Face Pose Estimation From Video Sequence Using Dynamic Bayesian Network

Shahrel A. Suandi[1,2]
[1]School of Electrical & Electronic Engineering
Engineering Campus, Universiti Sains Malaysia
14300 Nibong Tebal, Pulau Pinang
Malaysia
shahrel@eng.usm.my

Shuichi Enokida[2] Toshiaki Ejima[2]
[2]Department of Artificial Intelligence
Kyushu Institute of Technology
Kawazu 680-4, Iizuka City
Fukuoka Prefecture, 820-8502 Japan
{enokida,toshi}@mickey.ai.kyutech.ac.jp

Abstract

This paper describes a technique to estimate human face pose from color video sequence using Dynamic Bayesian Network(DBN). As face and facial features trackers usually track eyes, pupils, mouth corners and skin region(face), our proposed method utilizes merely three of these features – pupils, mouth center and skin region – to compute the evidence for DBN inference. No additional image processing algorithm is required, thus, it is simple and operates in real-time. The evidence, which are called horizontal ratio and vertical ratio in this paper, are determined using model-based technique and designed significantly to simultaneously solve two problems in tracking task; scaling factor and noise influence. Results reveal that the proposed method can be realized in real-time on a 2.2GHz Celeron CPU machine with very satisfactory pose estimation results.

1. Introduction

Estimating human face pose has become one of the important tasks in human-computer interaction systems. It is used as an additional factor to create current computer vision systems more intelligent. Examples of applications include monitoring driver vigilance [9, 15], best video shot selection for personal identification [7, 18], human-computer interaction [1, 4] and gaze estimation [2, 11].

Methods to estimate face pose can be classified into *model-based* and *face property-based*(or *appearance-based*) [8]. In model-based approach, a three-dimensional$(3D)$ model of face is assumed and the face pose is recovered by establishing $3D$ to $2D$ transformation relationship. Here, geometrical relationships between object shape in real-space(3D), camera specifications like lens parameter in order to project the object onto image plane and image(2D) are taken into account. Projection variations like full perspective, weak perspective or mirror(orthogonal) projections has to be specified before hand in order to compute the desired pose correctly. Usually, the output from this approach will be given simultaneously, which are pitch, yaw and roll angles(pitch, yaw and roll are the angles made about x, y and z axes, respectively.). Works reported by Horprasert *et al*. [5], Qiang Ji *et al*. [8] and, Gee and Cipolla [2] are among predecessor researchers who employ model-based approach in their face pose estimation task.

On the other hand, in property-based approach, assumption that a unique causal-effect relationship between 3D face pose and certain facial image properties is made. Image properties may include image intensity, color, geometrical relations, histograms and frequency domains(spectral). The face pose is given after a process of recognition which requires a learning process to create the classifier or recognizer. For this process, a large number of corresponding image properties data is required. Some famous reported works are presented in Refs. [6, 14, 18] using Support Vector Machine(SVM), eigenspace, and Boosting, respectively. Even though this approach is simpler than model-based approach, it is difficult to recognize two small different side by side pose due to the difficulties of getting small different angle data. As such, this approach is not suitable for applications that demand precise face pose like mouse control by gaze direction.

Our work to estimate face pose is done based on property-based approach, where two important cues defined as horizontal ratio and vertical ratio are used as the properties(evidences) to infer the pose using Dynamic Bayesian Network(DBN). DBN is used due to its simplicity compared to Hidden Markov Model(HMM) and Kalman Filter Model(KFM) [13], and it also takes into account time-series information, which may reduce noise influence attributed from tracking error. Apart from this, the horizontal and

vertical ratios are determined using model-based approach. Two different models to compute each ratio are introduced in this paper. One model is built with respect to horizontal pose and the other is built with respect to vertical pose. Ratios are considered due to its robustness towards noise. Both of these ratios are computed from face region, pupils and mouth center, which are our feature of interests(FOIs). As most of face and facial features trackers provide these information while tracking, no additional image processing algorithm is required. We also introduce "head cylindrical model", a model that is made based on the results of analyzing anthropometric statistics of American adults [19]. This model is used to determine head center, a quantity that is important for horizontal ratio computation. Having such model, we also show that head motions like right-left rotation and right-left side motion can be distinguished. Results are given in discrete manners from nine face pose classes, *i.e.*, $\pm 45^\circ$ horizontal pose and $\pm 30^\circ$ vertical pose.

For the rest of this paper, we first briefly explain our tracking module in Section 2. Following this, we describe our face pose estimation technique in Section 3. Experimental results and discussions are presented in Section 4. Finally, we conclude this paper with Section 5.

2. Tracking Features of Interest

In this paper, we use EMoTracker introduced in [16] to track the FOIs – face region, pupils and mouth center. The method incorporated in EMoTracker is simple but shown to be robust to face type and pose. However, since mouth corners are not tracked, tasks to detect and track mouth corners are added into this particular work. Mouth corners tracking contributes the mouth center. Since mouth area is given (shown in green rectangle in Figure 1), mouth corners are detected by first, performing vertical and horizontal integral projection within detected mouth area to estimate the corners, and then, template matching is applied at the estimated corners to accurately search mouth corners. Once the mouth corners are found, center of mouth is determined by taking the center of both mouth corners. Although center of detected mouth area from EMoTracker can be used as the mouth center, it is not robust to noise compared to the one determined from mouth corners. Figure 1 shows the results after mouth corners and mouth center have been detected.

