Scene Detection in *De Boer* Historical Photo Collection

Melvin Wevers
Department of History, University of Amsterdam, Amsterdam, The Netherlands

Keywords: Digital History, Computer Vision, Scene Detection, Digital Heritage.

Abstract: This paper demonstrates how transfer learning can be used to improve scene detection applied to a historical press photo collection. After applying transfer learning to a pre-trained Places-365 ResNet-50 model, we achieve a Top-1 accuracy of .68 and a Top-5 accuracy of .89 on our data set, which consists of 132 categories. In addition to describing our annotation and training strategy, we also reflect on the use of transfer learning and the evaluation of computer vision models for heritage institutes.

1 INTRODUCTION

Computer vision algorithms have become capable of locating and identifying objects represented in images with high accuracy. While this specific technology is commonplace in self-driving cars, drone technology, and the analysis of social media posts, its use for heritage institutes has only recently been growing (Bhargav et al., 2019; Bell et al., 2018; Mager et al., 2020; Niebling et al., 2020). The possible benefits for heritage institutions range from automatically enriching collections, to improving search capabilities, or enabling large-scale analysis of visual collections. The latter is of particular interest for historians with an interest in the visual representations of the past.

While computer vision technology advanced rapidly in the last decade, most computer vision research focuses on use cases that require contemporary data as training material. Exceptions include art-historical research that relies on computer vision (Madhu et al., 2019; Bell and Impett, 2019; Offert, 2018). Training on contemporary material results in models not tuned to detect objects in heritage collections. In other words, the models have difficulty detecting past representations of objects, which often looked noticeably different. Moreover, the list of categories of objects in existing models are not always existent or relevant to heritage collections. In addition to objects changing, the visual medium itself has often-times changed, granting contemporary visual medium a different materiality. Compare, for example, a grainy black and white image of a traffic situation in the 1950s to a high-resolution color image taken with a zoom lens of a highway in 2020. In this instance, the cars will look different, but the improved technological capabilities of the camera also shaped the materiality of the picture. Models that are trained on millions of contemporary images do not have the sensitivity to deal with the color schemes and object representations in older images.

One approach to counter this blind spot in models trained on contemporary data is to *fine-tune* their performance by feeding them with historical material. This method builds upon the categories existent in the modern data sets. However, these categories do not always map onto the categories present in the collections of heritage institutes, or they do not align with the search interests of users of such heritage collections. Another approach is *transfer learning*, which adds new categories to existing models, which are trained using the historical material that has been annotated with these new categories. Because the models have already been pre-trained with large collections of images, we often only require a small number of images to learn a new category. Of course, this depends on the visual complexity of the category and the diversity present in the training data for that category. In other words, an object that always looks the same is easier to learn than one that shared visual aspects but also differs considerably.

Existing research that applies computer vision methods to historical photographs looks into automatically dating images (Palermo et al., 2012), matching images taken from different viewpoints (Maiwald, 2019), photogrammetric analysis of images (Maiwald et al., 2017), and the classification of historical buildings (Llamas et al., 2017).

As part of this research, we set out to discover

Wevers, M.
Scene Detection in De Boer Historical Photo Collection.
DOI: 10.5220/0010288206010610
In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) - Volume 1*, pages 601-610
ISBN: 978-989-758-484-8

what type of computer vision tasks were seen as most relevant in the context of heritage institutions. For this purpose, we conducted fourteen interviews with digital humanities experts, visitors of historical archives, and heritage experts.[1] We discussed several computer vision tasks, from which the respondents showed the most interest in object and scene detection, with a specific interest in tracing buildings. Other tasks such as the estimation of group sizes, facial emotion detection, detection of logos, and posture estimation, the respondents deemed less relevant.

This paper focuses on one particular computer vision task: scene detection. This task tries to describe "the place in which the objects seat", rather than classifying the object depicted on an image, or locating particular objects in a picture by drawing bounding boxes around them (Zhou et al., 2018). More specifically, this paper applies transfer learning to Places-365, a scene detection model trained on contemporary data. We reflect on creating training data, the training strategy, as well as the evaluation of the model. Finally, we discuss the benefits of this type of computer vision algorithm for heritage institutes and visual archives. Rather than merely detecting objects, the ability to search for images based on the scene represented in them is a useful feature for heritage institutions, especially since this information is often not included in the metadata. Using such information, historians could examine the representations of particular scenes, for example, protests, at scale.

