Automatic Segmentation of Indoor and Outdoor Scenes from Visual Lifelogging

Buhagiar, Juan; Strisciuglio, Nicola; Petkov, Nicolai; Azzopardi, George

Published in:
Applications of Intelligent Systems

DOI:
10.3233/978-1-61499-929-4-194

Publication date:
2018

Applications of Intelligent Systems
N. Petkov et al. (Eds.)
IOS Press, 2018

doi:10.3233/978-1-61499-929-4-194

Automatic Segmentation of Indoor and Outdoor Scenes from Visual Lifelogging

Juan Buhagiar [a], Nicola Strisciuglio [b], Nicolai Petkov [b], and George Azzopardi [b]
[a] *Department of Artificial Intelligence, ICT Faculty, University of Malta*
[b] *Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, the Netherlands*

Abstract. Visual Lifelogging is the process of keeping track of one's life through wearable cameras. The focus of this research is to automatically classify images, captured from a wearable camera, into indoor and outdoor scenes. The results of this classification may be used in several applications. For instance, one can quantify the time a person spends outdoors and indoors which may give insights about the psychology of the concerned person. We use transfer learning from two VGG convolutional neural networks (CNN), one that is pre-trained on the ImageNet data set and the other on the Places data set. We investigate two methods of combining features from the two pre-trained CNNs. We evaluate the performance on the new UBRug data set and the benchmark SUN397 data set and achieve accuracy rates of 98.24% and 97.06%, respectively. Features obtained from the ImageNet pre-trained CNN turned out to be more effective than those obtained from the Places pre-trained CNN. Fusing the feature vectors obtained from these two CNNs is an effective way to improve the classification. In particular, the performance that we achieve on the SUN397 data set outperforms the state-of-the-art.

Keywords. Scene Classification, Indoor and Outdoor, CNN, Transfer Learning.

1. Introduction

Lifelogging is the process of tracking the activity of an individual by any means such as ledgers or technology. This process has been around for decades, but the data was neither always available nor was it automatically interpreted into relevant information. There are different types of sensors that contribute to lifelogging; e.g. the ones that measure the heart beat and the number of steps. Wearable cameras are examples of devices that can be used to perform visual lifelogging by capturing images with a certain frequency. Figure 1 shows an example of such a camera, which can obtain ego-centric images, meaning images from the subject's own point of view. Such images offer new opportunities in many fields such as military strategy, enterprise applications, tourism, intelligent surveillance, medicine, and others [1]. Experiments to characterize the relation between the images and the feelings perceived by the users have also been performed [2].

Here, we propose a method that classifies indoor and outdoor scenes taken from a wearable camera, which captures images (of size 2560×1920 pixels) every 30 seconds. The results of this classification problem may be used in other applications. For instance, one may quantify the time a person spends indoors and outdoors or the time a person

Figure 1. An example of a visual lifelogging camera.

spends socialising. This information can be used to build a journal of indoor and outdoor activities.

2. Related Works

A state-of-the-art method in many domains of computer vision such as object and scene recognition that has become popular in recent years is the adaptation of convolutional neural networks (CNN). Inspired by the visual cortex in animal brains, the neurons of a CNN are learned in such a way that it requires minimal amounts of preprocessing [3]. CNNs have been used in multiple areas such as recommender systems [4], natural language processing [5] and computer vision [6], [7].

The breakthrough of CNNs had, essentially, occurred when Krizhevsky et al [6] proposed a method that uses a CNN for object categorisation and trained it on the large ImageNet data set of around 1.2 million high-resolution images consisting of 1000 different classes. The method obtained an error rate of 37.5% when the most probable class is considered and an error rate of 17% when the top five probable classes are considered.

More related to the problem that we tackle in this work, Zhou et al [8], [9] presented a method for scene categorisation utilising a CNN and the Places data set of seven million labelled images of scenes. The CNN was trained with around 2.5 million labelled images from 365 categories and was tested on 5000 images per category. This resulted in a top 1 error rate of 50% and a top five error rate of 18.9%.

In order to learn all parameters of a CNN, tons of training data have to be used. Since, we do not have the resources to obtain such a large amount of images it is not practical to train a new CNN. To the best of our knowledge, there are no pre-trained CNNs for the application at hand, that of classifying indoor and outdoor scenes from egocentric images. We investigate the use of transfer learning from two pre-trained CNNs. Research has suggested that this method for obtaining features should be the basis of most visual recognition tasks [10]. It has also been shown that even when using CNNs trained for tasks with different type of images, generalisation remains strong [11].

