Single image pose estimation for automated photogrammetry

Joe Eastwood

Danny Simms-Waterhouse, Samanta Piano, Richard Leach

Joe.Eastwood@nottingham.ac.uk

Project Aims

Close range photogrammetry is a popular technique in optical form metrology, particularly for the relatively low equipment cost compared to competing techniques such as structured light projection. Despite this popularity the measurement pipeline (shown in Figure 1) is slow and user dependant. We aim to optimise the measurement process to decrease the measurement time, and to fully automate the process to remove any dependence on the user.

In a first step to achieving these larger aims, presented herein is a method for autonomously establishing the spatial relationship between the camera and the measurand without the need for specialised fixturing.

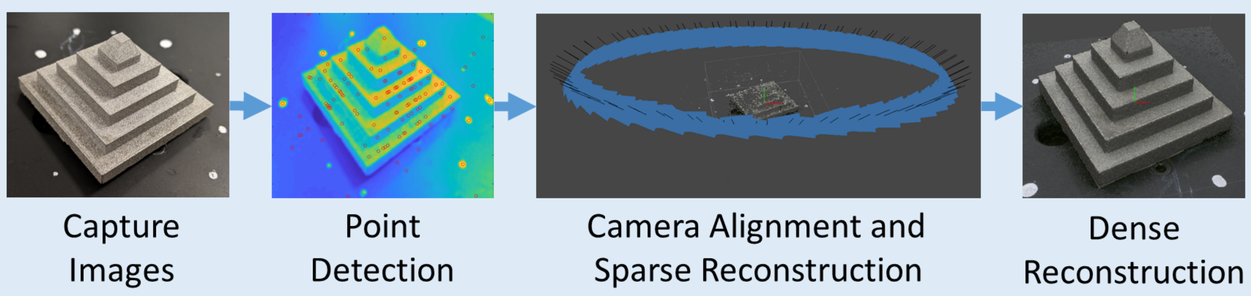

Figure 1: Photogrammetry pipeline

Approach

A form of Convolutional Neural Network known as a Residual Network (ResNet) is trained to first categorise an initial image by which of four test artefacts is present in the image. The ResNet then predicts the [X,Y] coordinates of each artefact on the rotation stage, as well as the artefact's rotation about the global Z axis [Θ].

The model architecture, as shown in Figure 2, has the following notable features; convolutional layers made of 'residual blocks' as shown, a set of convolutional layers shared between both the categorisation and regression tasks, a separate branch for categorisation which feeds-forward into the regression layers, and a combined loss function.

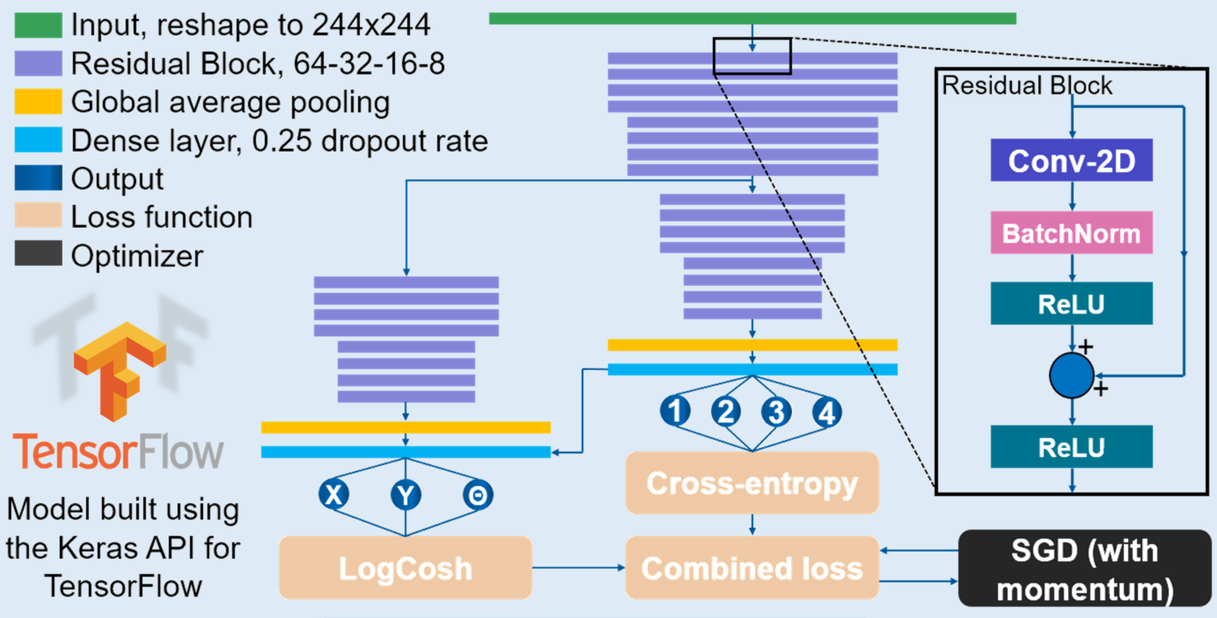

Figure 2: ResNet architecture

Conclusions

Using synthetic images to train a ResNet to predict the relative position of a camera to a given object has been shown to be an effective approach. In the future this work will be combined with a genetic pose optimisation algorithm to move toward a fully automated and optimised measurement pipeline.