2 DATA

This data set for this study is the photo collection of press agency *De Boer* for the period 1945-2004.[2] The *De Boer* collection focuses on national and regional events, although over time the collection's focus gradually shifted to the region Kennemerland. This region, just north of Amsterdam, includes cities such as Haarlem, Zandvoort, Bloemendaal, and Alkmaar. The value of the collection was recognized locally, nationally, and internationally. In 1962, the founder of the press photo agency, Cees de Boer, was awarded the World Press Photo and the Silver Camera (*Zilveren Camera*). He won the World Press Photo for a picture showing Dutch singer Ria Kuyken being attacked by a young bear in a circus. Seven years later,

Figure 1: Arrival of Martin Luther King Jr. at Schiphol Airport, August 15, 1964. Taken from *De Boer* Collection.

Cees' son Poppe won the national photojournalism award. The photo collection depicted a wide range of events, ranging from the first and only show of *The Beatles* in the Netherlands in Blokker—a small village just north of Amsterdam, the arrival of Martin Luther King in 1964 (see Figure 1), to the opening of restaurants, and sports events. The photo press agency was one of the main providers of pictures to the regional newspaper *Haarlems Dagblad*.

The photo collection consists of approximately two million negatives accompanied with metadata for the period 1945-2004. The metadata is extracted from physical topic cards and logs kept by the photo agency. On approximately nine thousand topic cards with over a thousand unique topics, the agency detailed what was depicted in the pictures. For the period 1952-1990, these logs have been transcribed using volunteers.[3]

The archive is currently in the process of digitizing the two million negatives and linking these to the already-transcribed metadata.[4] At the moment of this pilot study, the archive has only digitized a selection of images. This pilot study explores whether transfer learning can be applied to the data and how we should categorize and label the data. During the larger digitization project, we will use the annotation strategy developed in this study to label a selection of images to further improve the scene detection model. This

[1] Due to the Covid-19 pandemic, we had to conduct these interviews virtually. We originally intended to invite more people in person to the archive to conduct face-to-face interviews.

[2] This data has been kindly provided by the *Noord-Hollands Archief*.

[3] https://velehanden.nl/projecten/bekijk/details/project/ranh

[4] https://velehanden.nl/projecten/bekijk/details/project/ranh_tagselection_deboer

model will also be used to automatically label images or detect images that are more difficult to label and which require human annotators. The enriched photo collection will be used to improve the search functionality of the archive. The labeled data and resulting model will be made publicly available to serve as a starting point for other historical photo collections that want to apply computer vision to their collection.

For this pilot study, we relied on a subset of 2,545 images that had already been digitized by the *Noord-Hollands Archief*. Together with archivists and cultural historians, we constructed scene categories for these images.[5] We used the categorization scheme used in Places-365 as a starting point. As the name implies, this scheme contains 365 categories of scenes, ranging from 'alcove' to 'raft'.[6] We combined the categories in Places-365 with categories from the catalog cards that *De Boer* used. These catalog cards could not be linked to the categorization scheme directly, because these cards often contain categories that are too specific or categories that were represented visually. These categories could only be inferred from contextual information but not from the image itself. During the process, we encountered that the creation of categories is always a trade-off between specificity and generalization. In making these categories, we kept in mind whether there remained a historical and visual consistency in the categories and whether the category would be of use for users of the collection. At the same time, we had to make sure we were not too specific, which would impact the amount of training data.

During the annotation process, we noticed the difficulty in distinguishing between scenes that were characterized by a particular object and scenes defined by the location or the action performed in the image. For example, the images in Figure 4 contain specific objects, while the images in Figure 3 depict scenes. Still, the images in the latter also contain objects such as flags, wreaths, cranes, and trees that are not necessarily exclusive to a specific scene. This particular challenge is also discussed by developers of scene detection data sets. The developers of the SUN database, for example, describe how scenes are linked to particular functions and behaviors, which are closely tied to the visual features that structure the space. In the case of large rooms or spacious areas, these allow for different types of behavior, opening up the possibility for different scene categories (Xiao et al., 2010). Moreover, there are instances of large inter-class similarity between scenes, for instance, between the categories 'library' and 'bookstore'. At the same time, we find large intra-class variations, in which the depicted scenes that belong to one category are actually quite diverse (Xie et al., 2020).

