Computer Vision 2 - Assignment 3

Tim van Loenhout - 10741577
timvanloenhout@gmail.com
University of Amsterdam

1 INTRODUCTION

This report notes my findings for Assignment 3 of the course Computer Vision 2. First, a PCA model is used to construct a face model, which is then mapped to a 2D plane using a pinhole camera model. Subsequently, the 2D projection is compared to a ground truth image based on the retrieved facial landmarks. Then, the PCA model is optimized such that the predicted landmarks approximate the ground truth landmarks, resulting in a 3D reconstruction obtained from a singular monocular image. Furthermore, this reconstruction is textured based on the RGB values from the input image and finally, the model is extended to construct textured 3D reconstructions for each frame in a video.

2 MORPHABLE MODEL

In order to construct a 3D face, 30 principal components for facial identity and 20 principal components for facial expression are used. Furthermore, parameters α and δ are used to enforce the model to construct a diversity of facial geometries. The values of α and δ are sampled from a uniform distribution between -1 and 1, resulting in for instance the faces in figure 1. As can be seen in figure 2, increasing the range of this distribution causes the faces to become more extreme, which is not ideal for reasons that are discussed later in this assignment.

Figure 1: Face geometries based on α and δ, uniformly sampled from range -1,1.

Figure 2: Face geometries based on α and δ, uniformly sampled from range -3,3.

2.1 Pinhole Camera Model

The 3D point clouds are mapped to a 2D plane using the pinhole model, seen in figure 5.

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \\ \hat{d} \end{bmatrix} = \underbrace{[\mathbf{V}] \times [\mathbf{P}]}_{\Pi} \times \underbrace{\begin{bmatrix} \mathbf{R}(\boldsymbol{\omega}) & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}}_{\mathbf{T}} \times \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$$

Figure 3: Pinhole camera model.

The viewpoint and projection matrix in this model are constructed from the left/right, top/bottom and near/far ranges of the point cloud. The rotation matrix and translation vector are constructed from the learnable parameters *tau* and *omega*, of which the latter is first converted from degrees to radians. The results of such a mapping can be seen in figure 4, where the 3D point cloud is first rotated either 10 or -10 degrees around the y-axis and then projected onto a 2D plane.

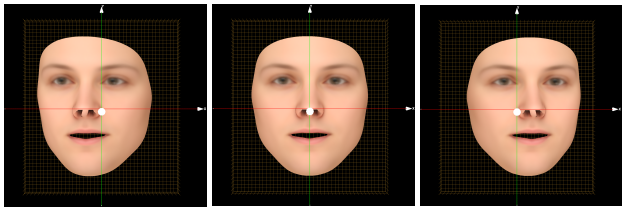


Figure 4: Mapping from 3D to 2D after rotating around the y-axis.

From the 2D mapping, 68 facial landmarks are obtained by taking the points corresponding to the the indices provided by the .anl file. The landmarks corresponding to a model that is rotated 10 degrees around the y-axis and translated -500 in the z-direction can be seen in figure 5.

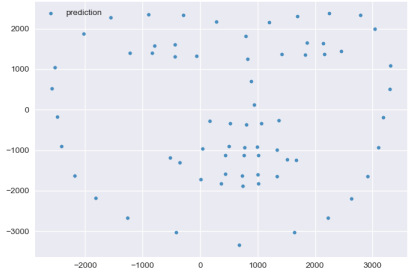


Figure 5: Facial landmarks after rotating 10 degrees around the y-axis and translating -500 in the z-direction.

3 LATENT PARAMETERS ESTIMATION

3.1 Ground Truth Facial Landmarks

By learning the parameters ω, τ, α and δ the model can construct a shape from a singular monocular image; in this report an image of Barack Obama.

Figure 6: Input image.

From this image, ground truth landmarks are obtained (figure 13a) and then flipped horizontally and up-scaled to match the size of the prediction (figure 13b and 13c).

