PAPER

Estimating Head Orientation Using a Combination of Multiple Cues

Bima Sena Bayu DEWANTARA[†a)], *Nonmember and* **Jun MIURA**[†b)], *Member*

SUMMARY This paper proposes an appearance-based novel descriptor for estimating head orientation. Our descriptor is inspired by the Weber-based feature, which has been successfully implemented for robust texture analysis, and the gradient which performs well for shape analysis. To further enhance the orientation differences, we combine them with an analysis of the intensity deviation. The position of a pixel and its intrinsic intensity are also considered. All features are then composed as a feature vector of a pixel. The information carried by each pixel is combined using a covariance matrix to alleviate the influence caused by rotations and illumination. As the result, our descriptor is compact and works at high speed. We also apply a weighting scheme, called Block Importance Feature using Genetic Algorithm (BIF-GA), to improve the performance of our descriptor by selecting and accentuating the important blocks. Experiments on three head pose databases demonstrate that the proposed method outperforms the current state-of-the-art methods. Also, we can extend the proposed method by combining it with a head detection and tracking system to enable it to estimate human head orientation in real applications.

key words: *human head orientation, Weber feature, gradient, intensity deviation, covariance, block importance feature, head detection and tracking*

1. Introduction

The human head and face are the most common parts of the human body used in computer vision applications such as detecting the presence of a person, identifying and verifying a person, and indicating one's attention. In order to maintain a good communication or a good interaction, the estimating of head orientation is important. Head orientation can be used to estimate an attentional awareness during an interaction process so that one may expect an appropriate response that is in-line with the degree of attention.

Many methods have been proposed to deal with head orientation estimation [1]. The methods can be categorized into three main groups: (1) facial features based method, (2) model-based method, and (3) appearance-based method. The facial features based methods detect facial components such as the eyes, mouth and nose and calculate the geometrical relationship between them for face orientation estimation. This approach normally requires a high precision facial component detector. However, the detector is usually sensitive to distance and changes in illumination levels. The model-based methods use a priori known 3-D models of a human's head or 2-D models of a face such as the face contour and the facial components that are matched with the unrecognized 3-D head or 2-D face models. This approach usually works well at limited angles only, e.g., pan angle is within ±50^o. The appearance-based methods assume there is a relationship between the head or face pose and changes in some properties in a 2-D facial image. This approach accommodates wider pan, tilt and roll angles in a plane. However, inaccuracy is the main drawback of this approach.

Some applications such as car driver awareness detection, human-robot interaction, and video surveillance involve a variety of view-angle ranges, from a tight range up to a full 360^o view. In human-robot interaction, the range of angle ±90^o is commonly used. Therefore, we focus our work on the appearance-based approach by proposing a novel descriptor that combines more features such as Weber, gradients and intensity deviation. This descriptor is named the Covariance of Weber-Gradient-Deviation Descriptor (CWGDD). We show that integrating more features with different capabilities will make our descriptor more robust in order to discriminate various head orientations.

1.1 Related Works

Many prior works on head pose estimation have been surveyed in [1]. In general, most of them used limited feature variations such as intensity and edge information.

Han *et al.* [2] proposed the Image Abstraction and Local Directional Quaternary Pattern (IA-LDQP). An edge-like image is extracted from a precisely segmented-image using a Difference of Gaussian (DoG) filter. A set of edges is then extracted into a histogram by following the LDQP technique. However, distribution of the histograms are frequently similar among different poses due to binary intensity. This issue potentially reduces accuracy.

The Gabor-Filter is frequently used for edge detection. The multiple-scale and multiple-orientation edge features that are extracted from an image using the Gabor-Filter have been proposed in the Covariance of Gabor Filter (CovGa) [3]. This feature is then combined with other information such as pixel positions and intensities using a covariance matrix. This method successfully achieves a good accuracy. However, this approach consumes excessive time due to large number of scales and orientations used.

One notable work is the Covariance of Oriented Gradient (COG) by Dong *et al.* [4]. The idea of using a covariance matrix for the head pose classification problem was first pro-

Manuscript received September 15, 2015.
Manuscript revised January 10, 2016.
Manuscript publicized March 3, 2016.

[†]The authors are with the Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan.
a) E-mail: bima@aisl.cs.tut.ac.jp
b) E-mail: jun.miura@tut.ac.jp

DOI: 10.1587/transinf.2015EDP7375

posed in this paper. They argued that a head pose can be assumed as changes in the edges information of a face image. This assumption can be realized in a low-resolution image extracted by a gradient approach. This descriptor is mainly composed of a gradient image and orientations of gradient combined with other information such as pixel positions and its intrinsic intensity when using a covariance matrix. The experimental results showed that this method achieves an excellent result for classifying head poses. Referring to the achievements of the last two methods [3], [4], a covariance matrix seems very promising and suitable to combine and to compact multiple features.

CovGA and COG are shown to be effective for human head pose estimation. However, we think that combining the advantage of gradient features in COG with other feature that is effective for analyzing different cue may improve the performance. Weber Local Descriptor (WLD) [5] is a robust descriptor for texture analysis. This descriptor works well for discriminating and classifying textures (Brodatz and KTH-TIPS2 texture databases) by considering two pieces of information at once; (1) differential excitation and (2) orientation of the gradient. The differential excitation elegantly detects the edges and is robust to the illumination change while the orientation of gradient is very powerful for characterizing a directional change in the intensity or color in an image. This descriptor is never used for head orientation estimation. Because of these advantages, we have adopted into our system.

There are two main contributions in this paper. First, we combine different types of features such as pixel position, intrinsic intensity, Weber, gradients and intensity deviations to create a more robust descriptor. The engagement of the last three features is proven to be effective for discriminating each pose/orientation. Secondly, we perform a deeper analysis to measure the degree of importance of each image's block. The degree of importance is determined by a weight. The application of appropriate weights can provide significant improvement in the results. Determination of the appropriate and optimal weights are our greatest contribution. The use of a Genetic Algorithm (GA) [6] to determine the values of the weights is a further additional contributions.

1.2 Paper Organization

The rest of this paper is organized as follows. Section 2 describes the development of our descriptor for estimating head orientation. Section 3 discusses an implementation of our descriptor for a real scene application of the head orientation estimation. Section 4 presents the experimental results and discussions. Section 5 concludes our work and possible future works.

2. Building a Descriptor for Estimating Head Orientation

2.1 Weber-Based Feature

The head orientation variations can be perceived as the change of image pattern. Characterizing the image patterns can be performed in most texture domains, e.g.: SIFT [9], LBP [10] and WLD [5]. WLD-based feature has been proven to perform well for texture analysis. The changes of head orientation are closely related to the changes of texture.

