TopLogo10 [Su16]	10	$10 \cdot 70 = 700$	863
Total	**80 (union)**	15 598	23 222

4.2 Logo Detection

There are a lot of possibilities to search for objects in an image as explained in section **??**. Keypoint detectors and external proposal systems are translation and rotation invariant, but usually these systems cannot be trained on a specific dataset. Girshick et.al. proposed the Faster R-CNN neural network, detailed in section **??**, for end to end learning to detect and classify objects on an image. This network has a bounding box regressor for each trained class, so it is capable to produce object type specific region proposals.

In the following sections, different detectors will be introduced.

4.2.1 Region Proposal Network

During the training of a Faster R-CNN network, a region proposal network will be trained to detect the objects, which it was trained on.

4.2.2 Class Agnostic Faster R-CNN

Faster R-CNN can be trained for two classes: background and logo. For this purpose the classes of every annotation box will be neglected.

4.2.3 Jointly Trained Detector and Classifier

Both a classifier and a detector can be built together in a faster R-CNN network. For this purpose, the feature extraction and the region proposal network can be shared between the two tasks. The easiest way to train such a network is a siamese like faster rcnn.

A siamese network [Had06] is basically used for calculating the similarity score between two inputs. If it is combined with an appropriate loss function e.g. contrastive loss [Had06] or max margin loss [Sim13] [Her16], the network can be trained with an image pair and a label indicating if the objects on the images are from the same category or not. The network then learns to project objects from the same category with a low-, otherwise with a large distance to each other, according to a metric. This is achieved by the sharing of weights of the feature extraction layers.

For faster RCNN, the parameters for region proposal network can also be shared. This setup has the advantage, that training of one task can also improve the performance of the other task, which is not currently trained. In particular for logo retrieval, the network can benefit from a bounding box annotated logo dataset, without specific brand indication. To annotate such a dataset needs much less human resources. A quantitative evaluation can be found in section **??**.

4.3 Logo Comparison

In order to retrieve as much objects from images as possible, the detectors should work with a high recall. Although, for difficult tasks, like open-set logo detection high recall value induces low precision, thus a lot of false positive possible object locations are produced. These examples should be eliminated by the classifier. RESNET: [He15] Krizhevsky?s CNN can be used (without fine- tuning) as a blackbox feature extractor, yielding excellent performance on several recognition tasks work by Donahue et al. [Don13]

5

Experiments

Precision and recall are favoured values in image retrieval. Precision is the fraction of relevant retrieved objects and all the retrieved objects. Recall is the ratio of relevant retrieved objects to all the relevant objects. Although, many retrieval system is capable to return a ranked list of the retrieved objects, precision and recall ignore this information. Thus, the trained models are evaluated with the nowadays very popular mean average precision (mAP) metric. In particular, a descending sorted list will be created for every company, based on the probabilities of being logos from the specific brand on given positions of the images. Firstly the precision curve as a function of recall is acquired for every list. The average precision is then calculated as the area under the precison-recall curve. The average of these values gives the mean average precision. All the mentioned results will be calculated with the evaluation implementation of py-faster-rcnn [Gir17] [Ren15].

5.1 Training with Synthetic Data

5.1.1 METU Trademark dataset

To try to increase the size of training dataset, a synthetic dataset was generated, where one logo from the METU Trademark dataset [Tur17] was placed on an image. As basis, images from Tripadvisor were used. There are some transformations, which were applied on the logo images before. The majority of the logo's background has a white color. Thus one third of the dataset is left original. The brightness of the rest of them was adjusted to the brightness of the image on which the logo is placed, and for one third of the logos the mean hsv value of the logo is calculated and rotated with 90 degree chosen randomly. The table **??** summarizes the applied transformations.

