

Dynamic Tracking of Facial Expressions Using Adaptive, Overlapping Subspaces

Dimitris Metaxas, Atul Kanaujia, and Zhiguo Li

Department of Computer Science, Rutgers University
{`dnm,kanaujia,zhli`}@`cs.rutgers.edu`

Abstract. We present a *Dynamic Data Driven Application System (DDDAS)* to track 2D shapes across large pose variations by learning non-linear shape manifold as overlapping, piecewise linear subspaces. The learned subspaces adaptively adjust to the subject by tracking the shapes independently using Kanade Lucas Tomasi(KLT) point tracker. The novelty of our approach is that the tracking of feature points is used to generate independent training examples for updating the learned shape manifold and the appearance model. We use landmark based shape analysis to train a Gaussian mixture model over the aligned shapes and learn a Point Distribution Model(PDM) for each of the mixture components. The target 2D shape is searched by first maximizing the mixture probability density for the local feature intensity profiles along the normal followed by constraining the global shape using the most probable PDM cluster. The feature shapes are robustly tracked across multiple frames by dynamically switching between the PDMs. The tracked 2D facial features are used to deform the 3D face mask. The main advantage of the 3D deformable face models is the reduced dimensionality. The smaller number of degrees of freedom makes the system more robust and enables capturing subtle facial expressions as change of only a few parameters. We demonstrate the results on tracking facial features and provide several empirical results to validate our approach. Our framework runs close to real time at 25 frames per second.

1 Introduction

Tracking deformable shapes across multiple viewpoints is an active area of research and has many applications in biometrics, facial expressions analysis and synthesis for deception, security and human-computer interaction applications. Accurate reconstruction and tracking of 3D objects require well defined delineation of the object boundaries across multiple views.

Landmark based deformable models like Active Shape Models(ASM)[1] have proved effective for object shape interpretation in 2D images and have led to advanced tools for statistical shape analysis. ASM detects features in the image by combining prior shape information with the observed image data. A major limitation of ASM is that it ignores the non-linear geometry of the shape manifold. Aspect changes of 3D objects cause shapes to vary non-linearly on a hyper-spherical manifold.

A generic shape model that would fit any facial expression is difficult to train, due to numerous possible faces and relative feature locations. In this work we present a generic framework to learn non-linear shape space as overlapping piecewise linear subspaces and then dynamically adapting the shape and appearance model to the Face of the subject. We do this by accurately tracking facial features across large head rotations and re-training the model specific to the subject using the unseen shapes generated from KLT tracking. We use the Point Distribution Models(PDM) to represent the facial feature shapes and use ASM to detect them in the 2D image. Our generic framework enables large scale automated training of different shapes from multiple viewpoints. The shape model is composed of the *Principal Components* that account for most of the variations arising in the data set. Our *Dynamic Data Driven framework* continuously collects different shapes by tracking feature points independently and adjusts the principal components basis to customize it for the subject.

2 Related Work

A large segment of research in the past decade has focused on incorporating non-linear statistical models for learning shape manifold. Murase et. al. [2] showed that pose from multiple viewpoint when projected onto eigenspaces generates a 2D hypersphere manifold. Gong et. al [3] used non-linear projections onto the eigenspace to track and estimate pose from multiple viewpoints. Romdhani et al. [4] proposed an ASM based on Kernel PCA to learn shape variation of face due to yaw. Several prominent work exist on facial feature registration and tracking, use appearance based models(AAM)[5,6]. [5] uses multiple independent 2D AAM models to learn correspondences between features of different viewpoints. The most notable work in improving ASM to learn non-linearities in the training data is by Cootes et. al[7] in which large variation in shapes is captured by parametric Gaussian mixture density, learned in the principal subspace. Unlike [5], our framework does not require explicit modeling of head pose angles. Although we use multivariate gaussian mixture model to learn initial clusters of the shape distribution, our subspaces are obtained by explicitly overlapping the clusters.

3 Learning Shape Manifold

An Active Shape Model(ASM) is a landmark based model that tries to learn a statistical distribution over variations in shapes for a given class of objects. Changes in viewpoint causes the object shapes to lie on a hyper-sphere and cannot be accurately modeled using linear statistical tools.

