EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2018)
R. Sablatnig and M. Wimmer (Editors)

Short Paper

An Image-based Approach for Detecting Faces Carved in Heritage Monuments

Yu-Kun Lai[1], Karina Rodriguez Echavarria[2], Ran Song[2], Paul L. Rosin[1]
[1]Cardiff University, UK
[2]Centre for Secure, Intelligent and Usable Systems,
University of Brighton, UK

Abstract

Heritage monuments such as columns, memorials and buildings are typically carved with a variety of visual features, including figural content, illustrating scenes from battles or historical narratives. Understanding such visual features is of interest to heritage professionals as it can facilitate the study of such monuments and their conservation. However, this visual analysis can be challenging due to the large-scale size, the amount of carvings and difficulty of access to monuments across the world. This paper makes a contribution towards this goal by presenting work-in-progress for developing image-based approaches for detecting visual features in 3D models, in particular of human faces. The motivation for focusing on faces is the prominence of human figures throughout monuments in the world. The methods are tested on a 3D model of a section of the Trajan Column cast at the Victoria and Albert (V&A) Museum in London, UK. The initial results suggest that methods based on machine learning can provide useful tools for heritage professionals to deal with the large-scale challenges presented by such large monuments.

CCS Concepts
•***Computing methodologies*** → *Neural networks; Mesh models;*

1. Introduction

The conservation and the study of visual imagery and narratives of large-scale monuments are usually challenging tasks for heritage professionals.

An example of such monuments is Trajan's Column in Rome. This monument is a 100-foot tall marble column containing carvings which depict the successful campaigns of the Emperor Trajan against the Dacians on the Danube frontier in AD 101–2 and 105–6. Coulston [Cou] suggests that groups of human figures carved in Trajan's Column present strong similarities in shapes and carving techniques in order to assign them to different roles, including Roman officers, musicians, citizen soldiers, infantryman and Dacians. Thus, the combination of both 3D geometrical information and semantic information of the narratives could offer numerous possibilities for conservation, preservation and scholarly research, as well as public communication.

This paper describes on-going work to develop image-based methods to automate the analysis of the shapes carved on monuments. Our current efforts focus on detecting faces, although the approach could be generalised to detect other types of shapes.

Face recognition techniques have previously been applied to the Cultural Heritage domain in order to enhance the understanding of portrait art [SRRC15]. Although these techniques have not been explored for sculptural work, they can potentially help curators and art historians to answer potential ambiguities concerning identity of the subject or to understand artists' styles.

The paper is structured as follows. Section 2 presents related work on visual analytic techniques for digital 3D models. Section 3 describes the methods used to create a 3D model of a section of the column. Section 4 presents the implementation and testing of a method for automatically identifying semantic objects, in particular faces, in the 3D model in order to improve the visual understanding of the carvings. Section 5 presents conclusions and further work.

2. Related work

Detection of human faces has attracted significant attention in the past few decades [SSBQ10, FDG*13]. Most work focuses on face detection from images, either using traditional methods which learn classifiers based on facial features or using deep learning to detect facial regions directly from images by going through a deep neural network learned using facial training images. The approach of Viola and Jones [VJ04] is a classic example of the traditional methods, which extracts facial features efficiently using integral images and trains cascaded classifiers to detect faces in real-time.

Deep learning methods (e.g. [YLLT15]) typically rely on training a deep convolutional neural network using a large number of training images with labelled faces. Some advanced face detectors are provided by open source libraries, including Open-

DOI: 10.2312/gch.20181365

Face [BRM16] which is based on identifying facial landmarks using Conditional Local Neural Fields [BRM13], and so not only detects faces but also localises facial landmark points. Another example is the general-purpose object detector provided in Dlib, a C++ deep learning library [Kin09], which is based on a convolutional neural network with the last layer being a maximum margin object detection (MMOD) loss layer [Kin15] to reliably detect objects, which shows good performance for face detection. With sufficient training data, deep learning methods achieve much better performance than traditional methods.

Faces on Trajan's Column are bas-reliefs rather than true 3D faces. Due to the well-known bas-relief ambiguity [BKY99], Lambertian surfaces under general bas-relief transformations can produce exactly the same image with appropriate lighting. These bas-relief images are typically compressed in terms of depth ranges but still look plausible.

