3D Morphable Model Parameter Estimation

Nathan Faggian[1], Andrew P. Paplinski[1], and Jamie Sherrah[2]

[1] Monash University, Australia, Faculty of Information Technology, Clayton
[2] Clarity Visual Intelligence, Australia
{nathan.faggain, andrew.paplinski}@infotech.monash.edu.au,
jsherrah@clarityvi.com

Abstract. Estimating the structure of the human face is a long studied and difficult task. In this paper we present a new method for estimating facial structure from only a minimal number of salient feature points across a video sequence. The presented method uses both an Extended Kalman Filter (EKF) and a Kalman Filter (KF) to regress 3D Morphable Model (3DMM) shape parameters and solve 3D pose using a simplified camera model. A linear method for initializing the recursive pose filter is provided. The convergence properties of the method are then evaluated using synthetic data. Finally, using the same synthetic data the method is demonstrated for both single image shape recovery and shape recovery across a sequence.

1 Introduction

To accurately determine the fine three-dimensional shape of an object using spatio-temporal information is a difficult and well studied task. Reffered to as "structure from motion" there are many methods both in the linear and nonlinear domain. This paper presents a recursive filter approach to the problem using Kalman filters. Through the use of two KFs we are able to track across variation in pose and identity. The requirement for real-time operation is a complication that is also addressed by our method since it is limited to extrapolation from a small number of salient features.

There are a number of popular methods to compute the motion and structure of the human face either from a single view or from a series of them. Recently there has been much development of methods that combine salient feature tracking and 3D structure estimation as one step [11,6,1]. Generally these methods lead to a search for the full set of parameters of a model and its camera matrices. Given this trait such methods are dubbed parametric methods. Such methods are well suited to tracking if the object is known. They restrict the object to a certain domain and can produce highly accurate results. This paper contributes another parametric fitting method that makes use of the popular KF as a way to fit the high dimensional 3DMM to images using a subset of salient feature points.

As it stands there are already a number of parametric methods. Baker and Matthews focus specifically on the extension of the popular Active Appearance Model (AAM) of Cootes et al [5]. They called the extension the 2D+3D AAM

A. Sattar and B.H. Kang (Eds.): AI 2006, LNAI 4304, pp. 519–528, 2006.

[11], where the key difference is a computationally efficient fitting strategy called Inverse Compositional Image Alignment [2]. In their method they first track a person specific AAM and then build a 3D Morphable Model using non-rigid factorization. The method requires a two step process of tracking and model construction before it can perform 3D tracking.

Dornaika and Ahlberk [6] also presented a modified AAM fitting algorithm. This was applied to the well known CANDIDE animation model using weak perspective projection. The solution demonstrated that a small generic model was capable of tracking accross pose end expression using their modified directed search. However the CANDIDE model is not a true statistical model of shape. Specifically it is not constructed from labelled training data.

In what could be considered a continuation of the AAM Anisetti et al. [1], extended the forwards additive approach of Lucas et al. [10] to fit the CANDIDE model to images. For the purpose of robust tracking they build a template from a single view, using a small number of points that correspond to vertices in their model. Using the defined template they are able to track across pose, illumination and expression. However in order for tracking to occur a 3D template needs to be defined.

In yet another development, Mittrapiyanuruk et al. [12] also demonstrated an accurate method for tracking rigid objects accross pose. Using a modified Gauss-Newton optimization to accommodate M-estimation they were able to demonstrate rotational accuracy within 5 degrees for estimates of yaw, pitch and roll. Similar to [1] a user labeled 3D template is required.

Our solution is similar to the mentioned methods but does not require hand labeling of a 3D template nor does it require a complex optimization framework. It only requires a somewhat accurate tracking of the feature points, which could be provided by an AAM or other tracking algorithm. Tracking can contain missing or inaccurate results because we make use of two Kalman filters, which allows for missing or bad data to be factored out of the fitting process. Our solution also exploits the relationship between the 3DMM and the subset of 2D feature points.

