Landmark-based Model-free 3D Face Shape Reconstruction from Video Sequences

Chris van Dam, Raymond Veldhuis, Luuk Spreeuwers
Faculty of EEMCS
University of Twente, The Netherlands
P.O. Box 217
7500 AE Enschede
c.vandam@utwente.nl
r.n.j.veldhuis@utwente.nl
l.j.spreeuwers@utwente.nl

Abstract: In forensic comparison of facial video data, often only the best quality frontal face frames are selected, and hence potentially useful video data is ignored. To improve 2D facial comparison for law enforcement and forensic investigation, we introduce a model-free 3D shape reconstruction algorithm based on 2D landmarks. The algorithm uses around 20 landmarks on the face and combines the structure information of multiple frames. Model based 3D reconstruction methods, such as Morphable Models, reconstruct a 3D face shape model that is strongly biased towards the average face. Therefore, we don't use statistical face shape models in our model-free approach. The 3D landmark reconstruction algorithm simultaneously estimates the shape, pose and position of the face, based only on the fact that all images in the sequence are recorded using a single calibrated camera. The algorithm iteratively updates the reconstruction by including new frames, while maintaining the consistency of the reconstruction. We demonstrate the convergence properties of the method reflected in the 2D reprojection error and the 3D error with respect to a ground truth model. We show that the quality of the reconstruction depends on the level of noise in the landmarks. In follow-up experiments we show that our method is able to reconstruct the 3D structure of a face, using a styrofoam head and real video data. The results of the real face data show the same behavior as the results of the simulated data, which indicates that our method is capable of reconstructing real facial structures, depending on the noise of the landmarks.

1 Introduction

One of the unsolved issues in forensic comparison of facial data is the comparison with 'wild' photo or video data. Law enforcement services are constrained to work with the case material provided, and unlike researchers, they are not able to use recordings from a controlled environment. Among the most difficult problems of 'wild' photo materials are the non-frontal poses of faces and low resolution facial images, because often material of overview cameras is used for facial comparison. Automatic face recognition software can only handle 2D facial data under a small pose angle. At the moment the accuracy of automatic face comparison algorithms degrades quickly for faces under large pose.

As a consequence often only the best quality frontal face frames are selected, and hence much video data is ignored. Law enforcement services are still in search of the 'tools' to compare non frontal faces. However, these 'tools' should treat the video data in such a way that no supplementary information is added to the video data. Reconstruction methods, such as Morphable Models [BV99], reconstruct a 3D face shape model that is strongly biased towards the average face. Such reconstructions could lead to unacceptable forensic conclusions. In the proposed method we try to avoid this situation caused by facial models.

In this paper we introduce a model-free 3D shape reconstruction algorithm based on 2D landmarks, so no additional statistical face models or average face models will be used. We assume that the calibration parameters of the camera, such as focal length, principal point and skew, are available. Any recording is assumed to contain a subset of frames with different views of a face without variation in facial expression. Our final goal is to reconstruct the face in 3D. We use around 20 landmarks on the face to estimate the shape of the face together with the pose and the position of the face for each view. We present three different experiments. In our first experiment we use simulated data to demonstrate the convergence properties of the method reflected in the 2D reprojection error and the 3D error with respect to a ground truth model. In the second experiment we continue our work in [DSV13] and we explore the strength of our method more extensively on realistic face shape data with a styrofoam head model. In our last experiment we use real video sequences for our reconstruction. Note that our reconstructed 3D models only contain shape information and no texture information. This paper continues with section 2 where we give a background on the methods and notations used in this paper. In section 3 we introduce and explain our proposed algorithm. In section 4 we show the performance of our algorithm in several experiments. Then we end up with the conclusion in section 5.

2 Background

Our problem, in which 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 view $i = 1..N$ we have to find the external camera parameters of that specific view. The static shape of the face can be described by $j = 1..M$ 3D landmarks. We will use M 2D landmarks with known correspondences to the 3D landmarks in all N views to obtain a 3D reconstruction of the landmarks on the face. Our camera is described by the pinhole camera model [HZ04], where a 3D point Q is projected on the image plane in 2D point q. The point projection equation is usually written as $q = P \cdot Q$, where P contains both the calibration parameters of the camera and the rotation and translation of a view.

