Anticipatory Matching Method for Query-Based Head Gesture Identification

Hidetoshi Nonaka and Masahito Kurihara

Graduate School of Engineering, Hokkaido University
N13W8, Kita-ku, Sapporo 060 8628, Japan, +81 11 706 6793
{nonaka | kurihara}@main.eng.hokudai.ac.jp

Abstract
This paper presents a matching method for head gesture identification in a query-based interface. Head gesture identification has been achieved by various methods; DP matching; hidden Markov model, and so on. The aims of these conventional methods are dealing with general gestures at arbitrary moment. On the contrary, in our purpose, the timing of a gesture or what gesture will be made are predictable to some extent. The main purpose of this research is to realize an interface under variety of conditions, in flexible situations every time, and everywhere, for example, during walking around. In our system, there are two levels of anticipation features. The one is modifying the reference patterns according to the context of the query and cognitive difficulty of the question before making query. And the other is incursive refinement of reference patterns during matching process with allowance of future revision.
Keywords: Head Gesture, DP Matching, Query, Anticipation

1. Introduction

We have developed a human computer interaction system using eye-gaze and head gesture [2]. The initial purpose of this research is affording a communication interface of computer systems to users with disabilities who cannot use their hands to operate a keyboard nor a mouse. In the system, head gesture is utilized for not only discriminating attentive gaze from inconscient seeing, and also compensating the gaze point suffered from fluctuation of head. Through this previous research, we have found that the head gesture is effective in wide applications: wearable computers; hands free input devices; ubiquitous computing; augmented reality; human-agent interaction systems; and so on. In this paper, we consider head-based query system based on yes/no questions without limitation to eye-controlled communication interfaces.

Head gesture identification has been achieved by various methods, which include DP matching; hidden Markov model; or their improved methods. The aims of these conventional methods are dealing with general gestures at arbitrary moment. And many refinements of these methods for spotting recognition have been proposed. On the other hand, in our purpose, the timing of a gesture is predictable to some extent. Moreover, even the fluctuation of time, variance of amplitude, latency period until the start of gesture, and which gesture is made are also predictable. It is because they depend on the context of the questions, the individual difference of physical response time, and

cognitive difficulty of query, which can be anticipated in advance of making query.

In our system, there are two levels of anticipation features. The one is modifying the reference patterns according to the context of the query and cognitive difficulty of the question before making query. And the other is incursive refinement of reference patterns during matching process with allowance of future revision. Identification of gesture is completed as the final condition of this refinement.

2. Hardware Configuration

The block diagram of hardware configuration is illustrated in Fig.1. *Pitch* of head movement is measured by accelerometer (ADXL202E, AD, 2000), and *yaw* of head movement is measured by gyroscope (ADXRS300, AD, 2002). These motion sensors are attached to the left frame of glasses. Head tracking unit is configured in a micro controller (PIC16F877 (QTFP)), Microchip, 2001), which is attached behind the head. The outputs (x-y) of accelerometer are duty cycles that are proportional to acceleration in a horizontal plane. It involves both the acceleration on a horizontal plane and that of gravity. The output of gyroscope is a voltage proportional to angular rate about the axis normal to the sensor. Head tracking data (about 128Hz) are sent to PC peripheral device by wireless communication: RF transmitter (AM-RT5, RF Solutions, 2000) and RF receiver (AM-HRR3, RF Solutions, 2000). The calibration of head tracking unit is initially achieved by in-circuit serial programming in wired condition. By these selections of devices, remarkable downsizing is realized. On the contrary, the data include various noises from posture and inclination of user's body, from spontaneous motion of user, and so on. Especially the data from accelerometer suffers from noises of steps while user is walking. We cope with them using time-frequency based decomposition discussed in Section 4.

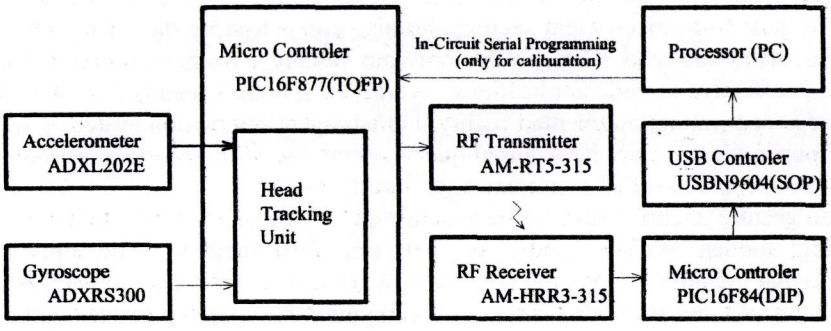

Fig. 1 Block diagram of hardware configuration.

