

Facial Features Tracking for Gross Head Movement Analysis and Expression Recognition

Dimitris Metaxas
Department of Computer Science
Rutgers University
Piscataway, New Jersey 08854
Email: dnm@cs.rutgers.edu

Atul Kanaujia
Department of Computer Science
Rutgers University
Piscataway, New Jersey 08854
Email: kanaujia@cs.rutgers.edu

Peng Yang
Department of Computer Science
Rutgers University
Piscataway, New Jersey 08854
Email: peyang@cs.rutgers.edu

Abstract—We present a real-time framework for Action Units(AU) and Expression recognition based on facial features tracking and Adaboost. Accurate feature tracking faces several challenges due to changes in illumination, subject’s skin color, large head rotations, partial occlusions and fast head movements. We use models based on Active Shapes to localize facial features on the face in a generic pose. Shapes of facial features undergo non-linear transformation as the head rotates from frontal view to profile view. We learn the non-linear shape manifold as multiple-overlapping subspaces with different subspaces representing different head poses. Further, we use the tracked features to accurately extract bounded faces in a video sequence and use it for recognizing facial expressions. Our approach is based on coded dynamical features. In order to capture the dynamical characteristics of facial events, we design the dynamical haar-like features to represent the temporal variations of facial events.

I. TRACKING SHAPES ON NON-LINEAR MANIFOLD

Recent research in shape analysis and registration have proposed improved methodologies for searching in highly non-linear Riemannian manifold for the globally optimal shape. In this work we propose several improvements over the past shape registration techniques. Unlike previous works, shape analysis is not performed in the common tangent space. This removes the restriction that all shapes should be in the vicinity of the mean shape. In addition, we propose a framework to learn non-linear shape manifold as overlapping subspaces. The number of subspaces is learned directly from the data using the normality test for clusters.

The active shape learning has been formulated as the posterior optimization with the global prior shape model and the local image likelihood model. For the shapes \mathbf{S} learned as a set of \mathbf{N} landmark point locations $\mathbf{S} = \{x_1, y_1, \dots, x_N, y_N\}$, a PCA subspace is learned that captures the relevant variance in shapes (95%) by projecting the data set onto eigenvectors \mathbf{P} with largest eigenvalues

$$\mathbf{X} = \bar{\mathbf{X}} + \mathbf{P} * \mathbf{b} + \epsilon \quad (1)$$

where $\mathbf{X} = \Phi(\mathbf{S})$, Φ being the linear transformation for global scaling, translation, rotation and linearizing the shape. Planar shape distribution lies on highly non-linear Riemannian manifold. The distance metric on the non-linear manifold is approximated as procrustes distance by projecting the shapes onto the tangent plane of the mean shape. The shape model

learned in the tangent space is an accurate representation of the shapes in the vicinity of the mean shape $\bar{\mathbf{X}}$. However for the shapes far away from the mean shape, the large scaling of the shape vectors causes the learned PCA subspace to distort and generate unrealistic shapes. Kernel methods [1] target this problem by projecting the shapes into feature space where linear methods can be applied. These methods suffer from two principal drawbacks that prevent their applications to large scale shape analysis. Firstly kernel methods are inclined to overfitting due to more parameters and hence are not robust to outliers. Secondly kernel methods require pre-image mapping for projecting the shapes back from the feature space to the image space. This introduces additional inaccuracies in the shape model. To address this problem for large set of shapes, we propose to learn the non-linear shape manifold as multiple overlapping linear subspaces. The original shapes are first projected to a global tangent space so that the euclidean distance can be used for clustering. The shapes are aligned to a reference shape iteratively by computing $\Phi_i(\mathbf{S}) = \gamma \mathbf{R} \mathbf{S} + \mathbf{T}$ where the γ is the scaling factor, \mathbf{R} and \mathbf{T} as the rotation and the translation matrices respectively. The aligned shapes are clustered using Gaussian Mixture Model. Based on the class responsibilities in the tangent space, the original shapes are grouped into multiple clusters with subspaces learned within each cluster independently.

In order to ensure smooth manifold during shape search, adjacent subspaces should overlap sufficiently. The amount of overlap can be controlled by variance flooring during the EM algorithm for clustering the data set. In addition we artificially add 15% of cluster points from the neighboring clusters. This overlapping in the global tangent space ensures that shapes in the original image space generate overlapping linear subspaces.

A. Shape Search and Tracking

Since the shapes do not lie on a common tangent space, the euclidean metric cannot be used to compare shapes. We maximize the posterior using alternating optimization of likelihood by sampling along the normal of the landmarks followed by shape regularization using the cluster based prior model. The shape regularization is done as follows - (1) project the shape onto global tangent space and compute class

