A SPARSE SAMPLING MODEL FOR 3D FACE RECOGNITION

Jun Yuan, Ashraf A. Kassim

Department of Electrical and Computer Engineering, National University of Singapore, Singapore

ABSTRACT

We propose a sparse sampling model as a feature selection tool for 3D face recognition, and compare its performance with the traditional dense subspace methods. The sparse LDA algorithm is applied to find the most discriminative features on range and texture images, meanwhile achieving the region selection purpose. The selected regions from both shape and texture are demonstrated. The classification remains accurate even at a high level of sparsity. To generalize the model, a probability density function is then estimated according to the selected region, and new samples are drawn accordingly to form new sparse features for classification. We also use the local coordinate system to make the sampling process more efficient, and insensitive to geometric transforms.

Index Terms— 3D face recognition, sparse sampling, LDA, feature selection

1. INTRODUCTION

Facial images are often represented as high-dimensional pixel arrays for recognition. However, they often belong to an intrinsically low dimensional manifold [1]. The subspace learning algorithms, such as PCA [2], LDA [3] are designed to find these low dimensional features for representation and classification purposes. Often, a PCA is applied to a face dataset (with proper pre-processing, such as registration and illumination correction) for dimension reduction, followed by an LDA which performs the classification task. The training of LDA is briefly reviewed as follows:

$$\boldsymbol{S_b p_i} = \mu(\boldsymbol{S_w} + \lambda \boldsymbol{I})\boldsymbol{p_i} \qquad (1)$$

where $\boldsymbol{p_i}$ is a set of generalized eigenvectors of inter-class and intra-class scatter $\boldsymbol{S_b}$ and $\boldsymbol{S_w}$. The regularization term $\lambda \boldsymbol{I}$ is applied to enhance the pixel independence and enables LDA to be applied in high dimensional spaces, with potential better classification results [4] [5].

Another problem related to the subspace methods is the interpretability, i.e. how each input variable affects the output. It is often that a subset of inputs is selected rather than the whole set of predictors, since many inputs can be correlated or simply irrelevant to the output. One popular approach is using the LASSO [6] or Elastic Net [7] to perform shrinkage and select the subset that best describes the output. The sparse PCA [8] and sparse LDA (SLDA) [4] are further developed to find sparse projections for representation and classification. In many applications including face recognition, much information in the image is redundant as the neighboring pixels are correlated; it is natural to use the selection techniques to seek a meaningful subset for analysis. Besides, the shrinkage/selection methods tend to reduce the model variance [9], especially in high dimensions; thus the performance can be better than the full model.

We applied the SLDA to the BU-3DFE dataset [10]. The dataset is registered and converted to range images and graylevel textures. We show from both range and texture information that the pixels/regions selected for classification have a tendency to cluster around the crucial regions (nose and eyes). These regions are more important than "flat" regions in classification. The Fisherfaces can be very sparse (1% ~ 2%) while the classification rate still compatible with the dense model.

A sparse sampling technique is also proposed based on the local regions selected. A probability density function (PDF) is fitted over these regions and the sparse features can be reproduced accordingly. The regular dense classification model can then be applied on these sparse sampled points. This process can be much faster than the SLDA in the selection process. Further, we bring in the triangular local coordinates for this sampling process. The coordinate system is invariant under affine transform, and avoids complex interpolation process in pose correction. This also makes the recognition model very efficient to execute.

2. PROPOSED METHODS

The LDA is reviewed as an optimal scoring problem [5]:

$$(\boldsymbol{p_i}, \boldsymbol{\theta_i}) = \underset{\boldsymbol{p_i}, \boldsymbol{\theta_i}}{\operatorname{argmin}} \left|\left|\boldsymbol{Y\theta_i} - \boldsymbol{Xp_i}\right|\right|^2,$$
$$s.t.\ \frac{1}{n}\boldsymbol{\theta_i^t Y^t Y \theta_i} = 1, \frac{1}{n}\boldsymbol{\theta_i^t Y^t Y \theta_j} = 0 \qquad (2)$$

where $\boldsymbol{X}$ is the data matrix, and $\boldsymbol{Y}$ is a matrix of dummy variables indicating the class of each data samples. The pair $\boldsymbol{p_i}$ and $\boldsymbol{\theta_i}$ form a set of successive discriminant projections

and the respective "scores". Optimizing (2) is equivalent to the LDA formulation (1).