As well as the issues arising from the intended non-linear compression in bas-reliefs, there are also challenges for automatic face detection due to the non-realistic nature of sculpture. In the related area of art analysis, it has been shown that, even when using state-of-the-art methods, the performance levels for the detection of faces in artworks (e.g. paintings) are considerably poorer than for the same task in regular photographs [CPZ15, WCH16].

In this work, based on the observation that the Trajan's Column bas-relief faces have a perceptually similar visual appearance to normal faces, we propose to render the column as images, and perform a comparative study of applying image-based face detectors to such images. Moreover, in addition to training with normal face images, we also train face detector models using shading images (which ignore albedo) and renderings of 3D face models, which are visually closer to the faces on the rendered Trajan's Column.

The following section will present the digitisation of a section of the Trajan's Column cast which produced the 3D model for the subsequent research.

3. Digital 3D model of the Trajan's Column cast

A section of the Trajan Column cast at the Victoria and Albert (V&A) Museum in London (UK) was digitised using a white light scanner, the Aicon 3D SmartScan. The size of the section is approximately 24 m^2, which represents approximately 7% to 8% of the total area of the column.

The digitisation took place over 5 days involving approximately 25 hours of scanning. In total, 250 individual scans were produced. The scanned model is shown in Figure 1. In order to test the face recognition methods on the column, individual areas on the 3D model were selected. These areas do not follow the Cichorius [Cic96] convention. Instead, they focus on areas which present interesting visual details such as a group of human figures.

4. Face detection methods

We compare representatives from three types of face-detection methods, namely a traditional method [VJ04], a deep learning method based on landmark localisation [BRM13, BRM16] and a deep learning method based on object detection [Kin09, Kin15].

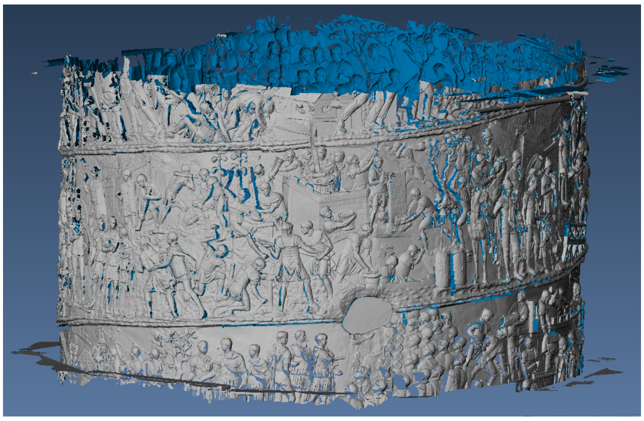
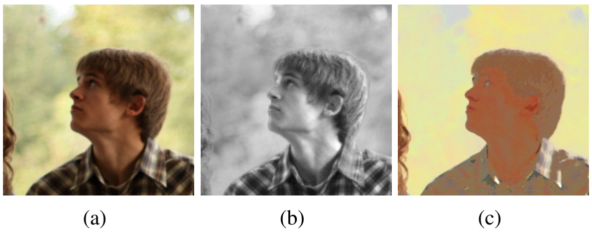


Figure 1: *3D scanned data for a section of the Trajan Column cast.*



(a) (b) (c)

Figure 2: *Image intrinsic decomposition of a training face image. (a) input image, (b) shading image, (c) reflectance image. (b) and (c) are obtained automatically from (a) using the intrinsic image decomposition method [BBS14].*

We use the Phong lighting model, with ambient, diffuse and specular strengths set to 0.17, 0.75 and 0.15, respectively, and the camera is set to face the section of interest with frontal lighting for each of the scenes selected. These settings were empirically set to avoid under-exposure and over-exposure to make the 3D faces well presented. An example is shown in Figure 4; it can be seen that it preserves details of shapes, making face detection possible.

For the traditional method, we use the OpenCV implementation of [VJ04] pre-trained with the Haar basis functions. For landmark-based face detection, we use the OpenFace [BRM16] pre-trained model along with default settings.

For the deep object detection method [Kin15], we use the neural network architecture provided by Dlib for face detection. It is a deep convolutional neural network which involves multiple downsampling layers of 5 × 5 convolutions with the ReLU activation function and batch normalization, followed by multiple 5 × 5 convolutional layers, also with ReLU and batch normalisation. The last layer uses MMOD loss [Kin09] to provide reliable face detection. We use the default settings, except for setting the batch size to 100 to fit in the available GPU memory.

