Head Pose Estimation Based on Extended Non-negative Matrix Factorization

Ce Zhan, Wanqing Li, and Philip Ogunbona

School of Computer Science and Software Engineering
University of Wollongong, Australia
Email: {cz847, wanqing, philipo}@uow.edu.au

Abstract

One popular solution to head pose estimation is to formulate it as a pattern classification problem, and treat the holistic facial appearance as the input to classifiers. However, since the face appearance contains all kinds of information, the variation caused by other factors such as identity, expression and lighting may be larger than that caused by different head poses. Thus, the key challenge of these appearance based methods lies in constructing a feature subspace that could successfully recovers head pose while ignoring other sources of image variation. In this paper, following the intuition of combining parts to form a whole face, non-negative matrix factorization (NMF) is extended to learn a localized non-overlapping subspace representation for head pose estimation. To emphasize the appearance variation in head poses, one individual extended NMF subspace is learned for each pose. The head pose of a given face image is then estimated based on its reconstruction error after being projected into the learned pose subspaces. Experiments based on benchmark face database demonstrate the efficiency of the proposed method.

Keywords: Pose Estimation, Similarity Comparison, NMF

1 Introduction

Head pose estimation is a very useful front-end processing for handling pose variations in automatic face analysis such as face recognition, face detection and expression recognition. The orientation of a user's head relative to the view of camera is also important information for applications like passive navigation and human-computer interface. Therefore, more and more techniques are investigated to realize the reliable head pose estimation [1].

Existing methods for head pose estimation can be roughly categorized either as geometric-based or appearance-based. The geometric-based methods utilize the location of facial points such as the eye corners, mouth corners, and nose tip to determine head pose from their relative configuration [2, 3]. These methods are fast and simple. However, one obvious drawback of the methods is that they are very sensitive to accurate localization of the facial points. Furthermore, these geometric-based methods assume that the configuration of facial points do not change significantly under different facial expressions.

Appearance-based methods avoid the problem of facial points localization, they typically use holistic face appearance as input. Some of these methods employ regression tools such as neural networks to develop a functional mapping from the face image to a head pose measurement. According to the survey conducted by Murphy-Chutorian et al. [1], the regression-based methods give some of the most accurate head pose estimates. However, these methods require a large number of training data of all the head poses. In practice, it is difficult to collect such training sets with the head poses precisely labeled.

Other appearance-based methods formulate the head pose estimation as a pattern classification problem. The range of head orientations is divided into a limited number of classes and classifiers for each class are trained. Rather than directly using the whole face image as the input feature vector to classifiers, these methods usually employ subspace analysis (manifold learning) methods such as principal component analysis (PCA), linear discriminant analysis (LDA), independent component analysis (ICA) [4] and isometric feature mapping (Isomap) [5] to extract a low-dimensional feature vector for classification. Since the face appearance contains all kinds of information, the variation caused by other factors such as identity, expression and lighting may be larger than that caused by different head poses. Thus, the key challenge of appearance-based methods lies in constructing a feature subspace that could successfully recovers head pose while ignoring other sources of image variation.

Local features have demonstrated their efficiency in many face related applications. Compared with global features, local features are generally more robust to facial appearance variations caused by local deformations (facial expressions), lighting and partial occlusion, since most of the variations affect only part of the face. However, due to the holistic property of head poses, it is difficult to estimate the poses purely based on the appearance of local facial areas. Thus in this paper, following the intuition of combining parts to form a whole face, non-negative matrix factorization (NMF) is extended to learn a localized non-overlapping subspace representation for head pose estimation. To emphasize the appearance variation in head poses, one individual extended NMF subspace is learned for each pose. The head pose of a given face image is then estimated based on its reconstruction error after being projected into the learned pose subspaces.

The rest of the paper is organized as follows: In Section 2 a brief introduction is given on nonnegative matrix factorization and its major extensions. Detail of the proposed method is described in Section 3. Section 4 presents the implementation of the proposed methods and the experimental results. Conclusions are drawn in Section 5.

