Automatic Hand Gesture Based Television Control System

PRAVEEN KUMAR DUBEY, NEHRU INTITUTTE OF ENGINEERING AND TECHNOLOGY, COIMBATORE

Abstract — *Hand gesture based television (TV) control is attracting more and more researchers. Most of existing works focus on the hand gesture recognition algorithm and the corresponding user interface, while the power consumption or computational cost is not considered carefully. In practice, keeping the camera device and the gesture recognition module running all the time often costs much energy. Till now, few methods have been reported to solve this problem. This paper proposes an automatic user state recognition scheme to recognize the TV user's state and activate the camera-based gesture recognition module only when the user is trying to control the TV. Specifically, the user's behavior active or not is detected by low-cost sensors, the user's gaze watching TV or not is tracked by the face-based view detection, and the user's state is then decided according to a finite-state machine composed of four states: Absent, Other Actions, Controlling, and Watching. The prototypes based on an ultrasonic distance sensor array, a red-green-blue (RGB) camera, and a depth camera are implemented and tested. The results show that the proposed scheme can effectively reduce the power consumption or computational cost of the original hand gesture based control schemes*[1].

Index Terms — TV control, user state recognition, finite state machine, hand gesture.

I. Introduction

Television (TV) is widely used all around the world. Till now, the TV display screen has been innovated for several generations while the TV controller keeps nearly unchanged during a long period. Recently, with the enrichment of TV programs, more and more frequent controlling operations such as channel switching and program searching are required, which makes the traditional point-click-style hand-held TV controller inconvenient. Taking the channel switching for example, it requires the user to lower his/her head to first see and then press the small buttons. Next, he/she needs to look up at the TV screen to see whether the program is the expected one. Otherwise, he has to repeat this process to switch the channel. According to Fitts' Law [1], the interaction system's execution time depends on the easiness for the user to press the button. The frequent looking-up-and-down reduces the control speed and affects user experiences to a large extent.

Recently, some new controlling apparatuses have appeared, e.g., the wearable or hand-held space controller. These kinds of controllers activate command signals by gesture recognition based on space sensors, e.g., the controller based on a space mouse [2], the one based on a data glove and a Heads up Display (HUD) [3], the wearable one like a wristwatch [4], the one based on a touch pad [5], the one with a low-cost data glove [6], and the ring-type controller [7]. Although these controllers improve user experiences in some extent, it requires the user to hold or wear a device which may be difficult to found sometimes in a big room.

Additionally, some hand-free controlling means have also been studied, e.g., voice control [8][9]. With the speech recognition module, the TV can interpret the user's voice command and then response accordingly. However, there are still difficulties in voice control. Firstly, the speech recognition algorithm is not mature enough due to large number of vocabularies and various dialects. Secondly, the interference including the environmental noise and the voice coming from the TV's speaker will affect voice recognition to a large extent and are not easy to eliminate [8]. Additionally, the recognition module is always kept on in order to detect when and what kind of voice command is activated, which leads to much computational cost and power consumption.

Another kind of hand-free controlling is the hand gesture based TV control [10][11]. The detected and recognized hand gestures are used as the command signals for TV controlling. Some user interfaces, e.g., icon-based interface [10] or motion-based interface [11], are adjusted accordingly in order to support natural hand control. Free-hand control makes the controlling much easier for the user to master and use, and some new sensors capable of depth detection have been developed for more accurate hand gesture recognition [12][13], which makes it a potential future for TV control. However, similar with the speech recognition module in voice control, the hand gesture recognition module is kept running all the time in order to detect the activation gesture for launching the gesture control. Thus, the means to reduce the computational cost or power consumption should be considered carefully.

Generally, with respect to the energy consumption, two aspects are considered: the sensor device, e.g., a red-green-blue (RGB) camera or a depth camera, and the gesture recognition module. Intuitively, it is not necessary to make these sensors always run especially when the user is not controlling the TV. Additionally, it is not reasonable for the user to turn on/off the gesture recognition module frequently and manually during TV watching. Thus, it is a key challenge to select the timing for automatically turning on

[1] This work was supported by the Central Research Institute, Huawei Technologies.
Electronic version published 03/20/14.

and off the device or recognition module, which has not been carefully considered in existing works.

