LOCATING NOSETIPS AND ESTIMATING HEAD POSE IN IMAGES BY TENSORPOSES

Jilin Tu, Thomas Huang

Beckman Institute
Electrical and Computer Department
University of Illinois at Urbana-Champaign

ABSTRACT

This paper introduces a head pose estimation system that localizes nose-tip of the faces and estimate head poses in images simultaneously. After the nose-tip in the training data are manually labeled, the appearance variation caused by head pose changes is characterized by tensor model. Given images with unknown head pose and nose-tip location, the nose-tip of the face is localized in a coarse-to-fine fashion, and the head pose can be estimated simultaneously. We evaluated our system on the Pointing'04 head pose image database with 50% of the data as training set and the rest as testing set. With the nose-tip location provided, our head pose estimators can achieve 94% head pose classification accuracy(within $\pm 15^o$). With nose-tip automatically localized, we achieves 85% nose-tip localization accuracy(within 3 pixels from the ground truth), and 81% head pose classification accuracy(within $\pm 15^o$).

Index Terms— Head Pose, Tensor, nose-tip localization, pointing04

1. INTRODUCTION

Locating human faces and determining head pose from image is one of the most important components for human computer interaction systems. Knowing the location of human face and its orientation allows the computer to determine human identity and focus of attention in the scene. In [1], the importance of nose detection and tracking is discussed for Human Computer Interaction(HCI) purpose.

For images(videos) in which human head/face has high resolution, the facial features(i.e., eyes, noses, and mouth) can be extracted, and head poses can be estimated with assumption of human face symmetry[2]. For faces in lower resolution images, facial features are hard to detect and track. People tend to consider the pose estimation as a classification problem. In [3], a comparative study of several coarse head pose estimation algorithms was carried out. It was found the neural network head pose classifier outperforms MAP classifiers in image of very low resolution, i.e. head images of size 8×8.

In [4], the Pointing'04 head pose database was made public. A number of research groups reported the performance of their head pose estimation systems on this set of data. Wu[5] proposed a two-level approach for estimating the head pose but with nose tip manually localized for the purpose of removing errors from misalignment. At lower level, the image is down-sampled and Gabor wavelet features are computed. The head poses are then classified by majority voting on the classification results based on KDA and PCA subspace models. 90% accuracy was achieved for head pose estimation error less than 15 degree. At higher level, the head pose is further refined in a window of 3 by 3 neighboring poses(within 15 degree) by Bunch Graph Analysis. In [4], the face area is obtained by color segmentation and head poses (only pan angles) are estimated based on face symmetry after the eyes are localized based on robust features. They achieved mean pose error less than 15 degree when the absolute head pan angle is less than 45 degree. In [6], faces are first localized by skin color segmentation and edge detection, and the poses are estimated by ANN classifier. The data is divided into 80% for training, 10% for cross-validation and 10% for testing. Their system achieved average pan error 9.5 degree, average tilt error 9.7 degree, 52% for pan accuracy and 66% for tilt accuracy. Similar performance of a similar ANN head pose classifier was also reported in [7]. Along the direction of ANN classifier for head pose estimation, Gourier[8] developed a auto-associative memories based on Winodrow-Hoff learning rule. They achieved a precision of 10.3 degrees in pan and 15.9 degrees in tilt with Jack-Knife(leaving-one-out) partition of the training and testing data.

In this paper, we introduce our system for automatic localizing the nose-tip of human face in studio quality images and estimate the head poses using tensor techniques[9]. With this tensor technique, we are able to do coarse-to-fine search in the image and locate the nose-tip and estimate the head pose simultaneously.

In section 2, we describes the basics of Tensor analysis and the Tensorposes model generated from pointing04 head pose database. In section 3, we describe the framework of our system. In section 4, we shows the performance of our system on Pointing04 image data set for nose-tip localization and for head pose estimation. Section 8 concludes with discussion of

This research was funded by the U. S. Government VACE program.

the performance of our system.