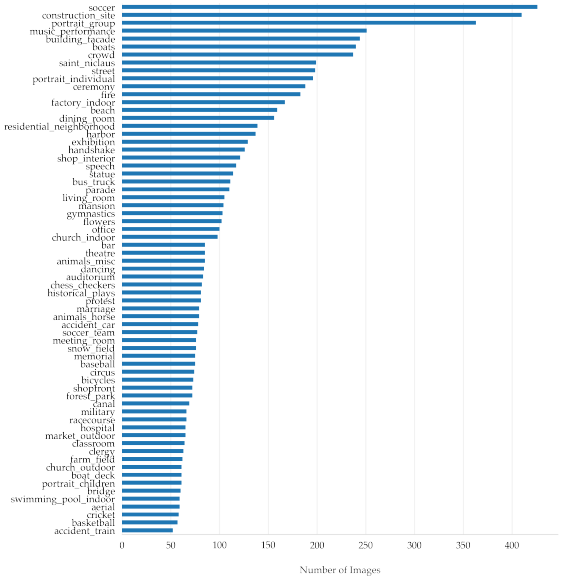

Figure 2: Distribution of categories ($N \geqslant 50$) in the training set.

After annotation, we used an initial baseline model to correct incorrectly annotated images. During the annotation process of the larger dataset, we will iteratively evaluate the training categories and possibly add new categories and merge or remove existing ones. Our resulting dataset includes 115 unique categories that can be found along with their descriptions in the Appendix. The distribution of the categories in our training set is heavily skewed and long-tailed (Figure 2). For training, we removed categories with less than twenty images, this included categories such as: 'houseboat', 'castle', and 'campsite'. However, we intend to again include these categories when we are annotating the newly digitized photographs. The categories 'soccer' and 'construction site', for instance, are very well represented, while most others appeared much more infrequently. Even though transfer learning can produce accurate results with few training examples, more training data, especially in categories with greater visual variations, is to be preferred.

[5] In the Appendix, we added annotation guidelines, which will be used for the annotation of the further data set.

[6] An overview of the categories can be found here: http://places2.csail.mit.edu/explore.html

3 SCENE DETECTION

The authors of the Places-365 data set describe a scene as an "environment(s) in the world, bounded by spaces where a human body would fit" (Zhou et al., 2018). Scene detection is aimed at detecting such an environment, or scene, from visual material. Humans can quickly recognize the functional and categorical properties of a scene while ignoring the objects and locations in a scene (Oliva and Torralba, 2001). Still, this does not mean that humans recognize scenes in an unequivocal manner. Contextual factors that are necessarily part of the image aid humans in describing scenes depicted in the image. Users of heritage collections rely on search not only to discover for representations of particular objects (Petras et al., 2017; Clough et al., 2017). Especially in the context of press photography, photographs often captured a particular historical event or setting. Press photos are highly diverse and more often than not contain more groups of people or objects placed in a particular setting.

In the *De Boer* collection, we find depictions of scenes, such as memorials, construction sites, or parades (see Figure 3). Even though these events are characterized by the presence of particular objects, in many cases, they cannot be described by just a single object or person. At the same time, the collection also includes images that can accurately be described by a single object (see Figure 4). Scene detection, however, is also able to learn such representations.

We decided to not focus on object detection and the segmentation of possible objects, but rather to focus on the image as a whole using scene detection. The category schemes in existing object detection models were not directly applicable to the photographs in this collection. Using existing object detection models did not yield useful categories. For example, for a picture of a shopping street, object detection would identify people, bags, and perhaps a traffic light. To be able to detect objects represented in our photo collection, we would need to draw bounding boxes around these objects and annotate them, which would prove to be a very time-consuming task. Nevertheless, in future work, we would like to explore the development of a framework that would provide historical descriptions based on the relationship between different objects in an image.

3.1 Transfer Learning

For the adaptation of existing scene detection models to our data set, we turn to transfer learning (Rawat and Wang, 2017). This method refers to using "what has been learned in one setting [...] to improve generalization in another setting" (Goodfellow et al., 2016). In our case, we use what has been learned about scenes captured in the Places-365 models to learn how to detect the scenes in our training data using our categorization scheme. Rather than training a model from scratch, transfer learning is known for reaching good accuracy in a relatively short amount of time. Basically, we build upon the information already stored in the places model, which is based on millions of labeled images.[7]

(a) Memorial.

(b) Parade.

(c) Construction Site.

Figure 3: Three photographs with scene labels from *De Boer* collection.

As a starting point, we use a ResNet-50 model—a convolutional neural network of fifty layers— trained on the Places-365 dataset.[8] This dataset consists of 1.8 million images from 365 scene categories. The Places-365 data set builds upon the SUN database,

[7]for code and data see: https://github.com/melvinwevers/hisvis

[8]https://github.com/CSAILVision/places365

(a) Artwork.

(b) Castle.

(c) Train.