2 3D Morphable Models

A 3D Morphable Model (3DMM) is a statistical representation of both the 3D shape and texture of an object in a certain domain. 3DMMs are popular in the face domain [15,13] and are used for this task in our work. A 3DMM is built from 3D laser scans of human faces, which are then put into dense correspondence [3]. The 3D scans are then packed to form $\mathrm{M} \times \mathrm{N}$ shape and texture matrices. The dimensionality of the shape and texture matrices are very high, in our case we have 75 3D heads and 150,000 shape and texture points per head. Using the aligned heads PCA is used to construct a 3DMM and provide equations for shape and texture variation:

$$\hat{\mathrm{s}} = \bar{\mathrm{s}} + \mathrm{S} \cdot \mathrm{diag}(\sigma_\mathrm{s}) \cdot \mathrm{c_s} \quad \hat{\mathrm{t}} = \bar{\mathrm{t}} + \mathrm{T} \cdot \mathrm{diag}(\sigma_\mathrm{t}) \cdot \mathrm{c_t} \tag{1}$$

where $\hat{s}$ and $\hat{t}$ are novel $3N \times 1$ shape and texture vectors, $\bar{s}$ and $\bar{t}$ are the $3N \times 1$ mean shape and texture vectors, S and T are the $3N \times M$ column (eigenvectors) spaces of the shapes and textures, σ are the corresponding eigenvalues, c_s and c_t are shape and texture coefficients. These linear equations describe the variation of shape and texture within the span of 3D training data. The coefficients c_s, c_t are scaled by the corresponding eigenvalue σ_s, σ_t of the eigenvector S, T. In comparison 3DMMs are similar to the 2D Active Appearance Model (AAM) [5], although 3DMM model sizes are larger and pose and illumination changes can be addressed in the 3DMM fitting process.

2.1 3DMM Fitting

One of the most accurate fitting methods for 3DMMs is the gradient descent approach [15]. This is analysis by synthesis where the coefficients are inferred from a difference between a rendered head (image) and the input image. Effectively it is the minimization of the cost function:

$$\epsilon = ||F(c_s, c_t) - I||^2 \tag{2}$$

where F is a rendering function that when provided with shape and texture coefficients produces an image that is aligned with the input image, I. Speed is an issue for such a method and generally model coefficients are determined in a number of minutes.

2.2 Regularized 3DMM Point Based Fitting

In our work we would like to avoid optimization as much as possible to determine the shape coefficients for a 3DMM. Instead of fitting a 3DMM using texture information we use only 2D point features. Such features could be provided by any reliable face tracking algorithm, such as the AAM. We use the recently proposed method by Blanz et al [4] which is a concise and mathematically optimal method to reconstruct a 3DMM from a sparse set of either 2D or 3D feature points. The method has two advantages: 1) it relies on only linear operators and 2) operates in real-time (at the expense of model accuracy.) Using the assumption that only a small set of corresponding 2D feature points are available the method minimizes the cost function:

$$\epsilon = ||L \cdot V \cdot S \cdot \mathrm{diag}(\sigma_s)c_s - r||^2 \tag{3}$$

where L is a camera matrix containing the projection parameters, V is a $2P \times 3N$ subset selection matrix, S is the $(3N \times M)$ column (eigenvectors) space of the training shapes, σ are the corresponding eigenvalues, c_s are shape coefficients and r is a $2P \times 1$ set of feature points. Solving for the coefficients is done in a statistically optimal way using the pseudo-inverse operation and a regularization term [4]:

$$c_s = U^T \frac{s_i}{s_i^2 + \eta} Vy \tag{4}$$

(a) $\hat{s}$ (b) $\eta = 10^{-3}$ (c) $\eta = 10^{1}$ (d) $\eta = 10^{3}$ (e) $\bar{s}$

Fig. 1. Variation of η, using only 46 Farkas [7] points, on a random 3D head, $\hat{s}$. Showing that as η increases the estimate approaches the mean, $\bar{s}$.

where both U, V are orthogonal matrices provided by Singular Value Decomposition [4]. However, when using this approach, the regularization term (η) is not intuitively set. For our selection of η we performed an experiment (in section 4), which shows that for our 3D data $\eta = 1$ was optimal for tracking. Figure 1 demonstrates the effect of changing η, as it increases the shape estimate approaches the mean of the 3DMM.

