AN INTELLIGENT GUIDING BULLETIN BOARD SYSTEM WITH REAL-TIME VISION AND MULTI-KEYWORD SPOTTING MULTIMEDIA HUMAN-COMPUTER INTERACTION

Cheng-Yu Chang, Chung-Hsien Yang, You-Sheng Yeh, Pau-Choo Chung, Jhing-Fa Wang, and Jar-Ferr Yang

Department of Electrical Engineering
National Cheng Kung University
No.1, Ta-Hsueh Road, Tainan 701, Taiwan

ABSTRACT

This paper presents an intelligent guiding bulletin board system (iGBBS), which is based on vision-interactive and multiple keyword-spotting technology. The system is aimed to provide different kinds of multimedia human-computer interaction (MMHCI) for users under different requirements. At first, a real-time front-view face detection using Harr-like features is used to decide when iGBBS should wake up and become interactive with the user. After system initialization, some feature points within the detected face area are going to be found. Then the orientation of user's head will be estimated via pyramidal Lucas-Kanade optical flow tracking. In addition, spotting the keyword from user's utterance with some related augmented reality responses would be provided as well. The performance of vision-interaction in iGBBS could be reached to 20 fps under Pentium IV 1G Hz PC. The error rate of multiple keyword-spotting interaction in iGBBS is about 36.2% and people can get the right response in 2.76 times search averagely. With the comparison to the traditional guiding system, bulletin board, or other non-vision-based input devices system, such like gloves or markers, our system offers a simple, useful and economical solution for the real-time interaction between the user and computer.

1. INTRODUCTION

During past decades, guiding systems and bulletin boards are widely existing at many places, especially prevalent at universities. However, traditional guiding system and bulletin boards (see Fig. 1) certainly have several drawbacks: a) *Inefficient Reusability*; b) *Space Consuming*; c) *Without Real-time Interaction with Users* and d) *Monotonous*.

In the recent modern stage, computerized presentation of multimedia has been discovered for many clear advantages over paper media and multimedia human-computer interaction/interface (MMHCI) becomes an active research area for engineering of computer science. Many researchers have paid more and more attention on the MMHCI domain for the purpose of making an ease-of-use environment between we human beings and machines. Our objective is to develop an intelligent guiding bulletin board system (iGBBS), which is based on vision-interactive and multiple keyword-spotting technologies. Users could interact with iGBBS through their head pose orientation and do NOT need any gloves or markers. For some users who are not familiar with top-down data searching, they could also interact with iGBBS by some keywords. The system will recognize these keywords and return some related data.

(a) (b)

Fig. 1. (a) traditional guiding system and (b) traditional bulletin boards.

The article is organized as following: In Section 2, we are going to discuss our iGBBS framework. Later, technology of real-time vision-interaction will be drawn in Section 3. Following that, we are going to discuss our methodology of multiple keyword-spotting interaction in Section 4. Finally, Section 5 evaluates the performance of iGBBS and concludes this paper.

2. SYSTEM OVERVIEW

Our system is aimed to provide different kinds of multimedia human-computer interaction (MMHCI) for users under different requirements. The overview of our system architecture is shown as Fig. 2. iGBBS is composed of several major parts:

- *Real-time Vision Interaction*: responsible for dealing

This work is supported by the National Science Council, Taiwan, under Grant NSC-94-2218-E-006-043 and developed in Department of Electrical Engineering, National Cheng Kung University.

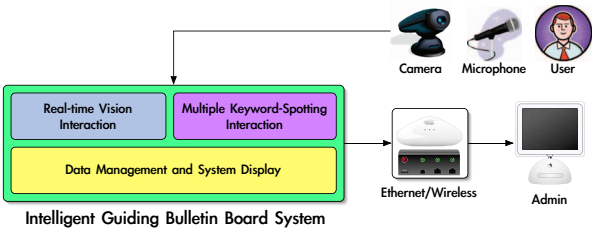

Fig. 2. The overview of our system architecture.