This paper proposes an automatic user state recognition scheme to recognize the TV user's state efficiently before activating or sleeping the camera device and hand gesture recognition module. The considered user states, including *Absent*, *OtherAction*, *Controlling*, and *Watching*, are initialized and updated according to the results of action detection and presentation detection based on low-cost sensors. Thus, the camera device and hand gesture recognition module are activated adaptively to reduce the system's computational cost and power consumption.

The rest of the paper is arranged as follows. In Section 2, the related work is introduced. The automatic user state recognition scheme and the corresponding hand gesture based TV control system as well as the detailed implementations are presented in Section 3. In Section 4, the proposed scheme is evaluated by implemented prototypes and various comparative experiments. Finally, the conclusions are drawn in Section 5.

II. Related Work

There exist some works on hand gesture based TV control. With respect to the gesture's features, there are two types, i.e., static gesture recognition and dynamic gesture recognition. With respect to the recognition sensors, there are two types, i.e., the recognition based on RGB camera and the one based on depth camera.

Static hand gestures are identified by the shape of hand, including position, orientation and number of fingers. A straightforward static gesture control method is hand-based mouse that changes hand gestures into mouse manipulations [10]. Specifically, the palm is detected and tracked for implementing the mouse moving, the change from palm to fist is detected for mouse clicking, and the fist is detected and tracked for mouse dragging. Some other controlling methods use a specific number of fingers and their angles [14], a predefined sequence of gesture state transitions [15], or the index finger and thumb [16][17]. Generally, during static gesture controlling, hand poses have to keep static for a certain period of time. Although the correct recognition rate can be increased, the controlling action is not natural enough.

Differently, the dynamic hand gesture methods, which recognizing hand motion patterns as command signals, seem more natural for TV control [18]. For example, the gesture drawing based on motion tracking is used for command signals [11]. In this scheme, hand gestures are recognized by capturing the motion path when the user draws different symbols in the air. These gestures are used to interact with the TV. For dynamic hand gesture recognition, it is important to define a suitable set of motion patterns easy to remember and recognize. Currently, two kinds of devices are popular: the RGB camera and the depth camera. Especially, the latter provides three-dimension information of hand or finger position, and could thus improve the hand motion recognition. With a RGB camera, the two-dimension motion information of hand is decided [18][19]. Generally, due to the camera's high resolution, the color information can be used to accelerate the detection of hand regions. This kind of camera is now inexpensive and acceptable by TV manufactures or consumers. However, the camera is often sensitive to light conditions in room, which degrades hand recognition.

The depth camera is now attracting more and more researchers and engineers, although the available resolution of the depth map is still low [12][20][21]. Now, three kinds of depth camera are often used: the camera array for stereo vision, the camera based on time of flight (TOF), and the one based on structured light. For example, various motions are recognized from the spatiotemporal trajectory features composed of horizontal, vertical, temporal, and depth information captured by a TOF camera [12]. In another scheme, the TOF camera is combined with RGB camera to improve the hand gesture recognition [20]. The depth camera based on TOF or structured light is often robust against light changes in room, and is thus suitable for various environments.

As can be seen, with the improvement of sensing devices and recognition algorithms, the free-hand TV control scheme can provide higher recognition rate and more natural user experiences. Nevertheless, the system's energy consumption is a concern and still not considered carefully, e.g., the computational cost on continuous gesture recognition, or power consumption on continuous device running. Some novel methods are expected to activate or sleep the devices and recognition modules adaptively according to the user state. This paper will propose a solution for this problem.

III. The Proposed Hand Gesture Based Low-Cost TV Control System

The proposed TV control system, targeted for low-cost hand gesture control in an automatic manner, composed of both hardware and embedded software components. Fig. 1 shows the hardware components, including the TV, set-top box, and the other sensors connected to the TV system, including the RGB or depth camera module and ultrasonic distance array module. The embedded software components which connect the sensors with the TV, process the sensor data, and drive the TV control process will be presented as follows.

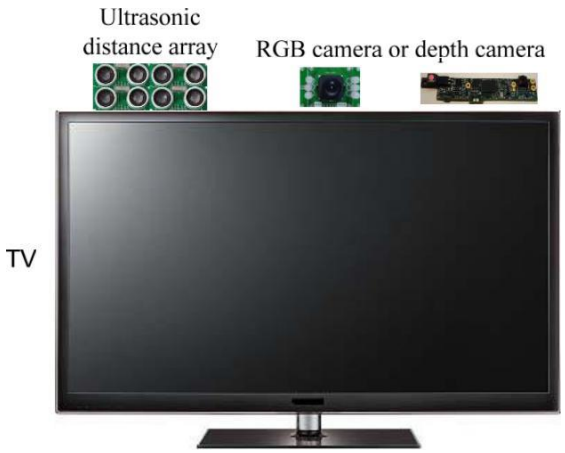


Fig. 1. The proposed TV control system's hardware components include the TV, set-top box, and some sensors, including the RGB or depth camera module and ultrasonic distance array module.