From the FOIs, horizontal ratio is computed using face region and pupils, while vertical ratio is computed using face region, pupils and mouth center. Finally, face pose estimation is performed utilizing DBN as the estimator. The overview of our proposed method is illustrated in Figure 2.

3. Face Pose Estimation

To estimate the face pose, we first determine two important cues that are used as the evidence for DBN inference.

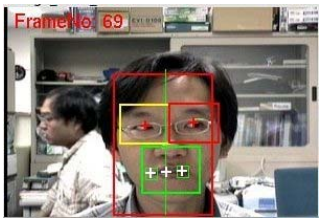

Figure 1. Results of tracking features of interest(FOIs).

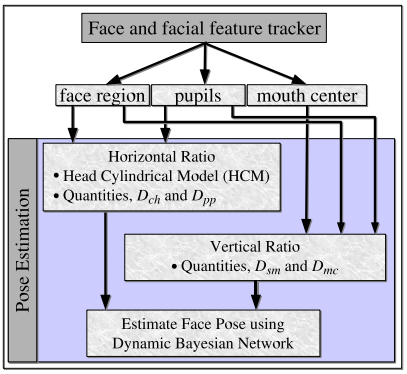

Figure 2. Overview of proposed method.

As being explained in preceding section, these cues are the horizontal ratio and vertical ratio. Although other items like facial features triangle's angles and ratio may be considered as well, it is shown empirically that these quantities are weak to noise.

3.1. Evidence For Face Pose Estimation

The positions of pupils and mouth center located from our tracking module provide useful cues for face pose estimation. Although the relative positions of each of these three points change as the head changes its pose, relying solely on this change in position to estimate pose makes the estimation highly dependent on the accuracy of the facial parts detection, which is not plausible for low resolution input. Instead, we also need to consider the position of each facial features with respect to a reference point whose position is invariant for any face pose. In this case, we have considered head center and bottom of skin region as the reference points(Figure 3) to compute horizontal and vertical ratios, respectively. However, due to large variations in human clothings, we use the initial vertical ratio as the reference while tracking. In the next section, we describe horizontal and vertical ratios in detail.

3.1.1 Horizontal Ratio, $\mathcal{H}$

Horizontal ratio, denoted with $\mathcal{H}$ in Eq. (1), is defined as the ratio of D_{ch} to D_{pp}. D_{ch} is horizontal distance from pupils center to head center and D_{pp} is horizontal distance between both pupils. It actually explains that a profile face

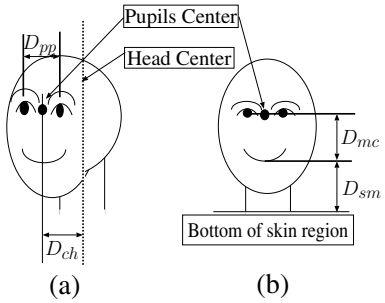

Figure 3. Quantities used to compute observable evidence, $\mathcal{H}$ and $\mathcal{V}$ for face pose estimation. (a) and (b) show the quantities to compute horizontal and vertical ratio, respectively.

will provide an infinite value of $\mathcal{H}$, whereas, a frontal face will provide value of 0.

$$\mathcal{H} = \frac{D_{ch}}{D_{pp}}\,. \tag{1}$$

Despite of being simple to be computed, it is impossible to determine D_{ch} directly from the tracking module due to head center is unknown from the tracker. On top of that, as all available information from the tracker are retrieved from a monocular camera, we do not have head depth information which might be useful to compute the head center. In our work, head center is defined as a center that works similarly like human head joint, *i.e.*, all head rotations are made about this joint. Therefore, it shall be invariant regardless what the face pose is. However, all available information from the tracker are variant. To solve this problem, we first set this head center limitations only for horizontal head rotation, then, we determine this quantity in each frame using a model known as "*head cylindrical model(HCM)*". As head is a non-rigid object, we manage HCM position by using non-linear least square regression method. This is done by observing two quantities; both pupils distance and distance from the pupil on the side face is facing to skin edge. A brief introduction to HCM is given below while the details on HCM including the proof for $\mathcal{H}$ can be found in [17]. In the same paper, we also show that horizontal human face pose can also be determined from the horizontal ratio that we have proposed.