3 KF 3DMM Point-Based Fitting

In this paper we extend Blanz' method to work on a video sequence using two Kalman filters (KFs). Specifically we combine both an EKF and a KF. Using an EKF we first solve the non-linear problem of exterior orientation and then with a linear KF we determine the model coefficients assuming that the Blanz shape estimate is optimal given a reduced set of salient features (this is shown in [4]). With this assumption in hand our use of the KF means that we infer the state parameters, which are the 3DMM shape coefficients, directly from feature measurements. This also relies on the assumption that identity should not vary across pose changes. During our work we found the assumption that identity should remain constant is highly dependant on two key factors: 1) accurate measurements of salient features and 2) accurate estimates of 3D to 2D model alignment (exterior orientation), both of which are hard to guarantee. Using KFs we can effectively factor out these inaccuracies as process noise. Using KFs our technique models the human face using a 3DMM under the assumption that there is one true identity state (plus some process noise) across a video sequence.

Figure 2 shows the filter series for one image in the sequence. Given a measurement, r, a prior estimate of the camera matrix, and and a prior estimate of the identity parameters, the pose filter, Ω_{P}, estimates the current camera matrix, $\mathrm{L_n}$. Using r and $\mathrm{L_n}$, the identity filter, Ω_{I}, estimates the current identity parameters. Its output are the coefficients, $\mathrm{c_s}$, estimated using the Blanz et al. method, Ω_{V}, for a given regularization term, η. From this Ω_{I} estimates the shape coefficients assuming it is measuring the state directly. The estimated camera matrix and shape coefficients then become the prior for the next image in the sequence.

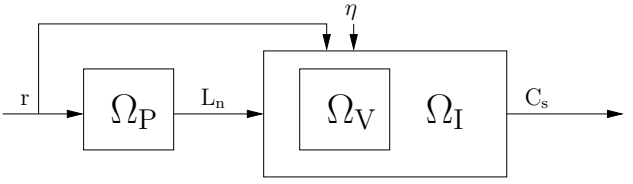


Fig. 2. DUAL KF Framework: Ω_P - pose filter, Ω_V - 3DMM fitting, Ω_I - identity filter

3.1 Kalman Filtering

Since its introduction by Rudolph Kalman in 1960 [8], the Kalman filter has been applied to a great many parameter estimation taks. This work demonstrates another application for the discrete filter. The Kalman filter is composed of two components, where the first is a set of time update equations:

$$\hat{x}_n = \hat{x}_{n-1} + Bu_{n-1} \tag{5}$$

$$P_n = AP_{n-1}A^T + Q \tag{6}$$

The time update equations project forward the current state $\hat{x}_n$ and error covariances P_n. This leads to the estimate of the system state which is used in the measurement update equations:

$$K_n = P_n H^T (HP_n H^T + R)^{-1} \tag{7}$$

$$\hat{x}_n = \hat{x}_n K_n (z_n - H\hat{x}_n) \tag{8}$$

$$P_n = I - K_n H P_n \tag{9}$$

The measurement update equations are responsible for feedback. The equations change the apriori error covariance estimate (P_n) and obtain an improved posteriori state estimate ($\hat{x}_n$) of the system. The Kalman gain (K_n) is also calculated, a variable that could be considered as how much to "trust" the estimate versus the measurement.

For the task of 3D face tracking the relationship between the model and projects are non-linear in measurement. Both the projection and rotations of the 3DMM introduce the non-linearity. It is for this reason that the Extended Kalman Filter (EKF) is used. The EKF is a linearisation of the measurement and time update equations for the Kalman filter [9]. It is applied when a non linearity is present in the update or measurement equations. The difference in equations is subtle and well covered in the literature.