We assigned nodding head to "yes", and shaking head to "no". The former is detected mainly by accelerometer, while the latter is mainly detected by gyroscope.

3. Estimation of User's Response Time

The response time of user can be estimated with cognitive difficulty of the question precedent to making query. The information processing of the human can be described as the integration of separate processors: perceptual processor; cognitive processor; and motor processor, as an approximation (Card et al. [3]).

Following the above reference, in the perceptual processor, the cycle time τ_P can be estimated on the order of

$$\tau_P = 100[50\sim200] \text{ ms},$$

with the notation τ_{typ} [$\tau_{low} \sim \tau_{upp}$], which means typical value τ_{typ} is 100 ms, while lower and upper bounds τ_{low} and τ_{upp} are 50 ms and 200 ms respectively. In the motor processor the cycle time τ_M is

$$\tau_M = 70[30\sim100] \text{ ms},$$

and in the cognitive processor, τ_C is

$$\tau_C = 70[25\sim170] \text{ ms}.$$

The cycle time of cognitive processor is further classified, for example, matching word against working memory: 47[36~52] ms, perceptual judgment: 92 ms, choice reaction: 153 ms, and so on.

Using these typical values of cycle time, we can estimate the reaction time for various cognitive tasks:

- Simple reaction time: $T = \tau_P + \tau_C + \tau_M = 240$ [105 ~ 470] ms
- Physical matches: $T = \tau_P + 2\tau_C + \tau_M = 310$ [130 ~ 640] ms
- Name matches: $T = \tau_P + 3\tau_C + \tau_M = 380$ [155 ~ 810] ms
- Category matches: $T = \tau_P + 4\tau_C + \tau_M = 450$ [180 ~ 980] ms

In our system, choice reaction time is always added as a cognitive cycle, because it is a head-based query system based on yes/no questions. Moreover, decision time increases with uncertainty about the judgment. It is known as Hick's law: $T = I_c H$, where H is the entropy of the decision and I_c is a constant. I_c depends on individual difference, and H depends on the context of queries. We use T_{low} and T_{upp} for estimation of starting point of gesture, which is mentioned in Subsection 4.4.

4. Matching Methods for Gesture Identification

4. 1 Successive DP Matching Method

We proposed a Successive DP matching method modifying DP matching for our purpose [2]. A brief description is as follows.

Let

$$\begin{aligned} X = &\quad \cdots, x_i, \cdots, x_1, x_0 \\ Y = &\quad \cdots, y_j, \cdots, y_1, y_0 \end{aligned} \tag{1}$$

be two discrete time series. The time-normalized distance D between these two time series is defined as

$$D(X,Y) = \min_{\substack{i=i(k) \\ j=j(k)}} \left[\sum_{k=0}^{-K} d(i,j) \right] \quad , \tag{2}$$

where $K>0$ is a constant and $d(i, j)$ is the distance between x_i and y_j. Warping function $i(k)$ and $j(k)$ are subject to following conditions

$$\begin{aligned} & i(k-1) \le i(k) \\ & j(k-1) \le j(k) \\ & i(k-1) + j(k-1) < i(k) + j(k) \\ & |\, i(k) - j(k) \,| < r \quad , \end{aligned} \tag{3}$$

where $r > 0$ is a constant .

The minimization problem (2), (3) can be solved by dynamic programming:

$$\begin{aligned} & D_0 = d(0,0) \\ & D_{k-1} = d(i(k), j(k)) + D_k \quad k = 0, -1, -2, \cdots, -K+1 \\ & D = D_{-K} \quad , \end{aligned} \tag{4}$$

where the pair $(i(k), j(k))$ is chosen from $\{(i(k)\text{-}1, j(k)), (i(k), j(k)\text{-}1), (i(k)\text{-}1, j(k)\text{-}1)\}$ such as to minimize $d(i(k), j(k))$ under the condition of $|i(k) - j(k)| < r$.

