Neural detection of spheres in images for lighting calibration

Laurent Fainsin, Jean Mélou, Lilian Calvet, Axel Carlier, Jean-Denis Durou

Neural detection of spheres in images for lighting calibration

Laurent Fainsin, Jean Mélou, Lilian Calvet, Axel Carlier, and Jean-Denis Durou

IRIT, UMR CNRS 5505, Université de Toulouse, France

ABSTRACT

Accurate detection of spheres in images holds significant value for photometric 3D vision techniques such as photometric stereo.[1] These techniques require precise calibration of lighting, and sphere detection can help in the calibration process. Our proposed approach involves training neural networks to automatically detect spheres of three different material classes: matte, shiny and chrome. We get fast and accurate segmentation of spheres in images, outperforming manual segmentation in terms of speed while maintaining comparable accuracy.

Keywords: lighting calibration, object detection, DETR, transformers.

1. INTRODUCTION

The digitization of real objects and environments has become a key element in many fields: archaeology, post-production, medicine, etc. Each of these fields has its own constraints and objectives. For example, the 3D digitization of heritage allows to preserve the object of study while simplifying its access to the general public, as well as to professionals who, by having a digital copy faithful to the original, will be able to use it for their analyses.[2]

Photogrammetric methods, which are well known, provide a surface reconstruction from the correspondence between images taken from different viewpoints. However, photogrammetry does not explicitly use the photometric characteristics of the scene. As a result, the quality of the results is highly dependent on the presence of texture on the surface of the objects to be reconstructed. In addition, these methods do not allow the recovery of the real color of the scene.

Photometric methods relate the appearance of a 3D point to the angle between the surface normal at that point and the direction of the incident light.[1] However, these methods require a very precise knowledge of the lighting of the scene. It is therefore necessary to calibrate the lighting during the shooting. To do this, it is customary to position a sphere in the scene, which is visible in each of the images.[2] If the 3D reconstruction method requires changing the camera pose, or if the sphere must be held at the end of a pole when access to the scene is limited, this implies that the position of the sphere varies from one image to another, as illustrated in the example of Figure 1. Manual trimming of the sphere in each image then becomes very tedious. It is also notable that, depending on the lighting estimation algorithm that will be used, the sphere may be matte or glossy.

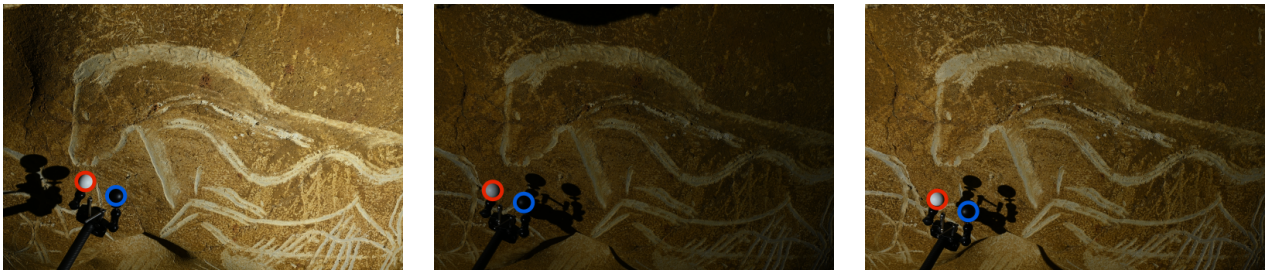

Figure 1. Three images of the "Panneau du cheval gravé" (Chauvet-Pont-d'Arc cave, Ardèche, France), used for 3D reconstruction by photometric stereo. Two calibration spheres are positioned at the end of a pole: a white matte sphere (indicated in red) and a black shiny sphere (indicated in blue). Although the camera pose is constant, the position of the spheres in the image differs from one image to another.

Once the 3D model is obtained, it may need to be re-lit, either for heritage display in a museum or for use in film post-production. Image-based lighting (IBL)[3] is a method that uses a real environment as a light source to re-light the 3D model in a realistic way. To capture this environment, it is customary to use a chrome sphere.

The detection of spheres in images is therefore a recurrent problem in the field of 3D reconstruction. We propose a new approach to this problem, based on the latest advances in neural networks. Our approach also provides a classification of the detected spheres (matte, shiny or chrome), which allows us to use the appropriate algorithm afterwards.

