QUERY DRIVEN LOCALIZED LINEAR DISCRIMINANT MODELS FOR HEAD POSE ESTIMATION

Zhu Li[1], Yun Fu[2], Junsong Yuan[3], Thomas S. Huang[2], and Ying Wu[3]

[1]Multimedia Research Lab, Motorola Labs, Schaumburg, IL 60196, USA
[2]Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[3]Dept. of EECS, Northwestern University, Evanston, IL 60208, USA

ABSTRACT

Head pose appearances under the pan and tilt variations span a high dimensional manifold that has complex structures and local variations. For pose estimation purpose, we need to discover the subspace structure of the manifold and learn discriminative subspaces/metrics for head pose recognition. The performance of the head pose estimation is heavily dependent on the accuracy of structure learnt and the discriminating power of the metric. In this work we develop a query point driven, localized linear subspace learning method that approximates the non-linearity of the head pose manifold structure with piece-wise linear discriminating subspaces/metrics. Simulation results demonstrate the effectiveness of the proposed solution in both accuracy and computational efficiency.

1. INTRODUCTION

Appearance based manifold modeling and subspace learning approaches have been found to be very effective in face recognition and head pose estimation applications. Unsupervised approaches like Eigenfaces [15], learn the subspace for recognition via the Principle Component Analysis (PCA) of the face manifold, supervised approaches like Fisherfaces [1] learn the metric for recognition from labeled data via the Linear Discriminant Analysis (LDA). The incorporation of the labeling data improves the performance by finding subspaces where discriminating features are preserved, while non-discriminating features are dropped. Examples of linear approaches in head pose estimation can be found in [4] [14] [3].

The PCA/LDA approaches for head pose estimation are fundamentally limited because of the non-linearity of the underlying manifold structure, and richness in local variations. In recent years, non-linear methods for high dimensional non-linear data modeling, LLE [13], Graph Laplacian [2], have achieved very good results in finding manifold structure through embedding a graph structure of the data derived from the affinity modeling. However these solutions are non-linear functions dependent on the training data, and can not directly handle unknown query data. For example, it is difficult to embed a new query point into the learned non-linear manifold, without recalculate the embedding with the whole dataset.

LEA [5] and Laplacian faces [8] partially solves this problem by finding a compromise by linearizing the solution to the original graph embedding problem. Even though the solutions have better performances than pure Euclidean metric based approaches like Eigenfaces and Fisherfaces, the solution is still a global linear solution. When the problem space is large, e.g, large population face recognition, head pose estimation, the discriminating power of the subspace/metric learnt decreases.

To overcome the non-linearity of the problem, kernel method [12] has been employed to model non-linearity through a kernel mapping of training data to a higher dimension space with richer structure for discriminating metric learning. This approach has been found to be effective in face detections/recognitions [10], however, the solution typically involves a quadratic optimization with an $n \times n$ Hessian matrix, where n is the size of the training sample, which can be prohibitive in complexity.

To address these issues, we developed a piece-wise linear subspace/metric learning method to map out the global non-linear structure for head pose estimation. This approach has been applied successfully with video indexing/retrieval problem [11] with good results, where the hierarchical structure among each local neighborhood is characterized by a kd-tree.

In this work, each head pose appearance local neighborhood is identified by the query point, and there is no hierarchy in the global structure. By localizing, the problem size has been reduced from the original size n to some $n' << n$. This allows for better modeling for a given model and the reduction in problem size can also makes kernel method computationally more tractable.

The paper is organized into the following sections: In section 2 we lay out the formulation of the problem and give it our solutions. In section 3 the data set is explained and simulation results presented. In section 4, we draw the conclusion and outlying future works.