Figure 4: Three photographs characterized by a central object from *De Boer* collection.

consisting of 899 categories with short descriptions.[9] For our annotation guidelines, we used the SUN descriptions as a starting point. Given that our data set consists primarily out of black and white images, it is worth noting that this type of image was excluded from the SUN data. Next, we create a random validation set, containing twenty percent of the training images. We use this validation set, to estimate the model's performance in terms of accuracy and its fit. In the context of a heritage institute, scene detection is a meaningful task.

Applied to the Places-365 validation data set, the Places-365 ResNet-50 model reaches a 54.74% top-1 accuracy and 85.08% top-5 accuracy. Top-1 accuracy refers to the percentage of images where the predicted label with the highest probability matches the ground-truth label. Top-5 accuracy calculates whether the ground-truth label is in the top-five predicted labels. Because of the ambiguity between scene categories, top-5 accuracy is generally seen as a more suitable criterion than top-1 accuracy (Zhou et al., 2018). The Places-365 model outperforms the widely-used ImageNet-CNN on scene-related data sets, underscoring the benefit of using tailor-made models for scene detection over more generic models, such as ImageNet. Due to its performance on the places-365, we also turned to the ResNet-50 implementation. We have also experimented with simpler model architectures, but these were less able to capture the complexity of the images.

For our study, we load a pre-trained Places-365 model, which we then tune to our categories using the deep learning framework Fast.AI (Howard et al., 2018). We transfer learn using the One-Cycle method, an efficient approach to setting hyper-parameters (learning rate, momentum, and weight decay), which can lead to faster convergence (Smith, 2018).

To account for the unevenly distributed and often small number of training data and to prevent overfitting, we experiment with different data augmentation methods, including image transformation, label smoothing, MixUp, and CutMix. Overfitting refers to the model adapting too closely to the training data, making it less suitable for working with images the model was not trained on. The image transformations, in this case, include, flipping, rotating, and zooming of the source image. These transformations increase the number of training images the model sees and makes it harder for the model to overfit to a particular image. Labels Smoothing adds a small amount of noise to the one-hot encoding of the labels, making it more difficult for the model to be confident about the prediction. This reduces overfitting and improves the generalizability of the model (Szegedy et al., 2014).

MixUp and CutMix are two relatively recent data augmentation techniques. MixUp basically overlays two different images with different labels. For example, one can have an image of a cat with an overlay of a dog. This makes it more difficult for the model to determine what's in each image. The model has to predict two labels for each image as well the extent to which they are 'mixed' (Zhang et al., 2018). CutMix is an extension of a random erasing technique that

[9] https://groups.csail.mit.edu/vision/SUN/

has often been used in data augmentation. In random erasing, a random block of pixels is removed from the image, making it harder for the model to learn the image. CutMix cuts and pastes random patches of pixels between training images. The labels are also mixed proportionally to the area of patches in the training image. CutMix pushes the model to focus on the less discriminative parts of an image, making it suitable for object detection (Yun et al., 2019). The downside of MixUp and CutMix can be that they make it too difficult for the model to detect features, causing underfitting.

We train the model for a maximum of 35 epochs, with early stopping at five epochs monitoring changes in validation loss. To account for underfitting, we experimented with varying the α parameter of MixUp and CutMix, which controls how much of the augmentation is applied. Too much of the augmentation would make it too difficult for the model to extract generalizable features. Training a model is finding a balance between underfitting and overfitting, which is dependent on the size and complexity of the training data. In the domain of cultural heritage, we are often working with limited sets of annotated training material, making it worthwhile to examine to what extent transfer learning can be applied and how we can cope with overfitting to these limited sets of data. In our use case, we want to model to classify unseen data, which requires a model that can generalize.

4 RESULTS

Our baseline model which only uses image transformations achieves a top-1 accuracy of 0.62 and a top-5 accuracy of 0.88 (see Table 1). Adding MixUp with a low α (0.2) and Label Smoothing slightly improves the top-1 accuracy to 0.68, but these augmentations have no effect on the top-5 accuracy. The addition of CutMix and Label Smoothing has a similar effect with a Top-1 accuracy of 0.67. It is noteworthy that the baseline model already achieves good results, which underscores the power of transfer learning and the use of the Places-365 model as a starting point for more specific scene detection tasks. We can see that in almost nine out of ten cases, the correct result can be found in the top 5 results. For our larger data set, we will again explore to what extent MixUp and CutMix can boost the performance of the model.

There was one category that was never correctly identified, namely 'accident stretcher'. This category depicts people carried away on a stretcher. The category consists of only 6 training images, which are also quite diverse. The category 'funeral' also scores

Table 1: Training Results.

