MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan
8-15

Head Pose Estimation using Adaptively Scaled Template Matching

Miki Yamada, Osamu Yamaguchi,
Akiko Nakashima and Takeshi Mita
Multimedia Laboratory
Corporate R&D Center, Toshiba Corporation
Kawasaki 212-8582 Japan
{miki.yamada, osamu1.yamaguchi,
akiko.nakashima, takeshi.mita}@toshiba.co.jp

Kazuhiro Fukui
Department of Computer Science
Graduate School of Systems
and Information Engineering
University of Tsukuba
1-1-1 Ten-noudai, Tsukuba 305-8573 Japan
kfukui@cs.tsukuba.ac.jp

Abstract

This paper proposes a real-time head pose estimation system using a new image matching technique. The system consists of a training stage, in which subspace dictionaries for classifying head poses are computed using template matching and the factorization method, and a recognition stage, in which head poses are estimated using the subspace method.

The system uses the method of adaptively scaled template matching, in which the expansion ratio of the template is adapted to the size of the tracked object and a search distribution with high center density is used for position and expansion ratio search. It is effective for accurately tracking regions and feature points on a face. The new method also increases the number of successful detections of the object because it rarely misses the point of maximum similarity.

A head-centered coordinate system (H-coordinates) is also proposed for representing head poses independently of camera position. Using H-coordinates, we can classify head poses by two variables (horizontal and vertical angles) independently of inclining head motion (tilt angle), which does not change the gaze.

Numerical experiments show the effectiveness of the proposed method.

1 Introduction

There are several applications of head pose estimation in image sequences. For example, we consider the application of an alarm system that detects whether or not a driver looks aside while driving a vehicle.

Head pose estimation by image processing using a single camera, which is discussed in this paper, is safer than that using a range finder, and it is faster and its hardware is simpler than methods using stereo image matching[1].

Head pose estimation systems need to use geometrical methods such as the factorization method[2] using face feature points, and image matching methods such as template matching[3, 4]. The template matching methods can be used for detecting a face or face feature points. Many conventional template matching methods compute pyramidal images to perform image matching for size-changing image objects such as a face.

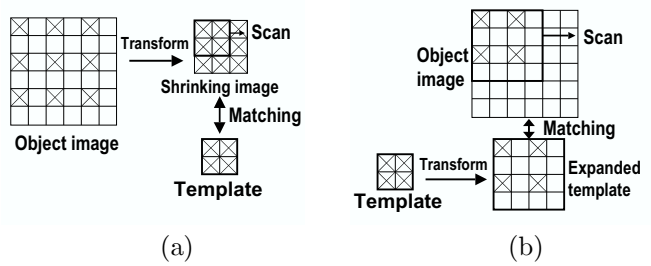
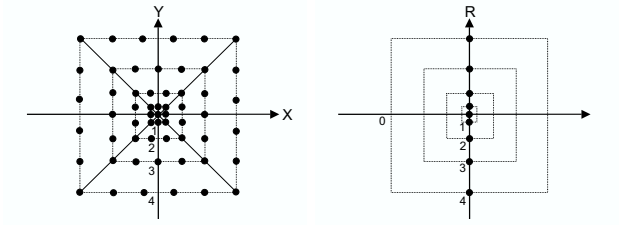


Figure 1. (a) A conventional scanning method in case that the template size is smaller than the object in the image. (b) The adaptively scaled template matching method. The resolution of matching position is equivalent to that of the object image. The expansion ratio of the template is a continuous value adapted to the object size.

Image matching for images of a three-dimensional object must handle performance degradation caused by displacement and size change of an object image. This paper proposes a method of adaptively scaled template matching, which is more accurate and robust than a method using pyramidal images because it calculates the expansion ratio of a template, which is adapted to the size of the object.

The proposed system consists of two stages: the first is the training stage, in which subspace dictionaries for classifying head poses are computed off-line using template matching and the factorization method. In the second stage, face images are classified online into one of seven head pose classes using the subspace method[5] (the CLAFIC method[6]). Adaptively scaled template matching is effectively used for accurately tracking regions and feature points on a face in both stages.



(a) (b)

Figure 2. Two types of search distributions with high center density, which are used in the adaptively scaled template matching method. Search points are indicated by "•". (a) A two-dimensional search distribution used for position search. (b) A one-dimensional search distribution used for expansion ratio search.

