Towards 3D Facial Reconstruction from Uncalibrated CCTV Footage

Chris van Dam Raymond Veldhuis Luuk Spreeuwers

University of Twente [*]
Faculty of EEMCS
P.O. Box 217, 7500 AE Enschede, The Netherlands

{c.vandam, r.n.j.veldhuis, l.j.spreeuwers}@utwente.nl

Abstract

Facial comparison in 2D is an accepted method in law enforcement and forensic investigation, but pose variations, varying light conditions and low resolution video data can reduce the evidential value of the comparison. Some of these problems might be solved by comparing 3D face models: a face model derived from CCTV camera footage and a reference face model acquired from a suspect. In our case we will assume uncalibrated CCTV footage, because the original camera setup may be destroyed or replaced after the incident, so precise camera information is no longer available. In contrast to other statistical methods, like Morphable Models, we would like to use no additional statistical information at all. Our method is based on a projective reconstruction of landmarks on the face and an auto-calibration step to obtain a 3D face model in a Euclidean space. In our experiment the effect of the number of frames and noise on the landmarks is explored for 3D face reconstruction based on landmarks. An estimation of the 3D face shape can already be obtained using 25 points in 30 frames.

1 Introduction

In forensic research anno 2012 most of the law enforcement services still use 2D frontal facial comparison. Although this can give good results for frontal or near-frontal reference faces, many problems still arise due to pose variations, varying light conditions and low resolution video data. One way to improve facial comparison would be to compare 3D facial models instead of 2D models, since in most cases there is much more information available in CCTV (Closed-Circuit Television) camera footage. The use of 3D face models requires a change in the technical infrastructure of the law enforcement services and their current methods, but we think that this method can improve the facial comparison results by taking advantage of more information available in the original evidence.

Next to eye witnesses, the most common source of evidence in street crime, burglary and robbery cases is CCTV camera footage. In this paper we will take a specific case into account: fraud at an ATM with an uncalibrated camera installed. The suspect is close to the camera and therefore there is much perspective distortion in the frames of the camera footage. We assume that there is no information available of the original camera, because in many cases the original camera setup may be destroyed or replaced after an incident. So the only data available is CCTV camera footage of the suspect, mainly containing footage of the suspect's face. Our goal in the Person Verification 3D project is to create a 3D facial reconstruction of the suspect, which can be used

[*] In cooperation with the Netherlands Forensic Institute.

for 3D facial comparison. In this paper we will use landmarks in multiple 2D frames to obtain an initial estimation of the camera parameters and the 3D shape of the face. In our experiments we determine the minimum number of points and frames needed to obtain an accurate reconstruction of a simulated face model. Next we determine the maximum noise that will still allow us to obtain a precise reconstruction. Finally we do some experiments with auto-calibration of the reconstruction to validate if the methods described in this paper can be applied on face models.

2 Background

Our problem, where the face of the suspect is moving in front of a static camera, is equivalent to a problem where the camera is moving and the suspect is static. So for each frame $[i = 1..M]$ we have to find the internal and external camera parameters of that specific frame. The static shape of the face can be described by $[j = 1..N]$ 3D landmarks. We will use N 2D landmarks with known correspondences to 3D landmarks in all M frames to obtain a 3D reconstruction of the face. Sturm and Triggs provide a method to obtain a projective structure X and projective motion P by factorization of the projections x of all frames [1]:

$$\lambda_{ij} x_{ij} = \hat{P}_i \cdot \hat{X}_j = P_i \mathcal{H} \cdot \mathcal{H}^{-1} X_j \tag{2.1}$$

Where $\hat{P}_i$ is a 3×4 projection matrix of frame i, $\hat{X}_j$ is a 4×1 homogeneous 3D vector of point j, x_{ij} is a homogeneous 2D vector of the projection i of landmark j and λ_{ij} is a scalar representing the projective depth of x_{ij}. If the projective depths λ_{ij} are known, the system of equations is of rank 4. The projective depths can be estimated using epipolar geometry on pairs of frames, see [1] for details. A rank 4 approximation of the system can be found using the Singular Value Decomposition (SVD) of the system. For details about the linear algebra or SVD see [2]. Noise or imprecise measurements on the landmarks can lead to a system with a higher rank. The error minimized by Sturm and Triggs in equation 2.2 is based on both the estimated projective depths and the image coordinates, but has no geometric meaning, see [3]:

$$\sum_{i=1}^{N} \sum_{j=1}^{M} \| \lambda_{ij} x_{ij} - \hat{P}_i \cdot \hat{X}_j \|_F^2 \tag{2.2}$$