	Top 1-Acc	Top 5-acc
Baseline + Aug	**0.62**	**0.88**
Label Smoothing (LS)	0.61	0.87
MixUp (0.2)	0.62	0.86
MixUp (0.2) + Aug	0.63	0.87
MixUp (0.2) + Aug + LS	**0.68**	**0.89**
MixUp (0.3)	0.63	0.88
MixUp (0.3) + Aug	0.61	0.87
MixUp (0.3) + Aug + LS	0.60	0.86
MixUp (0.4)	0.67	0.89
MixUp (0.4) + Aug	0.61	0.86
MixUp (0.4) + Aug + LS	0.67	0.89
CutMix (0.2)	0.63	0.87
CutMix (0.2) + Aug	0.63	0.87
CutMix (0.2) + Aug + LS	0.64	0.87
CutMix (0.5)	0.62	0.87
CutMix (0.5) + Aug	0.66	0.89
CutMix (0.5) + Aug + LS	0.67	0.89
CutMix (1.0)	0.61	0.86
CutMix (1.0) + Aug	0.63	0.87
CutMix (1.0) + Aug + LS	0.67	0.89

low on precision (0.25) and recall (0.1). This category contains ten training images, but again they are quite diverse and show a strong resemblance to 'church indoor', 'parade', and 'memorial'. We expect this accuracy to improve with more training material. The categories that the model most often confuses with each other includes the categories: 'ceremony', 'handshake', 'portrait children', 'residential neighborhood', 'harbor', and 'portrait group'. These categories were commonly predicted as respectively: 'handshake', 'ceremony', 'portrait group', 'street', 'boats', and 'ceremony'. The predictions offered by the model are sensible since they are for the most part closely related to the correct labels. For example, the distinction between the category street and a residential neighborhood is difficult, and actually, it might be more appropriate to attach both labels to the image. Also, the former might be better described as an object and not a scene, while in some contexts a street can also be an environment in which objects are housed. This example foregrounds the conceptual overlap between an object and scene. Figure 8 shows the images with the top losses, which indicates that among the predictions the correct answer had a low probability. Upon closer inspection, we have to conclude that these predictions are not necessarily wrong. We see, for example, an image that shows a handshake but also people in military attire. This image was labeled as 'military' but classified as 'handshake'. This leads us to conclude that it might be worthwhile for the annotation of the larger dataset to allow multiple labels and creating a multi-label classifier. This type of classifier requires more training

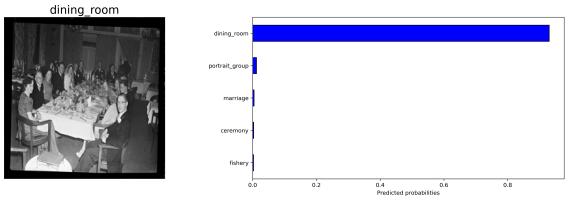

Figure 5: Top-5 predictions for a picture with the label 'Dining Room'.

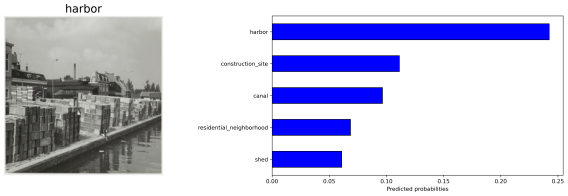

Figure 6: Top-5 predictions for a picture with the label 'Harbor'.

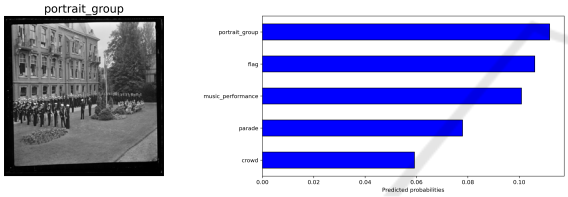

Figure 7: Top-5 predictions for a picture with the label 'Portrait Group'.

data, but in future work, we will explore how much more training data is required in order to reach accurate predictions.