For training the face detection, we use the 7,213 face images provided by Dlib [dli] from various sources, which have labelled ground truth facial regions. The training is performed on a desktop

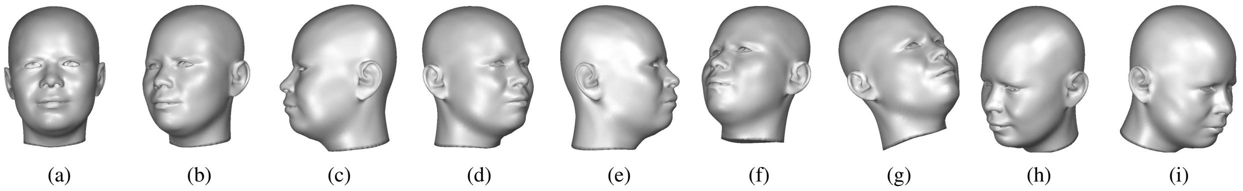


Figure 3: *An example of 9 poses generated for each 3D face (for a given subject and expression).*

Table 1: *Statistics of different face detectors on 7 scenes (4). The numbers shown are TP (true positives) / FP (false positives). The methods compared are: Viola & Jones [VJ04], OpenFace [BRM16], and Dlib MMOD [Kin09, Kin15] trained with normal face images (Dlib-N), shading images (Dlib-Sh) and synthetic images, i.e. rendered 3D faces (Dlib-Syn).*

Scene	[VJ04]	[BRM16]	Dlib-N	Dlib-Sh	Dlib-Syn
1	1/2	1/0	4/0	10/0	1/2
2	2/1	3/0	8/0	8/0	3/4
3	0/1	2/1	6/0	10/0	7/3
4	1/4	0/1	3/0	4/1	4/8
5	0/2	0/1	2/0	2/0	2/2
6	1/4	0/1	6/0	7/0	0/3
7	0/2	0/0	2/0	4/0	1/2
Total	5/16	6/4	31/0	45/1	18/24

computer with a Titan Xp GPU, and it takes about 13 hours to train the model. Once the model is trained, face detection is real-time.

Since the technique [Kin15] is a general object detection method, we propose to train the model not only with normal face images, but also images which have closer visual characteristics to the faces on the rendered Trajan's Column image. The basic observation is that rendered images from the shapes on Trajan's Column represent the geometry and lighting, but not reflectance (albedo). Two approaches are used, namely intrinsic image decomposition and synthetic images produced by rendering 3D faces.

Firstly, we take the same training images, and apply intrinsic image decomposition [BBS14], which automatically decomposes an input colour image into its shading and reflectance components. An example is shown in Figure 2, where the face in the shading image (b) is visually closer to the faces in the rendered column because reflectance has been removed. To achieve this, the method [BBS14] learns a conditional random field (CRF) model using a large number of manual annotations obtained via crowdsourcing. Alternative intrinsic decomposition methods can also be used.

For the second approach, we render 3D faces to obtain synthetic shading images. To cover various facial shapes, expressions and poses, we use 3D faces from the 3D FaceWarehouse [CWZ*14] which includes 150 subjects, each with 47 facial expressions. To have a good coverage of facial poses, we render 9 poses for each subject and expression (see Figure 3 for an example). We then used all these face images to train the deep model.

4.1. Comparative results

We choose 7 representative scenes with diverse figures from our 3D scan of the Trajan's Column (Figure 4).

We first compare the results of different face detection methods, all trained with *normal* face images. The statistics of TP (true positives) and FP (false positives) are reported in Table 1. Both Viola & Jones [VJ04] and OpenFace [BRM16] perform quite badly, significantly worse than their typical performance on normal face images. In comparison, the results of Dlib MMOD [Kin09, Kin15] are significantly better. It successfully finds 31 correct faces and no incorrect faces, despite substantial differences of visual characteristics between the rendered faces on the column and the faces in normal colour images. This shows that a convolutional deep network with the MMOD loss layer has good generalisability.