3.2 Parameter Optimization

The learning objective is to minimize a loss function that comprises the ground truth/prediction distance and a regularization term based on the size of α and δ. Each iteration, a 3D point cloud is generated, resulting in a loss over which the gradients are computed to update the parameters in the right direction. Once the average loss over the previous 10 iterations is no longer smaller than the prefacing 10 losses, convergence is assumed.

From experimenting, I observed that the model learns better when it emphasizes more on translation and rotation during the early stages of training. Hence, I set the learning rates for the rotation and translation parameters to 1 opposed to the 0.01 rate used for the facial feature values parameters.

Figure 13 shows this training process where first the landmarks are mainly forced into the right orientation after which the features are slowly fine-tuned to approximate the ground truth. The corresponding loss curve can be seen in figure 8 and the final result in figure 7.

Figure 7: Constructed 3D face model after convergence.

3.3 Hyperparameter Tuning

In order to increase or decrease the constraints to the facial features, the hyperparameters λ_{alpha} and λ_{delta} can be tuned. Experimentation shows that when using the image in figure 6, no constraints are required, however, when using other faces, the value of λ_{alpha} and λ_{delta} appear to play a more important role. That is, Barack

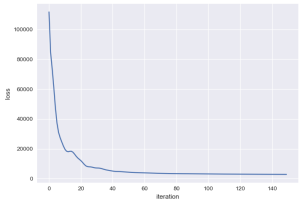


Figure 8: Training loss curve.

Obama has a fairly general face shape. This combined with his basic expression in the used image causes the default facial landmarks too be already relatively close to the ground truth. For other faces on the other hand, the model has to start from a larger error. A large initial discrepancy does not only result in a false start which lengthens the optimization duration, but could also cause the parameter updating to take off into the wrong direction when this discrepancy is too big. It is therefore important to prevent the model from constructing faces that become too extreme. By properly initializing λ_{α} and λ_{δ}, a balance is obtained where the model is free to pursue the correct face identity and expression while being forced to do so in a constraint space. However, as mentioned before, for the particular face used in this report, setting λ_{α} and λ_{δ} to 0 suffices.

4 TEXTURING

For texturing the 3D model, the projection information can be used to trace all points back to their original location on the input image. However, these locations are float values, hence bilinear interpoliation is used to obtain the corresponding color value for each of the 3 RGB channels, based on the surrounding neighborhood pixels. The resulting textured model can be seen in figure 9.

Figure 9: Textured 3D model.

5 USING MULTIPLE FRAMES

As can be seen in figure 9, the model is not able to texture parts of the face that are not shown in the image. Nor will it be able to accurately model the shape of these parts. By extending the model to learn multiple frames from a video, each showing the face from a different angle, a more all-round 3D model can be constructed. That is, for each frame, a different layer of the model parameters is optimized, except for α as a person's face identity is assumed constant. To show that the algorithm is able to cope with variations in rotation, translation and emotion, four frames (figure 10) are modelled and textured (figure 11). When a face portrays a constant expression throughout all frames, they could be merged into a single 3D model using for instance the iterative closest point algorithm.

Figure 10: Frames.

Figure 11: Textured 3D frames.

6 DISCUSSION

Fist of all, the iteratieve closest point algorithm is a simple approach to finding the relative camera position and angle between two point clouds. Using this transformation information, multiple point clouds can be combined into a single 3D reconstruction. The main advantage of this algorithm lies in its simplicity; it does not require any feature extraction, such as is the case in structure from motion and it can be extended to any number of dimensions. Nonetheless, despite its simplicity it is not very fast, as the used point-matching techniques are not the most efficient. This efficiency can be increased up to a certain degree by using parallelisation. Another drawback is that while the algorithm converges to a minimum, it is not guaranteed to be a global minimum. Also, the algorithm fails when the point clouds are too far apart, and finally, ICP requires an expensive depth-camera. The SFM algorithm on the other hand can construct a 3D model from solely 2D images for which a simple camera suffices. By matching the keypoints from these images, many angles and distances can be combined into a single model. However, the algorithm requires a minimum number of such features, which furthermore should be well distributed across the frames.