The two data sets mentioned above, ImageNet and Places, have been used to design different CNN architectures, such as AlexNet [12], ResNet [13], VGG16 [14] and GoogleNet [15] and have been made available online. The pre-trained CNNs have also

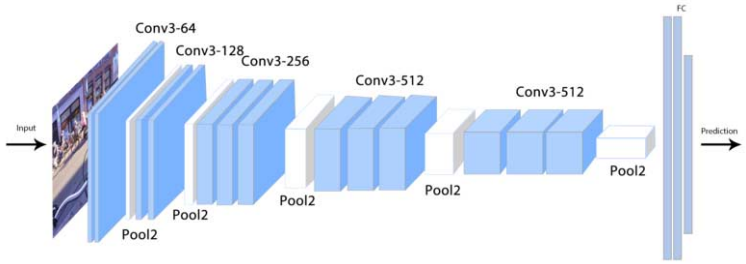


Figure 2. An example of the VGG-16 CNN structure [17].

been evaluated in-depth and compared using different metrics in [16], [9]. The CNN which achieved the highest performance on both data sets is VGG16 [14].

3. Method

3.1. Overview

We propose a method that utilises features, from pre-trained CNNs, to train a classifier that is able to recognise if a given image consists of an indoor or an outdoor scene. We use a VGG-16 structure pre-trained on either the ImageNet data set [6] or the Places [18] data set, which is illustrated in Fig. 2. We extract features by constructing vectors from the FC7 and softmax layers in the CNNs. Different tests were developed to compare the two pre-trained CNNs, the different feature vectors constructed and two different classifiers, namely SVMs and Random Forests.

3.2. Transfer Learning

Transfer learning, also called inductive transfer, is the ability of a learning mechanism to apply a previously learned skill or concept to an existing task. We use transfer learning to avoid the need of training a new CNN. This method allows us to use information learned from solving a previous problem, object classification and scene classification in our case, to solve the problem at hand (i.e. indoor and outdoor image classification). To achieve this we feed images into the pre-trained CNN and form a feature vector from the FC7 layer and the softmax layer. These feature vectors along with indoor and outdoor tags are used as features to train classifiers.

3.3. Forming a feature vector

We initially pre-process all images by resizing them to 256×256 pixels. Next, we normalize each image by subtracting the average image computed on the training set. For each image we apply the VGG-16 CNNs trained on ImageNet and Places data sets and use the output of the fully connected layer (FC7) and of the softmax layer as feature vectors. We extract four different feature vectors: ImageNet softmax layer, ImageNet FC7 layer, Places softmax layer and Places FC7 layer.

Furthermore, we explored different strategies to combine the extracted representations in order to improve the classification accuracy. In the first approach we simply merge the feature vectors into a longer vector and in the second approach we learn a stacking classification model. The first layer of the latter approach consists of two classifiers, one that takes as input the features from the ImageNet-based CNN and the other classifier takes as input the features from the Places-based CNN. Finally, the output of both classifiers is used as input to another classifier. The three classifiers are of the same type; either all SVM or all Random Forest.