Face shape variation across multiple aspects is different across human subjects. It is therefore inaccurate to use a static model to track facial features for different subjects. Our approach to dynamically specialize the learned shape manifold to a human subject provides an elegant solution to this problem. However tracking shapes across multiple aspects requires modeling and synthesis of paths between the source and target shapes lying on a non-linear manifold. In

our framework non-linear region is approximated as a combination of multiple smaller linear subregions. For the first frame, we search the shape subspace iteratively by searching along the normals of the landmark points and simultaneously constraining it to lie on the shape manifold. The path between the source shape and the target shape is traversed by searching across multiple subspaces that constitute the non-linear shape surface. For the subsequent frames, we track the facial features independent of the prior shape model. The tracked shapes are used to learn Principal Components of the shape and appearance models that capture the variations specific to the human subject face. As a pre-requisite for shape analysis, all the 2D planar shapes are aligned to the common co-ordinate system using Generalized Procrustes Analysis[8]. The tangent space approximation \mathbf{T}_s projects the shapes on a hyper-plane normal to the mean vector and passing through it. Tangent space is a linear approximation of the general shape space so that the Procrustes distance can be approximated as euclidean distance between the planar shapes. The cluster analysis of shape is done in the global tangent space. We assume a generative multivariate Gaussian mixture distribution for both the global shapes and the intensity profile models(IPMs). The conditional density of the shape \mathbf{S}_i belonging to an N-class model $p(\mathbf{S}_i|\text{Cluster}) =$

$$\sum_{j=1}^N \gamma_j (2\pi)^{-\left(\frac{N}{2}\right)} \|\mathbf{C}_j\|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{S}_i - (\mu_j + P_j b_j))^T \mathbf{C}_j^{-1} (\mathbf{S}_i - (\mu_j + P_j b_j))\right\} \quad (1)$$

We also assume a diagonal covariance matrix \mathbf{C}_j . γ_j are the cluster weights and (μ_j, P_j, b_j) are the mean, eigen matrix and eigen coefficients respectively for the principle subspace defined for each cluster. The clustering can be achieved by the EM algorithm with variance flooring to ensure sufficient overlapping between the clusters. For each of the N clusters we learn a locally linear PDM using PCA and using the eigenvectors to capture significant variance in the cluster(98%). The intensity profiles for the landmark points also exhibit large variation when trained over multiple head poses. The change in face aspects causes the profiles to vary considerably for the feature points that are occluded. The multivariate Gaussian mixture distribution(1) is learned for the local intensity profiles model(IPM) in order to capture variations that cannot be learned using a single PCA model.

Overlapping Between Clusters: It is important that the adjacent clusters overlap sufficiently to ensure switching between subspaces during image search and tracking. We can ensure subspace overlap by using boundary points between adjacent clusters to learn the subspace for both the clusters. These points can be obtained as nearest to the cluster center but not belonging to that cluster.

4 Image Search in the Clustered Shape Space

Conventional ASM uses an Alternating Optimization(AO) technique to fit the shape by searching for the best matched profile along the normal followed by constraining the shape to lie within the learned subspace. The initial average



Fig. 1. Iterative search across multiple clusters to fit the face. The frames correspond to iteration 1(Cluster 1), iter. 3(Cluster 5), iter. 17(Cluster 7), iter. 23(Cluster 6) and final fit at iter. 33(Cluster 6) for level 4 of the Gaussian pyramid.

shape is assumed to be in a region near to the target object. We use robust Viola-Jones face detector to extract a bounding box around the face and use its dimensions to initialize the search shape. The face detector has 99% detection rate for faces with off-plane and in-plane rotation angles $\pm 30^\circ$. We assign the nearest Cluster_{*i*} to the average shape based on the mahalanobis distance between the average shape and the cluster centers in the global tangent space. The image search is initiated at the top most level of the pyramid by searching IPM along normals and maximizing the mixture probability density (1) of the intensity gradient along the profile. The model update step shifts the shape to the current cluster subspace by truncating the eigen coefficients to lie within the allowable variance as $\pm 2\sqrt{\lambda_i}$. The shape is re-assigned the nearest cluster based on the mahalanobis distance and the shape coefficients are re-computed if the current subspace is different from the previous.

The truncation function to regularize the shapes usually generates discontinuous shape estimates. We use the truncation approach, due to its low computational requirement and faster convergence. The above steps are performed iteratively and converges irrespective of the initial cluster of the average shape.