5 DISCUSSION

This paper demonstrated how we annotated a data set of historical press photographs and applied transfer learning to leverage existing scene detection models to a new domain. We highlighted the difficulties of categorization and possible solutions for working with skewed data sets. While our baseline model with basic image transformation already reached good accuracy scores, further augmentations including MixUp and Label Smoothing improved our top-1 accuracy (from 0.62 to 0.68). Our top-5 accuracy was only very slightly improved by these additional augmentations (0.88 to 0.89). The latter indicates that the correct answer was often among the top 5 answers given. While there still exists some ambiguity about labels, in some instances one of these five labels was understandable from a visual perspective, but it might cause some ethical concerns. For example, images of a funeral with lots of flowers were occasionally labelled as 'parade'. Such mistakes are

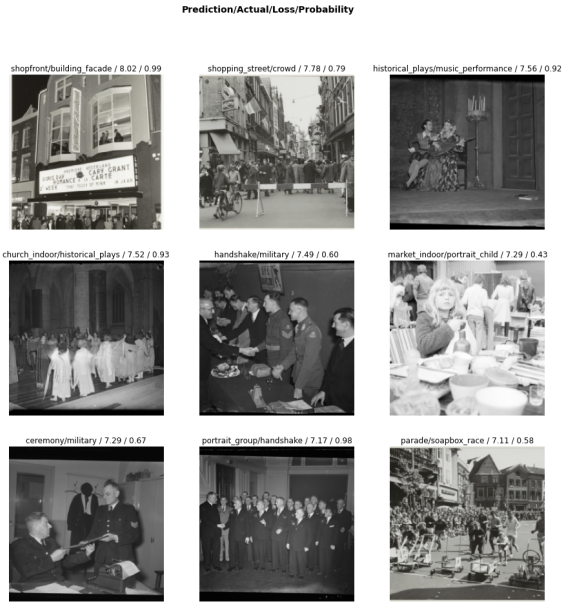

Figure 8: Top Losses.

more impactful than mistaking, for instance, an automobile for a truck. In future work, we want to explore how we can penalize such mistakes; hence, improving the learning process. Furthermore, once more images of the collection have been digitized, we can further refine the presented model and improve and expand our presented categorization scheme.

REFERENCES

Bell, P. and Impett, L. (2019). Ikonographie und interaktion. computergestützte analyse von posen in bildern der heilsgeschichte. *Das Mittelalter*, 24(1):31–53.

Bell, P., Ommer, B., Ommer, B., and Ommer, B. (2018). *Computing Art Reader: Einführung in Die Digitale Kunstgeschichte, P. Kuroczyński et al. (Ed.)*.

Bhargav, S., van Noord, N., and Kamps, J. (2019). Deep Learning as a Tool for Early Cinema Analysis. In *Proceedings of the 1st Workshop on Structuring and Understanding of Multimedia heritage Contents*, SUMAC '19, pages 61–68, Nice, France. Association for Computing Machinery.

Clough, P., Hill, T., Paramita, M. L., and Goodale, P. (2017). Europeana: What users search for and why. In *International Conference on Theory and Practice of Digital Libraries*, pages 207–219. Springer.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. The MIT Press, Cambridge, Massachusetts.

Howard, J. et al. (2018). fastai. https://github.com/fastai/fastai.

Llamas, J., M. Lerones, P., Medina, R., Zalama, E., and Gómez-García-Bermejo, J. (2017). Classification of Architectural Heritage Images Using Deep Learning Techniques. *Applied Sciences*, 7(10):992.

Madhu, P., Kosti, R., Mührenberg, L., Bell, P., Maier, A., and Christlein, V. (2019). Recognizing characters in art history using deep learning. In *Proceedings of the 1st Workshop on Structuring and Understanding of Multimedia heritAge Contents*, pages 15–22.

Mager, T., Khademi, S., Siebes, R., Hein, C., de Boer, V., and van Gemert, J. (2020). Visual Content Analysis and Linked Data for Automatic Enrichment of Architecture-Related Images. In Kremers, H., editor, *Digital Cultural Heritage*, pages 279–293. Springer International Publishing, Cham.

Maiwald, F. (2019). Generation of a benchmark dataset using historical photographs for an automated evaluation of different feature matching methods. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*.

Maiwald, F., Vietze, T., Schneider, D., Henze, F., Münster, S., and Niebling, F. (2017). Photogrammetric analysis of historical image repositories for virtual reconstruction in the field of digital humanities. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:447.

Niebling, F., Bruschke, J., Messemer, H., Wacker, M., and von Mammen, S. (2020). Analyzing Spatial Distribution of Photographs in Cultural Heritage Applications. In Liarokapis, F., Voulodimos, A., Doulamis, N., and Doulamis, A., editors, *Visual Computing for Cultural Heritage*, Springer Series on Cultural Computing, pages 391–408. Springer International Publishing, Cham.

Offert, F. (2018). Images of image machines. visual interpretability in computer vision for art. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175.

Palermo, F., Hays, J., and Efros, A. A. (2012). Dating historical color images. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C., editors, *Computer Vision – ECCV 2012*, pages 499–512, Berlin, Heidelberg. Springer Berlin Heidelberg.