Head Cylindrical Model(HCM) HCM contributes in providing a reliable method to compute head center. It carries two main properties;(1)invariant to head motions such as rotations and side-to-side motion – regardless what the face pose is, the head center shall remain at the same position with respect to face position;(2)invariant to scale – when the face moves near or far from the camera, the head center shall remain at the same position with respect to face size. With these two properties, head center is determined as follows:

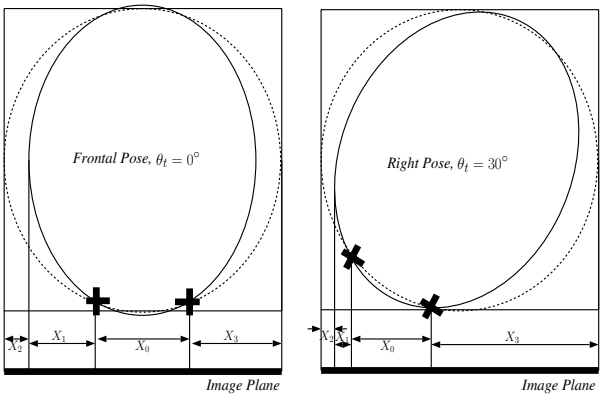

Figure 4. Examples of HCM (shown with dotted line) viewed from top in two different horizontal poses, $0°$ and right $30°$. The observable quantities are only X_0 and X_1. X_2 and X_3 are determined using nonlinear least square regression method. Notice that $X_0 = D_{pp}$.

1. As the only observable quantities are pupils and face region, we first compute these two quantities, X_0 and X_1. X_0 is the distance between both pupils on image plane, which equals to D_{pp}, while X_1 is the distance from the pupil (on which side the face is facing) to the edge of face region on the same side. X_2 and X_3, which are the distances from the face region edge to HCM edge, and from the other side's pupil to its side HCM edge, respectively, are determined indirectly from X_0 and X_1. This is depicted in Figure 4.

2. To handle scaling problems, next, we compute R_1, R_2 and R_3, where each of these is defined as follows: $R_1 = \frac{X_1}{X_0}$, $R_2 = \frac{X_2}{X_0}$ and $R_3 = \frac{X_3}{X_0}$.

3. Since only R_1 are determinable, we determine R_2 and R_3 by deriving each of them out using *nonlinear least square regression(NLLSR)* after R_1 is known.

4. To establish the relationship between R_1, R_2 and R_3, the observations made from mean model (a model made from anthropometrical statistics mean data [19]) which have been rotated to $15°, 30°$ and $45°$ are utilized. The relationships observed are between R_1 to R_2, and R_1 to R_3. These relationships are given in Eq. (2) and (3), respectively.

$$R_2 = 0.3346 - 0.0503R_1 - 0.0253R_1^{-1} \tag{2}$$

$$R_3 = 7.2353 - 25.4088R_1 + 33.7887R_1^2 - 14.6841R_1^3 \tag{3}$$

These equations show that when R_1 is known, then R_2 and R_3 may also be determined, which will consequently provide the values of X_2 and X_3 because $X_2 = R_2X_0$ and $X_3 = R_3X_0$.

5. When X_2 and X_3 are known, both edges of HCM are determined and finally, D_{ch} is computed.

Besides providing the head center, HCM is actually has the advantage to distinguish motions of rotations or side-to-side motion, that is, when changes in D_{ch} are observed, it means that rotation is happening, whereas, when changes in $x-$ or $y-$axes are observed while at the same time there is no changes in D_{ch}, then it is a side-to-side motion. Such functions might be very useful for monitoring one's vigilance systems, and we put this as our future work.

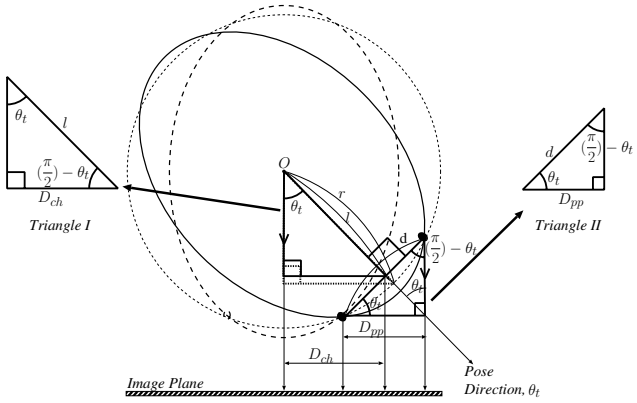

Figure 5. HCM (shown with dotted circle line) viewed from top. In the model are shown two triangles that can be drawn when the head moves horizontally to an arbitrary pose on the left. These triangles are utilized to compute $\mathcal{H}$.

Model used to define $\mathcal{H}$ Figure 5 shows the model used to derive $\mathcal{H}$. The bold dotted line shows the frontal pose, connected line shows an arbitrary left pose, and the normal dotted line shows the horizontal circular motion of head. At an arbitrary horizontal face pose denoted with θ_t, two right triangles can be drawn from the model. They are Triangle I and Triangle II. $\mathcal{H}$ is derived from these two triangles. From Triangle I and Triangle II, θ_t can be computed like shown in Eq. (4) and (5), respectively. However, due to tracking noise that might cause $D_{ch}(D_{pp})$ bigger than $d(l)$ in Eq. (5)(Eq. (4)), it is better to use the ratio which gives tangent in Eq. (6).