The SLDA adds a constraint to the projection vector $\boldsymbol{p_i}$, with the cost function modified as shown below [4]:

$$L(\boldsymbol{p_i}, \boldsymbol{\theta_i}) = \left|\left|\boldsymbol{Y\theta_i} - \boldsymbol{Xp_i}\right|\right|^2 + \lambda J(\boldsymbol{p_i})$$

$$J(\boldsymbol{p_i}) = (1-\alpha)\cdot\frac{1}{2}\left|\left|\boldsymbol{p_i}\right|\right|_2^2 + \alpha\left|\left|\boldsymbol{p_i}\right|\right|_1 \qquad (3)$$

The L_1 and L_2 penalty on $\boldsymbol{p_i}$ forms an Elastic Net problem [7]. The problem becomes LASSO if L_2 penalty is absent [6]. The solution of (3) is given by alternating optimization of between $\boldsymbol{p_i}$ and $\boldsymbol{\theta_i}$; the projection vectors $\{\boldsymbol{p_i}\}$ are solved via Elastic Net, and the scores are updated accordingly by a normalization process. The parameter λ controls the shrinkage/selection, while α controls the proportion of L_1 and L_2 penalty.

The Elastic Net can select a group of correlated variables while the LASSO tends to randomly pick one [7] [11]. The effect is visualized in the following figures, which consists of superimposed Fisherfaces (top 20) from SLDA with both range and texture images.

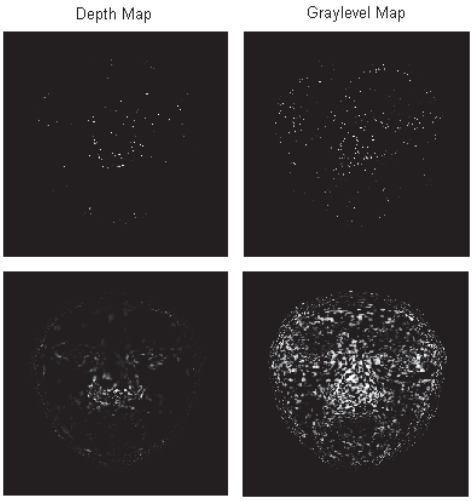


Fig. 1 Fisherfaces from LASSO and Elastic Net
Top: Lasso, $(\lambda, \alpha) = (0.05,1)$;
Bottom: Elastic Net, $(\lambda, \alpha) = (0.05, 0.1)$

The grouping effect of Elastic Net can be seen clearly. The pixels selected cluster around the nose and eyes, indicating the information in these regions are more crucial for classification. This conclusion is coincides with [12].

It is natural to use these regions at the beginning of the classification process, rather than applying SLDA to the full vector space, given the prior knowledge of these regions. It also alleviates the computational cost imposed by LDA in high dimensional space; since PCA cannot be applied in advance otherwise we again get a full model.

A common way is to fit a probability density function (PDF) to the regions selected by SLDA, and sample pixels accordingly. We use kernel density estimation (Parzen Window) to fit the PDF with a Gaussian smoother. The result from LASSO can be used for convolution, with proper radius where the resulting PDF matches the Elastic Net, since the latter groups correlated variables. The PDFs of range images and textures are shown below.

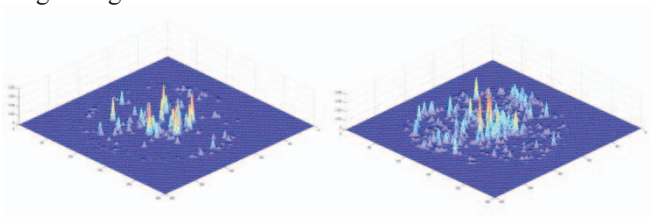
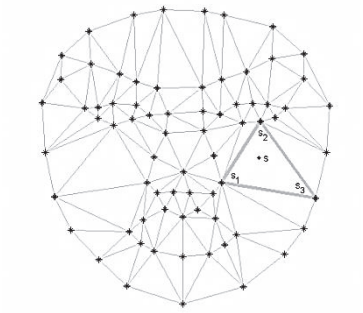


Fig. 2 PDF maps Generated by LASSO
Left: Range Image; Right: Texture; $(\lambda, \alpha) = (0.05,1)$