After a brief review of existing methods in Section 2, we describe our approach by first presenting our learning data and then the DETR network, respectively in Sections 3 and 4. We then present our results and applications in Section 5. Conclusion and perspectives are eventually presented in Section 6.

2. STATE OF THE ART

The literature on ellipse detection is extensive. In addition to the fact that ellipses are naturally present in many real scenes, they also appear in images due to the perspective projection of spheres. Many applications, such as camera calibration,[4] medical imaging diagnosis[5] or robotics[6] have made ellipse detection a fundamental problem in computer vision.

Methods based on Hough transform – The Hough transform is used to determine the parameters of ellipses by a voting procedure. This approach is very computationally and memory intensive. Various improvements have been proposed by introducing, for example, randomness.[7] The approximation of the silhouette of a sphere by a circle reduces the number of parameters from five to three, which considerably simplifies the use of the Hough transform and its variants. However, these methods are very sensitive to image noise and to the quantification of the parameters to be estimated.

Edge following methods – Edge tracking methods exploit the connectivity between edge pixels to detect ellipses. Studying the gradients of the image allows arc segments to be detected. Many methods actually detect line segments, from which arc segments can be deduced. Improvements can therefore be obtained in the detection of line segments, as well as in the post-processing. Other methods aim to detect arc segments directly, in order to reduce the computation time.[8]

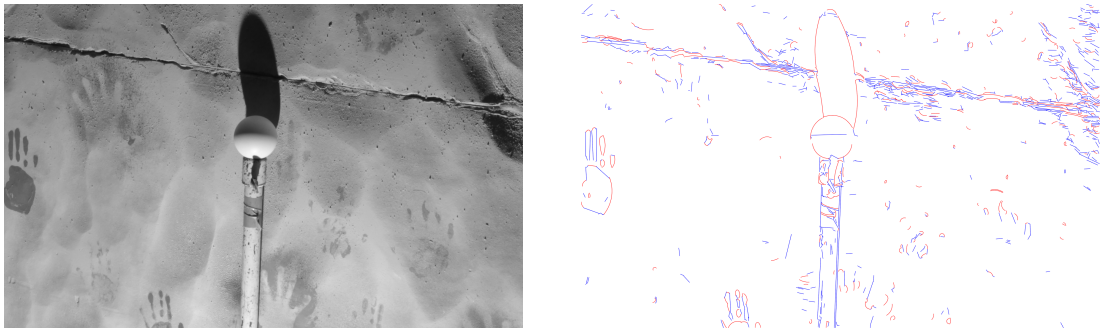

Figure 2. Result of the ELSD[8] detector. The input image (left) needs to be cropped or reduced in order to be processed. The detection (right) contains many false positives.

We recall that we ultimately aim to provide a very accurate 3D reconstruction. Since photometric stereo allows for accuracy of the order of the pixel size, the images used as input are generally very large, and cannot be processed by the algorithms we have tested. Moreover, once the images have been reduced, or cropped, many false positives appear and the results call for post-processing, as shown on the right of Figure 2.

Deep learning methods – Object detection and classification are problems for which neural networks have proven to be very effective, especially convolutional neural networks (CNNs). Region-based CNNs perform an object detection step, the result of which is a bounding box, followed by a classification step. Resulting from a series of improvements, Faster R-CNN[9] shows very good results but this two-step approach makes it rather slow. Ellipse R-CNN[10] uses Mask R-CNN,[11] an improvement of Faster R-CNN allowing instance segmentation,

for elliptical object detection. The well-known YOLO[12] and its successors offer one-step object detection. They are faster but much less accurate than the previous methods.[13] Recently, new approaches based on transformers show results that are quite comparable in terms of accuracy to Faster R-CNN.[14] They also perform better on larger objects, thanks to the use of global information, through the self-attention process.

Finally, it is notable that the detection of reflective objects such as chrome spheres is generally problematic and is an active area of research.[15]

3. DATASETS

Our first dataset includes archaeological photographs, which are used for 3D reconstruction by photometric stereo for heritage preservation. This dataset consists of 1013 images of lithic objects on dark, rugged backgrounds. Although it serves as a suitable starting point, we identified several limitations to this dataset. First, it contains exclusively white matte spheres and red or black shiny spheres, but no chrome sphere. In addition, the scenes presented are all relatively similar, making the models trained on this dataset likely to overfit.