5 Dynamic Data Driven Tracking Framework

We track the features independent of the ASM by Sum of Squared Intensity Difference(SSID) tracker across consecutive frames[9]. The SSID tracker is a method for registering two images and computes the displacement of the feature by minimizing the intensity matching cost, computed over a fixed sized window around the feature. Over a small inter-frame motion, a linear translation model can be accurately assumed. For an intensity surface at image location $\mathbf{I}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{t}_k)$, the tracker estimates the displacement vector $\mathbf{d} = (\delta\mathbf{x}_i, \delta\mathbf{y}_i)$ from new image $\mathbf{I}(\mathbf{x}_i + \delta\mathbf{x}, \mathbf{y}_i + \delta\mathbf{y}, \mathbf{t}_{k+1})$ by minimizing the residual error over a window \mathcal{W} around $(\mathbf{x}_i, \mathbf{y}_i)$ [9].

$$\int_{\mathcal{W}} [\mathbf{I}(\mathbf{x}_i + \delta\mathbf{x}, \mathbf{y}_i + \delta\mathbf{y}, \mathbf{t}_{k+1}) - \mathbf{g} \cdot \mathbf{d} - \mathbf{I}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{t}_k)] d\mathcal{W} \quad (2)$$

The inter-frame image warping model assumes that for small displacements of intensity surface of image window \mathcal{W} , the horizontal and vertical displacement of the surface at a point $(\mathbf{x}_i, \mathbf{y}_i)$ is a function of gradient vector \mathbf{g} at that point.

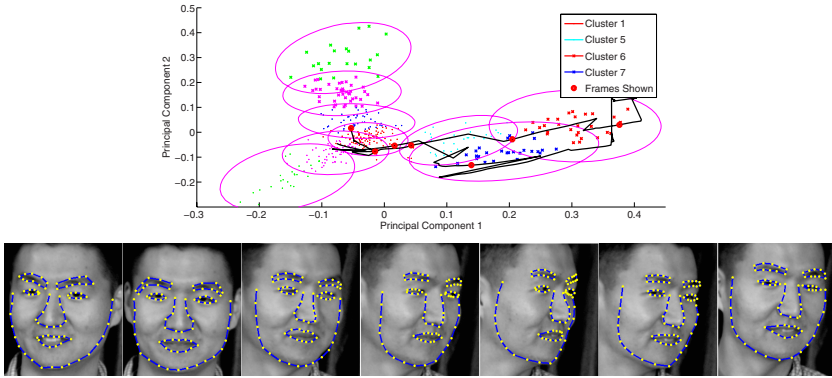


Fig. 2. (Best Viewed in Color) Tracking the shapes across right head rotation. **(Top)** The cluster projections on 2D space using 2 principal modes (for visualization) and the bounded by hyper-ellipsoid subspace. The right head rotation causes the shape to vary across the clusters. The red circles correspond to the frames 1, 49, 68, 76, 114, 262 and 281. The entire tracking path lies within the subspace spanned by the hyper-ellipsoids. **(Bottom)** The images of the tracking result for the frames shown as red markers in the plot.

The tracking framework generates a number of new shapes not seen during the training for ASM and hence provides independent data for our dynamic data driven application systems. Both the appearance (IPMs) and the shape models are composed of *Principal Vector* basis that are dynamically updated as we obtain new shapes and IPMs for the landmark points. For the shape \mathcal{X}_{i+1} at time step $(i + 1)$, the covariance matrix \mathbf{C}_i , is updated as

$$\mathbf{C}_{i+1} = ((N + i) - \frac{K}{N + i}) * \mathbf{C}_i + \frac{K}{N + i} * \mathcal{X}_{i+1}^T \mathcal{X}_{i+1} \quad (3)$$

where N is the number of training examples and i is the current tracked frame. The updated covariance matrix \mathbf{C}_{i+1} is diagonalized using power method to obtain new set of basis vectors. The subspace corresponding to these basis vectors encapsulates the unseen shape. The sequence of independent shapes and IPMs for the landmarks are used to update the current and neighboring subspaces, and the magnitude of updates can be controlled by the predefined learning rate K . The number of PCA basis vectors (eigenvectors) may also vary as a result of updation and specialization of the shape and the appearance model. Fig. 3 illustrates the applicability of our adaptive learning methodology to extreme facial expressions of surprise, fear, joy and disgust (not present in training images). For every frame we align the new shape \mathbf{Y}_t to the global average shape $\bar{\mathbf{X}}_{\text{init}}$ and re-assign it to the nearest Cluster _{i} based on mahalanobis distance. Finally after every alternate frame we ensure that the shape \mathbf{Y}_t obtained from tracking is a plausible shape by constraining the shape to lie on the shape manifold of the current cluster. Fig. 2 shows the path (projection on 2 principal components) of a shape (and