2. ESTIMATING HEAD POSE BY TENSORPOSES MODEL

Vasilescu [10] proposed to do higher order multi-linear analysis of image ensembles. The multi-linear analysis of facial image ensembles leads to so-called TensorFaces representation. While classification by tensor of unlabeled testing data was based on brutal-force nearest neighbor search of candidate coefficients across different modes in previous works, [9] proposed a rank-1 factorization approach so that the coefficient vectors in different modes of an unlabeled test image can be inferred simultaneously.

2.1. Tensor Basics[10]

An order-N tensor is the N dimensional generalization of a 1-D vector and a 2-D matrix. Denote a order-N tensor as $\mathcal{A}(\mathcal{A} \in R^{I_1 \times I_2 \times ... \times I_N})$, and the element of $\mathcal{A}$ as $a_{i_1 i_2 ... i_N}$, where $1 \leq i_n \leq I_n$, the mode-n vector $A_{(n)}$ is a matrix of size $I_n \times (I_1 ... I_{n-1} I_{n+1} ... I_N)$ flattening from the tensor by concatenating the column vector in mode-n row-wise. The $n - rank$ of $\mathcal{A}$ is defined as the rank of the mode-n vectors: $R_n = rank(A_{(n)})$.

The product of a tensor $\mathcal{A} \in R^{I_1 \times I_2 \times ... \times I_N}$ and a matrix $U \in R^{J_n I_n}$ in mode-n can be denoted as $\mathcal{A} \times_n U$, which is a tensor in $R^{I_1 \times I_2 \times ... J_n ... \times I_N}$. The following properties are worth noting:(1) $\mathcal{A} \times_m U \times_n V = \mathcal{A} \times_n V \times_m U$; (2)$(\mathcal{A} \times_n U) \times_n V = \mathcal{A} \times_n VU$; (3) $\mathcal{B} = \mathcal{A} \times_n U \iff B_{(n)} = U A_{(n)}$.

Similar to matrix SVD, N-mode SVD orthogonalizes the subspaces in each mode and decomposes the tensor as the mode-n product of N orthogonal subspaces,

$$\mathcal{D} = \mathcal{Z} \times_1 U_1 \times_2 U_2 \times_3 ... \times_N U_N \tag{1}$$

Tensor $\mathcal{Z}$ is called core tensor. Unlike the singular value matrix in matrix SVD, $\mathcal{Z}$ does not have a simple, diagonal structure. Mode matrix U_n contains the orthonormal vectors spanning the column space of matrix $D_{(n)}$. Therefore an straightforward solution of N-mode SVD is as follows:

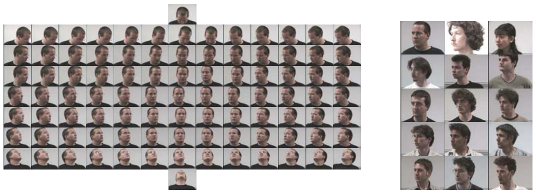
1. For n=1,2,...,N, compute the SVD of the flattened matrix $D_{(n)}$, let U_n be the row eigen-space.

2. Solve for the core tensor

$$\mathcal{Z} = \mathcal{D} \times_1 U_1^T \times_2 U_2^T \times_3 ... \times_N U_N^T \tag{2}$$

2.2. Tensorposes

Pointing04 head pose database[4] contains 2 sets of 93 head pose pictures for 15 subjects captured under different illuminations and scales. The 93 head poses includes combinations of 13 pan poses and 7 tilt poses, together with two extreme cases with 0^o pan angle and $\pm 90^o$ tilt angle respectively. Figure 1-(a) shows the pictures of one subject with the 93 head poses and Figure 1-(b) shows the 15 subjects with 45^o head pan angle and 0^o tilt angle. Figure 1-(b) also indicates that, there exists inconsistency between the appearances and the head poses, as the appearance of some subjects looks more like a pose of 90^o pan angle.