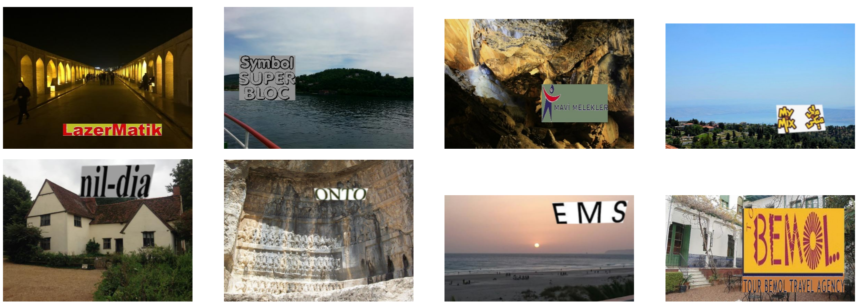

Figure 5.1: Generated synthetic logo images

5.1.2 FlickrBelgaLogos dataset

5.2 Logo Detection

5.3 Logo Retrieval

6 Conclusion

6.1 Summary

6.2 Future Work

Bibliography

[Bia15] Bianco, Simone; Buzzelli, Marco; Mazzini, Davide und Schettini, Raimondo: Logo recognition using cnn features, in: *International Conference on Image Analysis and Processing*, Springer, S. 438–448

[Bia17] Bianco, Simone; Buzzelli, Marco; Mazzini, Davide und Schettini, Raimondo: Deep learning for logo recognition. *Neurocomputing* (2017), Bd. 245: S. 23–30, URL `http://www.sciencedirect.com/science/article/pii/S0925231217305660`

[Dal05] Dalal, Navneet und Triggs, Bill: Histograms of Oriented Gradients for Human Detection, in: *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, IEEE Computer Society, Washington, DC, USA, S. 886–893, URL `http://dx.doi.org/10.1109/CVPR.2005.177`

[Don13] Donahue, Jeff; Jia, Yangqing; Vinyals, Oriol; Hoffman, Judy; Zhang, Ning; Tzeng, Eric und Darrell, Trevor: DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. *CoRR* (2013), Bd. abs/1310.1531, URL `http://arxiv.org/abs/1310.1531`

[Eve] Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J. und Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html

[Gir13] Girshick, Ross B.; Donahue, Jeff; Darrell, Trevor und Malik, Jitendra: Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR* (2013), Bd. abs/1311.2524, URL `http://arxiv.org/abs/1311.2524`

[Gir16] Girshick, Ross; Donahue, Jeff; Darrell, Trevor und Malik, Jitendra: Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* (2016), Bd. 38(1): S. 142–158, URL `http://dx.doi.org/10.1109/TPAMI.2015.2437384`

[Gir17] Girshick, Ross: py-faster-rcnn, `https://github.com/rbgirshick/py-faster-rcnn` (2017)

[Had06] Hadsell, Raia; Chopra, Sumit und Lecun, Yann: Dimensionality reduction by learning an invariant mapping, in: *In Proc. Computer Vision and Pattern Recognition Conference (CVPR?06*, IEEE Press

[He14] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing und Sun, Jian: Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *CoRR* (2014), Bd. abs/1406.4729, URL `http://arxiv.org/abs/1406.4729`

[He15] HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing und SUN, Jian: Deep Residual Learning for Image Recognition. *CoRR* (2015), Bd. abs/1512.03385, URL `http://arxiv.org/abs/1512.03385`

[Her16] HERRMANN, Christian; WILLERSINN, Dieter und BEYERER, Jürgen: Low-Quality Video Face Recognition with Deep Networks and Polygonal Chain Distance, in: *Proceedings of the 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, IEEE, Gold Coast, Australia, S. 244–250

[Jol09] JOLY, Alexis und BUISSON, Olivier: Logo retrieval with a contrario visual query expansion, in: *MM '09: Proceedings of the seventeen ACM international conference on Multimedia*, S. 581–584

[Kal11] KALANTIDIS, Y.; PUEYO, LG.; TREVISIOL, M.; VAN ZWOL, R. und AVRITHIS, Y.: Scalable Triangulation-based Logo Recognition, in: *in Proceedings of ACM International Conference on Multimedia Retrieval (ICMR 2011)*, Trento, Italy

[Kri12] KRIZHEVSKY, Alex; SUTSKEVER, Ilya und HINTON, Geoffrey E: ImageNet Classification with Deep Convolutional Neural Networks, in: F. Pereira; C. J. C. Burges; L. Bottou und K. Q. Weinberger (Herausgeber) *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc. (2012), S. 1097–1105, URL `http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf`