2 Nonnegative matrix factorization

Non-negative matrix factorization (NMF) [6] is a linear, non-negative approximate data representation. Given a non-negative data matrix $V = (v_{ij})_{m\times n}$, NMF finds the non-negative matrix $W = (w_{ij})_{m\times r}$, and the non-negative matrix $H = (h_{ij})_{r\times n}$, such that $V \approx WH$. The rank r of the factorization is generally chosen to satisfy $(n+m)r < mn$, so that the product WH can be regarded as a compressed form of the data in V. Let V represents a face database, each column of V contains n pixel values of one of the m face images in the database. Then, each face in V can be represented by a linear combination of r columns of W, the columns are called basis vectors (images). Each column of H is called a coefficient vector, that is in one-to-one correspondence with a face in V and describes how strongly each basis is present in the face. Since entries in W and H are all non-negative, only additive combinations of the basis vectors are allowed. Thus, NMF naturally leads to a part-based representation, the learned basis images tend to match intuitive facial features like mouth, nose and eyes.

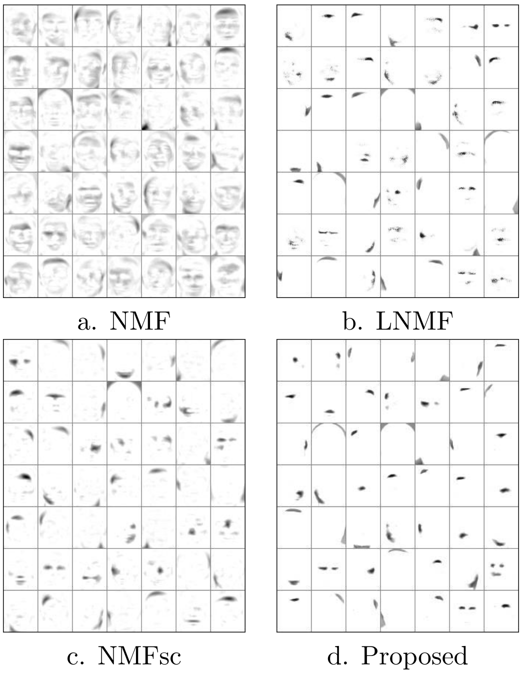

Figure 1: Basis images learned from ORL database using different methods ($r = 49$).

NMF can be taken as an optimization problem, where W and H are chosen to minimize the reconstruction error between V and WH. Various error functions (objective functions) have been proposed, one widely used is the Euclidean distance function:

$$E(W,H) = \|V - WH\|^2 = \sum_{i,j}(V_{ij} - (WH_{ij}))^2 \tag{1}$$

Although the minimization problem is convex in W and H separately, it is not convex in both simultaneously. Paatero and Tapper [7] proposed a gradient decent method for the optimization, Lee and Seung [8] devised a multiplicative algorithm to reach a local optimum.

One of the issues of NMF is that it does not always give a part-based representation. As suggested by Li et al. [9], when NMF is applied on ORL face database [10], in which faces are not well aligned, the learned basis images are holistic rather than local part-based (as can be seen in Figure 1a, the results are reproduced by us). To improve the performance of NMF in learning part-based representation, Li et al. proposed a local NMF method (LNMF) [9], that adds three additional constraints on NMF: Maximum Sparsity in H, Maximum Expressiveness of W, Maximum Orthogonality of W. Figure 1b shows the basis images learned from ORL database using LNMF. Comparing with NMF, we see that features gained by LNMF are more localized. However, some of the bases are still global and overlapped with each other. Furthermore, since more constraints are imposed, the convergence of LNMF is time consuming.

As an effect of part-based decomposition, NMF usually produces sparse representation. W is sparse since the learned bases tend to be non-global. H is often sparse due to that any given sample does not consist of all the available parts (bases). Hoyer [11] proposed a method called NMF with sparseness constraints (NMFsc), and suggested that by explicitly controlling the sparseness of W and H, NMF could give a more meaningful part-based representation. In NMFsc, the level of sparseness is measured based on the relationship between the L_1 norm and the L_2 norm:

$$sparseness(\mathbf{x}) = \frac{\sqrt{n} - (\sum |x_i|)/\sqrt{\sum x_i^2}}{\sqrt{n} - 1} \tag{2}$$

where n is the dimensionality of $\mathbf{x}$. Then NMFsc is defined as the following optimization problem:

$$\begin{aligned} \min_{W,H} E(W,H) \quad s.t.\ & W, H \geq 0, \sum_i W_{ij} = 1 \ \ \forall j \\ & sparseness(w_j) = S_w, \forall j, \\ & sparseness(h_j) = S_h, \forall j \end{aligned}$$

where w_j is the jth column of W and h_j is the jth row of H; S_w and S_h are the desired sparsenesses of W and H respectively. We show the basis images learned from ORL database using NMFsc in Figure 1c, where S_w is set to 0.75 and S_h is unconstrained as the best result achieved in [11]. As can be seen from the figure, NMFsc does not give a better part-based representation than LNMF. However, directly control the sparseness of the representation is very useful for many applications.