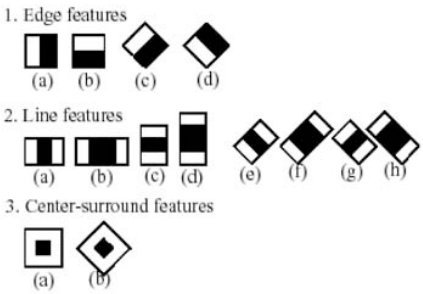

Fig. 3. The extened set of Harr-like features. The sum of pixels within the white rectangles could be substracted from the sum of pixels in the black area.

with when the system should wake up and become interactiveable with the user.

- *Multi-keyword Spotting Interaction*: spotting the keyword from user's utterance with some related augmented reality responses.
- *Data Management and System Display*: convenient administration for information providers.

3. REAL-TIME VISION INTERACTION

According to the visual attention psychology, a visual line reflects a direction or a place we human beings take care or not. So, we could suppose if we have detected a front-view face of someone, there must exist somebody who is interested in our system.

In iGBBS, we use an appearance-based and statistical approach for the front-view face detection. This approach was originally developed by Viola and Jones [1] and then analyzed and extended by Lienhart [2]. As Fig. 3 shown, we totally have 14 features which include 4 edge features, 8 line features and 2 center-surround features in order to reach the goal of generating a rich and over-complete feature set. These features are so called Haar-like features, which are based on the idea of the wavlet template, because they are computed similar to the coefficients in Haar wavelet transform[3].

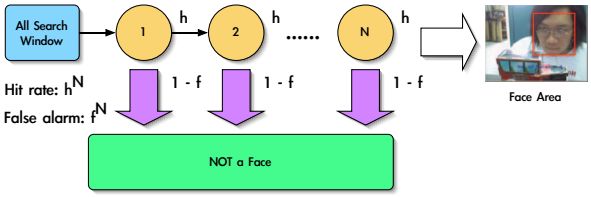

Fig. 4. Face detection cascade of classifiers with N stage. The classifier is trained to reach a hit rate of h, a false alam rate of f and rejection could happen at any stage.

Above caculated Haar-like feature value $\wp$ would then be used as an input of the decision weak classifier. Each weak classfier represents some simple feature within the input image that might be related to the face or not, as Eq (1) shown.

$$f_i = \begin{cases} +1, \text{if } \wp_i \geq t_i \\ -1, \text{if } \wp_i < t_i \end{cases} \quad (1)$$

Soon afterwards, a robust classifier made from multiple weak classifiers using boosting procedure would be generated. The robust boosted classifier F could be treated as a weighted sum of weak classifiers

$$F = sign(c_1 f_1 + c_2 f_2 + \cdots + c_n f_n) \quad (2)$$

In order to increasing the performance, Viola [1]] suggests constructing several cascades of classifier and each cascade is built from several boosted classifier F_n. During the detection stage, current search window would be analyzed by each classifier F_n and rejection could happen at any stage (see Fig. 4).

After we have got the front-view face area, and then all we have to do is orientation estimating and starting the interaction between iGBBS and the user.

Our first major step of orientation estimation is to find some feature points, such we call "Eigen Component" within the detected face area. "Eigen Component" represents the energies of a given window of an image after projecting according to its eigen-vectors. So, if the energies are not constant in all direction, it might have a corner or high texture information for us.

Later than feature points $\mathbf{u} = [u_x, u_y]$ are found, we would like to estimate the orientation of user's pose by feature tracking technology. However, traditional Lucas-Kanade optical flow methodology is available only when the pixel displacement is quite small. In order to increase the tracking accuracy and void the influence of large motion size, we use modified pyramidal Lucas-Kanade optical flow, which is different from traditional one[4][5], for the feature tracking.

Given a feature point $\mathbf{u}^L = [u_X^L, u_y^L]^T$ in frame I at pyramidal image level $L, L = 0, \ldots, L_m$, we would like to find its corresponding location $\mathbf{v}^L$ in frame J at pyramidal image