$$\sin\theta_t = \frac{D_{pp}}{l} = \frac{2D_{pp}}{d\sqrt{\gamma^2-1}}\,. \tag{4}$$

$$\cos\theta_t = \frac{D_{ch}}{d}\,. \tag{5}$$

From Eq. (4) and (5) and assuming $\gamma = 3.0$ [17], $\mathcal{H}$ becomes

$$\mathcal{H} = \frac{D_{ch}}{D_{pp}} = \sqrt{2}\tan\theta_t\,. \tag{6}$$

Therefore, horizontal human face pose can be computed from $\mathcal{H}$ using Eq. (7) [17].

$$\theta_t = \tan^{-1}\left(\frac{D_{ch}}{D_{pp}\sqrt{2}}\right). \tag{7}$$

3.1.2 Vertical Ratio, $\mathcal{V}$

Vertical ratio, denoted with $\mathcal{V}$ in Eq. (8) is defined as the ratio of D_{sm} to D_{mc}. D_{sm} is the vertical distance from bottom of skin region to mouth center, while D_{mc} is the vertical distance from mouth center to pupils center.

$$\mathcal{V} = \frac{D_{sm}}{D_{mc}}\,. \tag{8}$$

Unlike $\mathcal{H}$, $\mathcal{V}$ can be directly computed from the information given by the tracker. However, we have faced difficulties when using this cue directly as being defined in Eq. (8). The reason is that when the bottom of skin region to mouth center is too large due to user's clothings (while facing straight forward), for instance, $D_{sm} = 1.5D_{mc}$, then $\mathcal{V}$ becomes 1.5; suggesting an upper pose is made. As a result, high probabilities will be computed for upper pose 30° in the estimation module, which is not true. Therefore, for more intuitive approach, we set all $\mathcal{V}$ at initialization stage to be the reference, $\mathcal{V}_0$; and only consider the distinct amount from this value as the usable quantity to infer the face pose as defined in Eq. (9).

$$\mathcal{V} = \mathcal{V}_t - V_0\,, \qquad \text{where,} \tag{9}$$

$$\mathcal{V}_t = \frac{D_{smt}}{D_{mct}}\,, \tag{10}$$

$$\mathcal{V}_0 = \frac{D_{smt0}}{D_{mct0}}\,. \tag{11}$$

Here, $\mathcal{V}_t$ denotes the vertical ratio observed at time $t(t > 0)$, where its D_{smt} and D_{mct} are also computed at time t. $\mathcal{V}_0$ denotes the initial vertical ratio, where both its D_{smt0} and D_{mct0} are the values computed during initialization. By performing such calculation in Eq. (9), relationships as shown below are achieved:

$$\mathcal{V}\begin{cases} = 0 & , \text{ frontal pose} \\ > 0 & , \text{ upper pose} \\ < 0 & , \text{ lower pose} \end{cases}\,. \tag{12}$$

Model used to define $\mathcal{V}$ Model used to define $\mathcal{V}$ is shown in Figure 6. When view from the side, observable quantities are only D_{sm} and D_{mc}. R, R_0 and ψ_0 are constant values due to tracking is performed on the same person. At an arbitrary vertical pose, denoted by ψ_t, a, $D_{mc} = b$ and $D_{sm} = c$ can be expressed using the relations between Triangle I and Triangle II, like shown in Eq. (13), Eq. (14) and Eq. (15), respectively.

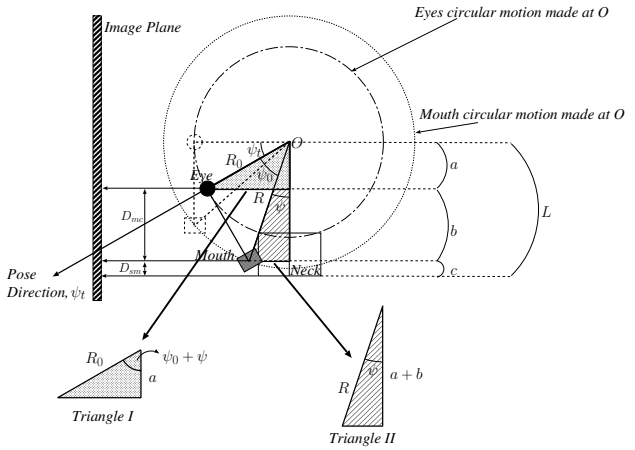

Figure 6. Head model constructed when viewed from side. Vertical pose ψ_t cannot be computed directly. Therefore, vertical ratio is considered.

$$a = R_0 \cos(\psi_0 + \psi) \tag{13}$$

$$D_{sm} = c = R(k - \cos\psi) \tag{14}$$

$$D_{mc} = b = R\{\cos\psi - \cos^2\psi_0 \cos\psi + \cos\psi_0 \sin\psi_0 \sin\psi\} \tag{15}$$

Here, k is a positive constant where when multiplied with R will give L, *i.e.*, $L = kR$. Accurate value of k is non-determinable due to the tracker only relies on skin region and we do not know the exact measurements for R and L, in which L is given as $L = a+b+c$. Therefore, we use ratio to represent vertical pose by taking into account the ratio of D_{sm} to D_{mc}. The ratio is defined in Eq. (16).