A. *Overview of the Proposed Control System*

Compared with traditional TV control systems based on hand gestures, the proposed system introduces a new scheme, named Automatic User State Recognition (AUSR). This scheme consists of several steps, e.g., the user state initialization or updating, the device activation or sleeping, the recognition module activation or sleeping, and the user input notification, as shown in Fig. 2. Based on the ultrasonic distance array and camera module, the user's action is detected and user state is initialized or updated automatically. The user state tells whether the user is controlling the TV or not, and will be defined in the following section. If the user is controlling the TV, the camera device and gesture recognition module will be turned on, and the user will be informed to begin hand gestures by e.g., displaying a message on the screen. Otherwise, the camera device or gesture recognition module is slept in order to save energy. When informed to begin, the user makes control gestures, which will be recognized by the system and then drive the TV. In the following content, the AUSR, including user state definition, initialization and transition based on automatic action detection and presentation detection, will be introduced in detail, and the performance evaluation on the implemented hand gesture based TV control systems will be presented.

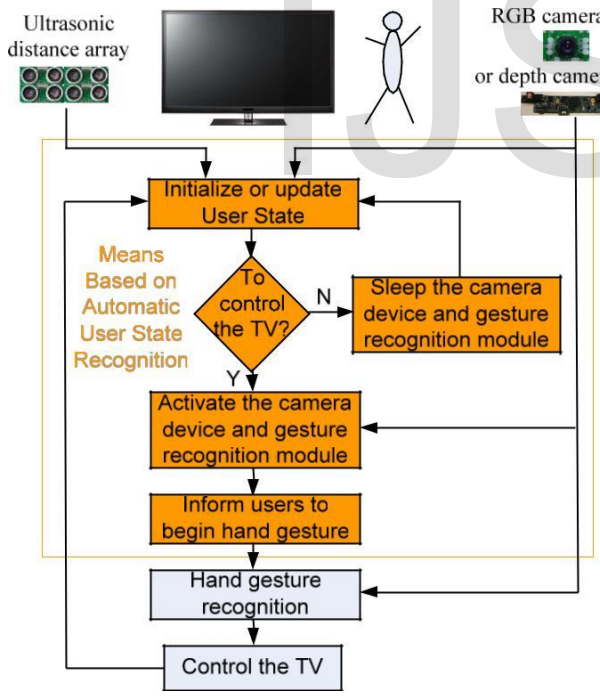


Fig. 2. The proposed TV control system's software components include user state initialization or updating, camera device and hand gesture recognition module activation or sleeping, user input informing, etc. The user's action is detected in order to decide the intention to control TV.

B. *User State Definition*

According to user's actions in front of TV, the user state is classified into four types, i.e., *Absent*, *OtherAction*, *Controlling*, and *Watching*. *Absent* means no user is sitting in front of the TV although it may be open. *OtherAction* means that some user is sitting in front of the TV and doing something, but he is not watching or controlling the TV. *Controlling* means that the user is controlling the TV, e.g., changing the channel or adjusting the volume. *Watching* means the user is watching TV. Based on Finite-State Machine (FSM), these states are denoted by (A=0,B=0), (A=1,B=0), (A=1,B=1), and (A=0,B=1), respectively. Here, if the user's intentional motion in front of the TV is detected, then A=1, otherwise A=0. If the user is looking at the TV screen, then B=1, otherwise B=0. These user states, together with the state transition, are defined in Fig. 3.

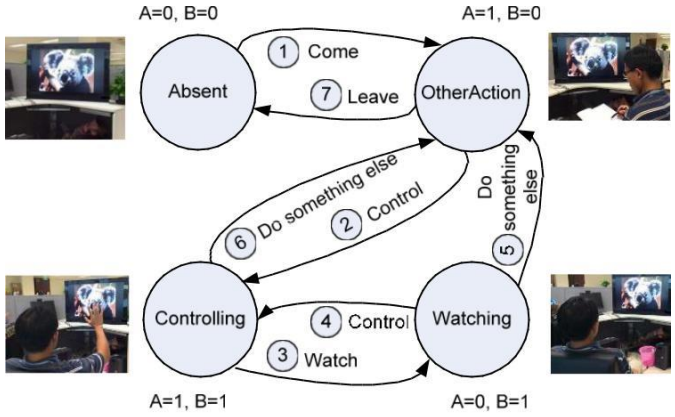


Fig. 3. The defined user states in TV watching, including *Absent*, *OtherAction*, *Controlling*, and *Watching*. Seven conditions may drive the state transition based on Finite-State Machine (FSM).