Subsequently, we acquired a second dataset presented by Murmann et al.[16] which includes over 1000 real scenes, each captured in high dynamic range (HDR) and high resolution, under 25 lighting conditions, for a total of 25375 pictures. This work provides an excellent learning basis for applications such as estimating the illumination of a single image or re-lighting an image. In order to measure ground truth of the incoming illumination, the authors place a chrome sphere and a gray sphere in the scene. Since the corresponding masks for each sphere are also provided, we can use this dataset for sphere detection.

Figure 3. Two images extracted from our datasets. Left: archaeological picture. Right: picture from.[16]

However, both these datasets present a set of scenes under different lighting: many images were virtually indistinguishable, except for variations in the direction of lighting. This may be a barrier to generalization. We proceed to train initial models on the aforementioned datasets; however, performance on entirely new images is suboptimal. These datasets alone are insufficient to allow our model to generalize. Thus, we introduce synthetic datasets in order to limit the risk of overfitting.

We use Blender and high-quality, high dynamic range images (HDRIs) obtained from PolyHaven* to generate a synthetic dataset featuring fully consistent lighting and reflections. This method involves rendering spheres in various colors (selected from the McBeth chart) and materials such as matte white, matte grey, shiny black, shiny red, shiny cyan, and chrome. The process of generating a scene requires loading a random HDRI, pointing the camera in a random direction, and generating multiple spheres within the viewing range at varying distances and scales while simultaneously checking for collisions or occlusions. Despite its relatively slow execution, with render times reaching up to 25 seconds, we were still able to quickly generate 36455 images using this method. Consequently, this facilitated the expansion of our learning domain, akin to a form of weak supervision.[17]

*https://polyhaven.com/hdris

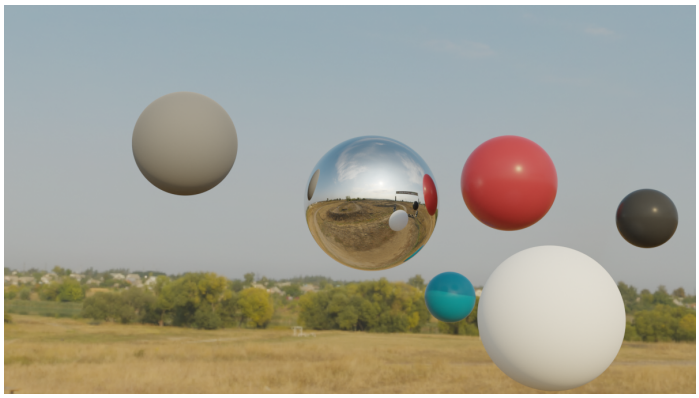

Figure 4. Image generated by Blender, extracted from our dataset. Rendering allows to generate chrome spheres consistent with the surrounding environment.

4. NEURAL NETWORK

We opt to experiment with the DETR model due to its simplicity of implementation, well-established state of the art performance, and broad framework support with pre-trained weights.[14] We use the Hugging Face implementation with a ResNet50 backbone pre-trained on ImageNet. DETR end-to-end transformer encoder-decoder architecture with a set-based global loss streamlines the detection pipeline by eliminating the need for many hand-designed components such as anchor generation and non-maximum suppression procedures.

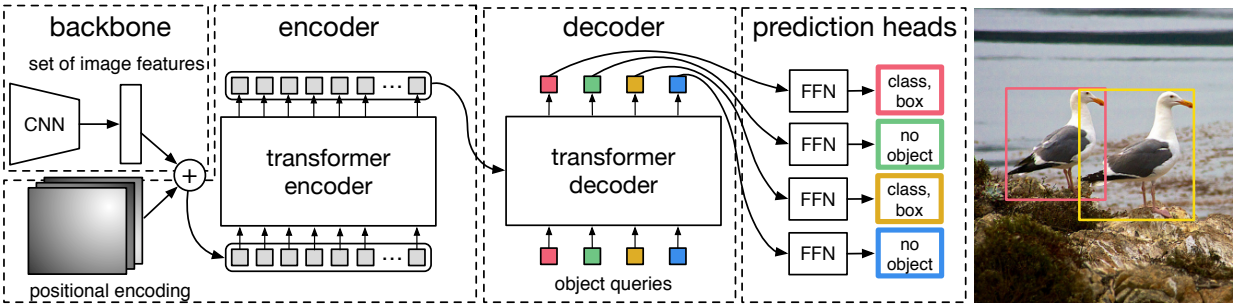

Figure 5. The DETR[14] architecture.