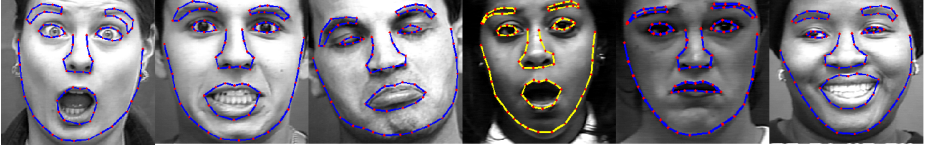


Fig. 3. 2D Tracking for extreme facial expressions

the corresponding cluster) for a tracking sequence when the subject rotates the head from frontal to full right profile view and back. The entire path remains within the plausible shape manifold spanned by the 9 hyper-ellipsoid subspaces.

6 Deformable Model Based 3D Face Tracking

Deformable model based 3D face tracking is the process of estimation, over time, of the value of face deformation parameters (also known as the state vector of the system) based on image forces computed from face image sequences. Our objective is to build a dynamically coupled system that can recover both the rigid motion and deformations of a human face, without the use of manual labels or special equipment. The main advantage of deformable face models is the reduced dimensionality. The smaller number of degree of freedom makes the system more robust and efficient, and it also makes post-processing tasks, such as facial expression analysis, more convenient based on recovered parameters. However, the accuracy and reliability of a deformable model tracking application is strongly dependent on accurate tracking of image features, which act as 2D image force for 3D model reconstruction. Low level feature tracking algorithms, such as optical flows, often suffer from occlusion, unrealistic assumptions etc. On the other hand, model based 2D feature extraction method, such as active shape model, has been shown to be less prone to image noises and can deal with occlusions. In this paper, we take advantage of the coupling of the 3D deformable model and 2D active shape model for accurate 3D face tracking. On the one hand, 3D deformable model can get more reliable 2D image force from the 2D active shape model. On the other hand, 2D active shape model will benefit from the good initialization provided by the 3D deformable model, and thus improve accuracy and speed of 2D active shape model. The coupled system can handle large rotations and occlusions. A 3D deformable model is parameterized by a set of parameters \mathbf{q} . Changes in \mathbf{q} causes geometric deformations of the model. A particular point on the surface is denoted by $\mathbf{x}(\mathbf{q}; \mathbf{u})$ with $\mathbf{u} \in \Omega$. The goal of a shape and motion estimation process is to recover parameter \mathbf{q} from face image sequences. To distinguish between shape estimation and motion tracking, the parameters \mathbf{q} can be divided into two parts: static parameter \mathbf{q}_s , which describes the unchanging features of a particular face, and dynamic parameter \mathbf{q}_m , which describes the global (rotation and translation of the head) and local deformations (facial expressions) of an observed face during tracking. The deformations can also be divided into two parts: \mathbf{T}_s for shape and \mathbf{T}_m for motion (expression), such

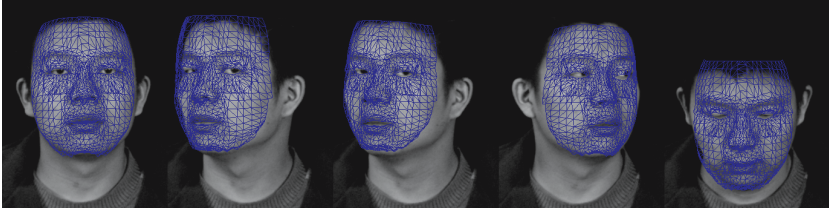


Fig. 4. 3D tracking results of deformable mask with large off-plane head rotations

that $\mathbf{x}(\mathbf{q}; \mathbf{u}) = \mathbf{T}_m(\mathbf{q}_m; \mathbf{T}_s(\mathbf{q}_s; s(\mathbf{u})))$. The kinematics of the model is $\dot{\mathbf{x}}(\mathbf{u}) = \mathbf{L}(\mathbf{q}; \mathbf{u})\dot{\mathbf{q}}$, where $\mathbf{L} = \frac{\partial \mathbf{x}}{\partial \mathbf{q}}$ is the model Jacobian. Considering the face images under a perspective camera with focal length f , the point $\mathbf{x}(\mathbf{u}) = (x, y, z)^T$ projects to the image point $\mathbf{x}_p(\mathbf{u}) = \frac{f}{z}(x, y)^T$. The kinematics of the new model is given by:

$$\dot{\mathbf{x}}_p(\mathbf{u}) = \frac{\partial \mathbf{x}_p}{\partial \mathbf{x}} \dot{\mathbf{x}}(\mathbf{u}) = \left(\frac{\partial \mathbf{x}_p}{\partial \mathbf{x}} \mathbf{L}(\mathbf{q}; \mathbf{u}) \right) \dot{\mathbf{q}} = \mathbf{L}_p(\mathbf{q}; \mathbf{u}) \dot{\mathbf{q}} \quad (4)$$

where the projection Jacobian matrix is

$$\frac{\partial \mathbf{x}_p}{\partial \mathbf{x}} = \begin{bmatrix} f/z & 0 & -fx/z^2 \\ 0 & f/z & -fy/z^2 \end{bmatrix} \quad (5)$$

which converts the 2D image forces to 3D forces. Estimation of the model parameters \mathbf{q} is based on first order Lagrangian dynamics [10], $\dot{\mathbf{q}} = \mathbf{f}_q$. Where the generalized forces \mathbf{f}_q are identified by the displacements between the actual projected model points and the identified corresponding 2D image features, which in this paper are the 2D active shape model points. They are computed as:

$$\mathbf{f}_q = \sum_j (\mathbf{L}_p(\mathbf{u}_j)^T \mathbf{f}_{image}(\mathbf{u}_j)) \quad (6)$$

Given an adequate model initialization, these forces will align features on the model with image features, thereby determining the object parameters. The dynamic system is solved by integrating over time, using standard differential equation integration techniques:

$$\mathbf{q}(t+1) = \mathbf{q}(t) + \dot{\mathbf{q}}(t)\Delta t \quad (7)$$

Goldenstein *et. al* showed in [11] that the image forces \mathbf{f}_{image} and generalized forces \mathbf{f}_q in these equations can be replaced with affine forms that represent probability distributions, and furthermore that with sufficiently many image forces, the generalized force converges to a Gaussian distribution. In this paper, we take advantage of this property by integrating the contributions of ASMs with other cues, so as to achieve robust tracking even when ASM methods and standard 3D deformable model tracking methods provide unreliable results by themselves.

7 Conclusion

In this work we have presented a real time DDDAS framework for detecting and tracking deformable shapes across non-linear variations arising due to aspect changes. Detailed analysis and empirical results have been presented about issues related to the modeling non-linear shape manifolds using piecewise linear models. The shape and appearance model updates itself using new shapes obtained from tracking the feature points. The tracked 2D features are used to deform the 3D face mask and summarize the facial expressions using only a few parameters. This framework has many application in face-based deception analysis and we are in the process of performing many tests based on relevant data.

Acknowledgement

This work has been supported in part by the National Science Foundation under the following two grants NSF-ITR-0428231 and NSF-ITR-0313184.

Patent Pending

The current technology is protected by patenting and trade marking office, "*System and Method for Tracking Facial Features*," Atul Kanaujia and Dimitris Metaxas, Rutgers Docket 07-015, Provisional Patent #60874,451 filed December, 12 2006. No part of this technology may be reproduced or displayed in any form without the prior written permission of the authors.

References

1. Cootes, T.: An Introduction to Active Shape Models. Oxford University Press (2000)
2. Murase, H., Nayar, S.: Learning and recognition of 3D Objects from appearance. IJCV (1995)
3. Gong, S., Ong, E.J., McKenna, S.: Learning to associate faces across views in vector space of similarities to prototypes. BMVC (1998)
4. Romdhani, S., Gong, S., Psarrou, A.: A Multi-View Nonlinear Active Shape Model Using Kernel PCA. BMVC (1999)
5. Cootes, T., Wheeler, G., Walker, K., Taylor, C.: View-Based Active Appearance Models. BMVC (2001)
6. Edwards, G.J., Taylor, C.J., Cootes, T.F.: Learning to Identify and Track Faces in Image Sequences. BMVC (1997)
7. Cootes, T., Taylor, C.: A mixture model for representing shape variation. BMVC (1997)
8. Goodall, C.: Procrustes methods in the statistical analysis of shape. Journal of the Royal Statistical Society (1991)
9. Tomasi, C., Kanade, T.: Detection and Tracking of Point Features. Technical Report CMU-CS-91-132 (1997)
10. Metaxas, D.: Physics-Based Deformable Models: Applications to Computer Vision, Graphics and Medical Imaging. Kluwer Academic Publishers (1996)
11. Goldenstein, S., Vogler, C., Metaxas, D.: Statistical Cue Integration in DAG Deformable Models. PAMI (2003)