2.1.1 Generating Features

The Weber-based feature [5] is mainly composed of two parts (see Fig. 1); (1) The differential excitation and (2) the orientation of the gradient. The differential excitation measures the intensity differences between a current pixel with its neighbors to find the salient variations within an image that is expressed as

$$\begin{aligned} v_1 &= \textstyle\sum_{i=0}^{p-1} (\Delta I_i) = \sum_{i=0}^{p-1} (I_i - I_c), \\ v_2 &= I_c, \end{aligned} \tag{1}$$

where I_c is the center pixel, I_i $(i = 0, 1, \ldots, p-1)$ denotes the $i-th$ neighbors of I_c and p is the number of neighbors. The differential excitation ξ can be expressed as:

$$\xi = \tan^{-1}\left(\frac{v_1}{v_2}\right) = \tan^{-1}\left(\sum_{i=0}^{p-1}\left(\frac{I_i - I_c}{I_c}\right)\right). \tag{2}$$

The orientation of the gradient shows a direction of

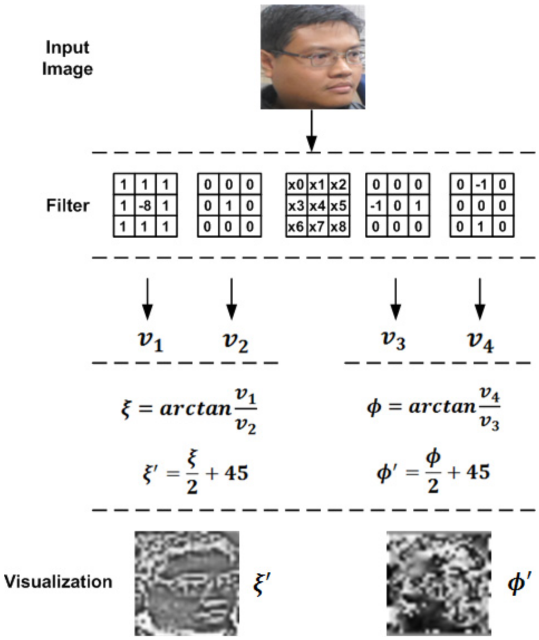


Fig. 1 Block diagram of Weber-based features generation

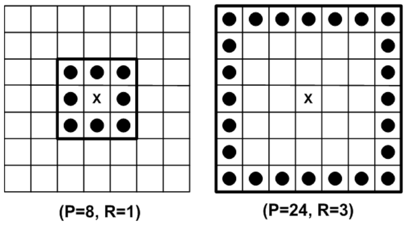


Fig. 2 Squared symmetric neighborhood for different (P, R)

pixels difference in vertical, $v_4 = x_7 - x_1$, and horizontal, $v_3 = x_5 - x_3$. x_1, x_3, x_5 and x_7 are the kernel filters shown in Fig. 1. The orientation of the gradient ϕ is expressed as:

$$\phi = \tan^{-1}\left(\frac{v_4}{v_3}\right) = \tan^{-1}\left(\frac{x_7 - x_1}{x_5 - x_3}\right). \tag{3}$$

The Arctangent function is applied since it can limit the output to prevent it from increasing or decreasing too quickly when the input becomes larger or smaller. For simplicity and to avoid a negative-value-effect to the *log* operator, both features are then quantized into positive values between 0^o-90^o using the following expression:

$$\begin{aligned} \xi' &= \tfrac{\xi}{2} + 45, \\ \phi' &= \tfrac{\phi}{2} + 45, \end{aligned} \tag{4}$$

where ξ' and ϕ' are the new quantized ξ and ϕ, respectively.

2.1.2 Analysis of Weber Scales

As described in [5], the size of kernel filters of Weber can be easily scaled. The scale can be generated by regulating the kernel filter size and the radii of the filter as shown in Fig. 2. The parameter P denotes the number of the neighbors, whereas R determines the radii of the operator. Based on our experiments, applying $R = 3$ gave us improved head orientation estimation results.

2.2 Gradients-Based Feature

The head orientation can be assumed as the changes of edges or shapes of particular parts of the face image [4]. Representation of the changes of edge or shape in the low-resolution image is suitable for characterizing image variations that are insensitive to illumination changes. To strengthen the image characterization, we apply the second order vertical ∇I_{yy} and horizontal ∇I_{xx} gradients to the image as follows.

$$\begin{aligned} \nabla I_{xx} &= \tfrac{\partial^2 I(x,y)}{\partial x^2}, \\ \nabla I_{yy} &= \tfrac{\partial^2 I(x,y)}{\partial y^2}. \end{aligned} \tag{5}$$

2.3 Intensity Deviation

The head orientation can also be defined as a change of intensities of a particular region in a given image. We mapped the pixel-based intensity deviation from an image to describe the changes of intensities. Using this map, the changes of location of specific intensities describe the head orientation changes. We apply normalized local and global intensity deviations to identify the distinctive map between each orientation. The specific transition area of the sigmoid function as shown in Fig. 3 is attracted our attention. We think there are two advantages when using this specific function: the linearity of the output for deviation up to a particular value and can hold the output at a particular significant deviation.

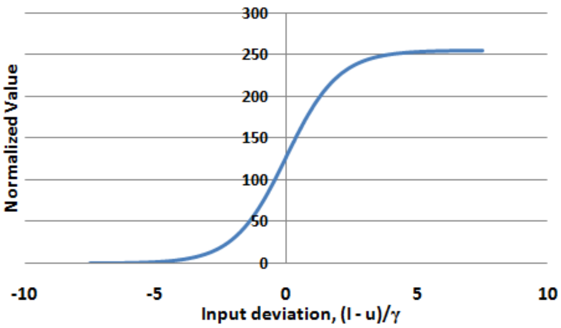


Fig. 3 Sigmoid function for normalizing intensity deviations ($\gamma = 34$)

The normalized local intensity deviation p is expressed as:

$$p^b_{x,y} = \frac{255}{1 + \exp^{(-(I^b_{x,y} - \mu^b)/\gamma)}}, \tag{6}$$

where $p^b_{x,y} (b = 1, 2, \ldots, B)$ is the normalized local intensity deviation of $b - th$ block, B is the number of blocks, $I^b_{x,y}$ is the pixel's intensity where $(x, y) \subset \mathbb{R}^b$, μ^b is the mean of intensity of the b-th block. γ is a scaling factor for the input deviation. We set $\gamma = 34$ for all experiments (see Sect. 4.1 for the detail).

The normalized global intensity deviation q is expressed as:

$$q_{x,y} = \frac{255}{1 + \exp^{(-(I_{x,y} - \mu)/\gamma)}}, \tag{7}$$

where μ is the mean of intensity of the image.

2.4 Contribution of Each Feature and the Purpose of Combination

Weber feature provides a good representation for texture analysis that is insensitive to illumination changes. Our preliminary evaluation shows that this feature is also insensitive to image noise. The gradient feature provides a good representation for object shape, that is also insensitive to illumination changes. Deviation feature can suppress significant intensity deviations, while keeping smaller ones which are useful for classification. The last two features are, however, sensitive to image noise. Combining the three features takes advantages of their strong points and is expected to compensate for the drawbacks, thereby exhibiting a better performance than using a single feature.

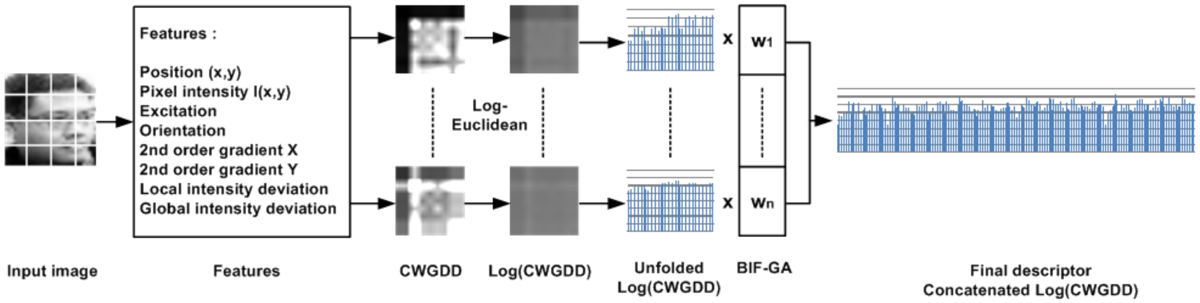


Fig. 4 Block diagram of our proposed descriptor

2.5 Covariance Matrix for Combining Features

We use a covariance matrix as proposed by Tuzel *et al.* [11] to combine all features. We extract a 9-dimensional feature from a grayscale image so that each pixel is composed as a vector of features as follows.

$$\mathbf{I}_i = [x\, y\, I_{xy}\, \xi'_{xy}\, \phi'_{xy}\, \nabla I_{xx}\, \nabla I_{yy}\, p^b_{xy}\, q_{xy}]^T, \tag{8}$$

where $\mathbf{I}_i$ is a vector of features of i-th pixel, x, y are the pixel positions, I_{xy} is the image intensity, ξ'_{xy}, ϕ'_{xy} are the new quantized excitation and orientation, respectively. ∇I_{xx}, ∇I_{yy} are the second order horizontal and vertical gradients, respectively. p^b_{xy}, q_{xy} are the normalized local and global intensity deviations, respectively.