$$\mathcal{V} = \frac{y}{Y} = \frac{k - \cos\psi}{\sin\psi_0 \sin(\psi_0 + \psi)}\,. \tag{16}$$

3.2. Estimation by Dynamic Bayesian Network

We adopt Dynamic Bayesian Network(DBN) as the estimator to determine the face pose. A DBN comes from a Bayesian Network(BN) [3, 12], but in the form of dynamic [10], where its explicitly model changes over time. It considers time-slice in which at each time-slice there exists a BN in which parent of the current node at time-slice t is the node at time-slice $t-1$. In our estimation module, we design a DBN with one hidden node $\mathcal{Q}$ for pose and two observation nodes, $\mathcal{H}$ and $\mathcal{V}$, as shown in Figure 7. There are all together nine possible poses, which represent poses between $\pm 45^\circ$ with 15° notch for horizontal pose, poses between $\pm 30^\circ$ with 30° notch for vertical pose, and frontal pose. Conditional probabilities table(CPT), $P(\mathcal{Q}_t|\mathcal{Q}_{t-1})$ for each pose other than $\pm 45^\circ$ are designed with the assumption that a head pose is more likely to be at the same or close to the detected head pose at previous time-slice. CPT for the sensor model denoted as $P(h|\mathcal{Q})$ and $P(v|\mathcal{Q})$ are designed with the assumption that all distributions are Gaussian, $P(h|\mathcal{Q}) \sim \mathcal{N}(\mu_\mathcal{H}, \sigma_\mathcal{H})$ and $P(v|\mathcal{Q}) \sim \mathcal{N}(\mu_\mathcal{V}, \sigma_\mathcal{V})$. Values for $\mu_\mathcal{H}$ and $\mu_\mathcal{V}$ are derived from the ideal interpretation of $\mathcal{H}$ defined in Eq. (6).

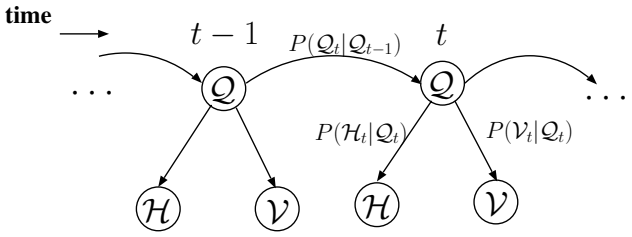

Figure 7. A DBN graph describing the transition and sensor model adopted in our work. Estimation is done at node $\mathcal{Q}_t$ with consideration of its parent, node $\mathcal{Q}_{t-1}$.

Our goal is to estimate the face pose by selecting the one with the highest posterior probability at each frame, given the temporal observations $h_{1:t}$ and $v_{1:t}$.

$$\hat{q}_t = \arg\max_{q_t} P(q_t|\, h_{1:t}, v_{1:t})\,, \tag{17}$$

where $\hat{q}_t$ denotes a pose given at time-slice t. Since the observations at a time-slice, *i.e.*, $\{h_t, v_t\}$, are only conditionally dependent to the face pose at that particular time-slice, the posterior probability of a face pose q_t is proportional to the product of the likelihood of q_t with respect to the each observation and the probability of the face pose given the posterior probabilities of face poses at the previous time-slice:

$$P(\mathcal{Q}_t|h_{1:t}, v_{1:t}) = \alpha P(h_t, v_t|\mathcal{Q}_t)\{\sum_{\mathcal{Q}_{t-1}} P(\mathcal{Q}_t|q_{t-1}) P(q_{t-1}|h_{1:t-1}, v_{1:t-1})\}. \tag{18}$$

In Eq. (18), the first term, $P(h_t, v_t|\mathcal{Q}_t)$ can be defined as $P(h_t|\mathcal{Q}_t)P(v_t|\mathcal{Q}_t)$ semantically, where $P(h_t|\mathcal{Q}_t)$ and $P(v_t|\mathcal{Q}_t)$ are directly obtainable from the sensor model [13].

4. Experimental Results

Three experiments to evaluate the proposed method have been carried out and they can be categorized into: the first one is to evaluate the validity of horizontal and vertical ratios as the cues to represent horizontal and vertical pose, respectively. The second experiment is to evaluate performance of the two ratios in a real tracking task. And finally, the third one is to make observation on the performance of our inference task using DBN with the two ratios as the observable evidences. All experiments are performed on a 2.2GHz Celeron CPU machine with LinuxOS 512MByte memory.

4.1. $\mathcal{H}$ and $\mathcal{V}$ Validity Check Experiments

In order to confirm the validity of $\mathcal{H}$ and $\mathcal{V}$ in estimating the face pose, two separate experiments have been performed. The purpose of the experiments is mainly to figure out if $\mathcal{H}$ and $\mathcal{V}$ can be used to represent horizontal and vertical poses, respectively, while simultaneously confirm that they can be used to infer the face pose. Using 150 male data from Softopia Japan for each of $0°$, $-15°$, $-30°$ and $-45°$, the positions of pupils, mouth corners and face region are recorded manually. For $+15°$, $+30°$ and $+45°$ data set, we make the data synthetically due to right pose images are always symmetrical to left pose images. In our work, "-" and "+" denote right and left pose, respectively. From the data that we have drawn manually, we compute the head center, D_{ch}, D_{pp}, D_{sm} and D_{mc} automatically using the algorithm described above and also compute θ_t using Eq. (7) from the first three information. The results are summarized in Table 1 and Table 2. These results are presented in terms of mean, variance and standard deviation.