The DETR model utilizes various loss functions to train the network. One of the loss functions used is the cross-entropy loss, which is applied to the class prediction outputs of the model. Another loss function used is the bounding box L1 loss, which measures the discrepancy between the predicted bounding boxes and the ground truth bounding boxes. Additionally, the Generalized Intersection over Union (GIoU) loss is used to calculate the accuracy of the predicted bounding boxes. Finally, the cardinality-based error is used to penalize the model for missing or erroneous object detection. Together, the weighted sum of these loss functions enable the model to learn and improve its object detection performance during training.

5. TRAINING AND RESULTS

Training data – On our various datasets, we separate training data and test data as follows:

- Archeological photographs: 1013 images, including 101 test images;
- Murmann et al.:[16] 25375 images, including 253 test images;
- Synthetic dataset: 36455 images, including 364 test images.

Training hyperparameters and results – For training we use the AdamW optimizer with a learning rate of 1.10^{-4}, a backbone learning rate of 1.10^{-5}, a weight decay of 1.10^{-4}, a cosine annealing scheduler,[18] set our class number to 3 and our queries number to 100. We also use a train batch size of 6, train for at most 2 epochs and employ a dataloader sampler to ensure the balanced utilization of all datasets. The main metric for evaluating the performance of the DETR model is the GIoU, our final trained model attained GIoU score of 0.9 on authentic test images, which accounts for 10% of the archaeological dataset, indicating a good detection of spheres.

It is important to highlight that the model exhibits a tendency to sometime "under-detect" spheres, as the detections tend to be slightly cropped. However, this has negligible impact in practice on the performance of the light detection task for matte and shiny spheres. On the other hand, this could pose a challenge in the case of chrome spheres, particularly if we intend to unwrap the scene within the reflection.

Figure 6. Deductions from our DETR. Left: the chrome and matt spheres are correctly detected (green and red squares). Right: the matt sphere is correctly detected (red square), while its shadow is not, whereas it has a circular silhouette, thus avoiding a false positive.

Comparison with manual detection – The process of manually outlining objects can be a cumbersome and time-consuming task, which typically requires between 7 and 40 seconds to complete, as indicated by Papadopoulos et al.[19] In contrast, automated detection techniques that leverage machine learning models offer a significant reduction in time and effort required for object detection. Although manual detection may yield higher accuracy in certain scenarios, the trade-off between speed and efficiency afforded by automated detection makes it a viable and practical option for many applications where perfect calibration is not strictly required.

Application to photometric stereo – As an illustration, we propose to use our sphere detection and classification algorithm for 3D reconstruction. Photometric stereo requires precise knowledge of lighting. In the absence of laboratory conditions, it is common to place a sphere in the scene to be reconstructed.

The "Panneau du cheval gravé" (Chauvet-Pont-d'Arc cave, Ardèche, France) cannot be approached without destroying valuable archaeological ground. Thus, the sphere is placed at the end of a pole, carried by hand, and moves in the scene from one shot to the next. Our algorithm allows us to detect and classify each sphere present in the images. We then apply a lighting estimation algorithm specific to each type of sphere. The albedo and normal maps obtained by photometric stereo are presented in Figure 7.

6. CONCLUSION AND PERSPECTIVES

In this paper, we present a new method for detecting calibration spheres using deep learning. This is a fairly straightforward task (the Hough transform does this very well), which is nevertheless characterized by the appearance of false positives in the presence of cast shadows and circular patterns. We propose an approach based on neural networks, which is much faster than manual detection, and even more accurate, in practice, when shadows are located near the boundary of a sphere silhouette.

Figure 7. Results of the photometric stereo method, with light estimated with auto-detected spheres. From left to right: zoom on one of the 17 pictures of the "Panneau du cheval gravé" presented in Figure 1, albedo map and normal map obtained by photometric stereo.

Lighting estimation using spheres could be carried out in a single step, but we prefer to limit ourselves to the first step, i.e. sphere detection, for two reasons. Firstly, we can remain very generic in the type of spheres used, since the network detects and classifies matt, shiny and chrome spheres. On the other hand, this allows us to take a more modular approach, inspired by the philosophy of some open-source 3D reconstruction software. Users are then free to use the output of our network for the application of their choice.