2.6 Symmetric Positive Definite and Distance Metric

The covariance matrix is one example of a symmetric positive definite (SPD) matrix. Measurement of the distance between two SPD matrices can be done using $Log-Euclidean$ metric [12]. Following the $Log-Euclidean$ metric, we treat our descriptor using the same manner as presented in [4]. The covariance matrix of block (m, n) is transformed to the matrix logarithm $\log(C_{m,n})$. Each $\log(C_{m,n})$ is then unfolded into a vector space by accommodating $d \times (d + 1)/2$ independent values only (half of the upper triangle or the lower triangle of the symmetrical matrix), where d is the number of dimensions of the feature vector.

2.7 Weighting Scheme for Accentuating Important Parts

A cropped image of a human head usually includes unnecessary parts such as background objects and clothes. Dividing an image into several blocks and controlling the feature's value of each block may improve the discrimination power of a descriptor. Therefore, we use a block importance feature (BIF) scheme, a set of weights that are heuristically predetermined using a Genetic Algorithm (GA) [6].

We chose GA due to the following reasons: (1) it is the most widely used algorithm and has matured as a robust optimization technique [7], and (2) it performs global search which is faster enough compared to the others [8].

For the GA training purpose, some individuals (vectors of genes that represent sets of weights) are initialized using normalized random values. We perform classifier-based optimization and use the hit rate (HR) of the classifier to calculate the fitness value. HR is expressed as

$$HR = \frac{\sum tp}{\sum tp + \sum fp}, \tag{9}$$

where tp and fp are the true positive and the false positive, respectively.

The fitness of each individual is evaluated using a combination of the HRs for three principal component analysis (PCA) based classifiers, i.e., PCA+ED (Euclidean distance), PCA+NC (nearest centroid) and PCA+LDA+NC (linear discriminant analysis) as follows.

$$fitness = HR_{(ED)} HR_{(NC)} HR_{(LDA+NC)}. \tag{10}$$

This fitness function is intended for learning block importance weights which are effective for various classifiers. Among five classifiers shown below, we use only PCA+ED, PCA+NC, and PCA+LDA+NC for this optimization because they do not need the learning phase, thereby reducing the optimization cost.

Mutation and crossover are then applied to the selected parent. Since the GA works iteratively, the optimization process is repeated using an iteration number as the stopping criteria.

2.8 Building a Complete Descriptor to Estimate Head Orientation

We have presented all of the features we used for supporting our descriptor. Figure 4 shows an input image that is preprocessed using histogram equalization to minimize illumination effects. Afterward, it is divided into B-blocks. A set of determined-features is extracted from each point in each block and processed using our method. By following the $Log - Euclidean$ metric, a $d \times (d + 1)/2$-dimensional of feature can be obtained from each block. This feature is multiplied by a BIF-GA weight. A final descriptor is then built using a simple concatenation. For B number of blocks, we have $B \times d \times (d + 1)/2$-dimensional of feature.

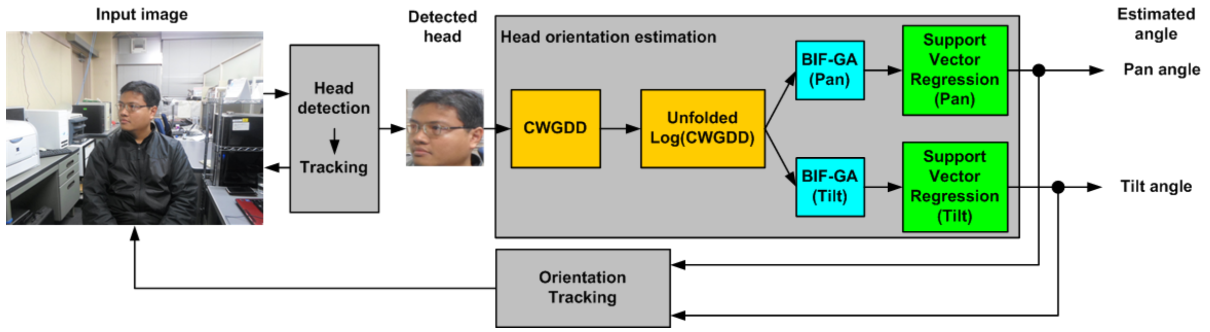


Fig. 5 Block diagram of an online head orientation estimation system

3. Estimating Head Orientation in a Real Application

3.1 Head and Body Detection

We extend our work for real application as shown in Fig. 5. In a real application, a robust head detector is required. We train our Viola's head detector [13] using 6,000 positive images (mix of original training images of Viola-Jones, Pointing'04 [14], FEI [15] and our AISL databases [16]) and 8,000 negative images. However, its performance frequently degrades in-line with a change in distance. We additionally use a human upper body detector [17] for setting an ROI (region of interest) for head detection.

3.2 Head Tracking and Orientation Smoothing

Our system is composed of two independent processes as shown in Fig. 5. The first process is head detection, while the second is head orientation estimation. To get a stable head detection result is very difficult because Viola's detector is specifically designed for detecting faces. Therefore, we apply a Kalman filter to assist our detector with tracking the detection result. Our state model is composed of the detected head position (x_t, y_t), its derivative $(\dot{x}_t, \dot{y}_t)$, and the bounding box size (w_t, h_t). A constant velocity model is utilized to model the head position in an image by considering a time interval Δt:

$$\begin{aligned} x_t &= x_{t-1} + \Delta t \dot{x}_{t-1} + \epsilon_{x_t}, \\ y_t &= y_{t-1} + \Delta t \dot{y}_{t-1} + \epsilon_{y_t}, \\ \dot{x}_t &= \dot{x}_{t-1} + \epsilon_{\dot{x}_t}, \\ \dot{y}_t &= \dot{y}_{t-1} + \epsilon_{\dot{y}_t}. \end{aligned} \tag{11}$$

The bounding box size is maintained as

$$\begin{aligned} w_t &= w_{t-1} + \epsilon_{w_t}, \\ h_t &= h_{t-1} + \epsilon_{h_t}. \end{aligned} \tag{12}$$

The state can be expressed as a tuple $\mathcal{X} = \{x_t, y_t, \dot{x}_t, \dot{y}_t, w_t, h_t\}$. Our Kalman filter model is expressed as $\mathcal{X}_t = \mathbf{F}_t \mathcal{X}_{t-1} + \epsilon_t$, where $\mathbf{F}_t$ is the state transition model and $\epsilon_t = N(0, Q_t)$ is the process noise which is assumed to be drawn from a zero mean multivariate normal distribution with covariance $Q_t = \{2, 2, 0.5, 0.5, 2, 2\}$.