The detected faces are shown in Figure 4 where results are obtained by training the same Dlib MMOD network architecture with normal colour face images, shading images and synthetic (rendered) 3D face images.

The face detector trained with rendered 3D faces however does not perform as well as expected, and the performance is significantly worse compared to training using normal face images (but still better than OpenFace or Viola & Jones). This is probably because the synthetic images do not contain sufficiently rich shapes and shading to cover the variations that appear in real-world bas-relief faces, especially after years of degradation.

In summary, none of the methods in the comparative studies require training data from the bas-reliefs from Trajan's Column for training, yet decent performance is obtained, especially for the Dlib MMOD face detector trained on shading images. Although there are possibilities to further improve the method, e.g. by developing specialised neural network architectures or loss functions, the promising results demonstrate the great potential of using machine learning to help assist automatic semantic analysis for cultural heritage artefacts, a topic which is largely unexplored, possibly due to limited training data.

5. Conclusions and further work

This paper presented research efforts to improve the analysis of visual imagery and narratives displayed on large-scale monuments. Further work will cover experimenting with detecting other elements in the monument such as animals, foliage and buildings. For this, image-based approaches can offer relevant solutions to deal with a wide variety of content.

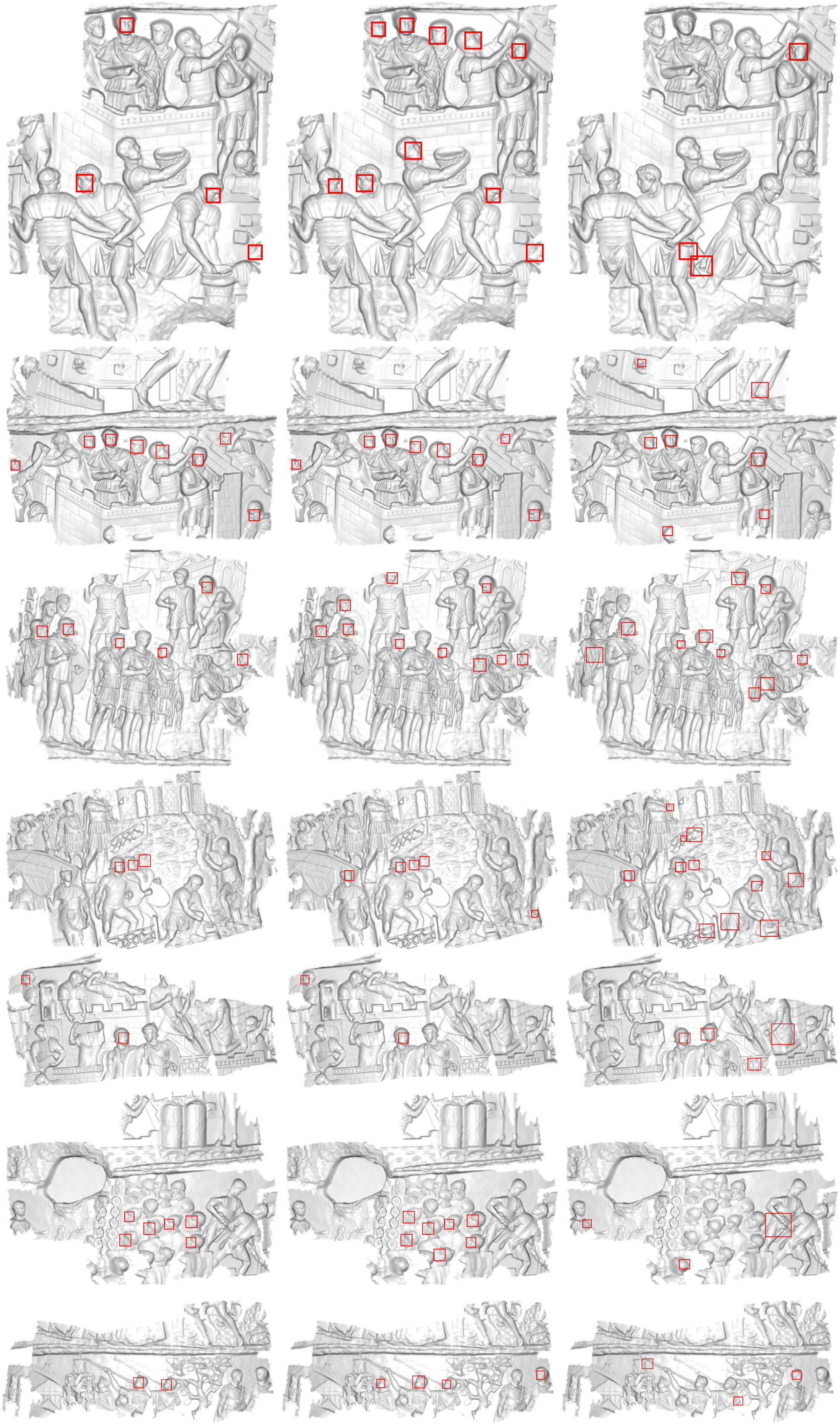

(a) Normal Face Images (b) Shading Images (c) Synthetic Images

Figure 4: *Comparison of results obtained using Dlib MMOD with different training images.*

Acknowledgements

The Titan Xp used for this research was donated by the NVIDIA Corporation. We also thank the V&A museum for allowing access to the column and for using the data for this research.

References

[BBS14] BELL S., BALA K., SNAVELY N.: Intrinsic images in the wild. *ACM Trans. on Graphics (SIGGRAPH) 33*, 4 (2014). 2, 3

[BKY99] BELHUMEUR P. N., KRIEGMAN D. J., YUILLE A. L.: The bas-relief ambiguity. *International Journal of Computer Vision 35*, 1 (1999), 33–44. 2