Petras, V., Hill, T., Stiller, J., and Gäde, M. (2017). Europeana–a search engine for digitised cultural heritage material. *Datenbank-Spektrum*, 17(1):41–46.

Rawat, W. and Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, 29(9):2352–2449.

Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay. *arXiv:1803.09820 [cs, stat]*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). Going Deeper with Convolutions. *arXiv:1409.4842 [cs]*.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE.

Xie, L., Lee, F., Liu, L., Kotani, K., and Chen, Q. (2020). Scene recognition: A comprehensive survey. *Pattern Recognition*, 102:107205.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. *arXiv:1905.04899 [cs]*.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018). Mixup: Beyond Empirical Risk Minimization. *arXiv:1710.09412 [cs, stat]*.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2018). Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464.

APPENDIX

Here we list our annotation guides lines per categories. Some of these categories have been not been included in the training, because they included fewer than twenty images.

Accident Car. Traffic accident involving an automobile

Accident Stretcher. Accident involving an person on a stretcher

Accident Train. Traffic accident involving an train

Aerial. Picture with an aerial perspective

Amphitheater. The collection contains many pictures taken at an outside amphitheater in Bloemendaal.

Animals cow. Pictures of live and dead cows.

Animals dog. Pictures of dogs

Animals horse. Pictures of horses

Animals misc. Animals that do not fit in the previous categories. For final dataset, this might be subdivided in more categories.

Artwork. Artwork without people, the focus is on the art work. There is also a separate category for statues.

Auditorium. Public building (used for speeches, performances) where audience sits. Pictures with and without audiences. Overlap with categories ‘conference room‘ and ‘speech’

Bakery. Photos inside bakery, baking bread, presenting bread indoor/outdoors Overlap with Kitchen

Bar. Area where drinks are served/consumed. Overlap with Dining Room/Game Room

Baseball. People playing baseball

Basketball. People playing basketball.

Beach. If the beach is the picture’s main focus, we select beach. Dunes is a separate category. Overlap with

crowd/horse/running. When there is only water visible, we still select Beach. Possibly 'sea' or 'ocean' as a category.

Beauty Salon. Images that feature hairdressers or beauty salons

Bicycles. Features bicycles or people riding bicycles. Overlap with cycling, category for professional cycling and street/

Boat Deck. Picture that focus on the deck of a boat, either with or without people. Not showing entire ship/boat from a distance. Overlap with boats category.

Boats. Category focusing on a boat or multiple boats. Overlap with harbor/shipyard. If boat is harboured and harbor is taking up large area of picture. Shipyard depicts construction area for boats, boats under construction Overlap with beach, canal, river, waterfront. Difference is focus on boat

Bookstore Library. Bookstore or library as a combined category, as they are often difficult to separate. The pictures feature rows of books and/or people reading. Overlap with office, room which also often features books.

Bridge. Picture should feature a bridge as a central element. Overlap with street/canal/river/boat

Building Facade. Depicting the facade of a building/rows of buildings. Not showing the full building from a distance. Overlap with residential area/mansion. Latter are single large house, former pictures of areas without a clear focus on the facade. Overlap shopping area/residential area/mansion

Bus/truck. Focus on large busses and trucks. Overlap with street

Butchers Shop. Butcher shop from inside, or people preparing meat. Overlap with animals/shopfront/shop/kitchen

Canal. Flow of water, also includes natural flow of water rivers. Overlap with bridge/river/boat/fishing/waterfront

Car. Pictures that focus on a car

Car Shop. Showroom in which cars are sold

Catwalk. Models on a catwalk

Cemetery. Pictures taken of a cemetery. Important element includes tombs, or tomb stones.

Ceremony. A group of people bundled together for a ceremony. This could be awards, flowers, or a medal. Note that there is also a specific category for hand shakes.

Chess Checkers. People playing either chess or checkers. These are visually quite similar, for larger dataset, this category might be divided into two, if there are enough images.

Church Indoor. Pictures taken inside a church. This could include masses, but also view of the church without people.

Church Outdoor. Pictures of the church/cathedral building.

Circus. Pictures taken of a circus, inside of a circus tent. The outside of the circus tent is categorized as tent.

City Hall Stairs. Pictures taken of a group of people on the stairs outside of the city hall of Haarlem. This is quite a specific categories, but there are quite a few pictures that fit this category.

Classroom. Students inside a classroom

Clergy. Pictures of clergy indoors or outdoors.