3.2 Pose Estimation

Solving for three-dimensional pose and structure is accomplished in our framework using two KFs. The first (pose, Ω_p) filter is used to solve the well-known exterior-orientation problem, minimizing the cost function:

$$\epsilon = ||(s \cdot R \cdot V \cdot (\bar{s} + S \cdot \mathrm{diag}(\sigma) \cdot f(c_s)) + t) - r||^2 \tag{10}$$

This is the task of solving the rigid body motion between the 3D model (given a set of model coefficients) and the 2D measurements, r. As a projection model we chose the scaled orthographic camera model. Such a model is a reasonable approximation of the true perspective case when the distance between the object and the camera is not small. We also chose to encode the rotation as a first order approximation, using the canonical exponential form, shown here as the Rodriguez equation:

$$\mathrm{R} = \mathrm{I}_3 + \hat{\mathrm{w}} \sin\theta + \hat{\mathrm{w}}^2(\cos\theta - 1) \tag{11}$$

where $\hat{\mathrm{w}}$ is a skew symmetric matrix. If it is assumed that rotations are relatively small then the first order approximation is a good estimate for rotation. This in combination with the scaled projection forms the measurement equation for the EKF:

$$\begin{bmatrix} \mathrm{x_r} \\ \mathrm{y_r} \end{bmatrix} = \begin{bmatrix} \mathrm{s} & 0 & 0 \\ 0 & \mathrm{s} & 0 \end{bmatrix} \begin{bmatrix} 1 & -\mathrm{w_z} & \mathrm{w_y} \\ \mathrm{w_z} & 1 & -\mathrm{w_x} \\ -\mathrm{w_y} & \mathrm{w_x} & 1 \end{bmatrix} \begin{bmatrix} \mathrm{X_m} \\ \mathrm{Y_m} \\ \mathrm{Z_m} \end{bmatrix} + \begin{bmatrix} \mathrm{t_x} \\ \mathrm{t_y} \\ \mathrm{t_z} \end{bmatrix} \tag{12}$$

where $(\mathrm{x_r}, \mathrm{y_r})$ are a measurement point which correspond to the current iterations 3D model point $(\mathrm{X_m}, \mathrm{Y_m}, \mathrm{Z_m})$, s is scale, $(\mathrm{t_x}, \mathrm{t_y}, \mathrm{t_z})$ is translation and $(\mathrm{w_x}, \mathrm{w_y}, \mathrm{w_z})$ encodes the incremental rotation of the model. After estimation, the parameters are packed into the camera matrix, L.

3.3 Structure Estimation

Structure is determined using the second identity KF to estimate the shape parameters for the 3DMM. These shape parameters span the facial structure variation that is present in our database of 3D heads. This is achieved in two steps. Firstly we use Blanz's [4] method to effectively minimize the cost function shown in (3), using the camera matrix L provided by the first pose EKF. We then use this estimate of shape coefficients, $\mathrm{c_s}$, as input to a the second identity KF. Assuming that we are measuring the state directly and that the process noise will encode the variation due to bad estimates of pose and error of feature extractions. After applying the KF to the measurement the new estimate of the shape coefficients, $\mathrm{c_s}$, is produced.

3.4 The Dual KF Algorithm

Our dual KF tracking is a combination of steps. Firstly a measurement is extracted from the video frame. The measurement, r, and the current estimate of 3D shape (for the first frame of a sequence this is the mean 3DMM shape) is used as an input to the the (pose) EKF. The pose EKF provides an estimate of the rigid body motion and projection parameters that align the projection of the estimated 3D shape to the measurement from the frame. This alignment information is then used to construct the camera matrix, L.

The rotation estimate of the EKF pose filter is then used to update a global estimate of rotation. It encodes the incremental change in object rotation. This

is required because we assume that the feature tracking is happening in real-time and measurement differences between frames are small. A first order estimate for pose is good enough to model this situation. After this update L contains the global estimate of rotation.

Secondly we use the camera matrix L and extracted measurement, r, as input to the second identity filter (Ω_{I}). This provides an estimate of the 3D structure for the object, $\mathrm{c_s}$. This is an estimate that is restricted to the span of the 3DMM and in the form of 3DMM coefficients. As the last step we use the new estimate of 3D shape to update the current estimate of 3D shape. The camera matrix, L and the estimate of 3D shape are ready for use in the next frame.