In the previous study [2], this method successfully coped with the identification of head gesture supposedly because it included eye-head cooperation. But in the current condition, there are more remarkable noises and artifacts, and it is difficult to dispense with additional strategies such as time-frequency analysis. The rest of this Section is devoted to our improvement using the notion of wavelet analysis and anticipation methodology.

4.2 Time-Frequency Based Decomposition: Forward Process

For a time series $v_{0,t}$, $t=0,\cdots,2^J-1\equiv n-1$, we define

$$\begin{aligned}
&w_{1,t}=\tfrac{1}{\sqrt{2}}(v_{0,2t+1}-v_{0,2t}),\quad v_{1,t}=\tfrac{1}{\sqrt{2}}(v_{0,2t+1}+v_{0,2t}),\quad t=0,\cdots,2^{J-1}-1\\
&w_{2,t}=\tfrac{1}{\sqrt{2}}(v_{1,2t+1}-v_{1,2t}),\quad v_{2,t}=\tfrac{1}{\sqrt{2}}(v_{1,2t+1}+v_{1,2t}),\quad t=0,\cdots,2^{J-2}-1\\
&\quad\vdots\\
&w_{j,t}=\tfrac{1}{\sqrt{2}}(v_{j-1,2t+1}-v_{j-1,2t}),\quad v_{j,t}=\tfrac{1}{\sqrt{2}}(v_{j-1,2t+1}-v_{j-1,2t}),\quad t=0,\cdots,2^{J-j}-1\\
&j=1,2,\cdots,J
\end{aligned}\tag{5}$$

We can rewrite (5) with $(2^{j-2}-1)\times(2^{j-1}-1)$ matrices:

$$\begin{aligned}
&W_1=H_0V_0,\quad V_1=G_0V_0\\
&W_2=H_1V_1,\quad V_2=G_1V_1\\
&\qqu\vdots
\end{aligned}\tag{6}$$

where $V_0=[v_{0,0},v_{0,1},\cdots,v_{0,2^{j-1}-1}]^T$, $V_1=[v_{1,0},v_{1,1},\cdots,v_{1,2^{j-2}-1}]^T$, $W_1=[w_{1,0},w_{1,1},\cdots,w_{1,2^{j-2}-1}]^T$ $W_2=[w_{1,0},w_{1,1},\cdots,w_{1,2^{j-3}-1}]^T$, $V_2=[v_{1,0},v_{1,1},\cdots,v_{1,2^{j-3}-1}]^T$, and

$$H_j=\frac{1}{\sqrt{2}}\begin{bmatrix}-1 & 1 & & & & \\ & & -1 & 1 & & \\ & & & & \ddots & \\ & & & & & -1\ \ 1\end{bmatrix},\quad G_j=\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1 & & & & \\ & & 1 & 1 & & \\ & & & & \ddots & \\ & & & & & 1\ \ 1\end{bmatrix},$$

which are $(2^{J-j}-1)\times(2^{J-j+1}-1)$ matrices. Hence, we have

$$\begin{bmatrix}W_1\\W_2\\\vdots\\W_J\\V_J\end{bmatrix}=\begin{bmatrix}H_0\\H_1G_0\\\vdots\\H_JG_{J-1}\cdots G_1G_0\\G_JG_{J-1}\cdots G_1G_0\end{bmatrix}V_0\ .\tag{7}$$

The transform $W_1,W_2,\cdots,W_J$ and V_J from V_0 is equivalent to *discrete Haar wavelet transform* (HWT). In our system J is set to be 7, then the matrix in (7) is 128×128. The computation requires $O(n)$ multiplications.