On the other hand, the output of the head orientation estimator also fluctuated. To reduce fluctuations of the orientation estimation results, we also utilize a Kalman filter to smooth the head orientation estimation. Our state model is composed of the estimated angles (α_t, β_t) and its derivative $\omega_{\alpha_t}, \omega_{\beta_t}$. Estimation of the pan and the tilt angles use a constant angular velocity model as follows.

$$\begin{aligned} \alpha_t &= \alpha_{t-1} + \Delta t \omega_{\alpha_{t-1}} + \varepsilon_{\alpha_t}, \\ \beta_t &= \beta_{t-1} + \Delta t \omega_{\beta_{t-1}} + \varepsilon_{\beta_t}, \\ \omega_{\alpha_t} &= \omega_{\alpha_{t-1}} + \varepsilon_{\omega_{\alpha_t}}, \\ \omega_{\beta_t} &= \omega_{\beta_{t-1}} + \varepsilon_{\omega_{\beta_t}}. \end{aligned} \tag{13}$$

The state can be expressed as a tuple $\mathcal{Y} = \{\alpha_t, \beta_t, \omega_{\alpha_t}, \omega_{\beta_t}\}$. The Kalman filter model is expressed as $\mathcal{Y}_t = \mathbf{G}_t \mathcal{Y}_{t-1} + \varepsilon_t$, where $\mathbf{G}_t$ is the state transition model and $\varepsilon_t = N(0, P_t)$ is the process noise which is assumed to be drawn from a zero mean multivariate normal distribution with covariance $P_t = \{2, 2, 1, 1\}$.

4. Experiments

4.1 Analysis of γ-Values

We performed extensive experiments using Pointing'04 head pose database (7 pan angles) to find γ which is suitable for our intensity deviation feature. We evaluated several values of γ within a range [1:128]. We found that $\gamma = 34$ achieved the highest classification accuracy and the lowest angle error in average of five classifiers as shown by Fig. 6.

4.2 Analysis of Different Block Size

To improve the discrimination power of the descriptor, we divided an input image into a set of blocks. However, finding the best size of blocks is rather difficult. To figure out this matter, we conducted a small experiment to find the best size of blocks by trying it on Pointing'04 database using 2 × 2, 3 × 3, 4 × 4, 6 × 6, 9 × 9, and 12 × 12. Based on our experiment, we found that the block size of 4 × 4 achieves the best accuracy and the block size of 6 × 6 achieves the second best accuracy.

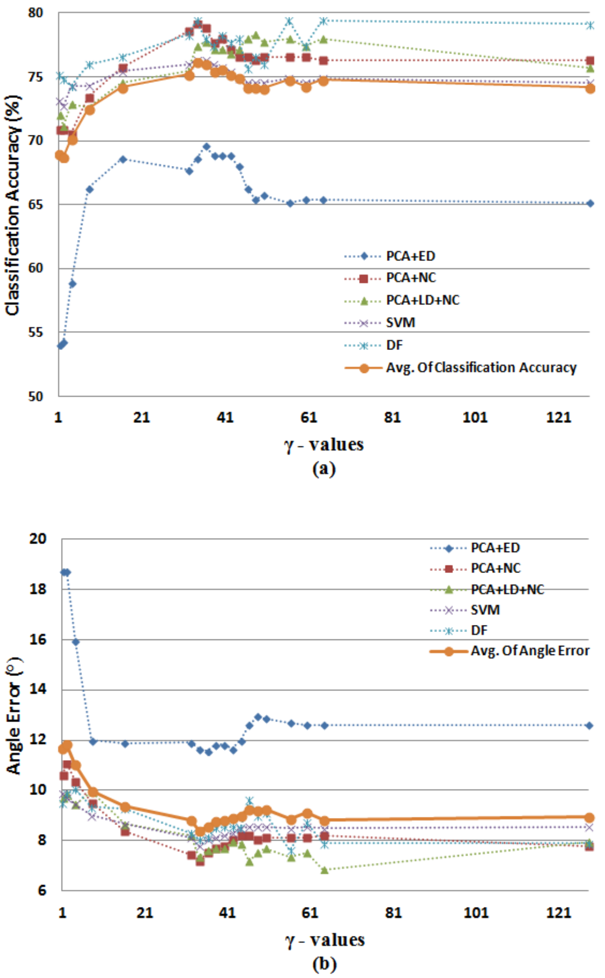


Fig. 6 Comparison of different γ-values; (a) the accuracy, and (b) the MAAE

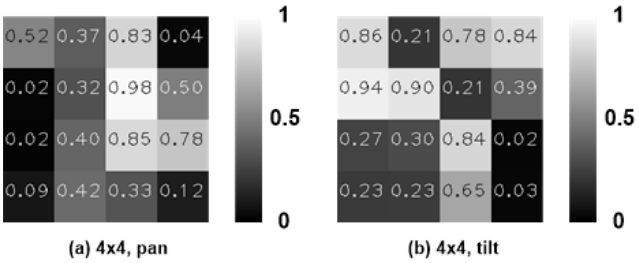
4.3 Optimizing Weights Using Genetic Algorithm

4.3.1 Dataset

We use 1,680 cropped-images of 105 subjects that are taken from two databases; Pointing'04 head pose database [14] and NCKU database [20]. The ranges of pan and tilt angles are ±90^o and ±60^o, respectively. We select a set of images with interval ±30^o only. The first 10 subjects of Pointing'04 and the first 60 subjects of NCKU are utilized as training data, while the remaining are used as a probe set. All images are resized into 36 × 36 pixel.

4.3.2 Optimization Results

Based on the result of Sect. 4.2, we only pay attention to analyze the block size of 4 × 4 and 6 × 6 in our system. The weights are obtained after running GA for 100 iterations with 20 generated-individuals. Figure 7 and Fig. 8 show the optimal weights of 4 × 4 and 6 × 6 blocks, respectively. Table 1 shows the comparison of the effectiveness using block size of 4×4 and 6×6. The Pointing'04 head pose database is used as the evaluation dataset. For some classifiers, increasing the block size is not give us better results for estimating pan orientation. However, opposite conditions occur for estimating tilt orientation.



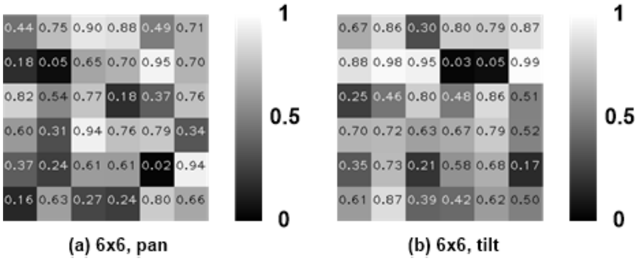
0.52	0.37	0.83	0.04
0.02	0.32	0.98	0.50
0.02	0.40	0.85	0.78
0.09	0.42	0.33	0.12

(a) 4x4, pan

0.86	0.21	0.78	0.84
0.94	0.90	0.21	0.39
0.27	0.30	0.84	0.02
0.23	0.23	0.65	0.03

(b) 4x4, tilt

Fig. 7 The optimal weights for 4x4 blocks; (a) pan and (b) tilt



0.44	0.75	0.90	0.88	0.49	0.71
0.18	0.05	0.65	0.70	0.95	0.70
0.82	0.54	0.77	0.18	0.37	0.76
0.60	0.31	0.94	0.76	0.79	0.34
0.37	0.24	0.61	0.61	0.02	0.94
0.16	0.63	0.27	0.24	0.80	0.66

(a) 6x6, pan

0.67	0.86	0.30	0.80	0.79	0.87
0.88	0.98	0.95	0.03	0.05	0.99
0.25	0.46	0.80	0.48	0.86	0.51
0.70	0.72	0.63	0.67	0.79	0.52
0.35	0.73	0.21	0.58	0.68	0.17
0.61	0.87	0.39	0.42	0.62	0.50

(b) 6x6, tilt

Fig. 8 The optimal weights for 6x6 blocks; (a) pan and (b) tilt

Figure 7 shows that the change of weights in the horizontal direction is relatively larger than in the vertical direction for pan and that in the vertical direction is larger for tilt. It seems that this coincides with our intuition. Figure 8 shows the weights for 6 × 6 blocks and exhibits a similar weight distribution, although the above tendency is less obvious.