Where x_{ij} are the image coordinates (which might include noise) and λ_{ij} the estimated projective depths corresponding to these points. The reconstruction we have now is a projective reconstruction of the cameras and shape. Before we can do any measurements of length or angles of the projective structure X, we need to find the 4×4 projective ambiguity $\mathcal{H}$, which is independent of the number of frames or the number of points, to update the projective space to Euclidean space, see Figure 1.

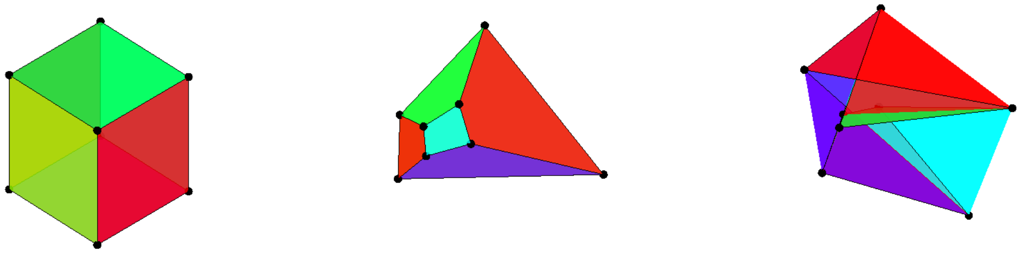

Figure 1: Three projective reconstructions of a cube with different ambiguities ($\mathcal{H}$).

The calibration can be achieved by adding extra information about the shape or internal parameters of the cameras, but since the intrinsics of the camera and the 3D shape of the face are unknown, there is no additional information available. A second method would be auto-calibration (self-calibration), in which case (almost) no additional information is needed for the calibration. The auto-calibration estimates the shape and camera parameters simultaneously. Two available methods for auto-calibration are the absolute dual quadric as described in [4] and Kruppa equations which can be found in [5]. According to Hartley and Zisserman [6]: *'The application of the Kruppa equations to three or more views provides weaker constraints than those obtained by other methods such as the modulus constraint or the absolute dual quadric'*. Since our purpose is to use as many frames (data) as possible, we choose the absolute dual quadric method for auto-calibration.

3 Auto-Calibration

Auto-calibration is a method to estimate the internal camera parameters from uncalibrated CCTV footage. The object itself is used to perform the calibration. Auto-calibration is based on the dual image of the absolute conic (DIAC), which is fixed under similarity transformations, so the internal camera parameters can be estimated despite of the unknown external parameters. The goal of the auto-calibration is to locate the plane at infinity and the absolute conic ω. For a projective reconstruction where the first frame contains no rotation and translation, $\mathcal{H}$ can be expressed in terms of the calibration matrix K and the plane at infinity v [7]. In our case K is the same for all frames.

$$\mathcal{H} = \begin{bmatrix} K & \mathbf{0} \\ v^\top & \lambda \end{bmatrix} \tag{3.1}$$

Since we can't determine the scale of the reconstruction without using additional input data, the scale factor λ can be chosen as $\lambda = 1$. The absolute dual quadric Q^*_∞ encodes both K and v in one mathematical entity. The null space of Q^*_∞ encodes the plane at infinity v. Without proof the following equation is given:

$$\omega^* = KK^\top = P_i Q^*_\infty P_i^\top \tag{3.2}$$

Equation 3.2 shows the relation between the projection of the absolute dual quadric and the calibration matrix K. Constraints on K can be transferred to the absolute dual quadric. The assumption of square pixels and a principle point close to the center of the camera are sufficient conditions to obtain linear equations for Q^*_∞, see [6] for more details.