2. HEAD POSE ESTIMATION PROBLEM

2.1. The Head Pose Estimation Problem

In a head pose estimation problem, typically a training data set of m subjects with n poses characterized by the tilt and pan angles, $\{P_k=[a_k, b_k]\ |_{k=1}^{n}\}$, is given as aligned and cropped w x h image luminance data, $X = \{X_j\ |_{j=1}^{n\times m}\}$, where $X_j \in R^{wxh}$, is the vectorized image data. A set of non-overlapping test data also with m subjects and n poses is also given, denoted as $Y = \{Y_j\ |_{j=1}^{n\times m}\}$, where $Y_j \in R^{wxh}$, The pose estimation consists of a subspace/discriminating metric learning phase, where the objective is to find a d-dimensional subspace basis A: dxD, where $D=wxh$, such that classifiers like SVM and nearest neighbor [7] classification achieves the highest accuracy.

In this work, we optimize on the metric learning part and use a nearest neighbor classifier though out the simulations. The objective can therefore be stated as,

$$\min_A \| AX_j - AY_i \|, \quad \text{if } P(X_j) = P(Y_i)\ \forall i, j$$

where $P()$ is the tilt and pan angle label function that returns the pose id P_k for data with labels. As discussed in the introduction section, A is to characterize a subspace where pose variations are captured, while inter-subject variations are minimized.

2.2. Global Linear Solution

An obvious solution to this problem is to use a global LDA model, where inter-pose appearances are mapped far-apart while intra-pose appearances scatters are kept constant,

$$A=\arg\max_A |A^T S_B A|, \text{ s.t. } |A^T S_W A|=1 \quad (1)$$

in which the between class scatter S_B is given as,

$$s_B = \sum_{k=1}^{n} n_k (\bar{X}_k - \bar{X})(\bar{X}_k - \bar{X})^T \quad (2)$$

In Eq. (2), $n_k=n$ is the number of samples in class k. The within class scatter S_W, is given as,

$$s_W = \sum_{k=1}^{n} \sum_{P(X_j)=k} (X_j - \bar{X}_k)(X_j - \bar{X}_k)^T \quad (3)$$

Notice that S_B and S_W are functions of all X_j's, and therefore the global subspace is a function of training data, i.e, $A(X)$. In the graph embedding interpretation, LDA embeds a graph with edges connecting all intra-class points [8].

As discussed in the introduction, a single model $A(X)$ contains only $wxhxd$ variables to characterize the subspace where lies the manifold spanned by n x m data points. For large size problems such as head pose estimation, as n x m grows, the number of edges in an affinity graph grows exponentially.

Indeed, in [8], the connections between LPP and PCA/LDA are explained as different graph construction strategy. In LPP, the objective is to,

$$\min_A \sum_{j,k} (Ax_j - Ax_k) S_{j,k}$$

Where $S_{j,k}$ is the affinity measure mapped from Euclidean distance between x_j and x_k via a heat kernel [8].

To reflect the discriminating model power to characterize inter and intra-class points relationships, it is necessary to characterize the tradeoffs between the complexity of the embedded graph, $G(X)$, and the expressive power of the model, $A(X)$.

Let the graph $G(X)$ be denoted by its vertices and edge set, $V(X)$ and $E(X)$. Let us define the model discriminating power coefficient (DPC) of X with linear model $A(X)$: dxD, as the ratio between the number of variables in the model and number of edges involved,

$$K(A) = \frac{w\times h\times d}{|E(X)|} \quad (4)$$

For a given model, as the number of embedded graph edges grow, the DPC decreases. To improve DPC of the model, graph embedding techniques like LPP[8], and LEA [4] remove edges with no significance for clustering/discriminating by k-NN search / ε-thresholding, or from ground truth. The $|E(X)|$ for n-point PCA, m-class LDA, and LPP/LEA with K neighbors per data are given below,

$$|E(X)| = \begin{cases} \binom{n}{2}, & PCA \\ \sum_{j=1}^{m} \binom{n_j}{2},\ s.t. \sum_{j=1}^{m} n_j = n, & LDA \\ nK, & LPP/LEA \end{cases}$$

Notice that PCA/LDA edges grow exponentially, while LPP/LEA edges only grow linearly.