4.4 Offline Experiment Using Head Pose Database

4.4.1 Dataset

We use three head pose databases that are set into datasets for different purposes as summarized in Table 2. Dataset DS1, DS2, DS3, DS4 and DS5 are used independently for offline evaluations. In this section, we compare our method with COG [4], IA-LDQP [2], WLD [5], and CovGA [3]. We re-implement COG, IA-LDQP, and WLD, while we used the result of CovGa directly from their paper. To make fair re-implementations and comparisons, we also use the same parameters settings, i.e., 32×32 pixel (COG and WLD) and 36 × 36 pixel (CWGDD), 4 × 4 blocks (COG, IA-LDQP, WLD and CWGDD).

We use Principal Component Analysis + Euclidean Distance (PCA+ED), Principal Component Analysis + Nearest Centroid (PCA+NC), Principal Component Analysis + Linear Discriminant Analysis + Nearest Centroid (PCA+LDA+NC), Support Vector Machines (SVMs) [18], and Decision Forest (DF) [19] as the classifiers. The param-

Table 1 Experimental results of CWGDD and CWGDD+BIF-GA using different size of blocks. The Pointing'04 head pose database is utilized to evaluate the performances.

Pose	Block	BIF	Accuracy %					MAAE (^{o})				
			+PCA			+SVM	+DF	+PCA			+SVM	+DF
			+ED	+NC	+LDA			+ED	+NC	+LDA		
pan	4x4	no	68.57	**79.14**	**77.43**	75.71	**79.43**	12.09	**7.20**	**7.37**	7.80	**8.06**
pan	6x6	no	**73.14**	77.43	74.29	**81.14**	77.14	**9.00**	8.66	9.26	**6.69**	**8.06**
pan	4x4	yes	69.43	**80.00**	**77.14**	78.57	**78.57**	11.74	**7.20**	**7.80**	6.94	**7.97**
pan	6x6	yes	**70.00**	72.86	75.14	**79.14**	77.43	**10.63**	9.43	8.49	**6.86**	8.14
tilt	4x4	no	58.67	63.29	65.29	**65.71**	74.57	**13.49**	**11.57**	**11.20**	**10.34**	7.80
tilt	6x6	no	**60.86**	**66.86**	**67.14**	64.00	**77.14**	14.83	12.09	11.49	12.17	**6.94**
tilt	4x4	yes	62.86	66.00	67.14	**70.40**	76.00	**12.06**	10.94	10.11	**9.37**	7.37
tilt	6x6	yes	**63.14**	**68.00**	**70.86**	65.71	**75.29**	13.20	**10.63**	**9.34**	11.14	**7.29**

Table 2 Experimental setup of head pose databases

Dataset	Database	Number of Image (Person)	ROI	Angles				Number of Image (Class) in Use	Cross Valid. (k-fold)
				Pan	Tilt	Range	Step		
DS1	CAS-PEAL [21]	4,200 (200)	face	7	3	±45^o	±15^o	4,200 (7 pans)	4
DS2	Pointing'04 [14]	2,790 (15 (x2))	head	13	9	±90^o	±15^o	1,050 (7 pans, ±30^o)	3
DS3	Pointing'04 [14]	2,790 (15 (x2))	head	13	9	±90^o	±15^o	1,050 (5 tilts, ±30^o)	3
DS4	AISL [16]	3,420 (20 (x3))	head	19	3	±90^o	±10^o	1,260 (7 pans, ±30^o)	4
DS5	AISL [16]	3,420 (20 (x3))	head	19	3	±45^o	±45^o	1,260 (3 tilts)	4
DS6	NCKU [20]	6,600 (90)	head	19	1	±90^o	±5^o	630 (7 yaws, ±30^o)	-

Table 3 Parameters setting of each classifier

Method	Parameter	Value
PCA	Number of eigenvalues	100
NC	Number of centroid	10
	Iteration	30
SVMs	Type	multiclass C-SVC
	Kernel type	linear
	C cost	1
DF	Number of tree	500
	Number of random features	5

eters used in each classifier are summarized in Table 3. The performance of each method is measured using the percentage of accuracy and the stability of estimated angle of orientation that is expressed as the mean absolute of angle error (MAAE). $k-fold$ cross validation is used for evaluation.

4.4.2 Experimental Results Using CAS-PEAL Head Pose Database

The CAS-PEAL dataset (DS1) is used to evaluate the performance of each method in terms of estimating face's pan angle. We reduce the effect of background by zooming up a face region that was precisely cropped using eye position. The tilt angles are grouped based on its pan angle. This database gives a distinctive challenge where the appearances of nearby classes are difficult to distinguish due to close angle. Examples of cropped-faces are shown in Fig. 9.

Our experiments show that our CWGDD outperforms the others for all classifiers that are proven by its accuracy and mean absolute angle error (MAAE) as shown in Table 4. On average, our method's accuracy is 2.69% higher than COG. The estimation result stability is also impressive. It is proven by its MAAE 0.42^o smaller than COG.

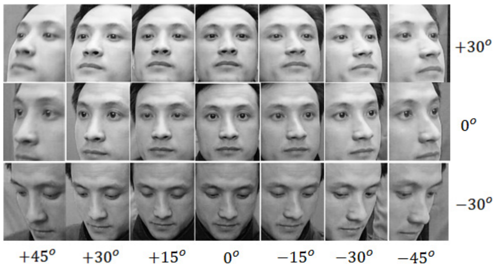


Fig. 9 Example of CAS-PEAL face pose database

4.4.3 Experimental Results Using Pointing'04 Head Pose Database

The Pointing'04 database is used to evaluate the performances of each method for estimating the head pan and tilt angles. Each image was manually cropped where we cannot guarantee its precision. Examples of cropped-heads are shown in Fig. 10. As in the other works, the pan and the tilt angles are estimated separately. For estimating pan, tilt angles are grouped into the same pan angle class while pan angles are grouped into the same tilt angle class for estimating tilt angle.

The first experiment is conducted to see the performances of each method to estimate the head pan using dataset DS2. Table 5 shows that our CWGDD and CWGDD+BIF-GA outperform the others for all classifiers for both the accuracy and the MAAE. On average, our method's accuracy is 7.45% higher than COG, while its MAAE is 4.36^o smaller than COG. CWGDD+BIF-GA successfully improves the accuracy by 0.69% higher and reduce its MAAE by 0.17^o smaller than CWGDD.

The second experiment was performed to see the per-

Table 4 Experimental results of estimating the pan angle using CAS-PEAL face pose database

Method	Accuracy (%)					MAAE (^{o})				
	+PCA			+SVM	+DF	+PCA			+SVM	+DF
	+ED	+NC	+LDA			+ED	+NC	+LDA		
COG [4]	70.30	90.00	95.10	85.52	90.62	5.46	1.61	0.74	1.89	1.41
IA-LDQP [2]	47.50	70.00	38.00	78.52	44.30	13.26	5.72	18.69	3.09	17.57
CovGa [3]	N/A	N/A	94.20	N/A	N/A	N/A	N/A	N/A	N/A	N/A
WLD [5]	73.10	82.50	85.80	83.33	60.30	4.75	2.86	2.24	2.87	7.69
CWGDD (Ours)	**74.86**	**91.10**	**95.71**	**91.19**	**93.00**	**4.41**	**1.29**	**0.70**	**1.34**	**1.05**

Table 5 Experimental results of estimating the pan angle using Pointing'04 head pose database