Generation of Training Data

In order to train the ResNet a vast amount of labelled training data is required. Rather than manually generating this data and associated labels, a simulation of the camera and system created in Blender was used to generate photorealistic synthetic images. As these images are generated in software, the ground truth label values are known explicitly. Once trained, the ResNet is used to make predictions on real photographic images.

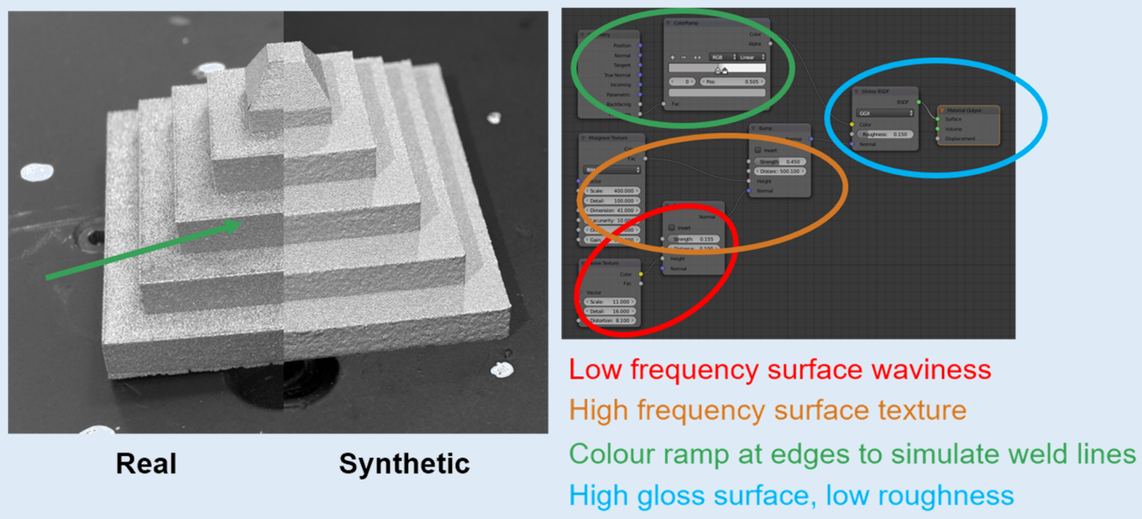

Figure 3: Synthetic data; comparison to real image, and texture generation

Training results

Figure 4 shows the model and validation losses over the training period. As can be seen the model converges to a loss of 0.02. Figure 5 shows visually the predictions made by the model on each artefact. In this Figure each original image has the predicted location overlaid in yellow wireframe. It can be seen qualitatively that the [X,Y] location prediction is of better quality than the rotation [Θ] prediction.

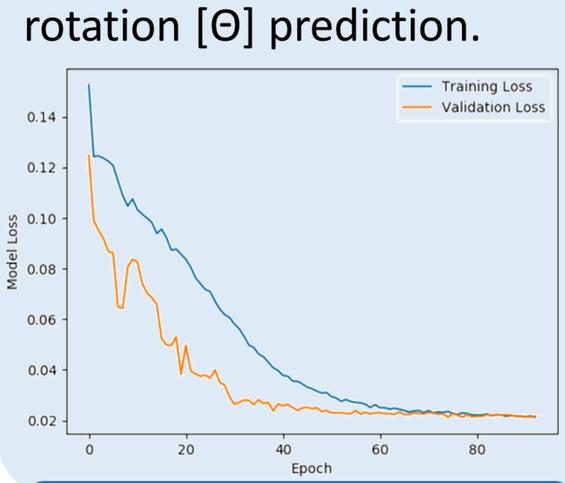

Figure 4: Model convergence

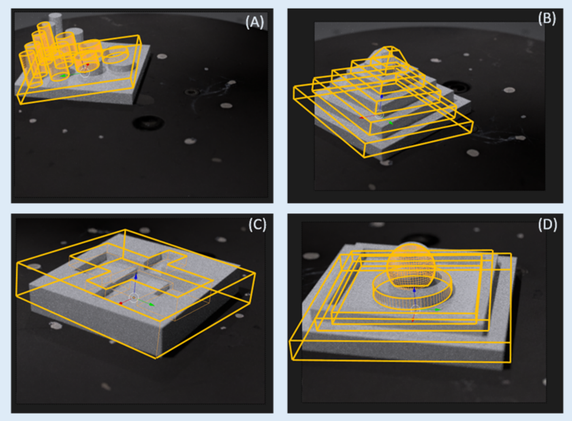

Figure 5: Example predictions