Further, we can introduce local coordinates to describe the PDF functions. A common choice is to use Delaunay Triangulation as in Fig. 3. Each point on the face can be represented as a convex combination of its three triangular vertices:



$$\boldsymbol{s} = \sum_{i=1}^{3}\lambda_i \boldsymbol{s_i},$$

$$\sum_{i=1}^{3}\lambda_i = 1, \lambda_i > 0$$

Fig. 3 Local Coordinates via Delaunay Triangulation

It can be seen that the above equations holds up to an piecewise affine transformation, i.e. the same coefficients $\{\lambda_i\}$ holds for transformed $\{\boldsymbol{s}, \boldsymbol{s_i}\}$. It is also insensitive under projective transforms. The points for classification can be sampled according to the transformed PDFs, or in a standard triangulation frame and then find the matching in the objective frame. This method offers an efficient solution for pose variations, since it only interpolates those sparse points selected by PDFs. The sampled points form a new feature set for low dimensional classifiers.

3. EXPERIMENT RESULTS

We use 81 subjects from the BU-3DFE datasets, each with 13 annotated faces. The original face image in 3D mesh form are resampled to decouple the range image and graylevel texture [13]. All the faces are registered with two eye centers and nose tip. The range/texture feature vectors are concatenated, forming a high dimensional vector (around 60, 000) for subspace learning. The number of training and testing samples is roughly 2:1. The performance of SLDA and sampling techniques are analyzed in this section.

3.1 Classification with SLDA

We vary the control parameter (λ, α) in SLDA and analyze the resulting sparsity, as well as classification performance. The sparse Fisherfaces (from texture) are shown in Fig.4.

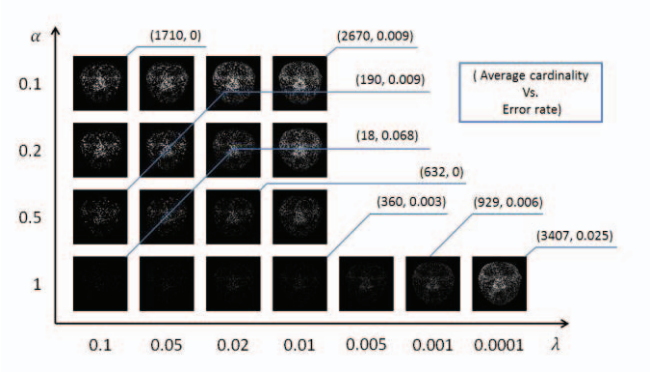


Fig. 4 Sparse Fisherfaces, with Varying (λ, α)

The sparsity level grows with larger λ and α, and the Fisherfaces are more concentrated around the nose and eyes, with higher density and larger weights. The Fisherfaces generated from range images follow the same pattern, and they are even more clustered in these regions as in Fig. 1.

The traditional PCA + LDA paradigm uses the full model and has an error rate about 0.003. Note the dimension is about 60,000, the SLDA can still achieve same classification rate using much less features! The cardinality of each sparse Fisherface can be less than 1% than those generated by dense model. If say 20 Fisherfaces are used for classification, the total cardinality would still be less than 5% of all the pixels.

Too sparse Fisherfaces can deteriorate the performance as evident from the lower-left corner case in Fig. 4. However, too "dense" sparse Fisherfaces generated by SLDA also leads to low classification rates. The reason might be that the SLDA is based on shrinkage (rather than best subset selection) which intends to trade bias with variance; seeking solution in high dimensional space, however, increases the model variance dramatically [9]. Note the dense LDA does not operate in high dimensional space since PCA is applied in advance for dimension reduction; but for SLDA model, PCA cannot be applied since we need to select sparse features.

We list the average cardinality vs. the error rate in the following tables.

	λ:0.1	0.05	0.02	0.01	0.005
α:1	(18, 6.8)	(76, 0.9)	(221, 0)	(360, 0.3)	(513, 0.3)
0.5	(190, 0.9)	(377, 0.3)	(632, 0)	(808, 0)	(967, 0.6)
0.2	(811, 0)	(1137, 0)	(1480, 0.3)	(1681, 0.9)	(1480, 0.9)
0.1	(1710, 0)	(2106, 0)	(2475, 0.6)	(2670, 0.9)	(2804, 1.2)

Table. 1. Average Cardinality vs. Error Rate (%)
(Low Sparsity Case, Coarse (λ, α))

	λ:0.1	0.09	0.08	0.07
α:1	(18, 6.8)	(22, 6.8)	(31, 3.8)	(40, 1.9)
0.9	(33, 3.1)	(42, 2.1)	(51, 1.9)	(67, 1.5)
0.8	(53, 3.1)	(64, 2.1)	(80, 1.5)	(98, 1.5)
0.7	(80, 1.9)	(95, 1.5)	(116, 1.5)	(138, 0.6)

Table. 2. Average Cardinality vs. Error Rate (%)
(High Sparsity Case, Fine Tuning (λ, α))

For the LASSO case $(\alpha = 1)$, we plot λ vs. Average Cardinality/Error Rate in Fig. 5. It can be seen that neither too sparse nor too dense projections help with classification. In this case, a sparsity of 50 ~ 1000 out of 60,000 is appropriate.