4. Experiments and Results

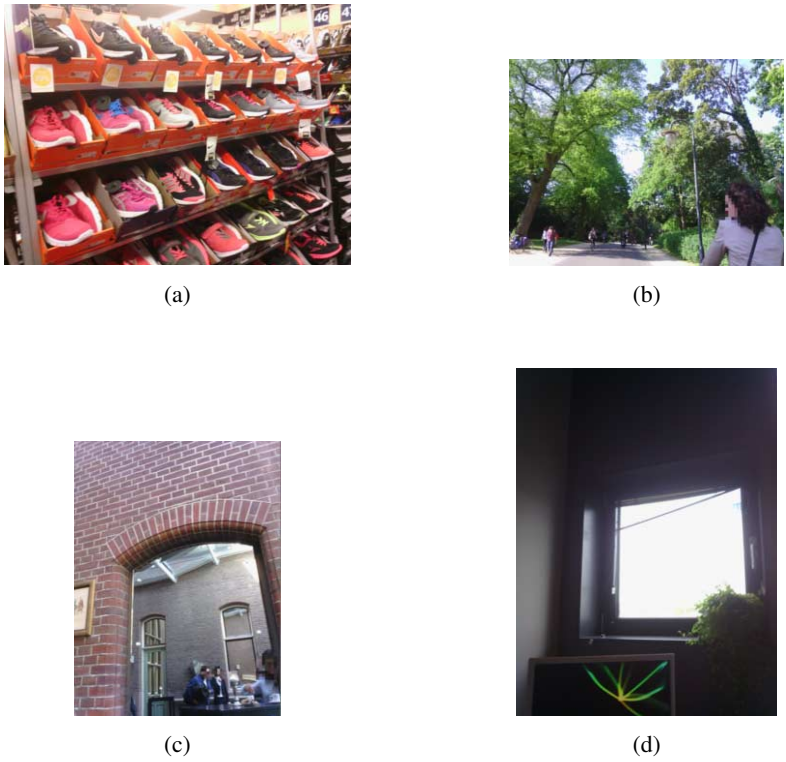
4.1. Data sets

We carried out experiments on two data sets, namely the UBRug and the SUN397 data sets. The UBRug data set contains 997 images with resolutions of 2592×1944 pixels, which have been manually divided into 497 indoor and 500 outdoor scenes. The images were obtained with the narrative clip 2 [19] wearable camera with an ego-centric perspective. The data set contains images of different level of difficulty. The top row of Figure 3 shows images which are easier to classify since they have a clear perspective of the environment. The bottom row of Figure 3 shows images which are harder to classify since the environment is not fully in view or is not clear. A typical situation is when an image that contains the view of a window with only the window pane visible and the outdoor scene.

The SUN397 data set [20] contains a comprehensive collection of annotated images covering a large variety of environmental scenes, places and the objects contained in them. Figure 4 shows some images from the data set and their respective labels. The data set contains 108,754 images in 397 different categories, of which 48,464 are labelled as indoor and 60,290 are labelled as outdoor.

4.2. Experiments

We designed eight experiments for each of the two data sets, four of which use SVM while the other use Random Forest as a classification model. All of the experiments are performed also using the FC7 and softmax feature vectors. Evaluation is performed by computing the average classification rate obtained from a 10-fold cross validation. For all tests we optimise the hyper-parameters of the classifier by repeating the experiment several times and choosing the best performing set of parameters. For each type of classifier we evaluate the performance of the method when only the FC7 layer of the ImageNet pre-trained CNN is considered, when only the FC7 layer of the Places pre-trained CNN is considered, when only the softmax layer of the ImageNet pre-trained CNN is considered, when only the softmax layer of the Places pre-trained CNN is considered, when joining the two sets of FC7 features, when joining the two sets of softmax features and when using a stacking classification approach.



(a) (b) (c) (d)

Figure 3. Example images from the UBRug data set. The bottom row shows examples that are more difficult to classify than the ones in the top row.

4.3. Hyper-parameter tuning for SVM and Random Forest

For the SVM classifier we determined the cost hyper-parameter by a 10-fold cross validation a the set of parameter values 0.001, 0.01,0.1,. . . ,1000. We selected the cost parameter with the highest cross-validation accuracy rate. Similarly, we determined the number of generated trees for the Random Forest classifier by investigating a range of values between 10 and 100 in intervals of 10.

4.4. Results

Table 1 shows the classification accuracies that we achieved. From these tests we see that the classifiers trained on ImageNet features perform generally better than the ones trained with Places features.

After performing concatenation, for every image, we use the feature vector to train an SVM and predict the classes of the images. The fifth and sixth rows of Table 1 show the results obtained from this test. We observe an improvement of classification accuracy, which is attributable to the complementarity of the features extracted with ImageNet and Places based CNNs.

Joining features by concatenation is a naive approach, therefore, we also investigated the possibility of using ensemble methods to further improve our classification. We have

Figure 4. Example images from the SUN397 data set.

identified stacking classifiers as a reasonable choice since it is a simple approach that has achieved positive results. The results, however, show that using stacking classifiers obtains a similar performance to the concatenation of features.

To further improve the results, we performed the same tests as above using the output of an intermediate layer of the CNN called FC7. Tests were run using the features individually as well as their concatenation. We observed that the additional features are more effective than those of the softmax layer.

Besides SVMs, we investigated the performance of a Random Forest classification model [21].

Table 1. Accuracy results using different features and classifiers on the UBRug data set

Method	**Layer**	**SVM Accuracy (%)**	**Random Forest Accuracy (%)**
ImageNet	Softmax	93.04 (±0.005)	96.95 (±0.0022)
ImageNet	FC7	97.49 (±0.005)	97.24% (±0.0035)
Places	Softmax	90.15 (±0.0044)	95.83 (±0.0045)
Places	FC7	96.52% (±0.003)	96.89% (±0.0014)
Vector Concatenation	Softmax	94.69 (±0.0055)	97.15 (±0.0032)
Vector Concatenation	FC7	97.66% (±0.0001)	**98.24%** (±0.0014)
Stacking classifiers	Softmax	94.27 (±0.0043)	97.39 (±0.004)
Stacking classifiers	FC7	**98.24%** (±0.0001)	97.07% (±0.0016)

4.4.1. Error Analysis

We perform an error analysis by looking at the commonly misclassified images on the UBRug data set. Figure 5 shows some images which have been misclassified multiple times by the considered methods. A common problem concerns the fact that these images have an occluded view of the environment because of various reasons such as objects obstructing the view, light reflections, among others.