4.3 Time-Frequency Based Decomposition: Backward Correction Process

For a time series $\tilde{v}_{0,t}$, $t=0,\cdots,2^J-1\equiv n-1$, we define

$$\begin{aligned}
&\tilde{w}_{1,t}=\tfrac{1}{2}(\tilde{v}_{0,t+1(\mathrm{mod}\,n)}-\tilde{v}_{0,t}),\quad \tilde{v}_{1,t}=\tfrac{1}{2}(\tilde{v}_{0,t+1(\mathrm{mod}\,n)}+\tilde{v}_{0,t})\\
&\tilde{w}_{2,t}=\tfrac{1}{2\sqrt{2}}(\tilde{v}_{1,t+3(\mathrm{mod}\,n)}-\tilde{v}_{1,t}),\quad \tilde{v}_{2,t}=\tfrac{1}{2\sqrt{2}}(\tilde{v}_{1,t+3(\mathrm{mod}\,n)}+\tilde{v}_{1,t})\\
&\quad\vdots\\
&\tilde{w}_{j,t}=\tfrac{1}{2^{j/2}\sqrt{2}}(\tilde{v}_{j-1,t+2^j-1(\mathrm{mod}\,n)}-\tilde{v}_{j-1,t}),\quad \tilde{v}_{j,t}=\tfrac{1}{2^{j/2}\sqrt{2}}(\tilde{v}_{j-1,t+2^j-1(\mathrm{mod}\,n)}+\tilde{v}_{j-1,t})\\
&j=1,2,\cdots,J
\end{aligned}\tag{8}$$

We can rewrite (5) in matrices forms:

$$\begin{aligned}
&\tilde{W}_1=\tilde{H}_0\tilde{V}_0,\ \tilde{V}_1=\tilde{G}_0\tilde{V}_0\\
&\tilde{W}_2=\tilde{H}_1\tilde{V}_1,\ \tilde{V}_2=\tilde{G}_1\tilde{V}_1\quad ,\\
&\qquad\vdots
\end{aligned}\tag{9}$$

where $\tilde{W}_j\equiv[\tilde{w}_{j,0},\tilde{w}_{j,1},\cdots,\tilde{w}_{j,n-1}]^T$, $\tilde{V}_j\equiv[\tilde{v}_{j,0},\tilde{v}_{j,1},\cdots,\tilde{v}_{j,n-1}]^T$ and

$$\tilde{H}_0\equiv\frac{1}{2}\begin{bmatrix}-1&1&0&0&\cdots&0&0&0\\0&-1&1&0&\cdots&0&0&0\\0&0&-1&1&\cdots&0&0&0\\\vdots&\vdots&\vdots&\vdots&&\vdots&\vdots&\vdots\\\vdots&\vdots&\vdots&\vdots&&\vdots&\vdots&\vdots\\0&0&0&0&\cdots&-1&1&0\\0&0&0&0&\cdots&0&-1&1\\1&0&0&0&\cdots&0&0&-1\end{bmatrix}\begin{bmatrix}\tilde{v}_{0,0}\\\tilde{v}_{0,1}\\\tilde{v}_{0,2}\\\vdots\\\vdots\\\tilde{v}_{0,n-3}\\\tilde{v}_{0,n-2}\\\tilde{v}_{0,n-1}\end{bmatrix},\quad \tilde{G}_0\equiv\frac{1}{2}\begin{bmatrix}1&1&0&0&\cdots&0&0&0\\0&1&1&0&\cdots&0&0&0\\0&0&1&1&\cdots&0&0&0\\\vdots&\vdots&\vdots&\vdots&&\vdots&\vdots&\vdots\\\vdots&\vdots&\vdots&\vdots&&\vdots&\vdots&\vdots\\0&0&0&0&\cdots&1&1&0\\0&0&0&0&\cdots&0&1&1\\1&0&0&0&\cdots&0&0&1\end{bmatrix}\begin{bmatrix}\tilde{v}_{0,0}\\\tilde{v}_{0,1}\\\tilde{v}_{0,2}\\\vdots\\\vdots\\\tilde{v}_{0,n-3}\\\tilde{v}_{0,n-2}\\\tilde{v}_{0,n-1}\end{bmatrix}$$