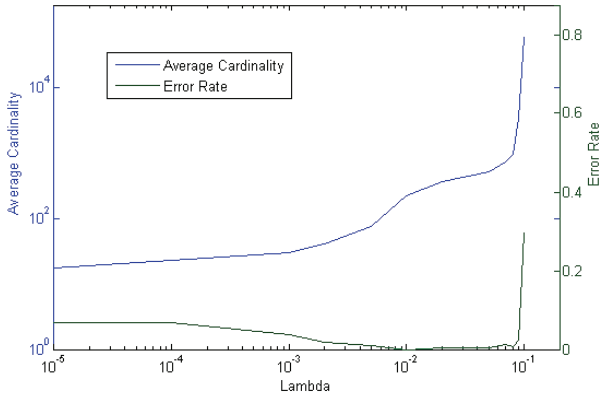


Fig. 5 λ vs. Average Cardinality/Error Rate, LASSO

3.2 Region Analysis

The SLDA selects points around nose and eyes in both range image and texture. These **selected points** have both larger density and weights than points from other regions, and they form discriminative features for classification. Since neighbor points are often correlated, **any points** from these regions should be more discriminative than others. We shall verify this proposition as follows:

1) Use Gaussian kernel to estimate PDF from LASSO.
2) Determine the cardinality and sample points from the estimated PDFs.
3) Train a classifier with the sampled sparse points.

We choose LASSO model with $\lambda = 0.05$, and generate PDFs as shown in Fig. 2. The cardinality is 26 and 50 for range image and texture respectively. The sampling is done according to 1) the PDFs in Fig. 2 and 2) uniformly (randomly without prior knowledge), with the same cardinality. The points sampled from the PDFs should be more discriminative than those sampled uniformly sampled. We use the 76 points to train a nearest center (NC), a nearest neighbor (NN) and an LDA classifier. The sampling – classification process are repeated 10 times for each classifier, and the average error rates are shown in Table. 3.

The classification can be very efficient for the low dimensionality. We repeat the above procedure 10 times and the resulting average error rate is shown in Table. 3.

(Error Rate)	Specified PDFs	Uniform Sampling
NC	0.142	0.339
NN	0.107	0.296
LDA	0.003	0.013

Table. 3 Comparison of Sparse Sampling Techniques

It can be seen that using specified PDFs outperforms random sampling. While this is more evident when using simple classifiers like NC or NN, there is still a significant performance improvement (about 4x in this case) for the LDA even with the low error rate achieved with random sampling. The success of the LDA in sparsely sampled low dimensional space gives us a more promising approach, than directly seeking in high dimensional space.

3.3 Sparse Sampling

As in section 2, we use the Delaunay Triangulation to build a local coordinate system. The coordinates of sparse feature points selected are represented by their triangular vertices. This framework provides a local position descriptor, which is convenient to deal with pose variations. Some faces with rotation and different view angles are shown in Fig. 6.

Fig. 6 Faces with Pose Variations

We repeat the experiments in section 3.2 with a new dataset synthesized by changing the pose of the faces. The feature points are sampled in a frontal view frame (serve as prior), and the respective matching points are located in the objective frames by interpolation. These points form features of each face and are classified by an ordinary LDA classifier. The error rate is around 0.005, which is similar to the previous results. Our method is computationally efficient since only a few points are used, which avoids interpolating the whole face.

Other local coordinate systems, such as the natural neighbor coordinates [16], can also be used. They can be more robust under affine transforms as more neighbors are used, but more sensitive to perspective effects. Smaller patches are used in triangular coordinates, and each patch can be approximated with a piecewise affine transform for a global perspective transform.