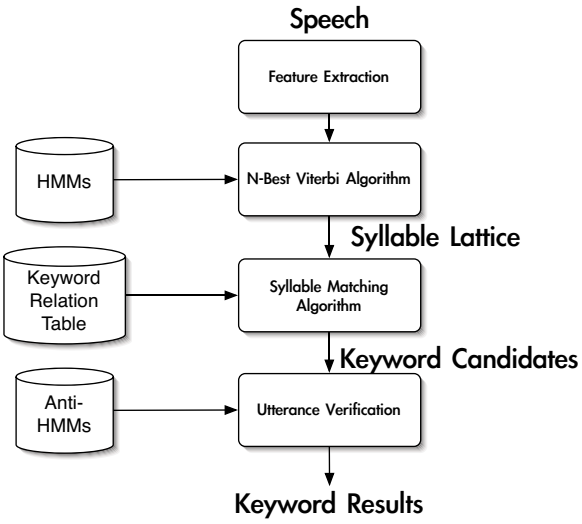

Fig. 5. The block diagram of the multi-keyword spotting system.

level L. The objective of feature tracking algorithm is to minimize the image mismatching function Eq (3), where $\mathbf{g}^L$ is the guess for optical flow at level L.

4. MULTI-KEYWORD SPOTTING INTERACTION

In our life, speech is the most natural way to communicate with people, so we also include speech recognition in our system. A multi-keyword spotting technique is introduced here to spot the keywords from a spoken utterance[6]. Proper responses therefore can be made in accordance with the spotted keywords.

For extending the single keyword spotting to multi-keyword spotting (see Fig. 5), we define a keyword relation table, which is used to decide the combination of two or more keywords in an utterance. The structure of the keyword relation is denoted as $(PK, \{SK\})$, where PK is the primary keyword and $\{SK\}$ is the secondary keyword set. In this approach, only one primary keyword is allowed and the secondary keyword is optional and has some relations to the primary keyword. Fig. 6 shows the diagram of possible keyword paths. According to the keyword relation table and the keyword position in an utterance, we can determine multi-keyword candidates.

After extracting keyword candidates in an utterance, we have to consider the combination of the primary and secondary keywords. If the secondary keyword is unreliable, that is, it has a high distance value; the effect of the secondary keyword should be greatly reduced. Thus, a sigmoid function is used as the weighting function to deal with this problem. Given a multi-keyword candidate $(PK, \{SK\})$, the weighting function for the primary keyword W_{PK_i} is defined as Eq

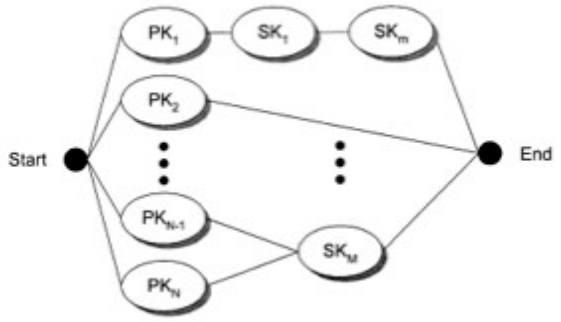

Fig. 6. A diagram of possible keyword paths

(4),

where c, ρ and λ are constants. N_i is the number of the secondary keywords in the utterance. $Dist_{SK}$ represents the distance of the secondary keyword SK and is defined in the following:

$$Dist_{SK} = \sum_{i=1}^{M} -\log P(O_i|s_i) \tag{5}$$

where M is the number of subsyllables and O_i is the observations corresponding to the subsyllable s_i in the secondary keyword SK. After the weighting function of the primary keyword is obtained, the weighted distance for the primary keyword is defined as

$$WD_{PK} = W_{PK} \times Dist_{PK} \tag{6}$$

Then, the weighted distance is used to determine the recognition result and the keyword rejection/acceptance decision is made by comparing WD_{PK} with a predefined threshold.

5. PERFORMANCE EVALUATION AND CONCLUSION

In this paper, an intelligent guiding bulletin board system (iGBBS), which is based on vision-interactive and multiple keyword spotting technology is proposed (as Fig. 8 shown). iGBBS has been built for office/lab search in Department of Electrical Engineering at National Cheng Kung University, Taiwan. We put this system on the first floor of the department building, and everyone can use it to search the office or lab that he wants to know. The performance of vision-interaction in iGBBS could be reached to 20 fps under Pentium IV 1G Hz PC. Fig. 7 shows the average response time of vision-interaction in our experiment. Different users could completely control iGBBS easily without a lot of training. The error rate of multiple keyword-spotting interaction in iGBBS is about 36.2% and people can get the right response in 2.76 times search averagely.