The user state will be changed from *Absent* to *OtherAction* if the user comes in front of the TV, from *OtherAction* to *Controlling* if the user makes hand gestures to control the TV, from *Controlling* to *Watching* if the user keeps silent and stays to watch TV, from *Watching* to *Controlling* if the user makes hand gestures to control the TV, from *Watching* or *Controlling* to *OtherAction* if the user does something else in front of TV, and from *OtherAction* to *Absent* if the user leaves the TV region. In the proposed scheme, these conditions for state transition are detected automatically based on the ultrasonic sensor array and the RGB or depth camera.

C. *User State Initialization*

To initialize the user state automatically, whether the user is intentionally acting or looking at the TV screen are detected by the proposed action detection and the presentation detection respectively. Firstly, based on the ultrasonic distance array, the user's action is detected and intentional action is decided (A=0 or 1). Then, based on the RGB or depth camera, the user head and face are detected and the gaze is decided (B=0 or 1). Finally, based on the decisions (on A and B), the user state is determined accordingly, as shown in Fig. 4. The action detection and presentation detection will be introduced in detail next.

Action detection. To identify the user state automatically, it is important to detect whether there exist user actions in front of the TV. Considering that the user may do anything including walking, standing, sitting, shaking the head, shaking the hand, etc., it is difficult to recognize the exact action type

with a machine. Fortunately, in the scheme proposed in this paper, it only needs to decide whether the user is actively moving, while whether he/she is watching TV will be decided by another way to be presented in the following content. To detect movements in an environment, the computer vision based method, which decides the movements by detecting the differences between adjacent images captured by a camera, is a possible solution. However, this method will introduce high computational cost caused by pixel-to-pixel comparison. Alternatively, in this paper, the ultrasonic sensor array composed of low-cost distance sensors is adopted to detect movements. Generally, the ultrasonic distance sensor [22][23] has certain sensing space denoted by the angle θ and maximal distance D, as shown in Fig. 5(a). The action detection is composed of several steps, including distance detection, movement detection, single sensor based action decision, and sensor array based action decision. In the sensing space, at moment t, the distance between an object and the sensor can be detected in real time and the outputted distance L(t) is

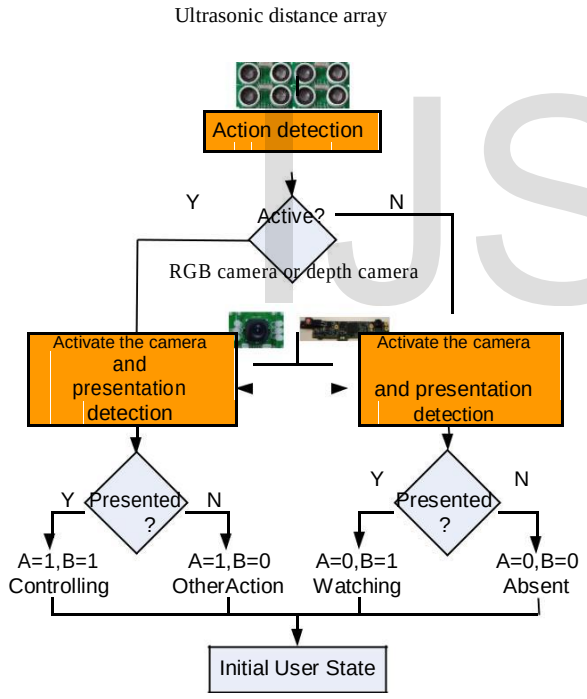


Fig. 4. The automatic user state initialization process consists of several steps, i.e., action detection, device activation, presentation detection, and state decision.