4 Experiments and Results

Using 75 laser scanned heads from the USF database [14], we constructed a 3DMM that retained 95% variance in shape and texture. This 3DMM was then used as the input to our DUAL KF fitting algorithm to test pose and identity estimates. We also determined the required 3D mapping for 47 Farkas [7] points and their projection to 2D, to generate the ground truth measurements for all experiments.

4.1 Optimizing η

To determine an optimal regularization term, η, we performed a simple optimization. Our experiment consisted of generating 100 random heads (using the 3DMM) and optimizing for the parameter η, minimizing the error function:

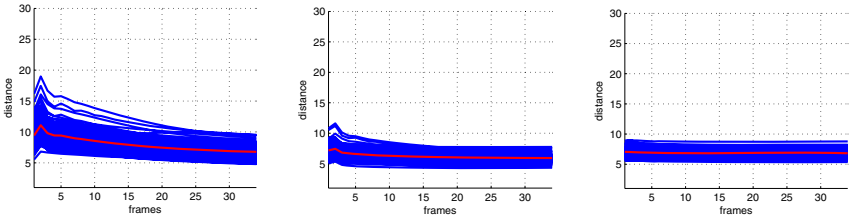
$$\epsilon = ||\Omega_{\mathrm{I}}(\eta, \mathrm{r}, \mathrm{L}) - \mathrm{R_s}||^2 \tag{13}$$

where $\Omega_{\mathrm{I}}(\mathrm{r}, \mathrm{L}, \eta)$ is the Kalman Filter estimate of identity, using the regularization term η, the camera matrix L and salient features r, and $\mathrm{R_s}$ is the set of ground truth shape coefficients for the individual random head. Across the sample of randomly generated heads we found consistently that the minimum error was attained at $\eta = 1$. This was surprising since unity did not appear to present the most perceptually pleasing shape estimate, visualy.

For another experiment and to demonstrate the influence of η on the tracking task we used a fixed identity rotating in a fixed motion and then varied η. This was repeated for 100 randomly generated identities. As expected the tracking performance improves as the value of η is accurately set to unity, this is with respect to the minimum Euclidean distance to the ground truth (figure 3). During this experiment we also found empirically that the DUAL KF identity estimates converged to a solution 95% of the time within as few as 25 frames.

4.2 Pose Estimate Accuracy

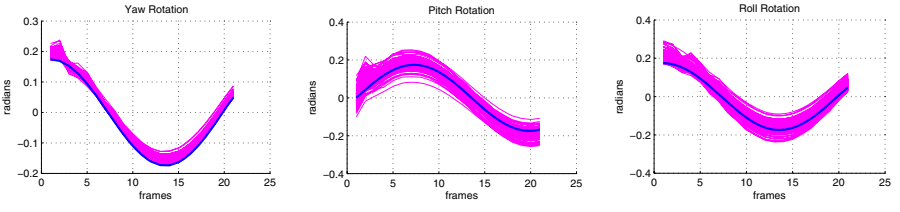
Using our 3DMM it was possible to accurately test pose; the 3DMM provides ground truth data. Thus using a specified set of yaw, pitch and roll functions



(a) $\eta = 10^{-1}$, mean = 9.94 (b) $\eta = 1$, mean = 5.98 (c) $\eta = 10^{1}$, mean = 6.77

Fig. 3. Identity KF convergence, distance versus measurements for different η

the method was examined by generating 100 random heads and using the dual KF approach to track the head motion. The accuracy was measured as the RMS error between the dual KF prediction and the actual measurement. When this error was below 1 pixel then the model was deemed to have accurately converged to the correct pose/identity.



(a) yaw (radians) / frames (b) pitch (radians) / frames (c) roll (radians) / frames

Fig. 4. Pose estimation convergence

The results in figure 4 show that the dual KF approach is capable of tracking the motion quiet successfully. In the case of the 100 random heads, the mean error in yaw pitch and roll was 2.21, −2.87 and −3.34 degrees respectively. We found empirically that the dual KF approaches pose estimates converged to a solution 95% of the time within as few as 4 frames.