2 Adaptively Scaled Template Matching

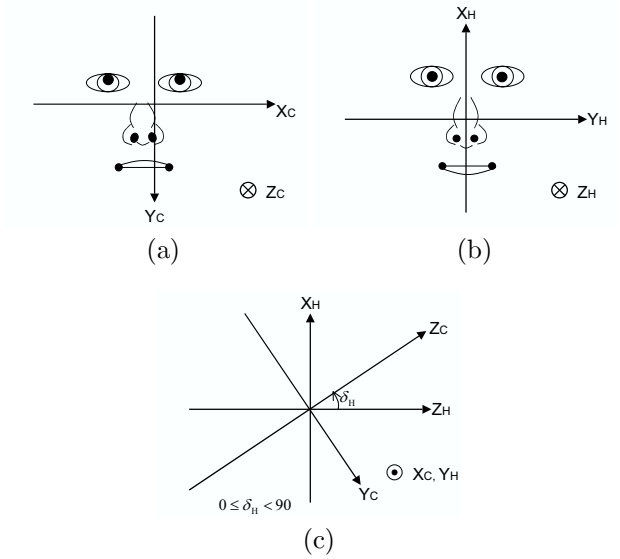
For dealing with changes of the object size, conventional template matching methods make use of pyramidal images, whose size ratios are generally fixed and have the same intervals. Each image layer of the pyramid is scanned in order to find the location of maximum similarity (Fig.1(a)).

Adaptively scaled template matching does not compute pyramidal images. Instead, the template is positioned on the object image and expanded dynamically with an arbitrary ratio (Fig.1(b)). The proposed method has higher accuracy in estimating the object position and scale compared to conventional methods.

The expansion ratio of the template is a continuous value adapted to the object size. We set the number of sampling points that are compared in template matching equal to the number of pixels of the original template for convenience of computation. There are pixels that are excluded in the expanded template in matching. The minimum intervals are set to one pixel for the position search and a ratio of about 0.005 for the expansion ratio search.

We use two types of search distributions with high center density (Fig.2). A two-dimensional search distribution is used for the position search. A one-dimensional search distribution is used for the expansion ratio search. For each frame, template matching using the search distribution is iterated until the matching position converges. Head movements are mostly smooth, so that a local search in the neighborhood of the matching position in the previous frame is feasible. For larger head motions, an accurate search needs more computing time.

Adaptively scaled template matching performs image matching with great accuracy of position and expansion ratio, but also increases the number of successful detections of the object because the adaptively scaled search rarely misses the point of maximum similarity.



(a) (b) (c)

Figure 3. (a) The relation between the camera coordinates and the object. (b) The relation between head coordinates and the object. (c) The relation between head coordinates and the camera coordinates.

Adaptively scaled template matching is used for face and face feature point tracking during the training stage(Fig.4(a)). The qualitative trinary representation (QTR)[7] is used for computing the matching measure. The QTR is based on the relative intensity magnitude of neighboring pixel values, and enables us to perform fast and accurate template matching. Adaptively scaled template matching is also used for face detection in the recognition stage(Fig.4(b)). The similarity defined in the subspace method[5, 6] is used here as a matching measure.

3 Head-Centered Coordinate System

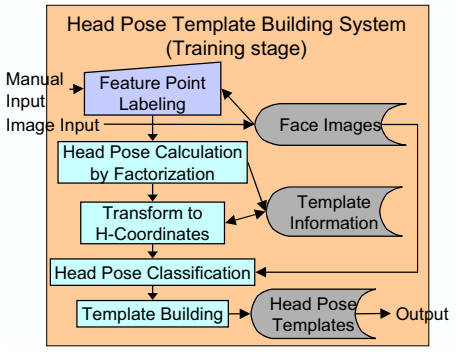
3.1 Rotation matrices of head poses

The result of shape and motion acquisition using factorization [2] is obtained for the camera coordinates. We suppose that face images are obtained for the camera layout of Fig.3(a). The head pose for the f-th image frame is obtained as the result of the factorization, which is described by the rotation matrix $R_{Cf\mathrm{C}}$ for the camera coordinates and defined as

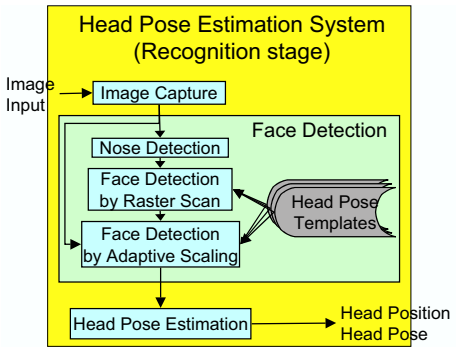
$$\boldsymbol{X}_{Cf} = R_{Cf\mathrm{C}}\boldsymbol{X}_{\mathrm{C}}, \tag{1}$$

where $\boldsymbol{X}_{Cf}$=$[X_{Cf}\ Y_{Cf}\ Z_{Cf}]^{\mathrm{T}}$ is the position of a feature point for the f-th frame, and $\boldsymbol{X}_{\mathrm{C}1}$ is represented as $\boldsymbol{X}_{\mathrm{C}}$.