$$\tilde{H}_1\equiv\frac{1}{2\sqrt{2}}\begin{bmatrix}-1&0&1&0&\cdots&0&0&0\\0&-1&0&1&\cdots&0&0&0\\0&0&-1&0&\cdots&0&0&0\\\vdots&\vdots&\vdots&\vdots&&\vdots&\vdots&\vdots\\\vdots&\vdots&\vdots&\vdots&&\vdots&\vdots&\vdots\\0&0&0&0&\cdots&-1&0&1\\1&0&0&0&\cdots&0&-1&0\\0&1&0&0&\cdots&0&0&-1\end{bmatrix}\begin{bmatrix}\tilde{v}_{1,0}\\\tilde{v}_{1,1}\\\tilde{v}_{1,2}\\\vdots\\\vdots\\\tilde{v}_{1,n-3}\\\tilde{v}_{1,n-2}\\\tilde{v}_{1,n-1}\end{bmatrix},\quad \tilde{G}_1\equiv\frac{1}{2\sqrt{2}}\begin{bmatrix}1&0&1&0&\cdots&0&0&0\\0&1&0&1&\cdots&0&0&0\\0&0&1&0&\cdots&0&0&0\\\vdots&\vdots&\vdots&\vdots&&\vdots&\vdots&\vdots\\\vdots&\vdots&\vdots&\vdots&&\vdots&\vdots&\vdots\\0&0&0&0&\cdots&1&0&1\\1&0&0&0&\cdots&0&1&0\\0&1&0&0&\cdots&0&0&1\end{bmatrix}\begin{bmatrix}\tilde{v}_{1,0}\\\tilde{v}_{1,1}\\\tilde{v}_{1,2}\\\vdots\\\vdots\\\tilde{v}_{1,n-3}\\\tilde{v}_{1,n-2}\\\tilde{v}_{1,n-1}\end{bmatrix}$$

$\cdots$.

The transform $\tilde{W}_1, \tilde{W}_2, \cdots, \tilde{W}_J$ and $\tilde{V}_J$ from $\tilde{V}_0$ is equivalent to *maximal overlap discrete Haar wavelet transform* (MOHWT). It is known that energy is decomposed properly as

$$\| \tilde{V}_0 \|^2 = \sum_{j=1}^{J} \| \tilde{W}_j \|^2 + \| \tilde{V}_J \|^2 \ , \tag{10}$$

where $\| \tilde{V}_0 \| \equiv \sum_{t=0}^{n-1} \tilde{V}_t^2 \cdots$.

Furthermore, V_0 can be reconstructed and partially reconstructed using inverse transformation, though it is not orthonormal but only orthogonal:

$$\tilde{V}_0 = \sum_{j=1}^{J} \tilde{H}_j{}^T \tilde{W}_j + \tilde{G}_J{}^T \tilde{V}_J \quad \text{(fully)} \tag{11}$$

$$\bar{V}_0 = \sum_{\{j\} \subset [1, J]} \tilde{H}_j{}^T \tilde{W}_j + \tilde{G}_j{}^T \tilde{V}_j \quad \text{(partially).} \tag{12}$$

The definition of wavelets and scaling functions are different from those of conventional MOHWT. It is only for the convenience of successive computation of transform in real time. The computation of transform and inverse transform require $O(n \log_2 n)$ multiplications.

4.4 Anticipatory Matching Method

The matching process is summarized as follows:

```
Boolean GestureDetection (void)
{
  Estimate T_low and T_upp;
  t = t - T_low ;  //Set the time origin to T_low .
  while (T_low < t < T_upp + 128 ){
     Calculate W_j and V_j successively by (5);
     if (W_j and V_j accord those of reference pattern){
        Calculate W̃_j and Ṽ_j by (8);
        Reconstruct V̄_0 by (12);
          //with selecting the subset{j}according to ref. pattern.
        Compare V̄_0 with reference pattern using S-DP matching;
        if (matched) return true;
     }
     return false;
  }
}
```

Fig. 2 C-like pseudo code of anticipatory matching method.

The estimations of T_{low} and T_{upp} are following the way described in Section 3. But actual cognitive process is more sophisticated, then the value is adjusted beforehand and empirically for each finite number of yes/no question.

The condition of the while-clause, $T_{low} < t < T_{upp} + 128$ means that the start point of gesture is supposed to be between T_{low} and T_{upp}. The process in the content of while-clause is illustrated in Fig. 3.