The detection of spheres in an image is, in fact, only the first step in the estimation of illumination, a task necessary for the implementation of 3D reconstruction techniques such as photometric stereo. So, is it really necessary to position spheres in the scene? A number of recent works aim to estimate lighting directly from photographs.[20] As the problem is ill-posed, deep learning approaches predominate. In the future, we also plan to estimate lighting without using a sphere, which would simplify the shooting protocol, but without necessarily using a neural network.

REFERENCES

[1] Durou, J.-D., Falcone, M., Quéau, Y., and Tozza, S., "A Comprehensive Introduction to Photometric 3D-reconstruction," in [*Advances in Photometric 3D-Reconstruction*], 1–29, Springer (2020).

[2] Mélou, J., Laurent, A., Fritz, C., and Durou, J.-D., "3D Digitization of Heritage: Photometric Stereo can Help," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **48**, 145–152 (2022).

[3] Reinhard, E., Ward, G., Pattanaik, S., and Debevec, P., [*High Dynamic Range Imaging*], Morgan Kaufmann (2005).

[4] Liu, Z., Wu, Q., Wu, S., and Pan, X., "Flexible and accurate camera calibration using grid spherical images," *Optics Express* **25**, 15269–15285 (June 2017).

[5] Gonçalves, W. and Bruno, O., "Automatic System for Counting Cells with Elliptical Shape," *Learning & Nonlinear Models* **9** (Jan. 2012).

[6] Jin, R., Owais, H. M., Lin, D., Song, T., and Yuan, Y., "Ellipse proposal and convolutional neural network discriminant for autonomous landing marker detection," *Journal of Field Robotics* **36**(1), 6–16 (2019).

[7] Basca, C. A., Talos, M., and Brad, R., "Randomized Hough transform for ellipse detection with result clustering," in [*Proceedings of EUROCON 2005*], **2**, 1397–1400 (2005).

[8] Patraucean, V., Gurdjos, P., and Grompone Von Gioi, R., "A Parameterless Line Segment and Elliptical Arc Detector with Enhanced Ellipse Fitting," in [*Proceedings of ECCV*], (Oct. 2012).

[9] Ren, S., He, K., Girshick, R., and Sun, J., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," arXiv preprint arXiv:1506.01497 (2016).

[10] Dong, W., Roy, P., Peng, C., and Isler, V., "Ellipse R-CNN: Learning to Infer Elliptical Object From Clustering and Occlusion," *IEEE Transactions on Image Processing* **30**, 2193–2206 (2021).

[11] He, K., Gkioxari, G., Dollár, P., and Girshick, R., "Mask R-CNN," in [*Proceedings of ICCV*], (2017).

[12] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., "You only look once: Unified, real-time object detection," in [*Proceedings of CVPR*], 779–788 (2016).

[13] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P., "Focal loss for dense object detection," in [*Proceedings of ICCV*], 2980–2988 (2017).

[14] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S., "End-to-End Object Detection with Transformers," arXiv preprint arXiv:2005.12872 (2020).

[15] Li, F., Ma, J., Tian, Z., Liang, H.-N., Ge, J., Zhang, Y., and Wen, T., "Mirror-Yolo: A Novel Attention Focus, Instance Segmentation and Mirror Detection Model," in [*Proceedings of ICFSP*], 76–80 (2022).

[16] Murmann, L., Gharbi, M., Aittala, M., and Durand, F., "A Dataset of Multi-Illumination Images in the Wild," in [*Proceedings of ICCV*], 4079–4088 (Oct. 2019).

[17] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I., "Robust Speech Recognition via Large-Scale Weak Supervision," arXiv preprint arXiv:2212.04356 (2022).

[18] Loshchilov, I. and Hutter, F., "SGDR: Stochastic Gradient Descent with Warm Restarts," arXiv preprint arXiv:1608.03983 (2017).

[19] Papadopoulos, D. P., Uijlings, J. R. R., Keller, F., and Ferrari, V., "Extreme clicking for efficient object annotation," arXiv preprint arXiv:1708.02750 (2017).

[20] Hold-Geoffroy, Y., Sunkavalli, K., Hadap, S., Gambaretto, E., and Lalonde, J.-F., "Deep outdoor illumination estimation," in [*Proceedings of CVPR*], 7312–7321 (2017).