We define the head-centered coordinate system (H-coordinates) and coordinate axes $\mathrm{X_{H}}f\mathrm{Y_{H}}f\mathrm{Z_{H}}f$, which are fixed on the object (Figs.3(b), (c)). H-coordinates have the advantage that we can consider head poses that do not depend on camera position.



(a)



(b)

Figure 4. (a) The head pose template building system (the training stage). (b) The head pose estimation system (the recognition stage).

The position of a feature point in the f-th frame is $\boldsymbol{X}_{\mathrm{H}f} = [X_{\mathrm{H}f}\ Y_{\mathrm{H}f}\ Z_{\mathrm{H}f}\ ]^{\mathrm{T}}$ for H-coordinates, and $\boldsymbol{X}_{\mathrm{H}1}$ is represented as $\boldsymbol{X}_{\mathrm{H}}$. The relation between H-coordinates and the camera coordinates is shown by Fig.3(c). The head pose for H-coordinates is described by $R_{\mathrm{H}f\mathrm{H}}$ defined as

$$\boldsymbol{X}_{\mathrm{H}f} = R_{\mathrm{H}f\mathrm{H}}\boldsymbol{X}_{\mathrm{H}}. \tag{2}$$

If the rotation matrix R_{CH} describes the relation between H-coordinates and the camera coordinates as $\boldsymbol{X}_{\mathrm{C}} = R_{\mathrm{CH}}\boldsymbol{X}_{\mathrm{H}}$, the head pose $R_{\mathrm{H}f\mathrm{H}}$ for H-coordinates is obtained using that for the camera coordinates obtained by the factorization as

$$R_{\mathrm{H}f\mathrm{H}} = R_{\mathrm{CH}}^{-1}R_{\mathrm{C}f\mathrm{C}}^{-1}R_{\mathrm{CH}}. \tag{3}$$

3.2 Pose representation in H-coordinates

We think that head poses are classified efficiently in H-coordinates, defined above, because it has the advantage that we do not depend on camera position and that we are independent of inclining head motion (tilt angle), which does not change the gaze. This coordinate system enables us to represent a nod or shake motion by changing one angle and fixing the other angles. H-coordinates are used for labeling images in the training stage and for estimation in the recognition stage.

Figure 5. Face regions (large rectangles) and face feature points (small rectangles) obtained by template matching.

The rotation matrix $R_{\mathrm{H}f\mathrm{H}}$ is represented with $\phi_{\mathrm{H}f}$(roll), $\theta_{\mathrm{H}f}$(pitch) and $\psi_{\mathrm{H}f}$(yaw). The rotation angles are set to ϕ_{H}=θ_{H}=ψ_{H}=0 for the first frame, and have the following roles

$$\begin{array}{lcl} \text{horizontal angle (right is positive)} & = & \psi_{\mathrm{H}f}, \\ \text{vertical angle (up is positive)} & = & \theta_{\mathrm{H}f}, \\ \text{tilt angle} & = & \phi_{\mathrm{H}f}, \end{array} \tag{4}$$

where a rotation matrix R is described as

$$R = \begin{bmatrix} \cos\phi\cos\theta & \cos\phi\sin\theta\sin\psi & \cos\phi\sin\theta\cos\psi \\ & -\sin\phi\cos\psi & +\sin\phi\sin\psi \\ \sin\phi\cos\theta & \sin\phi\sin\theta\sin\psi & \sin\phi\sin\theta\cos\psi \\ & +\cos\phi\cos\psi & -\cos\phi\sin\psi \\ -\sin\theta & \cos\theta\sin\psi & \cos\theta\cos\psi \end{bmatrix}. \tag{5}$$

4 Head Pose Estimation System

The head pose estimation is computed using the subspace method[5, 6]. We create subspace dictionaries of all head pose classes in advance during the training stage. At the recognition stage, the similarities between the subspaces and input images are calculated, and face detection and head pose estimation are executed simultaneously.

4.1 Training stage

At the training stage (Fig.4(a)), we first obtain the positions of face feature points (eyes, nose, mouth edges) for each image sequence (each person) by manually labeling regions in selected frames. Then templates of face and face feature points are cropped on the basis of the positions. Template matching using these templates is performed (Fig.5) to obtain face feature points for a number of image frames. The QTR is used for computing the matching measure, and adaptively scaled template matching is used here.