3 The proposed method

3.1 Extended NMF

In the proposed method, we extend the NMF for producing a localized, non-overlapping subspace representation. Inspired by LNMF and NMFsc, our extended NMF (ENMF) impose orthogonality constraint on basis matrix W while controlling the sparseness of coefficient matrix H. To reduce the overlapping between basis images, different bases should be as orthogonal as possible so as to minimize the redundancy. Denote $U = W^TW$, the orthogonality constraint can be imposed by minimizing $\sum_{i,j,i \neq j} U_{i,j}$. As introduced in Section 2, for learning localized bases, LNMF adds two more constraints to maximize the sparsity in H. Maximum sparsity in the coefficient matrix makes sure that a basis component cannot be further decomposed into more components, thus the overlapping between basis images is further reduced. However, a high sparseness in H forces each coefficient try to represent more of the image, and then the basis images tend to be global. Consider the extreme case when only one element in each column of H is allowed to be nonzero, then the NMF reduces to vector quantization (VQ), and all the basis images turn to holistic prototypical faces. Therefore, we chose to explicitly control the sparseness level of H, so that a compromise can be made between localization and overlapping and the value of the sparseness could be set based on different application scenarios.

The objective function of the ENMF is defined as:

$$E(W,H) = \frac{1}{2}\sum_{i,j}(V_{ij} - (WH)_{ij})^2 + \beta \sum_{i,j,i \neq j} U_{i,j} \tag{3}$$

where $U = W^TW$, β is a small positive constant. Then the ENMF is defined as following optimization problem:

$$\begin{aligned} \min_{W,H} E(W,H) \quad s.t.\ & W, H \geq 0, \sum_i W_{ij} = 1 \ \ \forall j \qquad (4) \\ & sparseness(h_j) = S_h, \forall j \end{aligned}$$

where h_j is the jth row of H; S_h are the desired sparsenesses of H; the sparseness is measured based on formula (2). A local solution to the above minimization can be found by using the following two step update rules:

1. $$W_{i\alpha} \leftarrow W_{i\alpha} \frac{(VH^T)_{i\alpha}}{(WHH^T)_{i\alpha} + \beta \sum_i W_{i\alpha}} \tag{5}$$

2. $$H_{\alpha\mu} \leftarrow H_{\alpha\mu} - \mu_H [W^T(WH - V)]_{\alpha\mu} \tag{6}$$

   Then project each row of H to be non-negative, have unit L_2 norm, and L_1 norm set to achieve desired sparseness S_h. (μ_H is a small positive constant. For the projection method, please refer to [11].)

Figure 1d shows an example of the bases learned from ORL database using the proposed ENMF, S_h is set to 0.1. As can be seen from the figure, more localized, less overlapped basis images are obtained, and limited bases contribute to each specific local facial area.

3.2 Head pose estimation

Let $V^{(k)}$ represents a set of face images that share one particular pose k, then a ENMF subspace $W^{(k)}$ learned from $V^{(k)}$ can be regarded as a specific feature space for the pose k. Given a new sample face image S (same size as the face images in the training set), its coefficient vector $L^{(k)}$ in the learned pose subspace $W^{(k)}$ can be obtained by:

$$L^{(k)} = W^{(k)-1}S \tag{7}$$

where $W^{(k)-1}$ is the pseudo inverse matrix of $W^{(k)}$. Based on the obtained coefficient vector $L^{(k)}$, the sample S can be reconstructed by:

$$S^{(k)} = W^{(k)}L^{(k)} \tag{8}$$

The reconstruction error $\epsilon^{(k)}$ between S and $S^{(k)}$ reflects the similarity between the sample and the training images that share the same pose k, smaller value of $\epsilon^{(k)}$ indicates a higher probability that the pose of S is k. Thus, after ENMF subspaces are learned for each of the pose, the qth pose will be assigned to the given sample S, if