Method	Accuracy (%)					MAAE (^{o})				
	+PCA			+SVM	+DF	+PCA			+SVM	+DF
	+ED	+NC	+LDA			+ED	+NC	+LDA		
COG [4]	63.10	65.40	69.70	69.14	75.71	15.26	13.29	12.26	14.49	9.00
IA-LDQP [2]	42.90	41.10	54.60	59.43	43.10	38.29	40.29	25.49	16.43	37.54
WLD [5]	45.10	50.60	67.40	66.57	48.90	21.77	31.80	11.57	11.13	21.69
CWGDD (Ours)	68.57	79.14	**77.43**	75.71	**79.43**	12.09	**7.20**	**7.37**	7.80	8.06
CWGDD+BIF-GA	**69.43**	**80.00**	77.14	**78.57**	78.57	**11.74**	**7.20**	7.80	**6.94**	**7.97**

Table 6 Experimental results of estimating the tilt angle using Pointing'04 head pose database

Method	Accuracy (%)					MAAE (^{o})				
	+PCA			+SVM	+DF	+PCA			+SVM	+DF
	+ED	+NC	+LDA			+ED	+NC	+LDA		
COG [4]	61.14	65.14	65.43	**71.71**	72.00	13.89	11.40	11.23	**8.66**	9.43
IA-LDQP [2]	43.40	44.00	47.10	52.86	20.00	25.29	23.40	25.49	13.25	36.00
WLD [5]	57.70	55.40	56.60	57.00	46.00	14.23	14.23	13.63	19.60	20.06
CWGDD (Ours)	58.67	63.29	65.29	65.71	74.57	13.49	11.57	11.20	10.34	7.80
CWGDD+BIF-GA	**62.86**	**66.00**	**67.14**	70.40	**76.00**	**12.06**	**10.94**	**10.11**	9.37	**7.37**

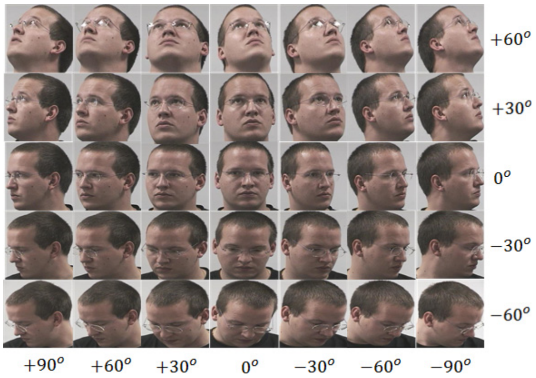


Fig. 10 Example of Pointing'04 head pose database

formances of each method to estimate the head tilt using dataset DS3. Table 6 shows that our CWGDD performances are bit worse than COG. In average, our CWGDD accuracy is 1.38% lower than COG, however, its MAAE is better than COG by achieving 0.04^o lower. Improving CWGDD with BIF-GA increases its accuracy by 2.77% and reduces the MAAE by 0.91^o.

From the experiments, gradient-based method is particularly effective to characterize a human head tilt. However, our BIF-GA scheme is able to gain the performance of CWGDD. Figure 11 shows examples of incorrect estimation. They are due to an incompletely cropped head area (Fig. 11 (a)) or a high portion of background (Fig. 11 (b)).

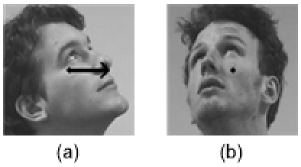


Fig. 11 Examples of failure cases on Pointing'04 database

4.4.4 Experimental Results Using AISL Head Orientation Database

The AISL head orientation database is used to evaluate the performance of each method in estimating the head pan and tilt angles when the background changes. This database simulates the changes of background for indoor environments which might confound the estimation results. In this experiment, again, we cannot precisely crop each image. The pan and the tilt angles are estimated separately. All background variations are combined together as shown in Fig. 12.

The first experiment is conducted to see the performances of each method to estimate the head pan using dataset DS4. Table 7 shows that our CWGDD and CWGDD+BIF-GA outperform the others for all classifiers. On average, CWGDD accuracy is 2.54% higher than COG. BIF-GA successfully improves the accuracy by 2.60%. The second experiment was performed to see the performances of each method to estimate the head tilt using dataset DS5.

Table 7 Experimental results of estimating the pan and the tilt angles using AISL head orientation database

Method	Pan - Accuracy (%)					Tilt - Accuracy (%)				
	+PCA			+SVM	+DF	+PCA			+SVM	+DF
	+ED	+NC	+LDA			+ED	+NC	+LDA		
COG [4]	67.62	77.78	79.05	80.63	77.46	65.40	64.76	41.59	**77.14**	**71.75**
IA-LDQP [2]	17.46	19.68	16.83	47.62	53.65	34.92	33.65	35.24	63.81	64.76
WLD [5]	34.29	60.95	67.30	65.08	58.73	56.51	54.60	33.97	66.44	61.90
CWGDD (Ours)	66.35	**79.37**	**83.49**	85.40	80.63	67.94	73.02	70.79	73.97	67.94
CWGDD+BIF-GA	**73.65**	79.05	81.59	**90.48**	**83.49**	**72.70**	**73.65**	**73.97**	74.29	68.25

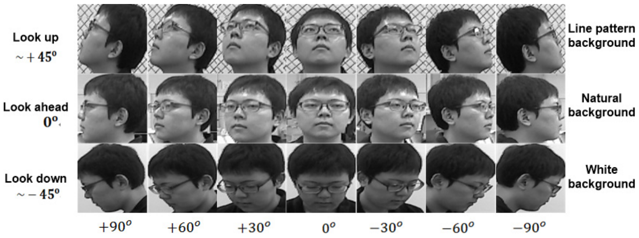


Fig. 12 Example of AISL head orientation database

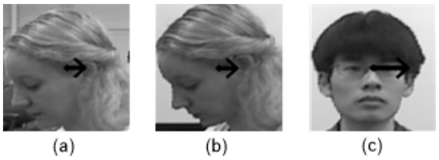


Fig. 13 Examples of failure cases on AISL database

Our CWGDD outperforms the other methods for PCA-based classifiers. On average, CWGDD accuracy is 6.60% higher than COG. BIF-GA successfully improves the accuracy by 1.84%. From the experiments, a single cue-based methods cannot get good results because of their sensitivity to the noises.

Figure 13 shows examples of incorrect results. Our method fails for Fig. 13 (a) and Fig. 13 (b) due to a different hair color from black, which causes a wrong intensity mapping when we used a dataset with only black hair persons for training. Figure 13 (c) shows an unbalanced cropped head area case.

4.4.5 Comparison with a Deep Learning-Based Method

Deep Learning has recently been gaining much popularity due to its high performance in image and speech processing. Cai *et al.* [23] developed a multi-class classification in head pose estimation based on a Deep Convolutional Neural Network (DCNN). They used an eight-layer DCNN.

We compared our CWGDD+PCA+LDA+NC with Cai's method using the CASPEAL head pose database. For our method, we used a subset containing 4,200 images of 200 subjects whose IDs range from 401 through 600, and performed 3-fold cross validation. They used the same 200 subjects for testing with using the remaining 840 subjects for training. They also enriched the variation of training dataset by shifting and scaling the training images, thereby generating 410,700 images for training. The 4,200 test images were divided into three subsets, each of which includes 1,400 images, and the results for these subsets are averaged to obtain the final accuracy.

Table 8 The comparison result against the Deep Convolutional Neural Network

Method	#Training samples	#Test samples	Classification Accuracy (%)
DCNN [23]	410,700	1,400	97.17
Ours	2,800	1,400	95.86

Table 8 shows the comparison result. DCNN outperforms ours with 1.31% higher accuracy. This result is, however, obtained by using about 147 times larger training data. It seems necessary to test the methods with more various conditions for more detailed comparison. This is one of our future work.