$$L(t) = C \cdot T(t)/2 , \quad (1)$$

where C is the sonic speed in air, generally 340 meters per second, and T(t) is the time between the sensor emits and receives the ultrasonic wave. Let P be the sampling interval, when the user's hand is moved timely, the distance at different sampling time can be sensed as shown in Fig. 5(b). Thus, during the period of P, the moved distance L_t at t moment is computed by

$$\Delta L_t = |L(t+P) - L(t)| = C \cdot |(T(t+P) - T(t))|/2 , \quad (2)$$

where |x| means the absolute value of x. Considering that the user may not always stay static even when he is watching TV, the threshold R should be considered to skip non-intentional movements according to

$$A_t = \begin{cases} 0, \Delta L_t \in R \\ 1, \Delta L_t \notin R \end{cases} . \quad (3)$$

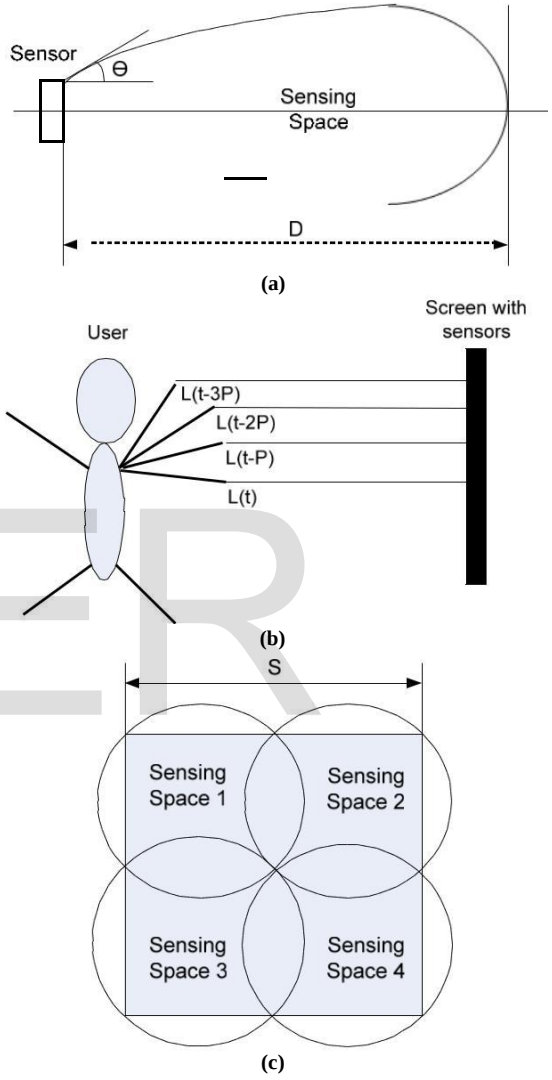


Fig. 5. The ultrasonic distance sensor array composed of four sensors. (a) is the sensor region of one sensor (side view), (b) is the measured distance between the moved hand and the screen at different sampling time, and (c) is the fused sensor region of four sensors (frontal view).

Considering that the user's action may not be continuous, the action sequence S_t composed of K components is adopted to identify actions. That is

$$S_t = [\Delta L_{t-K+1}, \Delta L_{t-K+2}, \ldots , \Delta L_{t-1}, \Delta L_t] . \quad (4)$$

Based on the threshold R for non-intentional movements, the filtered action sequence is computed by

$$S_{Rt} = [A_{t-K+1}, A_{t-K+2}, \ldots, A_{t-1}, A_t] . \quad (5)$$

Then, the action status A is decided by

$$A = \begin{cases} 0, & SUM(S_{Rt}) \in Q \\ 1, & SUM(S_{Rt}) \not\in Q \end{cases} \quad (6)$$

where Q is the threshold for deciding a continuous action and $SUM(S_{Rt})$ is defined as

$$SUM(S_{Rt}) = \sum_{i=0}^{K-1} A_{t-i} \quad . \quad (7)$$

Generally, the ultrasonic sensor's sensing angle is small, e.g., 15 degrees. In order to support a big sensing space, the sensor array composed of four sensors is constructed, as shown in Fig. 5(c). By combining the four sensors' sensing space, the sensing square is obtained, with the size S computed by

$$S = 2\sqrt{2} \cdot r = 2\sqrt{2} \cdot D \cdot tg\theta \,, \quad (8)$$

where r is the radius of the sensor's sensing circle, D the maximal sensing distance, and θ the sensing angle. Based on the sensor array, the action decision will be made according to

$$A = \begin{cases} 0, & \max\{A^0, A^1, A^2, A^3\} = 0 \\ 1, & \max\{A^0, A^1, A^2, A^3\} = 1 \end{cases} \quad . \quad (9)$$

where A^i (i=0, 1, 2, or 3) is the (i+1)-th sensor's action decision result, and max{x,y,z} means to get the maximal value from x, y and z.