Next we calculate the head pose of each image frame. We use the factorization method[2] for manually labeled frames, and use a method using the pseudo-inverse matrix of the shape matrix S, explained below, for frames labeled by template matching, because the latter method is fast and can prevent an effect of outliers on other frames after S is obtained.

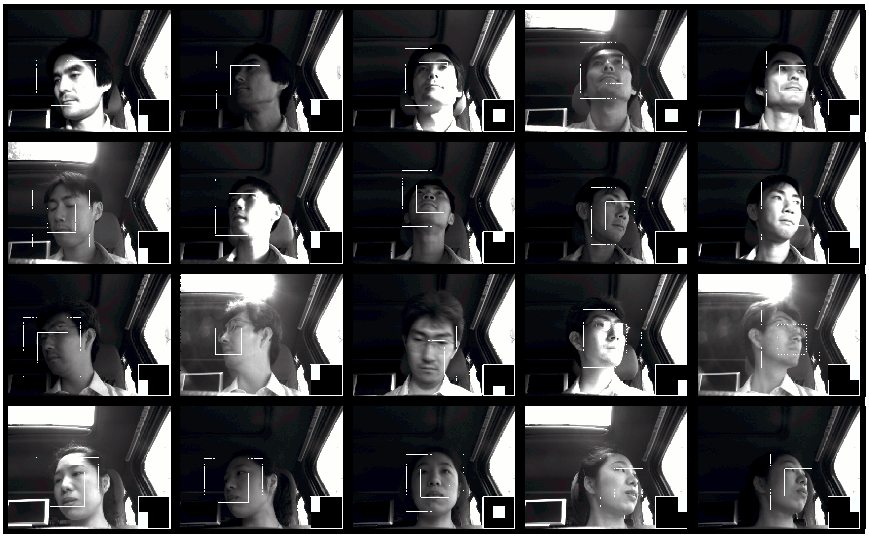
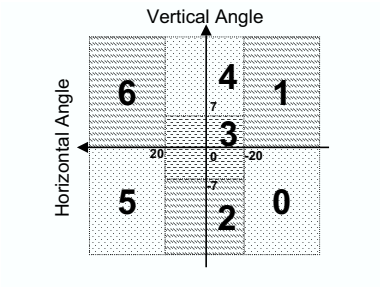


Figure 7. Results of the head pose estimation in various lighting conditions. The head pose estimates are shown graphically in the lower right corners.



(a)

(b)

Figure 6. (a) The seven classes of head poses (classes 0–6). (b) The results of head pose estimation. The head pose estimates are shown graphically in the lower right corners.

The measurement matrix W, the motion matrix M and the shape matrix S satisfy

$$W = MS, \tag{6}$$

where W is built from the positions of face feature points in the images, M contains object poses (head poses), and S represents the three-dimensional shape of the feature points. The object poses are obtained using singular value decomposition (SVD).

If we have already obtained S from some frames, an object pose M_{new} of a new frame is computed as

$$M_{\text{new}} = W_{\text{new}} S^{\dagger}, \tag{7}$$

where W_{new} contains feature points of the new frame, and the pseudo-inverse matrix $S^{\dagger}$ is calculated by $S^{\dagger} = S^{\mathrm{T}}(SS^{\mathrm{T}})^{-1}$.

Each head pose obtained above is transformed into H-coordinates. These are classified into seven classes (Fig.6(a)), which are shown graphically (Fig.6(b)) at the recognition stage. For example, class 0 belongs to the region whose horizontal and vertical angles (deg.) are $\psi > -20$ and $\theta < 0$, respectively. Finally, the subspaces of all head pose classes are computed by principal component analysis.

4.2 Recognition stage

At the recognition stage (Fig.4(b)), nostril detection and face detection using pyramidal images and a raster image scan are performed as the first step of face detection only when the detection in the previous frame is missed.

Nostril detection is used for accelerating the search and increasing the ratio of successful detection. The detection uses the separability filter and pattern verification using the subspace method[8].

The system detects the face more accurately and classifies the head pose into one of seven classes (frontal, upper frontal, lower frontal, upper left, lower left, upper right, lower right) (Fig.6) using adaptively scaled template matching in the second step. The face detection and estimation is based on the similarity between the subspace dictionary and the cropped face images. This system can process 20 frames per second on a Pentium4 3GHz PC.