Table 1. Results of evaluating $\mathcal{H}$ as an important cue to represent horizontal pose.

Data Class, θ_t	μ	σ^2	σ
$0°$	-0.60446	12.87483	3.58815
$15°$	15.13039	40.91540	6.39651
$30°$	30.56642	70.50935	8.39698
$45°$	38.09300	54.49600	7.38214
$-15°$	-16.72256	59.61362	7.72099
$-30°$	-32.25239	61.99329	7.87358
$-45°$	-38.29485	60.56723	7.78250

Table 2. Results of evaluating $\mathcal{V}$ as an important cue to represent vertical pose.

Data Class, ψ_t	μ	σ^2	σ
$30°$	0.20354	0.02032	0.14253
$-30°$	-0.25702	0.02886	0.16989

Analyzing the results from Table 1, the mean in each pose shows that $\mathcal{H}$ determined with our proposed method can be used to determine θ_t. In fact, while considering the standard deviation results, it promotes that using $\mathcal{H}$ in Eq. (7) is the best for frontal pose. But however, for other than frontal pose, the standard deviation values are bigger than the standard deviation given for frontal pose but as overall, they are smaller than 15, which ensures us that most of the results given are within the notch range($15°$). Hence, confirms that $\mathcal{H}$ can be used as the observation or evidence for face pose estimation and to compute θ_t as well.

On the other hand, analyzing the results from Table 2, it shows that upper and lower $30°$ pose may contribute ratios of approximately 0.2 and -0.2, respectively. This can also be observed from Figure 8 and 9. Thus, suggesting the validity of $\mathcal{V}$ in representing vertical pose.

4.2. $\mathcal{H}$ and $\mathcal{V}$ Performance in Tracking Task

While both cues are usable to represent horizontal and vertical pose, we further analyze these two cues performance in an actual tracking task. For comparison purpose, we have taken manually the position of face region, pupils and mouth corners. This information are taken from two different subjects in two video sequences which consist of about 400 frames each. These subjects were asked to start with a frontal pose and after a while, rotate their faces horizontally about one or two cycles(for example, left-right-left-right) followed by vertical motion about the same cycles. Once this is finished, we repeat the recording of these features again but this time it is done automatically using the tracker. While running, this tracker will simultaneously compute $\mathcal{H}$ and $\mathcal{V}$. From the truth data that have been taken manually, $\mathcal{H}$ and $\mathcal{V}$ are computed as well. The results for each subject are shown in Figure 8 and 9, respectively. Analyzing results for both subjects, we have observed that there is not much different between the ground truth and experiment data. This ensures us that ratios computed by our tracker are also reliable to compute $\mathcal{H}$ and $\mathcal{V}$.

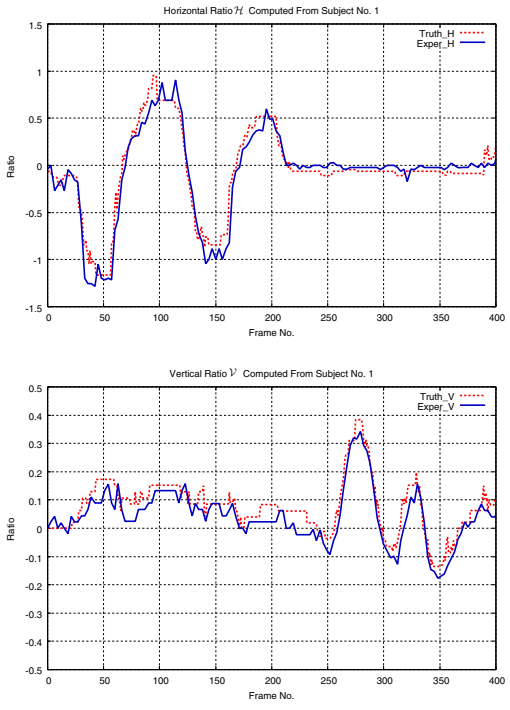

Figure 8. Horizontal ratio(upper graph) and vertical ratio(lower graph) plotted from subject 1. The terms 'Truth' and 'Exper' refer to the manually and automatically data set, respectively.