$$\arg\min_{d_x^L, d_y^L} \sum_{x=u_x^L-w_x}^{u_x^L+w_x} \sum_{y=u_y^L-w_y}^{u_y^L+w_y} (I^L(x,y) - J^L(x+g_x^L+d_x^L, y+g_y^L+d_y))^2 \quad (3)$$

$$W_{PK_i} = \begin{cases} 1 & \text{if } SK_j \text{ does not exist,} \\ (\prod_{j=1}^{N_i} \frac{c}{(1+\exp(-\lambda \times (Dist_{SK_j}, -\rho)))})^{1/N_i} + (1-c) & \text{otherwise.} \end{cases} \quad (4)$$

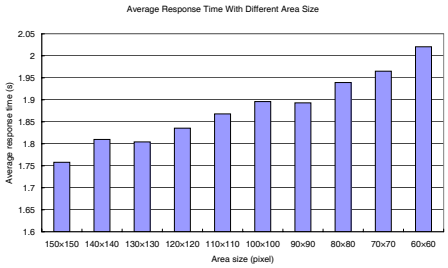

Fig. 7. Average response time of vision-interaction with different area size in our experiment. Our experimental program would generate 10 areas with different sizes on the monitor randomly and we ask users to use our system to control the mouse movement. The response time would be taken down only when the users move their mouse to the right position.

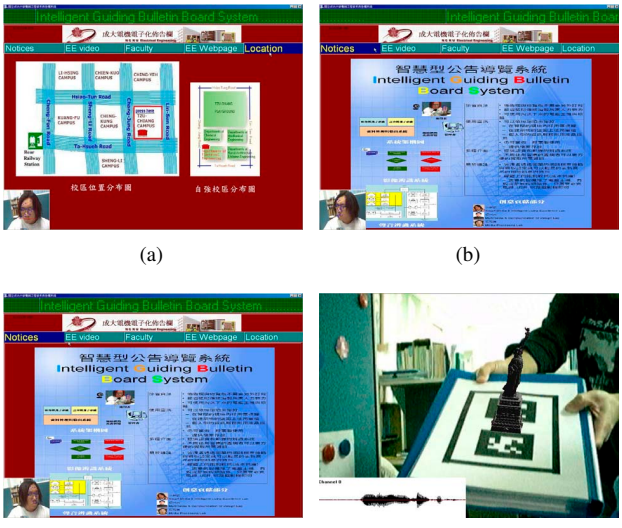

Fig. 8. Our intelligent guiding bulletin board system (iGBBS). (a) iGBBS in our EE building; (b)–(c) Real-time vision-interaction through feature points tracking and pose orientation estimation; (d) Fusion of keyword-spotting and 3D augmentations with ARTag[7].

Compared to the traditional guiding system or bulletin board, our contributions could be summarized as following: 1) *Re-usability*. No need for extra works in manufacturing bulletin boards. We could save manpower and material resources for putting up the traditional notice, and eliminated computers or monitors could be reused; 2) *Flexibility*. According to the surrounding environment, we could choose suitable display equipment; 3) *MMHCI*. iGBBS provides a vision- and keyword-based interactive system. Users could get information easily according to their customs; 4) *Maintainability*. The administrator could update related information and data through file transmission easily.

6. REFERENCES

[1] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[2] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, 2002, vol. 1, pp. I–900–I–903 vol.1.

[3] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates," 1997, pp. 193–199.

[4] Bruce D. Lucas and Takeo Kanade, "An iterative image registration technique with an application to stereo vision (darpa)," in *Proceedings of the 1981 DARPA Image Understanding Workshop*, April 1981, pp. 121–130.

[5] Jianbo Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, 1994, pp. 593–600.

[6] Chung-Hsien Wu and Yeou-Jiunn Chen, "Multi-keyword spotting of telephone speech using a fuzzy search algorithm and keyword-driven two-level cbsm," *Speech Communication*, vol. 33, pp. pp.197–212, 2001.

[7] Chris McDonald Gerhard Roth, Shahzad Malik, "Artag," http://www.cv.iit.nrc.ca/research/ar/artag/.