The landmarks can be localized by Active Appearance Models [14] [15], or combined with other detection and localization techniques [1]. With the grouping effect of Elastic Net and smoothed PDFs, slight changes in landmark positions still results in valid data samples, so this method can be quite robust. There are other sparsity based methods in face recognition such as [17], which focuses on classifier side; while ours focuses more on feature selection. It would be interesting to incorporate both in a recognition system, which serves as part of our future study.

3.4 Computation Time

We ran the experiments on an Intel Core i7 platform with 16GB RAM in Matlab. The computational cost of training SLDA is high, since the iterating rule solves the Elastic Net problem recursively until convergence in high dimensional space. Solving one Fisherface typically takes 150 to 360 seconds, with sparsity ranging from 100 to 1000, assume convergence needs 15 iterations. The testing process is very fast, about 0.02 ~ 0.3ms to recognize a face, depending on the dimension of sparse Fisherfaces.

To bring down the high computational cost of the SLDA is a major motivation for the sparse sampling model. Assuming that the PDFs are available; a sparse sampling model of 100 points takes only 0.6ms to sample one face and 75ms to train a regular LDA classifier.

4. CONCLUSION

We use the SLDA to perform feature selection in 3D face recognition. The selected discriminative features can be highly sparse, while maintaining competitive classification performance. The features cluster around discriminative regions, i.e. nose and eyes, and our proposed transform invariant sparse sampling model is able to reproduce those sparse features for efficient and robust classification.

5. REFERENCES

[1] Stan Z. Li and Anil K. Jain, *Handbook of Face Recognition*, Second Edition, Springer, London, pp.19-47, 2011

[2] M. Turk and A. Pentland. "Eigenfaces for recognition". Journal of Cognitive Neuroscience 3 (1): 71–86, 1991

[3] Belhumeur, V., Hespanha, J., Kriegman, D.: "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection". IEEE Trans. Pattern Analysis and Machine Intelligence. 19(7), 711–720, 1997

[4] Clemmensen, L., Hastie, T., Witten, D. and Ersbøll, B. "Sparse discriminant analysis", Technometrics, 53(4): 406-413, 2011

[5] Hastie, T., Buja, A., Tibshirani, R., "Penalized discriminant analysis". The Annals of Statistics 23 (1), 73-102.

[6] Tibshirani, R. "Regression shrinkage and selection via the lasso". Journal of Royal Statistical Society - Series B 58 (No. 1), 267-288, 1996

[7] Zou, Hui; Hastie, Trevor. "Regularization and Variable Selection via the Elastic Net". Journal of the Royal Statistical Society, Series B: 301–320, 2005

[8] H. Zou and T. Hastie and R. Tibshirani. "Sparse principal component analysis". Jcgs 2006 15(2): 262-286, 2006

[9] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning:Data Mining, Inference, and Prediction.* 2nd Edition, Springer, New York, pp.9-41, 2009

[10] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew Rosato, "A 3D Facial Expression Database For Facial Behavior Research". The 7th International Conference on Automatic Face and Gesture Recognition (FG2006). p211-216, 2006

[11] Efron, Bradley; Hastie, Trevor; Johnstone, Iain and Tibshirani, Robert. "Least Angle Regression". Annals of Statistics 32 (2): pp. 407–499, 2004

[12] Faltemier, T.C.; Bowyer, K.W.; Flynn, P.J., "A Region Ensemble for 3-D Face Recognition," Information Forensics and Security, IEEE Transactions on , vol.3, no.1, pp.62-73, 2008

[13] YV Venkatesh, AA Kassim, J Yuan, TD Nhuyen, "On the Simultaneous Recognition of Identity and Expression from BU-3DFE Datasets". Pattern Recognition Letters, Volume 33, Issue 13, pp. 1785 – 1793, 2012

[14] T.F. Cootes, G. J. Edwards, and C. J. Taylor. "Active appearance models". ECCV, 2:484–498, 1998

[15] Xinbo Gao, Ya Su, Xuelong Li and Dacheng Tao. "A Review of Active Appearance Models", IEEE Trans. on Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 40, No. 2, March 2010

[16] Sibson, R. "A brief description of natural neighbor interpolation (Chapter 2)". In V. Barnett. Interpreting Multivariate Data. Chichester: John Wiley. pp. 21–36, 1981

[17] Wright, J.; Yang, A.Y.; Ganesh, A.; Sastry, S.S.; Yi Ma, "Robust Face Recognition via Sparse Representation," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.31, no.2, pp.210,227, Feb. 2009