We prefer a method in which we can add additional views to the current solution to improve the reconstruction. To be able to find such a method, we should search for a method that starts with one pair of views and then provides an iterative solution or a solution that merges groups of views. The method described in [Har93] is able to estimate the rotation and translation parameters for one pair of views. This method expresses the relation between calibrated views in the essential matrix. The essential matrix can be estimated from corresponding landmarks in two views using a robust MSAC method (M-estimator SAm-

ple Consensus) method [TZ00]. Once we determined the relation between two views, the relative rotations and translation parameters can be estimated for both views. This method provides four solutions for the rotation and translation parameters, see Equation 2.1, but only one of these solutions is posing the points in front of the camera:

$$\begin{aligned} \hat{P}_1 &= \left[UWV^\top \mid +\mathbf{u_3}\right] \quad \hat{P}_2 = \left[UW^\top V^\top \mid +\mathbf{u_3}\right] \\ \hat{P}_3 &= \left[UWV^\top \mid -\mathbf{u_3}\right] \quad \hat{P}_4 = \left[UW^\top V^\top \mid -\mathbf{u_3}\right] \end{aligned} \tag{2.1}$$

where the rotation matrix defined by U, W and V is based on the result of a singular value decomposition of the essential matrix. The matrix W is a matrix that mirrors one of the axes. The translation u_3 is the last column of U, see [HZ04] [Har93]. This solution has 5 degrees of freedom, 3 for the rotation and only 2 for the translation, because the equation is determined up to an unknown scale. The rotation and translation parameters are extracted directly from the essential matrix of one pair of views. Then, we can estimate the structure by linear triangulation of one pair of views [HZ04]:

$$A = \begin{bmatrix} x\mathbf{p}^{3\top} - \mathbf{p}^{1\top} \\ y\mathbf{p}^{3\top} - \mathbf{p}^{2\top} \\ x'\mathbf{p}'^{3\top} - \mathbf{p}'^{1\top} \\ y'\mathbf{p}'^{3\top} - \mathbf{p}'^{2\top} \end{bmatrix} \tag{2.2}$$

where $\mathbf{p}^{i\top}$ are the rows of P in the first view and x, y are the x- and y-values of the projection of point Q in the first view. The other parameters are the corresponding values of the second view. The point Q can be found by solving $AQ = \mathbf{0}$. This method reconstructs only the visible points in one pair of views. The method can be extended to more than 2 views by including more equations from additional views in A. In our case we have a low number of landmarks, so the reconstruction based on two views gives a poor estimation of the shape. Therefore, we extend the algorithm using multiple views to overcome the problems of noise and the low number of landmarks. We introduce an algorithm that iteratively updates the reconstruction by including new views, while maintaining the consistency of the reconstruction for a low number of landmarks. The quality of the reconstruction can be determined by the 2D RMS reprojection error E_{2D}:

$$E_{2D} = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\|\mathbf{q_{ij}} - \hat{P}_i \cdot \hat{Q}_j\|^2} \tag{2.3}$$

where index i represents a view $1..N$ and j represents a point $1..M$, $\hat{P}$ contains the external camera parameters of each view and $\hat{Q}$ contains a collection of homogeneous 3D points. The homogeneous 2D vector $\mathbf{q_{ij}}$ represents the known projections including the noise on the landmarks.

3 Reconstruction Algorithm

In this section we describe the proposed algorithm for the reconstruction of the structure of the face based on 2D projections. In short the algorithm finds an initial pair of views

with a low reprojection error. Based on this pair of views we obtain a linear estimation of the structure. Then we start an iterative procedure in which we add one new view in every step of the procedure. After adding the new view, the current selection of views and the current structure estimation are optimized. The result of the reconstruction algorithm is an estimation of the 3D positions of the landmarks and an estimation of the rotation and translation parameters of each view.

The best initial estimate for the structure is found by calculating the reprojection error for every possible pair of views in the dataset and to select the pair with the minimum reprojection error. To calculate the reprojection error we need to know the rotation and translation parameters of each view. These values (except for the scale) can be extracted from the Essential Matrix, see Equation 2.1. The essential matrix can be estimated, in turn, from the projections using a robust MSAC method (M-estimator SAmple Consensus) [TZ00]. Knowing the rotation and translation parameters of a pair of views, allows us to estimate the structure for this pair of views. Based on this structure we can calculate the reprojection error for this pair of views. However, also the reprojection errors of the other views are important for consistency during the optimization. So, to find the best pair of views we choose a reference view and calculate all rotations and translation relative to the reference view. Then we calculate the reprojection error of the total set of views for every view as reference. A second criterion for the selection of the best pair of views is the number of landmarks that could be reconstructed, because not only the reprojection error is important, but also the number of visible corresponding landmarks in the initial pair of views. Our selection criterion is now to find the pair of views with the maximum number of corresponding landmarks in two views and a minimal reprojection error for the total set of views. We choose to obtain the subset of 25% of the solutions containing the most reconstructed points over all views. From this subset we select the pair with the lowest reprojection error over all views. This solution provides us a solution that is sufficient for initialization of our iterative optimization. We calculate an initial linear estimation of the structure based on the selected pair of views.