Figure 5. Examples of misclassified images.

4.4.2. SUN397

We performed tests on the SUN397 data set using the best performing methods on the UBRug data set. Table 2 shows the tests performed and their respective results. Due to the vast amount of images in this data set we did not optimise the hyper-parameters of the classifiers. Albeit, the results are very promising.

5. Discussion

The proposed approach outperforms the state-of-the-art methods on the SUN397 data set [20]. This is mainly attributable to more effective features that where determined from large and diverse data sets, which is in contrast to the hand-crafted features used by Xiao et al [20].

The way we used transfer learning consisted of using the output of pre-trained CNNs as input to other classification models. Even though this is a simple approach it resulted in high accuracy rates. Transfer learning has enabled us to utilise pre-trained CNNs without the need of training it from scratch. This was ideal due to the fact that we avoided the need to obtain millions of training images.

Apart from its effectiveness the method chosen has a very fast processing time where a single image takes less than one second on a computer with a quad core Intel i7 processor (up to 3.6GHz) and 16GB RAM. This is considered as a efficient method.

The methodology that we use is not only applicable for lifelogging scenes but can be applied to other domains, such as photography shots, and other classification problems. As a direction for future research, we aim to investigate the classification of more sub-classes within the indoor and outdoor environments. For instance, it would be useful to determine how much time an individual spends in a bedroom, kitchen or office. Similarly, it would be beneficial to understand the type of outdoor places a subject visits.

Table 2. Comparison of results between the methods of Xiao et al [20] and ours for the SUN397 data set.

Xiao et al. [20]	**Accuracy Rate**	**Ours**	**Accuracy Rate**
All	**94.2%**	ImageNet + SVM	97.06%
Overfeat	94.1%	Joined Concatenation + SVM	**98.13%**
All w/o Overfeat	93.5%	Joined Stacking SVM	95.15%
HOG 2×2	89.9%	Joined Concatenation + RF	97.97%

A major concern in egocentric visual lifelogging applications is the privacy of individuals visible in the scenes. One way of dealing with this challenge is to embed a system that automatically detects and blurs or masks all faces in the scenes, similar to what we did in the illustrations shown above.

6. Conclusions

In this paper, we proposed classification methodologies for indoor/outdoor scene recognition and evaluated the performance of various image representations and classification schemes. We observed that the ImageNet-based CNN features are more effective than those of the Places-based CNN for the problem at hand. Also, FC7 layer features have achieved better results than the softmax layer features. The results improved further when we combined the two sets of features. The two combination methods and the classification models achieved very similar results. For the benchmark SUN397 data set, the methods that we propose achieve better results than those reported in the literature.

References

[1] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, "The evolution of first person vision methods: A survey," *CoRR*, vol. abs/1409.1484, 2014. [Online]. Available: http://arxiv.org/abs/1409.1484

[2] E. Talavera, N. Strisciuglio, N. Petkov, and P. Radeva, "Sentiment recognition in egocentric photostreams," *CoRR*, vol. abs/1703.09933, 2017. [Online]. Available: http://arxiv.org/abs/1703.09933

[3] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[4] A. Van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Advances in neural information processing systems*, 2013, pp. 2643–2651.

[5] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[8] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva, "Places: An image database for deep scene understanding," *arXiv preprint arXiv:1610.02055*, 2016.

[10] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," *CoRR*, vol. abs/1403.6382, 2014. [Online]. Available: http://arxiv.org/abs/1403.6382

[11] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *CoRR*, vol. abs/1411.1792, 2014. [Online]. Available: http://arxiv.org/abs/1411.1792

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1–9.

[16] T. M. Team, "Matconvnet pre-trained imagenet ilsvrc cnns," http://www.vlfeat.org/matconvnet/pretrained/, 2017.

[17] H. El Khiyari and H. Wechsler, "Face recognition across time lapse using convolutional neural networks," *Journal of Information Security*, vol. 7, no. 03, p. 141, 2016.

[18] B. Zhou, A. Khosla, À. Lapedriza, A. Torralba, and A. Oliva, "Places: An image database for deep scene understanding," *CoRR*, vol. abs/1610.02055, 2016.

[19] "Narrative clip 2," 2016. [Online]. Available: http://getnarrative.com/

[20] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 3485–3492.

[21] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.