[BRM13] BALTRUSAITIS T., ROBINSON P., MORENCY L.-P.: Constrained local neural fields for robust facial landmark detection in the wild. In *IEEE International Conference on Computer Vision Workshops* (2013), pp. 354–361. 2

[BRM16] BALTRUSAITIS T., ROBINSON P., MORENCY L.: OpenFace: An open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision* (2016), pp. 1–10. 2, 3

[Cic96] CICHORIUS C.: *Die Reliefs der Traianssäule*. Berlin, G. Reimer, 1896. 2

[Cou] COULSTON J.: The human figure types. https://arts.st-andrews.ac.uk/trajans-column/the-project/the-human-figure-types/. Accessed: 2018-06-20. 1

[CPZ15] CROWLEY E. J., PARKHI O. M., ZISSERMAN A.: Face painting: querying art with photos. In *BMVC* (2015), pp. 65–1. 2

[CWZ*14] CHEN C., WENG Y., ZHOU S., TONG Y., ZHOU K.: FaceWarehouse: a 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics 20*, 3 (2014), 413–425. 3

[dli] Dlib face training data. http://dlib.net/files/data/dlib_face_detection_dataset-2016-09-30.tar.gz. 2

[FDG*13] FANELLI G., DANTONE M., GALL J., FOSSATI A., GOOL L. J. V.: Random forests for real time 3d face analysis. *International Journal of Computer Vision 101*, 3 (2013), 437–458. 1

[Kin09] KING D. E.: Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research 10* (2009), 1755–1758. 2, 3

[Kin15] KING D. E.: Max-margin object detection. *arXiv:1502.00046* (2015). 2, 3

[SRRC15] SRINIVASAN R., RUDOLPH C., ROY-CHOWDHURY A. K.: Computerized face recognition in renaissance portrait art: A quantitative measure for identifying uncertain subjects in ancient portraits. *IEEE Signal Processing Magazine 32*, 4 (2015), 85–94. 1

[SSBQ10] SEGUNDO M. P., SILVA L., BELLON O. R. P., QUEIROLO C. C.: Automatic face segmentation and facial landmark detection in range images. *IEEE Trans. Systems, Man, and Cybernetics, Part B 40*, 5 (2010), 1319–1330. 1

[VJ04] VIOLA P., JONES M. J.: Robust real-time face detection. *International Journal of Computer Vision 57*, 2 (2004), 137–154. 1, 2, 3

[WCH16] WESTLAKE N., CAI H., HALL P.: Detecting people in artwork with CNNs. In *European Conference on Computer Vision* (2016), pp. 825–841. 2

[YLLT15] YANG S., LUO P., LOY C. C., TANG X.: From facial parts responses to face detection: A deep learning approach. In *IEEE International Conference on Computer Vision* (2015), pp. 3676–3684. 1