$$\epsilon^{(q)} = min\{\epsilon^{(k)}\}(k = 1, \ldots, p) \tag{9}$$

where p is the total number of pose category, and the reconstruction error is calculated by mean square error (MSE):

$$\epsilon^{(k)} = MSE(S, S^{(k)}) = \frac{1}{n}\sum_{i,j}(S_{ij} - S_{ij}^{(k)})^2 \tag{10}$$

where n is the number of pixel in the face image.

4 Experimental results

Recently, Hu et al. [12] evaluated the performance of major subspace learning methods for head posed estimation based on CMU PIE face database [13]. In this paper, we compare the proposed method with the methods evaluated in [12] using similar experimental setup.

The CMU PIE face database contains 41,368 face images of 68 subjects, the images were captured by 13 synchronized cameras (thus under 13 different poses), under varying illumination and expression. A subset of the database is used in the experiment: for each of the head pose, 408 images are selected, 6 images per subject, with expression and illumination variations. Half of the images (from the first 34 subjects) are used for learning 13 pose subspaces and the rest of data are used for testing. Rather than cropping and aligning the face area based on key facial points as in Hu et al.'s work, the face region of the images is detected by the Viola-Jones face detection method [14] and cropped without alignment. For few images that are failed in automatic face detection, the face area are manually cropped. All the cropped face images are then converted to gray scale, and resized to 64 × 64. Some sample images are shown in Figure 2.

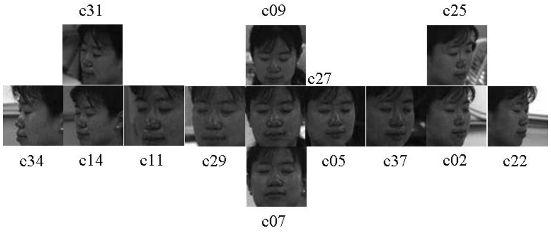

Figure 2: Sample face images of one person from CMU PIE database, the number in the figure is the pose number

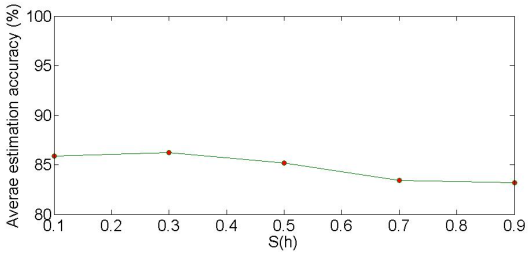

Figure 3: The testing results for different value of S_h while r is fixed to 49

While fixing the number of basis r to 49, the proposed method is first tested with different value of S_h (the sparseness of coefficient matrix). The results are shown in Figure 3. We can see that the best result is achieved neither with the highest nor with the lowest sparseness, but when S_h is set to 0.3. This observation justifies our analysis in Section 3.1, and demonstrates that a compromise made between localization and overlapping improves the efficiency of NMF for head pose representation.

Then, we set the value of S_h to 0.3 and test the proposed method by changing the number of basis r. The results are shown in Figure 4. As can be seen from the figure, higher estimation accuracy are obtained as the number basis increases. However, the increase is limited especially when the value of r exceeds 81, in those cases almost the same results are obtained.

Table 1 lists the best result we obtained together with the accuracies of other major subspace-based head pose estimation methods reported in [12]. The estimation accuracy for each pose are shown in Figure 5. Hu et al. evaluated 4 methods in their work: Eigenface method (PCA), Fisherface method (LDA), locality preserving projections (LPP), and Pose Specific Subspace (PSS). In the first three methods, a

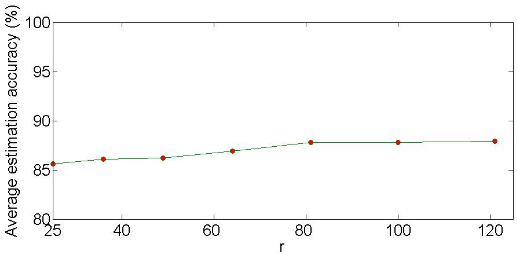

Figure 4: The testing results for different value of r while S_h is fixed to 0.3

Method	No. of basis	Accuracy
PCA	117	75.72%
LDA	13	75.37%
LPP	87	78.90%
PSS	80	83.19%
Proposed	121	87.82%