Presentation detection. To decide the user state, the presentation detection procedure which detects whether the user is looking at the TV screen will done first. Intuitively, gaze tracking may be a solution [24], except that it often needs a special device with an infrared ray emitter. Another way is to detect the face followed by the view angle detection, which often costs much computation. Here, to keep low cost, only face detection is adopted. To decide the view point, the face templates are restricted by frontal and side faces in certain angles, i.e., horizontal angle, vertical angle and roll angle, as shown in Fig. 6. Only the faces belonging to the restricted angle ranges are considered as watching the TV. Specifically, the horizontal view angle α is defined by

$$\alpha = arctg(W/2L) \,, \quad (10)$$

where arctg is an inverse function of a trigonometric function, W is the width of TV screen, and L the distance between the user's eyes and TV screen, as shown in Fig. 7. Similarly, the vertical view angle β is defined by

$$\beta = arctg(H/2L) \,, \quad (11)$$

where H is the height of TV screen, and L and arctg are same with the ones in (11). Differently, the roll angle Φ has no relation with TV screen's size and distance. Thus, the face template library L_F for presentation detection is composed of the faces F with horizontal view angle F_h changing from –α to α, vertical view angle F_v from –β to β, and roll angle F_r from –Φ to Φ. That is

$$L_F = \{F | — \alpha \subset F_h \subset \alpha, — \beta \subset F_v \subset \beta, \text{and} — \emptyset \subset F_r \subset \emptyset\} \,. \quad (12)$$

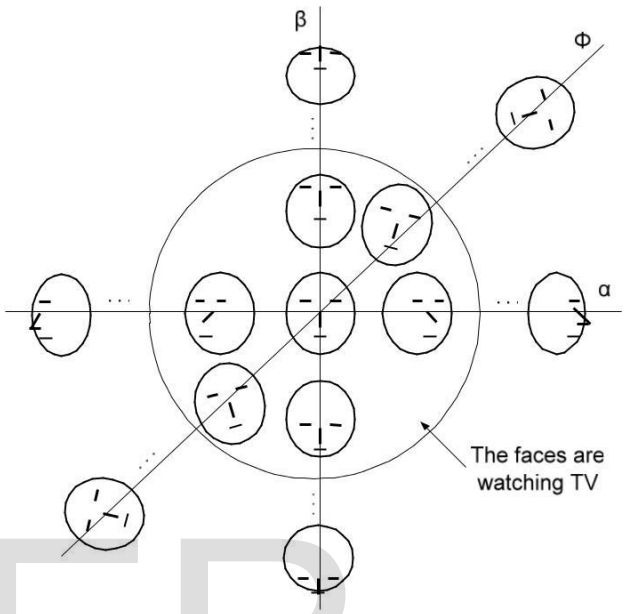


Fig. 6. The faces watching TV with different angles, i.e., horizontal angle α, vertical angle βand roll angle Φ. Only the faces among restricted angle ranges are regarded watching TV.

Based on the face library, the face detection is done to tell whether the user is looking at the TV screen. If at least one face is detected from the captured environment, the user is considered as watching the TV, otherwise not. Here, the face detection constructed on Haar feature extraction and AdaBoost classification [25][26] is adopted to obtain fast face detection. Based on the result, the presentation status is decided by

$$B = \begin{cases} 0, & \text{if the face is not detected} \\ 1, & \text{if the face is detected} \end{cases} \quad . \quad (13)$$

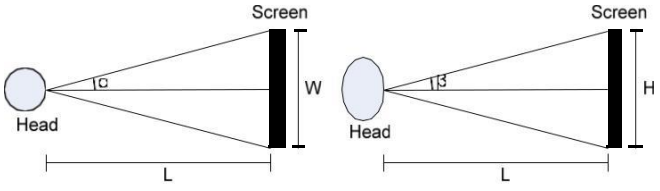


Fig. 7. The face's horizontal angle (left) and vertical angle (right) is in relation with the screen's width, height, and the user's distance.

User State Decision. Based on the results of action detection and presentation detection, the user state is decided according to the principles in Table 1. That is, if A=1 and B=1 then the user state is *Controlling*, if A=1 and B=0 then the user state is *OtherAction*, if A=0 and B=1 then the user state is *Watching*, and if A=0 and B=0 then the user state is *Absent*.