Instead of improving DPC by reducing edges of a global graph $G(X)$, in this work, we achieve higher DPC by also reducing the number of vertices in $G(X)$. As motivated by [11], we could partition the training data into a hierarchical structure via kd-tree, and for each data subset $X_{(t)}$ corresponding to sub tree t, we can compute its model via PCA, LDA, LPP or, LEA, as $A_t=A(X_{(t)})$. We end up with a set of linear models with hierarchical structure this way. In this case we have a problem similar to solutions like LLE [13], and Graph Laplacian [2], where it is difficult to select the right model / hierarchical levels that offer the best discriminating power for pose estimation, especially if query point lies on the boundaries of kd-tree partitions.

2.2. Localized Linear Solution

To solve this, instead of building models $A(X_{(t)})$ for each data partition node in kd-tree, a query point driven local neighborhood based model is computed. Let $q \in R^{w \times h}$ be an unknown head pose image, kNN neigbourhood of q is computed as $X_{(q)}$. The local linear discriminant model for this query point is computed as,

$$A(X, q) = \arg\max_{A} |A^T S_B A|, \text{ s.t. } |A^T S_W A|=1, \quad (5)$$

where,

$$s_B = \sum_{n_k \geq n_0} n_k (\overline{X}_k - \overline{X}_{(q)})(\overline{X}_k - \overline{X}_{(q)})^T \quad (6)$$

in which n_0 is the minimum number of sample per class requirement. This is used to remove trivial points with limited impact of graph structure. Similarly, the within class scatter is computed as,

$$s_W = \sum_{k: n_k \geq n_0} \sum_{P(X_j)=k, X_j \in X_{(q)}} (X_j - \overline{X}_k)(X_j - \overline{X}_k)^T \quad (7)$$

Notice that the model becomes a function of both training data set X and query point q in Eq. (5), and the DPC for this solution is given by,

$$K(A(X, q)) = \frac{w \times h \times d}{|E(X_{(q)})|}$$

where, the number of local graph edges, $|E(X_{(q)})|$, for K-NN localized PCA (l-PCA) and localized LDA (l-LDA) are given by,

$$\left\{ \begin{array}{ll} \binom{K}{2}, & l-PCA \\ \sum_{j=1}^{m'} \binom{n_j}{2} s.t. \sum_{j=1}^{m'} n_j = K, & l-LDA \end{array} \right\}$$

Notice that linearized graph embedding techniques like LEA [4] and LPP [8] can also be applied in this framework. The derivations are omitted.

The metric/subspace $A(X, q)$ offers better discriminating power than $A(X)$ in the sense that the model is well adapted to the local data and the DPC can be tuned to achieve better recognition performance.

3. DATA SET AND SIMULATION

For simulation we obtained head pose data from Pointing '04 data set [9], and [6]. The data set consists of *15* sets of images for *m=15* subjects, wearing glasses or not and having various skin colors. Fig. 1 shows some example Pointing'04 head-pose images.

Each set contains *2* series of *n=93* images of the same person at different poses. The first series, X, is used for training, and the second, Y, for testing. The pose or head orientation is determined by pan and tilt angles, which vary from -90° to +90°. Various poses with different pan and tilt angles for the same person is shown in Fig. 2.

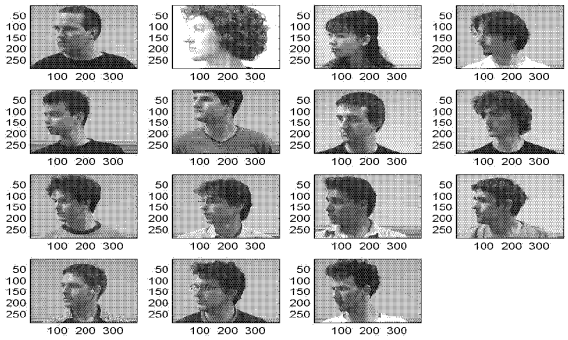

Figure 1. Point'04 head pose images

To demonstrate the discriminating metric performance changes with the model DPC, we set up some experiments with localized LDA and LPP. For each query point, q, a local neighborhood size K and dimensionality of subspace d are selected to compute local metrics, $A(X, q)$: d x D. The local LDA metric based pose estimation error-rate and its discriminating power coefficients (DPC) are plotted in Fig. 3 below for *d=32*.