4.5 Online Experiment Using Videos

We have successfully built a simple yet robust head orientation descriptor that works very fast and is applicable to real applications. To test our method online, we utilized two captured videos of a real scene at our campus. We combine DS2, DS3 and DS6 as the training dataset. COG is utilized as a baseline method because it is the closest competitor to ours. In this experiment, we utilize Epsilon Support Vector Regression (epsilon-SVR) [24] with polynomial kernel as the estimator. The tolerance of termination criterion is set to 0.001, coef0 is set to 0, and degree in kernel function is set to 3. We select block size of 6 × 6 due to the best estimation results of pan orientation using SVMs as shown in Table 1.

The main objective of this experiment was to compare the feasibility of our descriptor and COG for estimating the head orientation in a video sequence in both indoor and outdoor environments. We roughly divide the pan into three general directions: right ($\alpha > +10^o$), upright frontal ($-10^o \leq \alpha \leq +10^o$) and left ($\alpha < -10^o$), while the tilt is also divided into three general directions: up ($\beta > +10^o$), upright frontal ($-10^o \leq \beta \leq +10^o$) and down ($\beta < -10^o$).

The first video is taken in a corridor where a targeted person is walking and following a moving camera with a distance of 1.5 - 2 meters. In this experiment, human upper body detection is used to localize the targeted person's body due to the reduced stability of our head detector for a distant person. Our system performs head detection, estimates the pan and the tilt angles at once, and tracks them to pro-

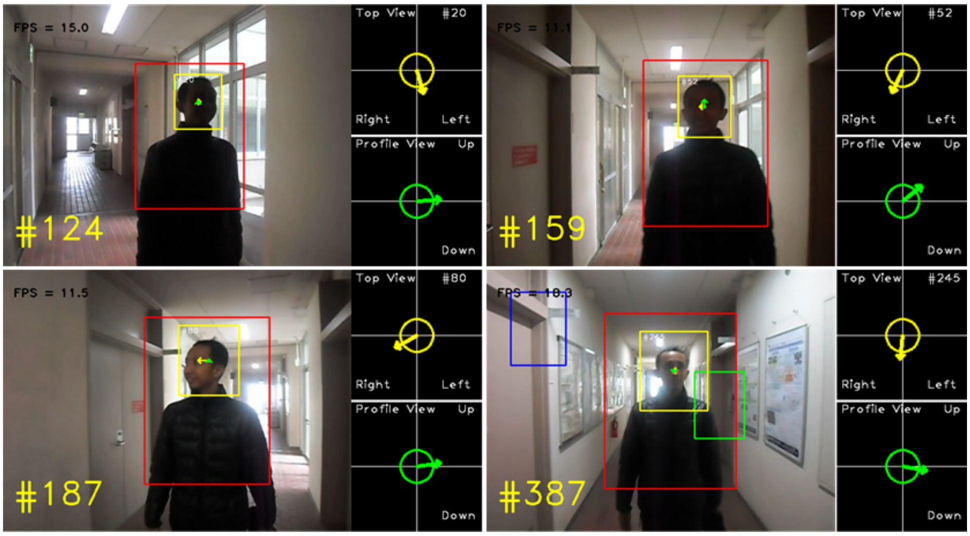


Fig. 14 Experimental results of an online indoor scene. First, the human upper body is detected and tracked (red bounding box). Head detection and head orientation estimation supported by tracking is performed within the body's bounding box. The estimated head orientation is shown by yellow (pan) and green (tilt) arrows within the bounding box. Miss-orientation estimation (#159) occurs due to a severe illumination, while false positive detections of the human upper body and miss-orientation estimation (#387) occur due to an ambient light and a less fit of the head tracking.

Table 9 Comparison result in the indoor experiment

Method	#Frame	#Detected	#Detected	Accuracy (%)	
		Body	Head	Pan	Tilt
COG	460	357	309	59.52	39.46
CWGDD+BIF-GA (Ours)	460	357	305	52.96	61.90

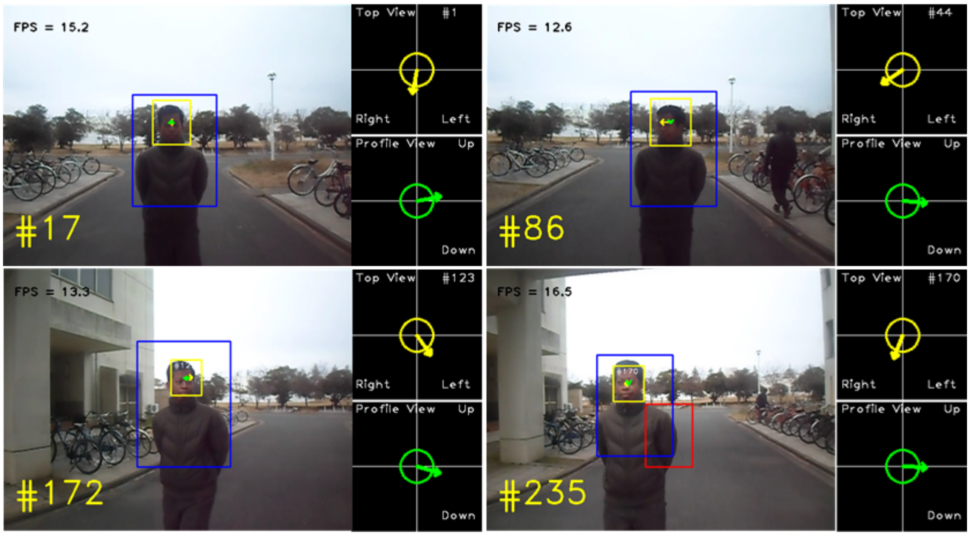


Fig. 15 Experimental results of an online outdoor scene. First, the human upper body is detected and tracked (blue bounding box). Head detection and head orientation estimation supported by tracking is performed within the body's bounding box. The estimated head orientation is shown by yellow (pan) and green (tilt) arrows within the bounding box. The human upper body detection, the head detection and the head orientation estimation achieve good performances in this experiment.

vide more robust estimation. Our system is able to estimate the target's head orientation in frames as shown in Fig. 14. However, the pan and the tilt errors sometimes occur during the experiment due to a severe illuminations (#159), an ambient lights and a less fit of the head tracking (#387).

We quantitatively compare the performance of our CWGDD+BIF-GA against COG by manually counting the number of frame, precision detected-body, precision

Table 10 Comparison result in the outdoor experiment

Method	#Frame	#Detected Body	#Detected Head	Accuracy (%)	
				Pan	Tilt
COG	345	273	260	68.56	62.33
CWGDD+BIF-GA (Ours)	345	270	255	74.33	78.43

detected-head, pan and tilt orientations accuracies as shown in Table 9.

The second video is taken outdoor where a targeted person is walking and following a moving camera with a distance of 2 - 2.5 meters. In this experiment, the performance of our head orientation estimation and the human upper body is better than it was in the first experiment as shown in Fig. 15. False body and head detection, and miss-orientation estimation are reduced. The comparison results of our CWGDD+BIF-GA and COG are shown in Table 10.

Based on the experimental results, the pan estimation accuracy of COG is slightly better than ours in the indoor experiment. However, COG fails to retain the accuracy for estimating the tilt. In general, our CWGDD+BIF-GA outperforms COG for indoor and outdoor experiments by achieving a better average in the accuracies.

The performance of our feature is stable because it utilizes a more variety of cues, that is, edge/shape (gradient), texture (Weber), and intensity patterns (deviation), while COG uses only gradient features. If a noise level of image is high, this affects the quality of gradient features, thereby degrading the performance of COG-based method. This is supported by our experiments that COG performs good for the off-line databases (CASPEAL and Pointing'04) but worse for AISL database and in real experiments.