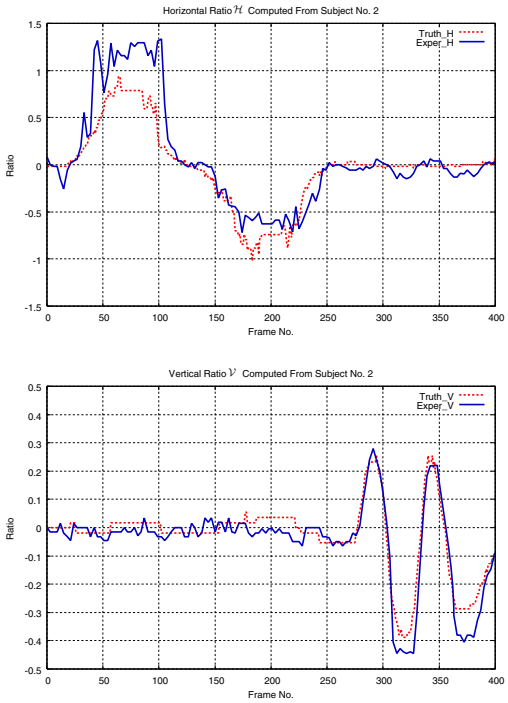

Figure 9. Horizontal ratio(upper graph) and vertical ratio(lower graph) plotted from subject 2. The terms 'Truth' and 'Exper' refer to the manually and automatically data set, respectively.

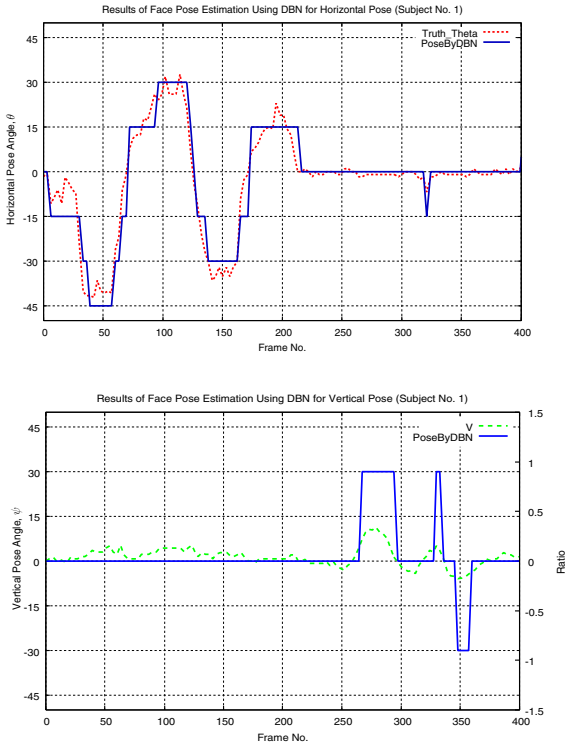

Figure 10. Face pose estimation results using DBN for subject one. The upper and lower graphs show the results for horizontal and vertical pose, respectively.

4.3. Face Pose Estimation Evaluation

To measure the performance of proposed method, we also execute the inference task while the tracker tracks and records the corresponding facial features. The results are given in discrete manner, where the highest probabilities given by an arbitrary pose will be taken as the pose. However, since there exist two types of 30° results; one represents the horizontal while the other represents the vertical pose; we separate the results into two pose types, horizontal and vertical pose. Graphs shown in Figure 10 and 11 are the observations resulted from subject one and two, respectively. In the evaluation method, as θ_t can be computed from $\mathcal{H}$ using Eq. (7), we use this value as comparison to the results given by DBN for horizontal pose. To evaluate the results given by DBN for vertical pose, a comparison with $\mathcal{V}$ curves computed along the tracking process is used.

Analyzing both results, it can be concluded that face pose estimation using DBN can be performed by merely using two evidences, $\mathcal{H}$ and $\mathcal{V}$. Results from subject one show better results compared to the other. This is due to false-positive detection of facial features and HCM high sensitivity at poses greater than 30° and less than -30°. The other reason is that, the sensitivity of each pose class relies on the Gaussian distributions that have been designed for each respective class. Although using DBN has actually solved the problems, in special cases where false-positive detection is inevitable, such results may be observed.

5. Conclusion

A novel method to estimate face pose which relies under the framework of probabilistic reasoning over time is described in this paper. In general, the task to estimate the face pose is analogous to solving the uncertainty problem of facial features positions relative to face region. The estimation task takes into account two observations as the evidence; horizontal and vertical ratios which are defined as $\mathcal{H}$ and $\mathcal{V}$, respectively. To compute $\mathcal{H}$, we need additional information which is not computable directly from the image due to limited information gained in a $2D$ image. To solve this problem, studies and analysis of anthropometrics statistics data have been carried out, in where a model known as head cylindrical model is designed. Through experiments, it has been shown that $\mathcal{H}$ and $\mathcal{V}$ can be used to represent horizontal and vertical pose; can be computed automatically; and appropriate to be used as the evidence. As overall, it can be concluded that face pose estimation from a monocular camera can be performed by only having in hand information like face region, pupils and mouth center. On top of that, since all proposed parameters are computational cheap, a real-time face pose estimation system can be realized. On the other hand, we have also faced problems