4 Experiments

In our experiments we first obtain a projective reconstruction from the projections and compare the reconstruction to a known ground truth. Our goal is to see if the quality of the reconstruction and the method are suitable for the reconstruction of facial models. The second step is the auto-calibration, which is completely separated from the projective reconstruction. To express the quality of the projective factorization, we use the 2D RMS reprojection error. The 2D reprojection error E_{2D} is defined as:

$$E_{2D} = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\|x_{ij} - \hat{P}_i \cdot \hat{X}_j\|^2} \tag{4.1}$$

In all experiments the internal camera parameters K are fixed. The generated projections are comparable with realistic face images. Therefore the camera rotations vary between -40 and 40 degrees and the translations vary between -10 and 10 units. All camera parameters are randomly chosen within their respective bounds. All projections fit within an image of 400×600 pixels. The 3D ground truth point cloud contains uniformly distributed random points within a bounding box of $100 \times 100 \times 10$ units.

4.1 Number of frames and number of points

In the first experiment we try to find the minimal number of frames and points needed to obtain a projective reconstruction. In theory 4 landmarks in 3 frames are enough to obtain a projective reconstruction, but if the image coordinates contain noise, more points and/or frames are necessary to average out the noise. The projection of each point in each frame is known, but we add Gaussian noise with a standard deviation of $\sigma = 1$ to both the x- and y-coordinates. For each combination of number of points and number of frames the reprojection error E_{2D} is calculated two times: with respect to the projections with noise and with respect to the ground truth image points. The experiment was repeated 1 000 times, with independent instances of noise for every combination of points and frames to get more stable results. The curves show the average value over all repetitions.

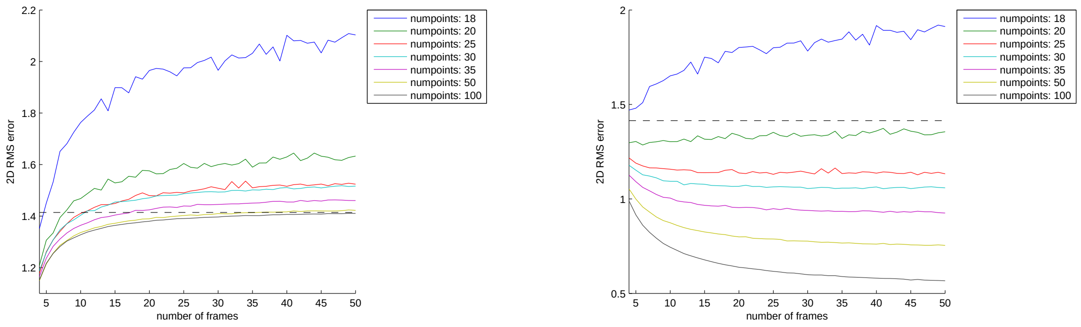

Figure 2: Variable number of point and frames. In the left graph the reprojection error is calculated with respect to the projections with noise and in the right graph with respect to the ground truth.

Notice that in the left graph at least 50 points are needed to approximate the value of the expected asymptote $\sqrt{2}$. Since the most consistent reconstruction is the reconstruction of the noise-free image points, the reprojection error approximates the $\sqrt{2}$ value for Gaussian noise of $\sigma = 1$. More points still improve the results, but a number of points around 50 seems to be the lower bound on an approximation of the asymptote. According to the number of frames at least 30 frames are needed to stabilize the lines. Adding more frames doesn't seem to offer a drastic improvement of the results.

In the right graph the error seems to be decreasing for the first time for around 25 points. Using more points leads to an even faster decreasing function. Using more frames seems to have less effect on the error for a given number of points. So adding more points has a stronger effect than adding more frames and can even lead to a switch from an increasing to a decreasing error function. Finding the lower bounds allows us to make an approximation of the number of points and frames needed for a 3D facial reconstruction. We would like to find the lowest number of points possible, because in CCTV footage usually plenty frames are available, but determining more landmarks is difficult. We choose 25 points as an acceptable lower bound on the number of points, since it provides a decreasing function when more frames are added.