TABLE I
USER STATE INITIALIZATION

The Algorithm of Decision of User State

If (the intentional motion is detected)
  If (the face watching TV is detected)
  Set (A=1,B=1) - Controlling
  Else
  Set (A=1,B=0) - OtherAction
Else if (the face watching TV is detected)
  Set (A=0,B=1) - Watching
  Else
  Set (A=0,B=0) - Absent

D. User State Transition

In the proposed scheme, the user state is updated automatically according to the following steps, shown in Fig. 8. Firstly, the action detection is done to decide whether the user is moving intentionally. If no movement is detected, the inactive time will be counted and compared with the threshold T_m. If the inactive time is bigger than T_m, then the user state will be changed, otherwise not. Here, T_m is used to decide the user state transition when the user does not move during this period. For example, the user state may be changed from *Controlling* to *Watching*, if the user watches TV without movements during the T_m period after several controlling actions. If the intentional movement is detected, the presentation detection will be followed with respect to the original user state, and the user state will be updated accordingly. Specially, the user state may be forced to be initialized in order to avoid deadlocks caused by unpredicted environment changes. Each kind of user state may be updated according to the principles shown in Table 2. As can be seen, if no intentional movement is detected, the camera device or presentation detection module will not be activated. Additionally, even though the movement is detected, if the original user state is *Absent*, the user state is transited to *OtherAction* directly with no need to activate the camera device or presentation detection module. Only when the movement is detected and the original user state belongs to {*OtherAction*, *Controlling*, *Watching*}, the camera device and presentation detection module will be activated.

E. Hand Gesture Recognition

The hand gesture recognition module and camera device are activated when the user state is *Controlling*. Firstly, the user is informed to begin or end hand gestures by through the message feedback on TV screen. Otherwise, the hand gesture detection will not be launched in order to save the system's computational cost or power consumption. Additionally, even during hand gesture period, the gesture is only detected continuously in T_f. T_f is the threshold for restrict the time period permitted for the user to do hand gestures. During this period, the recognition module will always be running. If no hand gestures are detected during this period, the gesture recognition module will be end. The goal here is to reduce the computational or power cost. For the hand gesture recognition, the algorithm based on the RGB or depth camera can be used.

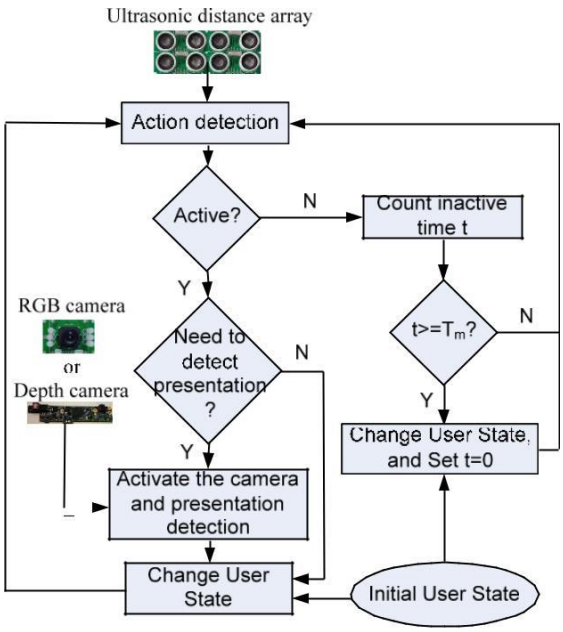


Fig. 8. The user state transition is composed of the steps: action detection, inactive time counting, camera activation, and presentation detection.

TABLE II
USER STATE TRANSITION

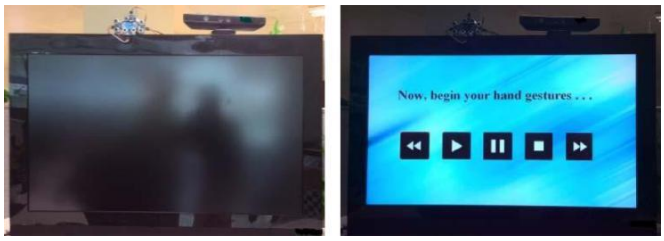
The Algorithm of User State Transition

If (the intentional motion is detected)
  If (A=1 and B=1)
  Detect hand gestures
  Else if (A=0 and B=0)
  Set (A=1,B=0)
  Else if (A=1 and B=0)
  if (the face watching TV is detected)
  Set (A=1,B=1)
  Else
  if (the face watching TV is detected)
  Set (A=1,B=1)
Else if (the intentional motion is not detected for a long time)
  If (A=1,B=0)
  Set (A=0,B=0)
  Else if(A=1,B=1)
  Set (A=0,B=1)