I do believe both approaches could be combined to obtain better 3D reconstructions. For instance, combining higher level depth information from the ICP algorithm with the lower level feature information from the SFM algorithm could increase the probability of finding an accurate transformation for the more information scarce area's. Also, ICP treats all points equally important, which makes it susceptible to outliers. By incorporating the key points as done in SFM you could filter out the most informative regions from the point clouds to obtain a robuster and more insensitive model, which moreover would be better able to cope with larger angular discrepancies between frames.

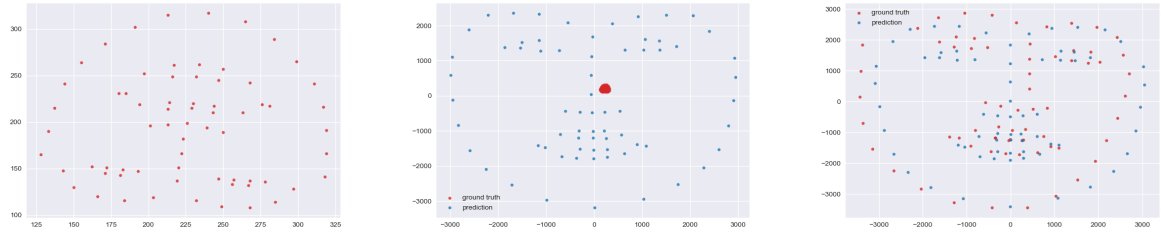
Opposed to both ICP and SFM, this report discussed a method for constructing a 3D model from only one 2D image. While this single-image dependency, and its relatively fast computation time are major advancements over the other two methods, it must be noted that this specific model is only applicable to faces. Extending the model to construct a 3D model from any type of image is very challenging as it requires a-priori information about the to be modelled objects.

All in all, each of the discussed approaches has its benefits and drawbacks, hence when combining the strengths of each model, better 3D reconstructions might be obtained. For instance, the model discussed in this report could be used to obtain textured 3D face models from multiple views in a time efficient manner without requiring any special equipment. The aforementioned conjunction of the ICP and SFM algorithms could then be used to stitch these multiple views together into a single textured 360 degree 3D model.

7 CONCLUSION

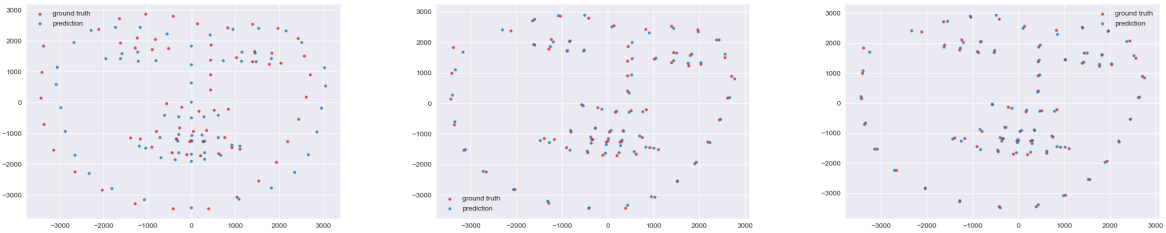
To conclude, the model discussed in this report is able to construct an accurate textured 3D model from a singular 2D image. However, the main drawback of this approach is that training takes place within a constrained optimization space for which a-priori information about the shape of the object is required. Also, while this model can be extended to learn reconstructions from multiple frames, another algorithm such as iterative closest point is required to stitch the multiple point clouds together. Nonetheless, when provided with information about the object shape, the model shows excellent results in terms of both reconstruction quality and speed.

..



(a) Raw ground truth. (b) Raw ground truth size comparison. (c) Transformed ground truth.

Figure 12: Ground truth facial landmarks, raw and transformed.



(a) Iteration 1. (b) Iteration 50. (c) Iteration 150.

Figure 13: Optimization process.