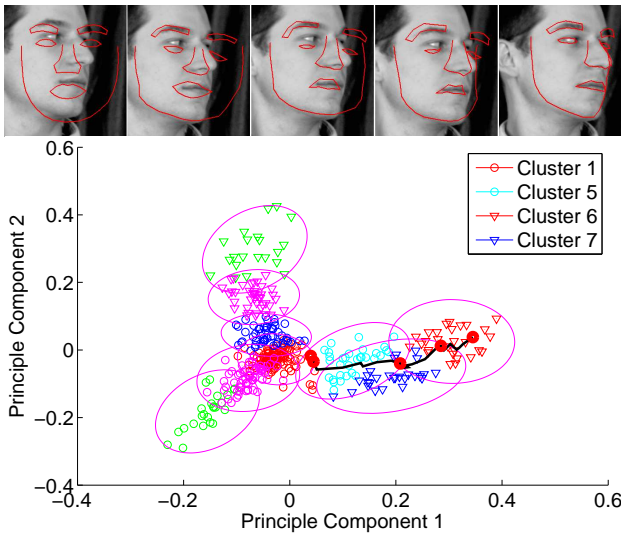


Fig. 1. Trajectory showing search for the optimal shape across clusters. The red circles denote the frames in the top row

conditionals, (2) project the shape to local tangent space of the cluster with maximum likelihood and constrain the shape to lie within its subspace. The overlapping between the clusters ensures smooth traversal across subspaces during search. Fig. 1 illustrates the algorithm showing the trajectory of the shape search. We track the features using Sum of Squared Intensity Difference(SSID) tracker across consecutive frames. 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. The inter-frame image warping model assumes that for small displacements of intensity surface of image window, the horizontal and vertical displacement of the surface at a point is a function of gradient vector at that point. During tracking, some features(ASM landmarks) eventually lose track due to blurring or illumination changes. To avoid this, at every frame we re-initialize the points which have lost track by searching along the normal and fitting the profiles.

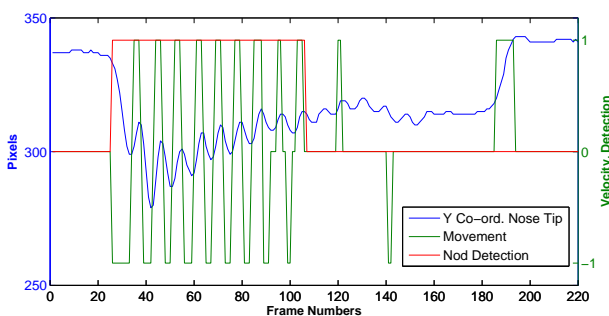


Fig. 2. Blue plot shows the y co-ord. of nose tip. Green plot shows the velocity direction with +1 indicating downward movement and -1 indicating the upward movement. The red plot shows the detection by thresholding distance between consecutive upward and downward movements, with 1 indicating presence of nodding. Also notice that small change in Y co-ordinate is ignored to keep false positives minimal.

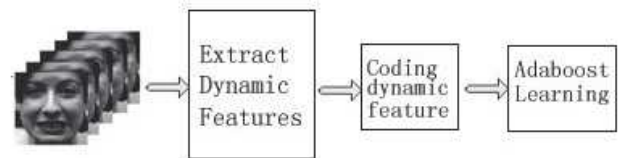


Fig. 3. Structure of the proposed framework

II. RECOGNIZING EMBLEMS

An Emblem denotes an event or movement that symbolizes an idea. Head nodding and shaking are the prominent emblems which are frequently encountered during any conversation and denotes agreement or dissent. Emblems associated with head movement like head tilt further convey different cognitive and emotional states of the subject during interrogation and interviewing. Robust tracking of facial features enables development of applications involving large scale video processing and high level event recognition. Detecting emblems during an interrogation provides useful cues about psychological state of the subject. Nose tip is the most stable tracked point as nose undergoes minimal deformation during tracking. Nose tip is approximated as average of lower 6 points of the nose feature. Head nodding and shaking can be recognized by detecting undulating patterns in y and x co-ordinates of nose tip respectively. Fig. 2 illustrates the nodding sequence detection plots.

III. RECOGNIZING ACTION UNITS

We propose a novel framework (fig. 3) for facial AU and expression recognition based on coded dynamical features. Our framework has three main components: dynamical feature extraction, coding dynamical features, and Adaboost learning. For the component of dynamical feature extraction, we design dynamical Haar-like features to capture the temporal variations of facial AUs and expressions. Inspired by the binary pattern coding [5], we analyze the distribution of each dynamical haar-like feature, and we create a code book for it. According to the code books, the dynamical haar-like features are further mapped into binary pattern features. Finally the Adaboost is used to learn a set of discriminating coded features for facial AU and expression recognition.

REFERENCES

- [1] S. Romdhani, S. Gong, and A. Psarrou, *A Multi-View Nonlinear Active Shape Model Using Kernel PCA*, BMVC, 1999.
- [2] A. Kanaujia, Y. Huang and D. Metaxas, *Emblem Detections by Tracking Facial Features*, International Workshop on Semantic Learning Applications in Multimedia, CVPR 2006
- [3] A. Kanaujia, Y. Huang and D. Metaxas, *Tracking Facial Features using Mixture of Point Distribution Models*, ICVGIP 2006
- [4] A. Kanaujia and D. Metaxas, *Large Scale Learning of Active Shape Models* ICIIP 2007
- [5] T. Ojala and M. Pietikainen, *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002
- [6] P. Yang, Q. Liu, D. Metaxas, *Boosting Coded Dynamic Features for Facial Action Units and Facial Expression Recognition* IEEE Computer Vision and Pattern Recognition, 2007