4.3 Structure Estimate Accuracy

To test the structure estimates we used the 3DMM to generate a random head and then attempted to fit the dual KF in two situations, 1) given only a single measurement and 2) as the random head was rotated in a similar fashion to the pose experiment. We used the Euclidean distance (as per first experiment) between the known and estimated heads as a measure of fitting error. In both cases we used only 46 Farkas [7] feature points for estimation.

The first experiment demonstrated that the DUAL KF could be used to reconstruct the dense 3D shape given only one frame. The fitting was achieved by repeating the single frame measurement. In this case the pose of the model

(a) Farkas points (b) Estimate (c) Ground truth

Fig. 5. Example of shape estimation from a single frame, distance from ground truth (8.76)

(a) Farkas points (b) Estimate (c) Ground truth

Fig. 6. Example of shape estimation from a sequence, distance from ground truth (5.39)

was estimated in only 3 frames of measurements. An example fitting is shown in figure 5. This fitting achieved a distance from the ground truth, 8.76 units

The second experiment showed that the DUAL KF could refine an identity estimate across pose changes. Figure 6 shows the converged identity after a successful tracking. Note the distance from the ground truth, 5.39 units, smaller than the single image case.

5 Conclusion

In this paper we have proposed a new method to estimate the parameters for a 3DMM from a sparse set of salient features tracked across a video sequence. We have combined the benefits of Blanz's optimal 3D parameter estimation and the Kalman filter to achieve exactly this. Our method shows that the EKF estimates 3DMM pose parameters to an (average) accuracy of below 3 degrees, in as few as 4 measurements. Our method also shows that a single identity is inferred using a linear KF in as few as 25 images.

Acknowledgments

The authors would like to thank Sami Romdhani from the Faculty of Computer Science at UniBas for putting the USF data [14] into correspondence and the Australian Research Council for its continued funding.

References

1. M. Anisetti, V. Bellandi, L. Arnone, and F. Bevernia. Face tracking algorithm robust to pose, illumination and face expression changes: a 3D parametric approach. In *Proceedings of Computer Vision Theory and Applications*, 2006.
2. S. Baker and I. Matthews. Lucas-kanade 20 years on; a unifying framework: Part1. Technical Report 16, Robotics Institute, Carnegie Mellon University, 2002.
3. C. Basso, T. Vetter, and V. Blanz. Regularized 3D morphable models. In *Workshop on: Higher-Level Knowledge in 3D Modeling and Motion Analysis*, 2003.
4. V. Blanz, A. Mehl, T. Vetter, and H. Seidel. A statistical method for robust 3D surface reconstruction from sparse data. In *Second International Symposium on 3D Data Processing, Visualization and Transmission*, pages 293–300, 2004.
5. T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. In *Proc. European Conference on Computer Vision*, volume 2, pages 484–498. Springer, 1998.
6. F. Dornaika and J. Ahlberg. Face model adaptation for tracking and active appearance model training. In *British Machine Vision Conference*, 2003.
7. L. G. Farkas. *Anthropometry of the Head and Face.* Raven Press, New York, 2 edition, 1994.
8. R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.
9. T. Lefebvre, H. Bruyninckx, and J. De Schutter. Kalman filters for non-linear systems: a comparison of performance. *Control*, 77(7):639–653, May 2004.
10. B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.
11. I. Matthews, J. Xiao, S. Baker, and T. Kanade. Real-time comined 2D+3D active appearance models. In *Computer Vision and Pattern Recognition*, volume 2, pages 535–542, 2004.
12. P. Mittrapiyanuruk, G. DeSouza, and A. Kak. Accurate 3D tracking of rigid objects with occlusion using active appearance models. In *Proceedings of the IEEE Workshop on Motion and Video Computing*, 2005.
13. S. Romdhani and T. Vetter. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
14. S. Sarkar. USF DARPA humanID 3D face database. University of South Florida, Tampa, FL.
15. T. Vetter and V. Blanz. A morphable model for the synthesis of 3D faces. In *Siggraph 1999, Computer Graphics Proceedings*, pages 187–194, 1999.