4.2 Effect of the number of frames on 2D and 3D error

In the following experiment we use a fixed number of points and explore the effect of the number of frames in 2D and in 3D. The 3D error is calculated based on the 3D ground truth landmarks, see Equation 4.2.

$$E_{3D} = \underset{\mathcal{H}}{\operatorname{argmin}} \sqrt{\frac{1}{N}\sum_{j=1}^{N} \|X_j - \mathcal{H}\hat{X}_j\|^2} \tag{4.2}$$

Where X_j is a normalized known 3D ground truth point of the shape and $\hat{X}_j$ is a normalized reconstructed 3D point. E_{3D} in Equation 4.2 can only be calculated for a known ground truth, but mostly finds the best auto-calibration possible independent of the auto-calibration method used.

Projective reconstructions are made from the projections of 25 points in a variable number of frames. Each subset of frames is randomly taken out of 100 000 frames. In each set zero-mean Gaussian noise with $\sigma = 1$ is added to both x- and y-dimension. The reconstructions are made with an increasing number of frames, ranging from 4 to 50 frames. The experiment is repeated 1 000 times with different subsets of frames. Outliers of more than 4 times the standard deviation were removed from the results. The results are shown in Figure 3.

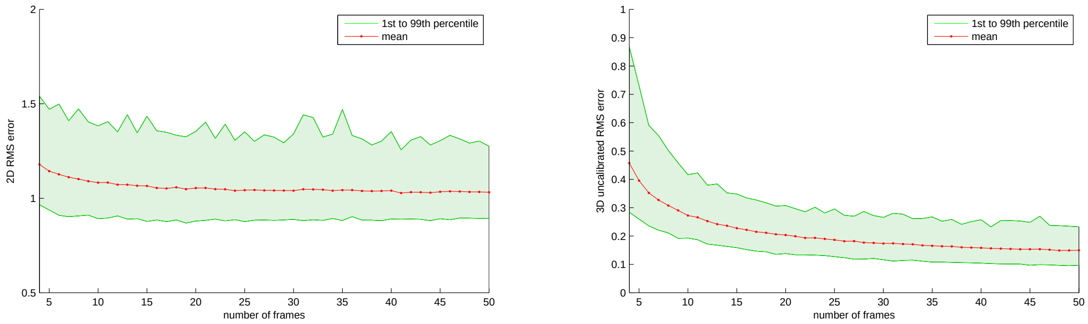

Figure 3: 2D RMS and 3D RMS error of projective reconstruction with a variable number of frames.

The 2D reprojection error with respect to the ground truth image points shows only a slight improvement of the quality between 10 and 50 frames, but the 3D error is reduced by 40% when comparing 10 to 40 frames (0.27 compared to 0.16). So more frames indeed improve the 3D reconstruction.

4.3 Gaussian noise

To validate our choice that 25 points are sufficient for a projective reconstruction, we add Gaussian noise with a higher σ to the projections in the next experiment. The Gaussian noise varies within the range of 0 to 2.5 times the standard deviation for both the x- and y-coordinates. The noise experiment is repeated 1 000 times. The results are shown in Figure 4.

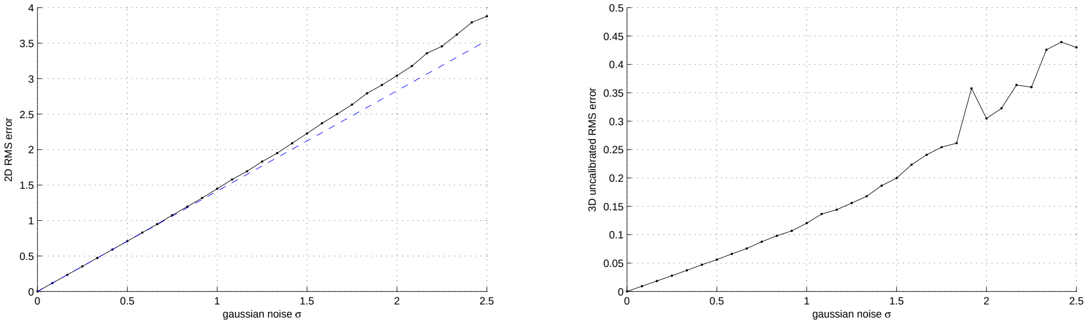

Figure 4: Variable Gaussian noise used for reconstruction.