(a)93 poses (b)15 persons

Fig. 1. Two views of the Pointing '04 head pose image database

We take the first set of images as training data. After the nosetip locations are manually labeled, we cropped image patches of size 18 by 18 at nose-tip after reducing illumination variations by Laplacian filtering[11]. As there are 15 subjects, if we ignore the two extreme head poses with tilt angle 90^o and -90^o (which can be modeled by PCA as special cases), the rest head pose images can be arranged into a tensor of size $314 \times 15 \times 13 \times 7$.

As the tensor data illustrates a multi-linear structure of the appearances resulting from the confluence of person identity, pan angle and tilt angle, the image tensor D can be decomposed into these factor coefficients by N-mode SVD as follows.

$$\mathcal{D} = \mathcal{Z} \times_1 U_{pixel} \times_2 U_{person} \times_3 U_{pan} \times_4 U_{tilt} \tag{3}$$

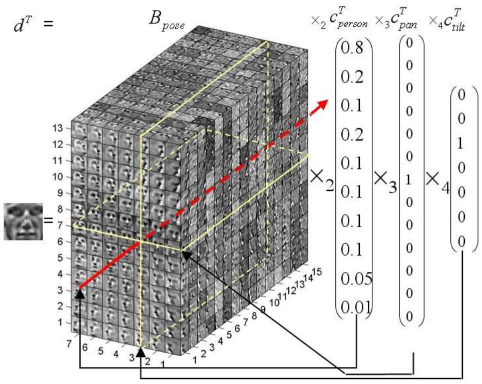
where $\mathcal{Z}$ is the core tensor, and U_{pixel}, U_{person}, U_{pan}, U_{tilt} are mode matrices that span the space of pixel, people identity, pan angle, tilt angle variations respectively. Given an new input image d, the coefficient vectors c_{person}, c_{pan}, c_{tilt} can be retrieved simultaneously by projecting d in the tensor subspaces $\mathcal{Z} \times_1 U_{pixel}$[9].

Consider a tensorposes subspace constructed as

$$\mathcal{B}_{pose} = \mathcal{Z} \times_1 U_{pixel} \times_3 U_{pan} \times_4 U_{tilt} \tag{4}$$

According to [10], $\mathcal{B}_{pose}$ is a multi-linear representation of PCA subspaces for the head poses. Intuitively, a new face d at certain pan and tilt angle can be reconstructed by the subspace in $\mathcal{B}_{pose}$ of the same pan and tilt angle, i.e., $d^T = \mathcal{B}_{pose} \times_2 c_{person}^T \times_3 c_{pan}^T \times_4 c_{tilt}^T$ where c_{pan} and c_{tilt} are boolean vectors with vector elements for the corresponding head pose being set to 1, and c_{person} contains the projection coefficients in the appearance subspace of the corresponding pan and tilt pose angle in $\mathcal{B}_{pose}$. This is illustrated in Figure 2. The left most layer of $\mathcal{B}_{pose}$ shows the mean faces for the PCA subspaces at each pan and tilt angles, and along the column toward the depth direction at each pose shows the eigenfaces that span the appearance variations caused by the 15 different subjects. In [9], a rank 1 tensor factorization for obtaining c_{person}, c_{pan} and c_{tilt} was proposed for Multilinear Independent Component Analysis(MICA), we found it works perfectly well for the Tensorposes model. With c_{pan} and c_{tilt}, the head pose of image d can be inferred as follows(proof in [12]):

$$\begin{aligned}\hat{p} &= argmax\{c_{pan}\}\\ \hat{t} &= argmax\{c_{tilt}\}\end{aligned}$$



d^T =
B_{pose}
$\times_2 c^T_{person}$ $\times_3 c^T_{pan}$ $\times_4 c^T_{tilt}$
0.8 0.2 0.1 0.2 0.1 0.1 0.1 0.1 0.05 0.01
0 0 0 0 0 0 1 0 0 0 0 0 0
0 0 1 0 0 0 0
$\times_2$ $\times_3$ $\times_4$
=

Fig. 2. Tensorposes factorization

We can further reconstruct $\hat{d}$ from the tensor model as $\hat{d}^T = \mathcal{B}_{pose} \times_2 c^T_{person} \times_3 c^T_{pan} \times_4 c^T_{tilt}$. The distance from the input image d to the pose tensor model is computed as

$$D(d, \mathcal{B}) = 1 - corr(d, \hat{d}) \tag{5}$$

We utilize this measure to localize the nose-tip in the test image.