In the optimization step one new view is added in each iteration to keep all views in our current estimation consistent with the estimated structure. The selection of the new candidate view is based on the convergence behavior of the candidate view. The view with the lowest reprojection error after 10 optimization iterations, is chosen as the next view. This candidate selection is necessary to prevent the algorithm from failing in the first few iterations. Based on the new selection of views, a linear estimation of the structure is obtained, see Equation 2.2. Then the reprojection error of both the rotation and translation parameters and the structure are minimized using the Levenberg-Marquardt algorithm. To prevent overfitting, we used only 30 Levenberg-Marquardt iterations for each optimization step, which performs properly for the minimization. Finally, the rotation and translation parameters of the views that were not in the selection set are optimized to maintain the consistency of the total set of views. The iterative optimization procedure continues until all views are added and optimized.

4 Experiments

The goal of the first experiment on simulated data is to determine the influence of the number of views on the reconstruction, and to investigate the convergence properties of our algorithm. We create a random point cloud of 25 3D points and obtain a set of 100 projections of this point cloud with variation in rotation and translation. The calibration information and a random selection of the projections are used in the reconstruction algorithm. We performed two experiments in which we added a different level of Gaussian zero-mean noise to the projections, with a standard deviation of 1.0 and 2.0 pixels respectively. The size of the face in each frame is around 250-350 pixels. The noise is added independently to the x- and y-coordinates of the projections. Finally we used a random mask to hide 30% of the data to imitate the hidden landmarks on a face. We use our reconstruction algorithm to estimate the 3D structure. The quality of the reconstruction will be determined by the 2D RMS reprojection error E_{2D}, see Equation 2.3. All landmarks that were not visible, were left out of the equation, so MN is defined as the total number of visible landmarks summed over all views. After reconstruction the 3D RMS error E_{3D} between the reconstruction and the ground truth point cloud can be calculated with:

$$E_{3D} = \underset{\mathcal{H}}{\operatorname{argmin}} \sqrt{\frac{1}{M} \sum_{j=1}^{M} \|Q_j - \mathcal{H}\hat{Q}_j\|^2} \tag{4.1}$$

where $\mathcal{H}$ is a rigid 3D transformation which aligns the ground truth point cloud Q with the reconstruction $\hat{Q}$ and j is the index of a point. The experiment is repeated 100 times with different instances of noise to investigate the robustness of the algorithm.

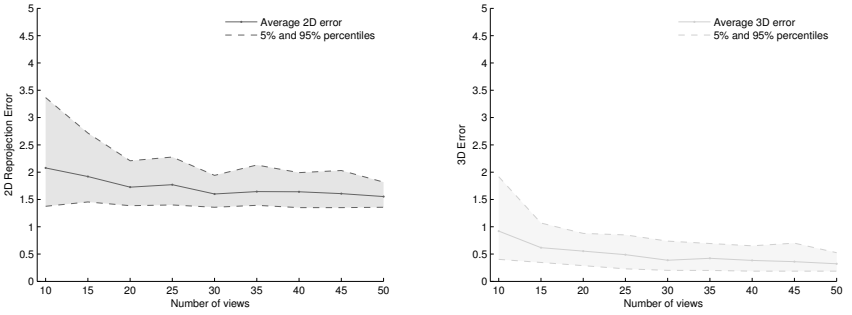

Figure 1: 2D and 3D error of the reconstructions using noise with a standard deviation of 1.0 pixels.

The graphs in Figure 1 show the expected behavior for Gaussian noise with a standard deviation of 1.0 pixels. The more views are added, the more robust the reconstruction is. If the shape is estimated perfectly, then we would expect the 2D reprojection error to converge to the level of noise added. The 2D reprojection error converges to an asymptote of $\sqrt{2} \approx 1.41$, which is the expected level of noise, see the left graph of Figure 1. Another observation we make is that the number of views above 30 has little influence on both the 2D and the 3D average error. The robustness of the algorithm is only slowly increasing for more than 30 views, see the right graph of Figure 1. So adding more than 30 views seems to have only a small impact on both the quality and robustness of the algorithm.