5 Experiments

We captured face images of 21 people in various lighting conditions while they were sitting in the

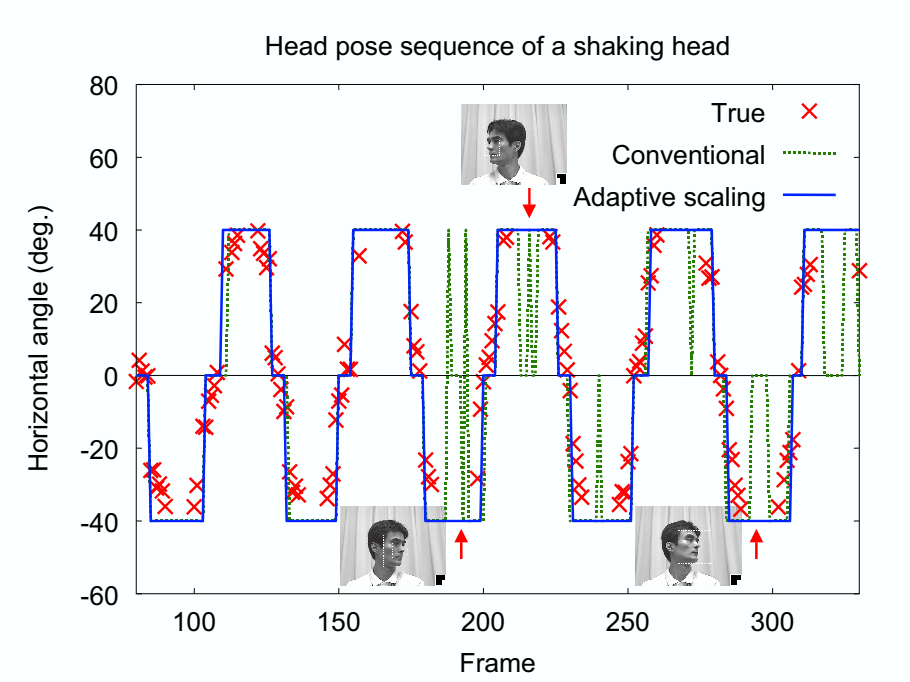


Figure 8. Head pose sequence of a shaking head. The estimates of the horizontal angles for the proposed method (solid line) and the system without adaptively scaled matching (dotted line) are compared. The head poses obtained by the factorization method are shown by "×". The error of the conventional method becomes larger at points of large horizontal angles.

driver's seat inside a car in the sunshine. We computed subspaces (dictionaries) of images of the face regions, and evaluated off-line the performance of the head pose estimation by the proposed system using these images.

The results of the head pose estimation are shown in Fig.7. The head pose in each frame is estimated correctly in various lighting conditions, and the estimate is shown graphically in the lower right corner of each image.

Figure 8 compares the estimates of the horizontal angles for the proposed method and the system without adaptively scaled matching for a head shaking motion. The angles calculated from feature positions by manual input and the factorization are assumed to be true values (mark ×). The error of the conventional method becomes larger than the proposed one at points of large horizontal angles.

6 Conclusion

A real-time head pose estimation system using a new template matching technique was proposed in this paper. The adaptively scaled template matching method is a fast and accurate image matching method for object regions of changing size in the image. The method was successfully applied to the problem of head pose estimation.

Acknowledgments

The authors would like to thank Björn Stenger for helpful comments.

References

[1] Y. Matsumoto and A. Zelinsky. An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. In *Proc. Int'l. Conf. on Automatic Face and Gesture Recognition (FG2000)*, pp. 499–504, Grenoble, France, March 2000.

[2] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, Vol. 9, No. 2, pp. 137–154, 1992.

[3] R. Pappu and P.A. Beardsley. A qualitative approach to classifying gaze direction. In *Proceedings of the 3rd International Conference on Automatic Face and Gesture Recognition*, pp. 160–165, Nara, Japan, April 1998.

[4] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 84–91, June 1994.

[5] E. Oja. *Subspace Methods of Pattern Recognition.* Research Studies Press Ltd., 1983.

[6] S. Watanabe, P. F. Lambert, C. A. Kulikowski, J. L. Buxton, and R. Walker. *Evaluation and Selection of Variables in Pattern Recognition.* Computer and Information Sciences II. Academic Press, New York, 1967.

[7] O. Yamaguchi and K. Fukui. Pattern hashing - object recognition based on a distributed local appearance model. In *Proc. IEEE Int'l. Conf. on Image Processing (ICIP-02)*, Vol. 3, pp. III–329–III–332, Rochester, New York, June 2002.

[8] K. Fukui and O. Yamaguchi. Facial feature point extraction method based on combination of shape extraction and pattern matching. *Syst. Comput. Jpn.*, Vol. 29, No. 6, pp. 49–58, June 1998.