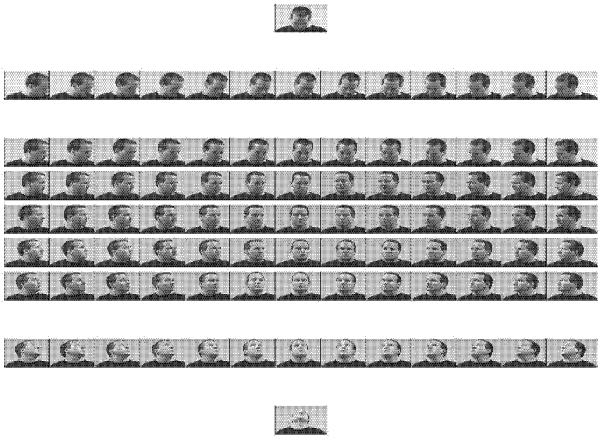

Figure 2. Tilt/Pan angles examples

Notice that in Fig. 3, the performance falls off as DPC decreases beyond certain points. When the local neighborhood is too small the metric learnt does not generalize well either, as also indicated by performances at very high DPC levels. The overall trend of recognition performance decreases with DPC increase is well demonstrated in Fig. 3 for both localized LDA and LPP cases.

The proposed solution performs well compared with state of art global graph embedding techniques like LPP. The error rates for pan and tilt angles recognitions are shown in Table 1. Notice that supervised methods, i.e, the graph pruning utilizes labeling information, perform better than non-supervised methods. Among them, l-LDA performs the best overall, and achieves the best results in 3 out of 4 cases, followed by another supervised approach, LPP$^{(1)}$, and also close with LDA performance. The localized LPP method does not perform as well as l-LDA.

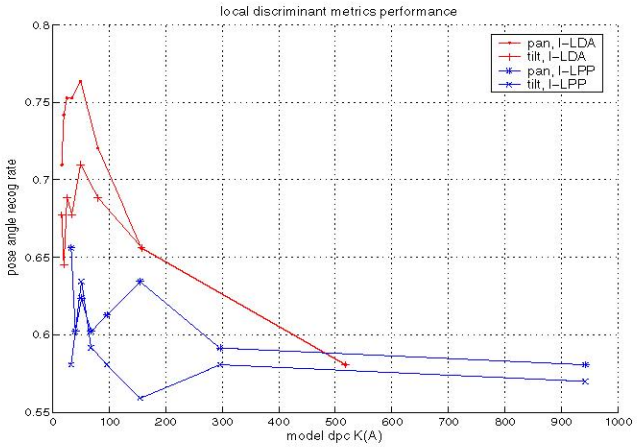

Figure 3. Model DPCs and pose angles recognition rates

Non-supervised methods do not perform as well. PCA and LPP[2] all have high error rates in estimation performance. The *l*-PCA method mitigates the lack of labeling information by localization, and rather surprisingly, performs well and close with supervised global methods like LDA and LPP[1]. This is another indication of benefit brought by localization.

Table 1. Pose estimation error rates

	Pan (d=16)	Tilt (d=16)	Pan (d=32)	Tilt (d=32)
PCA	*33.5*	*44.3*	*26.9*	*35.1*
LDA	*30.1*	*33.3*	*25.8*	*26.9*
LPP[1]	*30.1*	*31.2*	*24.7*	***22.6***
LPP[2]	*67.7*	*76.3*	*63.4*	*61.3*
l-PCA	*25.2*	*37.8*	*24.5*	*37.6*
l-LPP	*33.9*	*44.5*	*29.2*	*40.2*
l-LDA	***20.4***	***30.7***	***19.1***	*30.7*

Table 2. Computational complexity (sec) per recognition

	K=30	*K=60*	*K=90*
l-LDA, *d=16*	*0.105*	*0.132*	*0.121*
l-LDA, *d=32*	*0.145*	*0.146*	*0.176*
l-LPP, *d=16*	*0.094*	*0.122*	*0.104*
l-LPP, *d=32*	*0.132*	*0.116*	*0.144*

The computational complexity of the localized metric for pose recognition is summarized in Table 2, with various dimensions and sizes of neighborhood. Notice that the average speed of pan/tilt angles recognition is about 7 to 10 per sec, with un-optimized Matlab code running on an 2.0G Hz PC.