For noise $\sigma \leq 1$ the reprojection error with respect to the image points with noise seems to be linear with the Gaussian noise. Also the 3D error (for a perfect calibration) seems to be linear. If more noise is added, both errors rise above linear, see the blue line in Figure 4. This might be a problem of the depth estimation in the projective reconstruction method or more frames may be needed to get a consistent projective reconstruction.

4.4 Auto-calibration

The auto-calibration method of Section 3 is also calculated for a variable number of frames and for increasing Gaussian noise. The 3D error of the auto-calibration is calculated by finding the best simularity transformation $\mathcal{H}$, because after the auto-calibration the 3D points and the cameras should no longer be affected by an affine or a projective transformation, see Equation 4.2. The auto-calibration uses a projective reconstruction as a starting point. The results are shown in Figure 5.

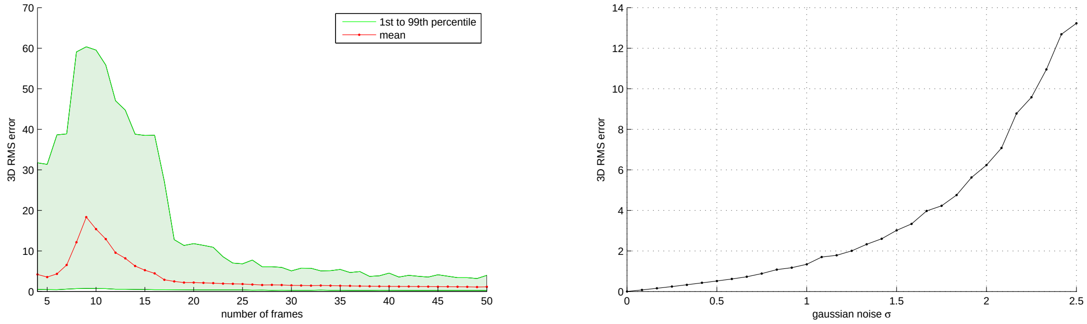

Figure 5: 3D error after auto-calibration for a variable number of frames and increasing Gaussian noise.

The left graph in Figure 5 shows that a minimum number of frames is required to perform the auto-calibration. Adding more frames improves the 3D result of the auto-calibration, but as can be seen in the right graph of Figure 5 adding noise ($\sigma > 1.5$) leads to an explosion of the 3D error function. So some work needs to be done on the robustness of the auto-calibration under Gaussian noise or more information needs to be added to the problem.

5 Conclusion

At least 25 image points in 30 frames are needed to obtain a precise perspective reconstruction. This is still valid for Gaussian noise with $\sigma \leq 1$. Accurate automatic or even manual labeling of the landmarks is needed to find such landmarks on a face. Though the number of frames certainly influences the quality of the reconstruction, the number of points seems more important to obtain an accurate reconstruction. Our experiment shows that there is no use adding additional frames for a small number of points. The number of points even has influence on the increasing or decreasing behaviour of the 2D error function. If more noise ($\sigma \geq 1.5$) is involved, the projective reconstruction deteriorates and auto-calibration might fail. The auto-calibration method or conditions need to be improved to obtain better results. Also more work needs to be done to determine if additional available information of the camera or the background scene can improve the 3D reconstruction.

References

[1] P. Sturm and B. Triggs, "A factorization based algorithm for multi-image projective structure and motion," in *4th European Conference on Computer Vision (ECCV' 96)*, pp. 709–720, Springer-Verlag, 1996.

[2] C. D. Meyer, ed., *Matrix Analysis and Applied Linear Algebra.* Society for Industrial and Applied Mathematics, 2000.

[3] G. Wang and J. Q. Wu, *Guide to Three Dimensional Structure and Motion Factorization.* Springer Publishing Company, Incorporated, 1st ed., 2010.

[4] B. Triggs, "Autocalibration and the absolute quadric," in *Conference on Computer Vision and Pattern Recognition (CVPR '97)*, pp. 609–614, IEEE Computer Society, 1997.

[5] R. I. Hartley, "Kruppa's equations derived from the fundamental matrix," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 133–135, 1997.

[6] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2nd ed., 2004.

[7] R. Gherardi and A. Fusiello, "Practical autocalibration," in *11th European Conference on Computer vision: Part I (ECCV '10)*, pp. 790–801, Springer-Verlag, 2010.