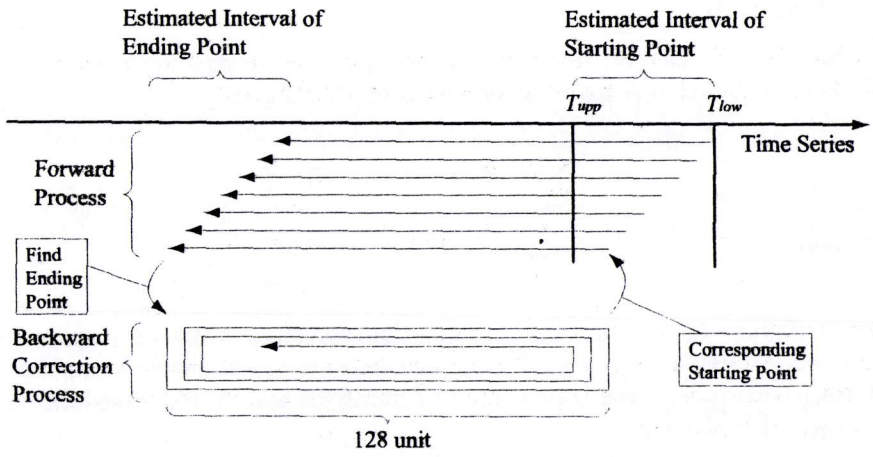

Fig. 3 Schematic of forward process and backward correction process.

The starting point of gesture is unknown, whereas it is confirmed when the ending point is decided. In the forward process described in Subsection 4.2, decomposition is executed with $O(n)$ multiplications. The process is unidirectional, therefore, it can be successively achieved. It is orthonormal transform and it is guaranteed to have inverse transform, but it suffers from sensitivity to the fluctuation of starting point. On the contrary, the backward process described in Subsection 4.3 requires $O(n \log_2 n)$ multiplications and involve bi-directional computation. But the inverse transform is robust against the time shift of origin. And it yields the appropriate reconstruction of original data.

Therefore, it is natural way starting with certain interval as the candidates of starting point leaving the possibility of the future determination, to process rough decomposition (forward process), and after determination of ending point (i.e. starting point), to process precise decomposition and reconstruction (backward process).

In the step of reconstruction, noises and artifacts are eliminated by selection of the subset $\{ j \}$ from $[0, J]$. The subset is accompanied with the reference pattern by a priori decomposition. Finally reconstructed $\bar{V}_0$ and reference pattern are compared with S-DP matching method.

5. Conclusions

This paper proposed a communication interface system based on head-gesture for yes/no question. The main purpose of this research is to realize an interface under variety of conditions, in flexible situations. In order cope with noises and artifacts in such a condition, we introduced a matching method for head gesture. In our system, there are two levels of anticipation features. The one is modifying the reference patterns according to the context of the query and cognitive difficulty of the question before making query. And the other is incursive refinement of reference patterns during matching process with allowance of future revision. The latter involves "forward process", "backward correction process", and "successive DP-matching".

The starting point of gesture is unknown, whereas it is confirmed when the ending point was detected. It is natural way starting with certain interval as the candidates of starting point leaving the possibility of the future determination. In forward process the data are roughly decomposed with the time complexity $O(n)$, where n is the size of data. After detection of ending point, precisely decomposed and reconstructed with the time complexity $O(n \log_2 n)$. The essential fact is the starting point depends on the ending point in the future.

Both of the transforms in the forward process and backward correction process are roughly equivalent to Haar wavelet transform. The former is essential for the unidirectionarity of processing. But the latter is replaceable by other wavelets, and it is possible to be improved. In this paper, we adopted it for simplicity. The optimal selection of wavelet and quantitative evaluation of the system is necessary for further research.

References

[1] Dubois, D. M. (2000) Review of Incursive, Hyperincursive and Anticipatory Systems —Foundation of Anticipation in Electromagnetism—, *AIP Conference Proceedings*: *CASYS'99*, **517**, pp.3-30.

[2] Nonaka, H (2003) Communication Interface with Eye-gaze and Head Gesture using Successive DP Matching and Fuzzy Inference, *Journal of Intelligent Information Systems*, **21**, 2, pp.105-112.

[3] Card, S., Moran, T, and Newell, A. (1983) The Human Information-Processor, in "The Psychology of Human-Computer Interaction", pp. 23-97, LEA, 1983.

[4] Haar, A (1910) Zur Theorie der Orthogonalen Funktionensystems, *Mathematische Annalen*, **69**, pp. 331-371.

[5] Daubechies, I. (1988) Orthonormal Bases of Compactly Supported Wavelets, *Communi- cations on Pure and Applied Mathematics*, **41**, pp. 909-996.