3. THE SYSTEM FRAMEWORK

Based on the tensor technique, we developed a system to automatically localize the nose-tip and estimate the head poses. As shown in Figure 3, the training stage involves building the tensorposes model using labeled data. When testing image is provided, we first do skin color segmentation to locate the face area, then hierarchical Laplacian filters are applied to reduce illumination variations. The nose-tip of the subject's face is then searched in a coarse-to-fine manner and the head pose can be estimated simultaneously.

The skin color segmentation model was built in RGB space based on [13]. We segment the skin color region using Ostu's adaptive threshold in both coarse and fine resolutions. Morphological open/close operations are carried out to eliminate outliers and the holes in the segmented area is filled. Finally the segmentation is obtained by AND fusion of the segmentation in both coarse and fine resolutions. Some random skin color segmentation results are shown in Figure 4.

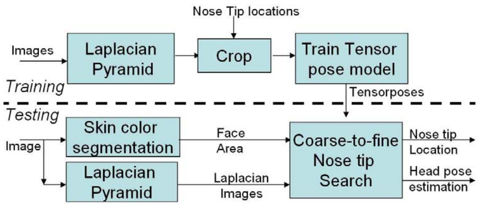


Fig. 3. The framework

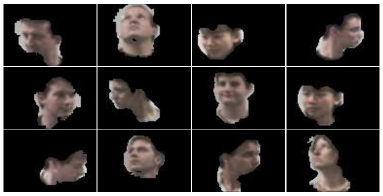


Fig. 4. Skin color segmentation

The Laplacian pyramid is computed by differencing the adjacent levels of a Gaussian pyramid that is built by repeatedly smoothing and down-sampling. By experiments, we empirically chose to do pose estimation at Laplacian pyramid level 2 and level 3 with image patch size 25×25 and 18×18.

We first scan through the face area segmented by skin color model pixel by pixel in Laplacian pyramid level 3 using Tensorposes model, obtain 10 nose-tip location candidates according to Equation 5. We then find the best match in the neighborhood of these nose-tip candidates in Laplacian pyramid level 2. Denote the candidate nose-tip locations l_i^c and the estimated head pose as $(p_i^c,\, t_i^c)$ with distance measure D_i^c in pyramid level c at location i ($c = 2, 3$; $i = 1..10$), the optimal nose-tip location, and head pose can be determined as

$$argmin_{l,p,t}\{C(|p_i^3 - p_i^2|, |t_i^3 - t_i^2|) + \alpha D_i^3 + \beta D_i^2 + \gamma \| l_i^3 - l_i^2 \|\} \tag{6}$$

with $C(\cdot)$ as a cost function for the pose differences, and α, β, and γ as blending coefficients. The intuition is that the head poses estimated at different Laplacian pyramid levels should be consistent at the true positive nose-tip location, and likewise, the distance between the nose-tips estimated across different resolution should be small, and the tensorpose distance measures at the two pyramid levels should be small.

4. EXPERIMENTS

4.1. Nose-tip Localization Accuracy

We trained our system with the first set of the Pointing'04 dataset[4] after the nose tips are manually marked. The nose tips in the second set are also marked as ground truth. The histogram of the nose tip localization error(in pixel at pyramid level 3) for the second dataset(1395 images) is shown in Figure 5. The algorithm achieved 89.38% localization accuracy with error tolerance of 3 pixels.