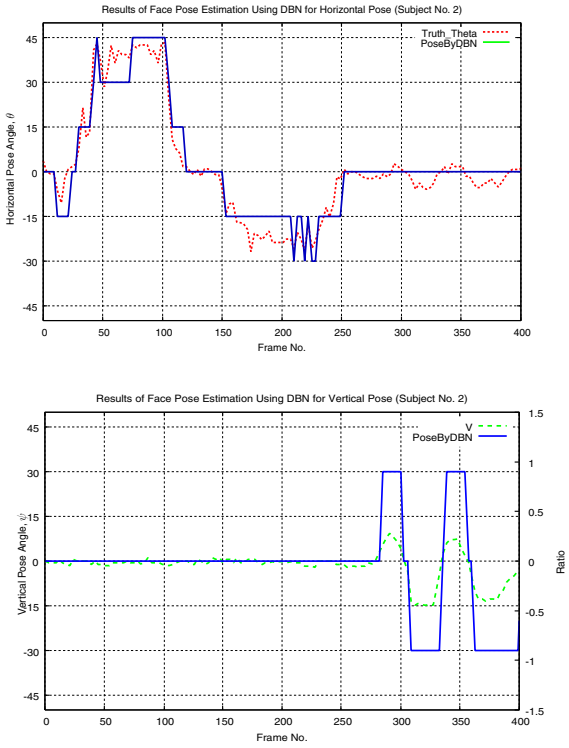

Figure 11. Face pose estimation results using DBN for subject two. The upper and lower graphs show the results for horizontal and vertical pose, respectively.

in our work, for instance, the sensitivity of head cylindrical model at poses smaller than -30° and bigger than $+30^\circ$ which affect head center computation, and facial features tracking accuracy. We will look into these problems in the nearest future. As a long term plan, we would like to integrate this system with personal identification system that has been developed at our laboratory.

Acknowledgment

This work is partially supported by Fundamental Research Grant Scheme (FRGS) from Ministry of Higher Education Malaysia under project code no. 203/PELECT/6071142.

References

[1] J. W. Davis and S. Vaks. A perceptual user interface for recognizing head gesture acknowledgements. In *ACM Person User Interface (PUI)*, 2001. 1

[2] A. Gee and R. Cipolla. Determining the gaze of faces in images. *Image and Vision Computing*, 12(10):639–647, December 1994. 1

[3] D. Heckerman. A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division, Microsoft Corporation, November 1996. 5

[4] J. Heinzmann and A. Zelinsky. Robust real-time face tracking and gesture recognition. In *International Joint Conference on Artificial Intelligence, IJCAI'97*, volume 2, pages 1525–1530, 1997. 1

[5] T. Horprasert, Y. Yacoob, and L. S. Davis. Computing 3-d head orientation from a monocular image sequence. In *IEEE International Conference on Automatic Face and Gesture Recognition (FGR'96)*, pages 242–247, 14–16 October 1996. 1

[6] J. Huand, X. Shao, and H. Wechsler. Face pose discrimination using support vector machines (svm). In *Proceedings of International Conference on Pattern Recognition*, 1998. 1

[7] K. S. Huang and M. M. Trivedi. Robust real-time detection, tracking and pose estimation of faces in video streams. In *IEEE International Conference on Pattern Recognition (ICPR'04)*, volume 3, pages 965–968, August 2004. 1

[8] Q. Ji and R. Hu. 3d face pose estimation and tracking from a monocular camera. *Image and Vision Computing*, 20(7):499–511, May 2002. 1

[9] Q. Ji and X. Yang. Real-time eye, gaze and face pose tracking for monitoring driver vigilance. *Real-Time Imaging*, 8(5):357–377, October 2002. 1

[10] K. P. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, Graduate Division, University of California, Berkeley, 2002. 5

[11] K. R. Park, J. J. Lee, and J. Kim. Gaze position detection by computing the three dimensional facial positions and motions. *Pattern Recognition*, 35(11):2559–2569, November 2002. 1

[12] I. Rish. Advances in bayesian learning. In *Proceedings International Conference on Artificial Intelligence*, 2000. 5

[13] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*, chapter Probabilistic Reasoning Over Time. Prentice Hall, 2 edition, Dec 2002. 1, 5

[14] H. Saito, A. Watanabe, and S. Ozawa. Face pose estimating system based on eigen space analysis. In *Proceedings of IEEE Internation Conferene on Image Processing*, volume 1, pages 638–642, 1999. 1

[15] P. Smith, M. Shah, and N. da Vitoria Lobo. Monitoring head/eye motion for driver alertness with one camera. In *IEEE International Conference on Pattern Recognition (ICPR'00)*, pages 4636–4642, September 3-8 2000. 1

[16] S. A. Suandi, S. Enokida, and T. Ejima. Emotracker: Eyes and mouth tracker based on energy minimizaton criterion. In *4th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'04)*, pages 269–274. IAPR, December 2004. 2

[17] S. A. Suandi, T. S. Tai, S. Enokida, and T. Ejima. Horizontal human face pose determination using pupils and skin region positions. In *Lecture Notes on Computer Science*. Springer-Verlag, December 2007. Accepted. To be published. 3, 4

[18] Z. Yang, H. Ai, B. Wu, S. Lao, and L. Cai. Face pose estimation and its application in video shot selection. In *IEEE International Conference on Pattern Recognition (ICPR'04)*, volume 1, pages 322–325, August 2004. 1

[19] J. W. Young. Head and face anthropometry of adult u.s. citizens. Technical Report R0221201, Beta Research Inc., July 1993. 2, 3