4.6 Computation Time

Besides the accuracy, we also measured the averaged processing time of each method to complete the descriptor generation. The average time for completing the descriptor generation using COG, IA-LDQP, WLD, CovGa, CWGDD, and CWGDD+BIF-GA are around 0.69 ms, 32.44 ms, 0.98 ms, more than 96.79 ms, 0.79 ms, and 0.79 ms, respectively. The evaluation is conducted using Microsoft Visual C++ running on a personal computer system equipped with 3.60 GHz Intel processor i7 supported by 16 GB of RAM. Our weighting scheme does not burden our method because the optimization process has been undertaken in advance. Increasing the block size from 4 × 4 to 6 × 6 increases the computation time by about 0.15 ms.

Online experiments using indoor and outdoor videos show that our method is fast enough by achieving 11 - 16 fps. It implies that all processes such as the human upper body detection and tracking, the head detection, and the head orientation estimation only take about 90.5 - 62.5 ms in total.

5. Conclusion

We have presented our novel descriptor for estimating human head orientation. The combination of many features such as Weber, gradient and intensity deviation collectively was proven to be more effective to characterize the differences of each head orientation than just using a single feature. This combination significantly strengthens our descriptor to estimate the head orientation. A covariance successfully reduces the dimension of the descriptor, so that it can work very fast while still maintaining a strong discrimination ability. Our Genetic Algorithm-based optimization of the block important feature also significantly improved the performance of the head orientation estimation. Based on the experiments, our descriptor outperforms the other baseline methods by reaching a high accuracy and better stability than the other methods for almost all classifiers. A comparison with the Deep Convolutional Neural Network method exhibits our method is also comparable.

The experiments in a real scene environment show that our method is very promising and is applicable for an online application. However, some incorrect results while estimating head pan and tilt angles remain. Finding the best solution to solve this problem is a major concern for us and may provide a focus for our future works.

Acknowledgments

This work is in part supported by JSPS Kakenhi No. 25280093.

References

[1] E. Murphy-Chutorian and M.M. Trivedi, "Head Pose Estimation in Computer Vision: A Survey," IEEE Transaction on PAMI, vol.31, no.4, pp.607–626, 2009.

[2] B. Han, S. Lee, and H.S. Yang, "Head Pose Estimation Using Image Abstraction and Local Directional Quarternary Patterns For Multiclass Classification," Pattern Recognition Letters, vol.45, pp.145–153, 2014.

[3] B. Ma, A. Li, X. Chai, and S. Shan, "CovGa: A Novel Descriptor Based On Symmetry of Regions For Head Pose Estimation," Neurocomputing, vol.143, pp.97–108, 2014.

[4] L. Dong, L. Tao, and G. Xu, "Head Pose Estimation Using Covariance of Oriented Gradients," IEEE ICASSP, pp.1470–1473, 2010.

[5] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikäinen, X. Chen, and W. Gao, "WLD: A Robust Local Image Descriptor," IEEE Transactions of Pattern Analysis and Machine Intelligence, vol.32, no.9, pp.1705–1720, 2010.

[6] J.H. Holland, "Genetic Algorithms," Scientific American, vol.267, no.1, pp.66–72, 1992.

[7] N. Yusup, A.M. Zain, and S.Z.M. Hashim, "Evolutionary Techniques in optimizing machining parameters: Review and recent applications (2007–2011)," Expert Systems with Applications, vol.39, no.10, pp.9909–9927, 2012.

[8] G. George and K. Raimond, "A Survey on Optimization Algorithms for Optimizing the Numerical Functions," Internal Journal of Computer Applications, vol.61, no.6, pp.41–46, 2013.

[9] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol.60, no.2, pp.91–110, 2004.
[10] T. Ojala, M. Pietikäinen, and D. Harwood, "A Comparative Study of Texture Measures with Classification Based on Feature Distributions," Pattern Recognition, vol.29, no.1, pp.51–59, 1996.
[11] O. Tuzel, F. Porikli, and P. Meer, "Region Covariance: A Fast Descriptor for Detection and Classification," ECCV, vol.3952, pp.589–600, 2006.
[12] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Log-Euclidean Metrics for Fast and Simple Calculus on Diffusion Tensors," Magnetic Resonance in Medicine, Wiley-Liss Inc., vol.56, no.2, pp.411–421, 2006.
[13] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," CVPR, pp.I-511–I-518, 2001.
[14] N. Gourier, D. Hall, and J.L. Crowley, "Estimating Face Orientation From Robust Detection of Salient Facial Features," Proceedings of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures, Cambridge, UK, pp.617–622, 2004.
[15] C.E. Thomaz and G.A. Giraldi, "A New Rangking Method for Principal Components Analysis and its Application to Face Image Analysis," Image and Vision Computing, vol.28, no.6, pp.902–913, June 2010.
[16] B.S.B. Dewantara and J. Miura, "The AISL Head Orientation Database and Preliminary Evaluations," IEEE International Electronic Symposium, pp.140–144, 2015.
[17] I. Ardiyanto and J. Miura, "Partial Least Squares-based Human Upper Body Orientation Estimation with Combined Detection and Tracking," Image and Vision Computing, vol.32, no.11, pp.904–915, 2014.
[18] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Transactions on Intelligent Systems and Technology, vol.2, no.3, pp.1–27, 2011.
[19] www.bavuwe-software.com, "Decision Forest Parallel Implementation," 2012.
[20] J.C.J. Chen and J.J.J. Lien, "Multi-View Face Detection and Pose Estimation," 18th IPPR Conference on Computer Vision, Graphics and Image Processing, pp.933–940, 2005.
[21] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The CAS-PEAL Large Scale Chinese Face Database and Baseline Evaluations," IEEE Trans. On System, Man, and Cybernetics (Part A), vol.38, no.1, pp.149–161, 2008.
[22] H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," Journal of Educational Psychology, vol.24, no.6, pp.417–441, 1933.
[23] Y. Cai, M. Yang, and J. Li, "Multiclass classification based on a deep convolutional network for head pose estimation," Frontiers of Information Technology and Electronic Engineering, vol.16, no.11, pp.930–939, 2015.
[24] H. Drucker, C.J.C. Burges, L. Kaufman, A.J. Smola, and V.N. Vapnik, "Support Vector Regression Machines," in Advances in Neural Information Processing Systems 9, NIPS 1996, pp.155–161, MIT Press, 1997.

Bima Sena Bayu Dewantara received the B.Eng. degree in Information Technology from Electronic Engineering Polytechnic Institute of Surabaya, Indonesia, and M.S. degree in Electrical Engineering from Sepuluh Nopember Institute of Technology, Indonesia, in 2004 and 2010, respectively. He joined the Department of Electronic Engineering at Electronic Engineering Polytechnic Institute of Surabaya, Indonesia, as lecturer in 2005. Then, he moved to the Department of Informatics and Computer Science in 2007. He is currently a Ph.D candidate in Graduate School of Computer Science and Engineering at Toyohashi University of Technology, Japan. His research interests include pattern recognition, computer vision, machine learning and robotics system.

Jun Miura received the B.Eng. degree in mechanical engineering in 1984, the M.Eng. and the Dr.Eng. degrees in information engineering in 1986 and 1989, respectively, all from the University of Tokyo, Tokyo, Japan. In 1989, he joined Department of Computer-Controlled Mechanical Systems, Osaka University, Suita, Japan. Since April 2007, he has been a Professor at Department of Computer science and Engineering, Toyohashi University of Technology, Toyohashi, Japan. From March 1994 to February 1995, he was a Visiting Scientist at Computer Science Department, Carnegie Mellon University, Pittsburgh, PA. He received several awards including Best Paper Award from the Robotics Society of Japan in 1997, Best Paper Award Finalist at ICRA-1995, and Best Service Robotics Paper Award Finalist at ICRA-2013. Prof. Miura published over 180 papers in international journals and conferences in the areas of intelligent robotics, mobile service robots, robot vision, and artificial intelligence.