Construction Site. Construction site, this also includes pictures of demolitions. It is quite difficult to separate the two visually. If enough images of both, we could separate the two.

Courtyard. Area between buildings, or outside yard in a group of buildings.

Circus. Pictures taken of a circus, inside of a circus tent. The outside of the circus tent is categorized as tent.

Cricket. People playing cricket

Crowd. Gathering of people, where individuals are not clearly discernible. When posing for picture put in portrait category.

Cycling. Professional cyclists

Dancing. People dancing Overlap with bar, music performance, plaza

Dining Room. Area where people eat, both in restaurants and houses

Drive Way. Drive way in front of buildings

Dunes. Photos of dunes. Overlap with parade/memorial/crowd

Excavation. People digging up something, archeological finds Overlap with construction site

Exhibition. Room with artwork(s) and people. The focus is more clearly on the setting

Factory Indoor. Pictures taken with factories/assembly lines/production facilities

Farm Field. People working in farm fields/pictures of crops/farm fields Overlap with animals

Field Hockey. People playing field hockey

Fire. Pictures of fires, or building destroyed by fires.

Fishery. Pictures of fishing industry

Flag. People holding or raising flags Overlap with parade in which people hold flags

Flowers. Pictures that include flowers

Forest/Park. Picture taken in a forest or park

Funeral. Pictures of a funeral. Similar to memorial and cemetery. Here we only choose pictures that show a casket or burial itself.Overlap with memorial, parade, cemetery, flowers, church indoor

Gymnastics. People performing gymnastics indoor and outdoor Overlap with circus/dancing

Handshake. People shaking hands, subcategory of ceremony.
Harbor. Ships docked at a harbor, focus is not ships, but context of the harbor. Overlap with boats and waterfront ¿ waterfront is focus is on waterfront/kade
Historical Plays. People dressed up in historical garment enacting historical plays
Hospital. Pictures taken in a hospital/dentist/medical lab setting
Ice Skating. People skating on ice
Kitchen. People in kitchen, preparing food Overlap with butcher_store
Living Room. People situated in a living room space. Overlap with portraits, which are often taken in a living room. To learn the living room category, I placed pictures in here taken in living rooms, with enough information on the living room. Overlap with portrait
Mansion. Large houses, separated from housing blocks
Market Indoor. Indoor market/shopping fair
Market Outdoor. Outdoor market Overlap with crowd
Marriage. Depicting a marriage couple, marriage ceremony Overlap with portrait group
Meeting Room. Setting features meeting table with people sitting around it
Memorial. Depicting a memorial site. Overlap with funeral/flowers/flag
Military. Pictures depicting military personnel or military equipment. Overlap with parade and bus/trucks
Motorcycle. Depicting motorcycle(s)
Musical Performance. People performing music Overlap with dance
Office. Pictures set in an office environment
Parade. People parading, marching bands
Parade Floats. Pictures with flower trucks/floats
Patio. People on patio's sitting
Playground. People/Children playing in a playground. Overlap with funfair and portrait children
Pond. Scenery of a pond. Overlap with canal/river/park
Portrait Child. One child
Portrait Children. More children
Portrait Individual.
Portrait Group.
Protest. People protesting, banners clear signal
Racecourse. Pictures taken on racecourse. Also category soapbox race Overlap with car, accident.
Residential Neighborhood. Living area. Overlap with street/building facade
Rowing. People rowing
Running. People running in a sports event
Saint Niclaus. Pictures of the Saint Niclaus festivities
Shed. Wooden buildings, beach houses, living trailers
Shipyard. Construction area for ships. Overlap with boats and harbor.
Shop Interior. Photos of different kinds of shop interiors.
Shop Front. Picture of a shop front, 'etalage'
Shopping Street. Depicting street of shops/shoppers
Sign. Signs, plaques, maps
Snowfield. Scenes set in snow, people skiing, sledding
Soapbox Race.
Soccer. people playing soccer
Soccer Indoor. People playing indoor soccer
Soccer Team. Portrait of soccer team
Speech. Focus on person specking
Statue. Pictures of sculptures/statues
Street. Depicting street sceneries. Overlap with car, residential neighbourhood
Swimming Pool Indoor.
Swimming Pool Outdoor.
Theatre. Plays performed in a theatre setting
Tower. Pictures of towers, not church towers.
Train. Pictures of trains. Overlap with train station
Train Station. Pictures of train stations, trains at stations etc..
Tram. Pictures of trams
Volley ball. Pictures of people playing volley ball
Water Ski. People on water skis
Water Front. Scenes that focus on the water front
Windmill. Pictures that contain a windmill