[Let12] LETESSIER, Pierre; BUISSON, Olivier und JOLY, Alexis: Scalable mining of small visual objects, in: *Proceedings of the 20th ACM international conference on Multimedia*, ACM, S. 599–608

[Lon14] LONG, Jonathan; SHELHAMER, Evan und DARRELL, Trevor: Fully Convolutional Networks for Semantic Segmentation. *CoRR* (2014), Bd. abs/1411.4038, URL `http://arxiv.org/abs/1411.4038`

[Low04] LOWE, David G.: Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision* (2004), Bd. 60(2): S. 91–110, URL `http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94`

[Ren15] REN, Shaoqing; HE, Kaiming; GIRSHICK, Ross und SUN, Jian: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, in: C. Cortes; N. D. Lawrence; D. D. Lee; M. Sugiyama und R. Garnett (Herausgeber) *Advances in Neural Information Processing Systems 28*, Curran Associates, Inc. (2015), S. 91–99, URL `http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-network pdf`

[Rom11] ROMBERG, Stefan; PUEYO, Lluis Garcia; LIENHART, Rainer und VAN ZWOL, Roelof: Scalable logo recognition in real-world images, in: *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, ACM, New York, NY, USA, S. 25:1–25:8, URL `http://www.multimedia-computing.de/flickrlogos/`

[Rus15] RUSSAKOVSKY, Olga; DENG, Jia; SU, Hao; KRAUSE, Jonathan; SATHEESH, Sanjeev; MA, Sean; HUANG, Zhiheng; KARPATHY, Andrej; KHOSLA, Aditya; BERNSTEIN, Michael; BERG, Alexander C. und FEI-FEI, Li: ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* (2015), Bd. 115(3): S. 211–252

[Ser13] Sermanet, Pierre; Eigen, David; Zhang, Xiang; Mathieu, MichaÃ«l; Fergus, Rob und LeCun, Yann: OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *CoRR* (2013), Bd. abs/1312.6229, URL `http://dblp.uni-trier.de/db/journals/corr/corr1312.html#SermanetEZMFL13`

[Sim13] Simonyan, K.; Parkhi, O. M.; Vedaldi, A. und Zisserman, A.: Fisher Vector Faces in the Wild, in: *British Machine Vision Conference*

[Siv03] Sivic, Josef und Zisserman, Andrew: Video Google: A Text Retrieval Approach to Object Matching in Videos, in: *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, IEEE Computer Society, Washington, DC, USA, S. 1470–, URL `http://dl.acm.org/citation.cfm?id=946247.946751`

[Su16] Su, Hang; Zhu, Xiatian und Gong, Shaogang: Deep Learning Logo Detection with Data Expansion by Synthesising Context. *CoRR* (2016), Bd. abs/1612.09322, URL `http://arxiv.org/abs/1612.09322`

[Tur17] Tursun, Osman; Aker, Cemal und Kalkan, Sinan: A Large-scale Dataset and Benchmark for Similar Trademark Retrieval. *CoRR* (2017), Bd. abs/1701.05766, URL `http://arxiv.org/abs/1701.05766`

[Uij13] Uijlings, J.R.R.; van de Sande, K.E.A.; Gevers, T. und Smeulders, A.W.M.: Selective Search for Object Recognition. *International Journal of Computer Vision* (2013), URL `http://www.huppelen.nl/publications/selectiveSearchDraft.pdf`

[Zit14] Zitnick, Larry und Dollar, Piotr: Edge Boxes: Locating Object Proposals from Edges, in: *ECCV*, European Conference on Computer Vision, URL `https://www.microsoft.com/en-us/research/publication/edge-boxes-locating-object-proposals-from-edges/`

List of Figures

List of Tables

Acknowledgment

»Physics is to mathematics as sex is to masturbation«

R.P. Feynman

»In der Informatik geht es genauso wenig um Computer wie in der Astronomie um Teleskope.«

Dijkstra