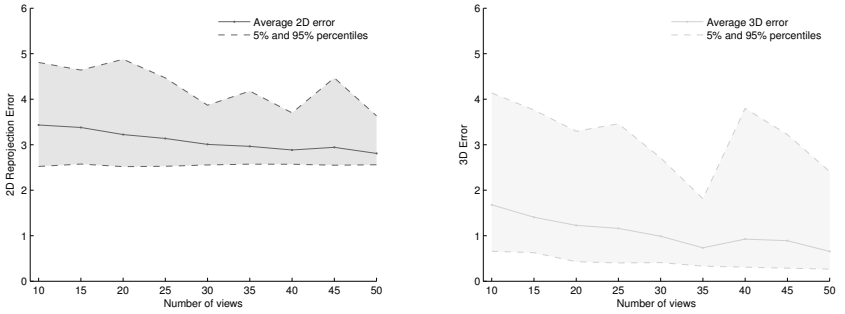

Figure 2: 2D and 3D error of the reconstructions using noise with a standard deviation of 2.0 pixels.

If the level of Gaussian noise is doubled to a standard deviation of 2.0 pixels, the behavior is similar to the previous experiment. The asymptote here is $\sqrt{8} \approx 2.83$, see the left graph in Figure 2. Adding more views has less effect on the robustness of the reconstruction algorithm, but it still has a decreasing effect on the average reprojection error. When more views are added, the average 3D error also decreases slowly, though the robustness of the algorithm seems not to increase. For more than 35 views, the system shows even more variation in the 3D errors than for 35 views, see Figure 2. This can be explained by the fact that the more views are added, the higher the change for heavy outliers in the projections. Since none of the selected views are skipped, outliers might severely decrease the result of the reconstruction. The reconstruction is assumed to be failed, if the reprojection error is above 5.0 pixels. For the experiment with Gaussian noise with a standard deviation of 1.0 pixels and more than 30 views, the algorithm converges to a solution in about 99% of the cases. In the case of a standard deviation of 2.0 pixels, the algorithm only converges in 75% of the cases. So the algorithm seems stable for Gaussian noise with a standard deviation of 1.0 pixels, but becomes less stable for Gaussian noise with a standard deviation of 2.0 pixels or above.

The goal of the second experiment with the styrofoam head is to determine whether the algorithm is capable of working with manually labeled face data. We acquired a 3D model of a styrofoam head with 22 colored pins located on the face. An orthogonal view of the styrofoam model can be seen in the left image in Figure 3. We choose a virtual camera and we extract the calibration data from this camera. We created 51 renderings of the model with different rotation and translation parameters, see Figure 3. All visible landmarks are labeled manually in all renderings. In contrast to the previous experiment, no noise was added to the projections, leaving us with only the noise of the manual landmarking. The reconstruction is based on the calibration data and subsets of the renderings. The 3D points of the ground truth model are also manually labeled on the 3D model of the styrofoam head, which, in contrast to the previous experiment, could influence the 3D error. The experiment is repeated 100 times for each number of views to determine the robustness of the algorithm.

The second experiment shows the same behavior as the experiment with Gaussian noise with a standard deviation of 1.0 pixels. Adding more views increases the quality of the 3D reconstruction, but for more than 40 views, in this case, the gain is very low for both the 2D and 3D error. The asymptote for the 2D error is around 2.0 pixels, which is somewhere

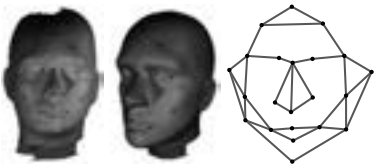

Figure 3: Left: Orthogonal view of the 3D styrofoam head model. Middle: One of the rendered 2D views of the model. Right: The reconstructed 3D landmarks with added edges for visibility.