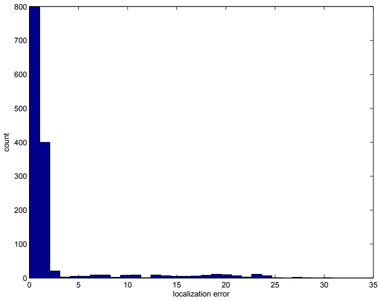


Fig. 5. The histogram of the nose-tip localization error

4.2. Pose estimation

We evaluated the pose estimation performance of our algorithm on second set of Pointing04 pose database with nose-tip localized both by hand and by the proposed system. Table 1 summarizes the results. Our performance is evaluated by using the first set of the Pointing04 pose data(50%) as training set and the second set(50%) as testing set with the subjects unknown. Comparing to the evaluation results of pose estimation algorithms based on ANN techniques, their performances were evaluated by splitting the data using leaving-one-out strategies[6] [7][8] (which usually take the majority of the data as training set, i.e. 80% of the data as training data, 10% as testing data and 10% as evaluation data in [6])or by supposing the subject identity is known[8]. This indicates that, our algorithm can generalize better for estimating head pose of unknown users in unknown environments.

Metric\Localization method	Manual	Automatic
Average Pan error	5.01^o	12.28^o
Average Tilt error	4.37^o	11.37^o
Pan Classification	73.04%	55.43%
Tilt Classification	81.47%	64.00%
Pan Classification (within $\pm 15^o$)	96.26%	87.76%
Tilt Classification (within $\pm 15^o$)	94.48%	82.77%

Table 1. Evaluation on head pose estimation by manual/automatic nose-tip locations

5. CONCLUSION

In this paper, we introduced our system for automatic nose-tip localization and head pose estimation in images based on Tensorposes model. The experiment results show that, our nose-tip localization algorithm achieves 88.28% accuracy(for within 3 pixels error tolerance), and 83%-87.76% pose estimation accuracy. Comparing to the past pose estimation algorithms based on ANN techniques, our system is evaluated with a more challenging data splitting strategy and achieves better generalization capability.

6. REFERENCES

[1] D.O. Gorodnichy, "On importance of nose for face tracking," in *5th Intern. Conf. on Automatic Face and Gesture Recognition*, Washington, DC, USA., 2002.

[2] Y.X. Hu, L.B. Chen, Y. Zhou, and H.J. Zhang, "Estimating face pose by facial asymmetry and geometry," in *the 6th Intl. Conf. on Auto. Face and Gesture Recog.*, 2004.

[3] L.M. Brown and Y.L. Tian, "Comparative study of coarse head pose estimation," in *Workshop on Motion and Video Computing (MOTION'02)*, 2002, p. 125.

[4] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial structures," in *Pointing 2004, ICPR*, 2004.

[5] J.W. Wu, J.M. Pedersen, D. Putthividhya, D. Norgaard, and M.M. Trivedi, "A two-level pose estimation framework using majority voting of gabor wavelets and bunch graph analysis," in *Pointing 2004, ICPR*, 2004.

[6] Rainer Stiefelhagen, "Estimating head pose with neural networks-results on the pointing04 icpr workshop evaluation data," in *Pointing 2004, ICPR*, 2004.

[7] M. Voit, K. Nickel, and R. Stiefelhagen, "Neural network-based head pose estimation and multi-view fusion," in *CLEAR EVALUATION WORKSHOP*, 2006.

[8] N. Gourier, J. Maisonnasse, D. Hall, and J.L. Crowley, "Head pose estimation on low resolution images," in *CLEAR EVALUATION WORKSHOP*, 2006.

[9] M.O. Vasilescu and D. Terzopoulos, "Multilinear independent components analysis," in *CVPR*, 2005, vol. 1, pp. 547–553.

[10] M. Alex O. Vasilescu and Demetri Terzopoulos, "Multilinear analysis of image ensembles: Tensorfaces," in *ECCV (1)*, 2002, pp. 447–460.

[11] P. J. Burt and E. H. Adelson, "The laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, 1983.

[12] J. Tu, Y. Fu, Y.X. Hu, and T. Huang, "Evaluation of head pose estimation for studio data," in *CLEAR EVALUATION WORKSHOP*, 2006.

[13] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *Int. J. of Computer Vision*, vol. 46, no. 1, pp. 81–96, 2002.