Table 1: The average accuracies of subspace-based head pose estimation methods

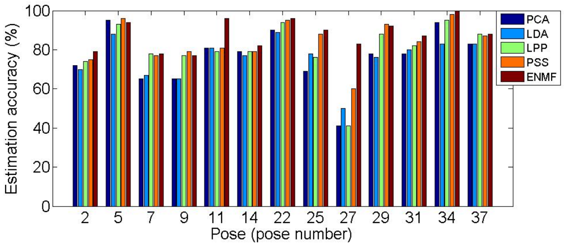

Figure 5: Estimation accuracy of each pose

given face is projected to one subspace for all the poses, the projected face is then compared to the stored pose representations and classified to one pose category. PCA seeks a projection that best represents the data in a least-squares sense, LDA looks for directions that are efficient for discrimination, while LPP tries to find a embedding that preserves intrinsic geometry of the data and local information. Similar as the proposed method, PSS models each pose with one individual eigenspace, the distance from the pose specific subspace is then exploited as the similarity measurement for pose estimation. We can see from Figure 5 that the proposed method almost outperforms the other methods for all the head poses, especially for pose 27 (frontal pose), which is relatively hard to be correctly classified since there are four near frontal views. As can be seen from Table 1, LPP and PSS perform better than the traditional subspace method PCA and LDA. By focusing on the local information and characterizing different pose by a subspace respectively, the proposed method combines the advantages of both LPP and PSS, thus achieves the best result.

5 Conclusion

In this paper, non-negative matrix factorization (NMF) is extended to learn a localized non-overlapping subspace representation for head pose estimation. To emphasize the appearance variation in head poses, one individual extended NMF subspace is learned for each pose. The head pose of a given face image is then estimated based on its reconstruction error after being projected into the learned pose subspaces. Based on the CMU PIE face database, the proposed method is compared with other major subspace-based head pose estimation methods. The experimental results demonstrate that the proposed facial representation can be effective for pose estimation. For future work, the ENMF based representation could be employed in other face related applications such as face recognition and expression recognition.

References

[1] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 4, pp. 607–626, 2009.

[2] T. Horprasert, Y. Yacoob, and L. S. Davis, "Computing 3-d head orientation from a monocular image sequence," in Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition, 1996, pp. 242–247.

[3] J.-G. Wang and E. Sung, "Em enhancement of 3d head pose estimated by point at infinity," Image and Vision Computing, vol. 25, no. 12, pp. 1864–1874, 2007.

[4] S. Z. Li, X. Lu, X. Hou, X. Peng, and Q. Cheng, "Learning multiview face subspaces and facial pose estimation using independent component analysis," IEEE Transactions on Image Processing, vol. 14, no. 6, pp. 705–712, 2005.

[5] B. Raytchev, I. Yoda, and K. Sakaue, "Head pose estimation by nonlinear manifold learning," in ICPR 2004, 2004, pp. 462–466.

[6] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, pp. 788–791, 1999.

[7] P. Paatero and U. Tapper, "Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values," Environmetrics, vol. 5, no. 2, pp. 1180–4009, 1994.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in Proceedings of NIPS'2000, 2000, pp. 556–562.

[9] S. Z. Li, X. Hou, H. Zhang, and Q. Cheng, "Learning spatially localized, parts-based representation," in Proceedings of CVPR'01, vol. 1, 2001, pp. 207–212.

[10] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," Proceedings of 2nd IEEE Workshop on Applications of Computer Vision, pp. 138–142, 1994.

[11] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," The Journal of Machine Learning Research, vol. 5, no. 5, pp. 1457–1469, 2004.

[12] Y. Hu and T. S. Huang, "Subspace learning for human head pose estimation," in ICME 2008, 2008, pp. 1585–1588.

[13] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression (pie) database," in AFGR 2002, 2002, pp. 1585–1588.

[14] P. Viola and M. Jones, "Robust real-time object detection," International Journal of Computer Vision, vol. 57, no. 2, pp. 137–154, 2004.