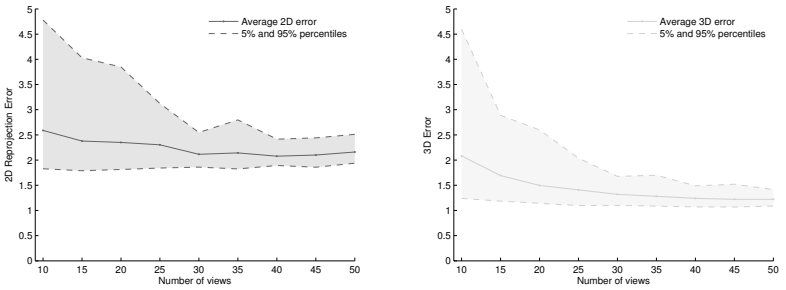

Figure 4: 2D and 3D error of the styrofoam face reconstructions.

between the results of the first two experiments. This noise level is similar to a $\sqrt{2} \approx 1.41$ pixels error in both x- and y-coordinates, which is probably the accuracy of the manual landmarking of the 2D dataset. A rough estimation gives us a head size of 300 mm and the size of the head in the frames is around 500 pixels. So each pixel represents 0.6 mm. Our method is able to estimate the landmarks with $(2.16 \cdot 0.6 =)$ 1.3 mm precision on average. The average 3D error is 1.22, which is around 0.7% of the size of the head. The results are in line with the results of the first experiment on simulated data. This second experiment shows that our algorithm has similar convergence properties and errors to the experiment on the simulated data, and can therefore be applied on manually labeled realistic face data.

In this last experiment we show that our algorithm can handle real video data using a calibrated camera. We acquired 100 frames of several volunteers in which they slowly moved and rotated their heads in front of a camera. We annotated 20 landmarks in each frame in a semi-automatic manner. Finally we calibrated our camera with 20 frames of a planar calibration board, which provided us the camera calibration data. In the next experiment we use a selection of 50 frames to reconstruct the structure of the face. Since we don't have 3D ground truth data of our landmarks, we will only use the 2D reprojection error and visual inspection to express the quality of the reconstruction. We ran the experiment two times, with different subsets of views: one using the 50 even frames and another using the 50 odd frames of the first volunteer.

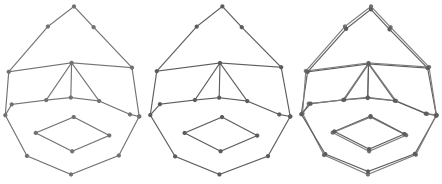

Figure 5: Left: Two 3D reconstructions of the first volunteer based on different subsets of views. Right: Aligned 3D models of the two reconstruction.

The 2D reprojection error for the even set was 2.13 and the reprojection error of the odd set was 2.54, where the size of the frames is similar to the styrofoam experiment. These results are completely in line with the results of the styrofoam experiment, see the left graph in Figure 4. There is a small variation in the 2D error, but nevertheless the variation seems acceptable compared to the previous results. Visual inspection of the 3D structure shows that both 3D structures are close to each other, see Figure 5. So even with real video data, including calibration, and landmarking, our algorithm is able to reconstruct the position of the 3D landmarks of the face.

5 Conclusion and Future Work

The experiment on the simulated point cloud shows that the quality of our reconstruction depends on the level of noise in the projections. For a small level of noise, around 1.0 pixels, the convergence and robustness of the algorithm seem sufficient. For a larger level of noise the system might become unstable, and even not converge to a useful solution. For Gaussian noise with a standard deviation of 2.0 pixels the algorithm only converges in 75% of the cases. The minimum number of views needed to get sufficient quality for the reconstruction is around 30 views. More views can improve the reconstruction, but this will only give a small improvement. In the second experiment, we showed that manual landmarking leads to an error comparable to a Gaussian noise with a standard deviation of 1.4 pixels. The results of the styrofoam experiment were in line with the simulated reconstructions with Gaussian noise. The third experiment with real video data shows results similar to the styrofoam experiment. The visual inspection of the 3D structure and the 2D reprojection errors indicate that the algorithm is capable of reconstructing real facial structures. In future work we will include the texture information in the reconstruction to get a full 3D model of the face. The full reconstruction allows us to perform facial recognition experiments on 2D faces under pose.

References

[BV99] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999.

[DSV13] C. van Dam, L.J. Spreeuwers, and R.N.J. Veldhuis. Model-free 3D Face Shape Reconstruction from Video Sequences. In *34th WIC Symposium on Information Theory in the Benelux*. WIC, 2013.

[Har93] Richard I. Hartley. An Investigation of the Essential Matrix, 1993.

[HZ04] Richard I. Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2004.

[TZ00] P. H. S. Torr and A. Zisserman. MLESAC: A New Robust Estimator with Application to Estimating Image Geometry. *Computer Vision and Image Understanding*, 78, 2000.