4. CONCLUSION

In this work we developed a query point driven, piece-wise linear local subspace learning method for head pose estimation. The discriminating power of the local metric is enhanced through pruning embedded graph edges by limiting the model to an appropriate local neighborhood. Simulation results demonstrate the advantage over some existing state-of-art solutions.

In the future, we will apply diffusion distance metrics in embedded graph vertices/edges pruning, and also apply kernel method to the subspace/metric modeling, taking advantage of the reduced problem size through localization.

5. REFERENCES

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", *IEEE Trans. on PAMI*, vol. 19(7), pp. 711-720, Jul. 1997.

[2] M. Belkin, and P. Niyogi, "Laplacian Eigenmap and Spectral Tech for Embedding and Clustering", *Proc. of NIPS*, Sep. 2001.

[3] L.B. Chen, L. Zhang, Y.X. Hu, M.J. Li and H.J. Zhang, "Head Pose Estimation Using Fisher Manifold Learning," *IEEE Workshop on AMFG'03*, pp. 203-207, 2003.

[4] Y. Fu and T.S. Huang, "Graph Embedded Analysis for head Pose Estimation," *IEEE Conf. on FG'06*, Southampton, UK, pp. 3-8, 2006.

[5] Y. Fu and T.S. Huang, "Locally Linear Embedded Eigenspace Analysis,"http://www.ifp.uiuc.edu/~yunfu2/papers/LEA-yun05.pdf, IFP-TR, UIUC, 2005.

[6] N. Gourier, D. Hall, J. L. Crowley, "Estimating Face Orientation from Robust Detection of Salient Facial Features," *IEEE ICPR Pointing'04 Workshop*, 2004.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *Elements of Stat Learning*, Springer Series in Stats, 2002. .

[8] X. He, S. Yan, Y. Hu, P. Niyogi, and H.- J. Zhang, "Face recognition using Laplacianfaces", *IEEE Trans. on PAMI*, vol. 27(3), pp. 1-13, Mar. 2005.

[9] J. Letissier, and N. Gourier, "The Pointing'04 Data Sets," *International Workshop on Visual Observation of Deictic Gestures (POINTING'04)*, 2004.

[10] S.Z. Li, Q.D. Fu, L. Gu, B. Scholkopf, Y.M. Cheng, H.J. Zhang. "Kernel Machine Based Learning for Multi-View Face Detection and Pose Estimation". *Proc. of 8th IEEE International Conference on Computer Vision.* Vancouver, Canada. July 9-12, 2001.

[11] Z. Li, L. Gao, and A. K. Katsaggelos, "Locally Embedded Linear Subspaces for Efficient Video Indexing and Retrieval", *Proc of IEEE Int'l Conf on Multimedia & Expo* (ICME), 2006.

[12] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An Introduction to Kernel-Based Learning Algorithms", *IEEE Trans. Neural Networks*, vol. 12(2), Mar. 2001.

[13] S.T. Roweis, and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding", *Science*, vol. 290, Dec. 2000.

[14] J.L. Tu, Y. Fu, Y.X. Hu, and T.S. Huang, "Evaluation of Head Pose Estimation For Studio Data," R. Stiefelhagen and J. Garofolo (Eds.): Multimodal Technologies for Perception of Humans, *CLEAR 2006*, LNCS 4122, pp. 281-290, 2007.

[15] M. Turk and A. P. Pentland, "Face recognition using Eigenfaces", *Proc. of IEEE CVPR*, 1991.