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

The TV control prototypes are composed of a LCD TV (Liquid-crystal-display televisions), an ultrasonic sensor array, a RGB or depth camera, and the low-cost computing module embedded in a Personal Computer (PC), as shown in Fig. 9. In detail, the TV screen is 42-inch (0.93-meter height, and 0.52-meter width) with the optimal controlling distance ranging from 1.0 meters to 3.5 meters, and thus the face library is composed of faces with the angle (α=25°, β=15°, Φ =30 °). The ultrasonic sensor array is composed of four distance sensors, each of which has the sensing angle of θ=15° and maximal sensing distance D=4.5m. At the distance of 3.5 meters, the sensing square's size S=2.7m. The RGB camera outputs the picture with resolution 1024x768, and depth camera with resolution 640x480. In experiments, the low cost means based on various hand gesture recognition algorithms

will be evaluated and discussed, including the study on user behavior, the user state detection accuracy, and the energy consumption with respect to different gesture algorithms.

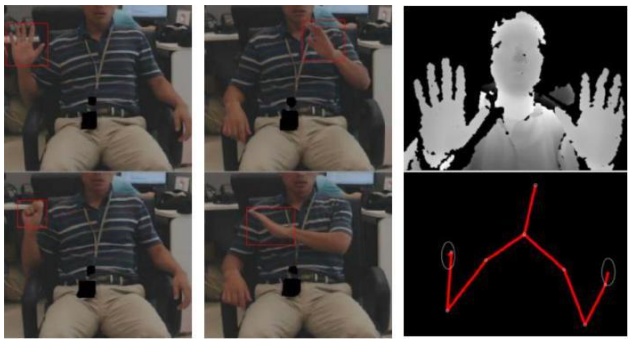


(a) (b)

Fig. 9. The implemented TV control system (a), composed of TV, sensor array, camera and embedded computing module, and the implemented user control interface (b). In this implementation, "Now, begin your hand gestures …" appears for about 3 seconds only when the user turns to the state, *Controlling*, from another state. But, this note is not a must.

A. The Considered Hand Gesture Recognition Algorithms

Three kinds of hand gesture recognition algorithms are considered, i.e., the icon-based hand gesture [10], motion-based hand gesture [11], and depth camera based hand gesture [12]. In the icon-based hand gesture, the palm and fist are detected to simulate the mouse operations, such as clicking and double clicking. In the motion-based hand gesture, the hand is tracked, and the drawing path is recognized as command signal. Here, the recognized dynamic hand gestures include swift left, right, up, and down. In depth camera based hand gesture, the hands, as parts of the human body, are detected and tracked by making use of the depth information. Here, some dynamic hand gestures are recognized, including swift left, right, up, down, back, and forward. Some hand detection and tracking results are shown in Fig. 10.



(a) (b) (c)

Fig. 10. Results of hand detection and tracking, including the hand (up) and fist (bottom) detection (a), hand tracking (b), and depth based hand detection (c) (up, the depth image, and bottom, the human skeleton with hands detected).

B. User Study

Four types of user states have been considered for TV watching, i.e., *Absent*, *OtherAction*, *Controlling*, and *Watching*. A user study has been made to see these user states' frequency. In this study, 6 persons are considered: 2 young users (younger than 30 years), 2 middle-age users (between 31 years and 50 years), and 2 old users (older than 50 years). For each user, the actions during one hour (continuously) are recorded, and the corresponding user states are classified manually. The four user states' percentages corresponding to young, middle or old user are summarized and averaged, as shown in Fig. 11. As can be seen, during TV watching, old people often pay more attention (about 70% percents) to TV screen than young and middle user do, while they also leave the TV for other things for more time than other users. This is because old users prefer quiet environment, do only one thing at certain time, but may not sit for a long time due to health condition. On the other side, young and middle users often spend more time (more than 15% percents) to do something else during watching TV, e.g., Internet access with the tablet or pad, and message talking with the smart phone. For all the users, they spend little time (less than 10% percents) to control the TV, although young users may switch the channel frequently. This user study proves the reasonability to reduce the control cost automatically by considering of user behavior.

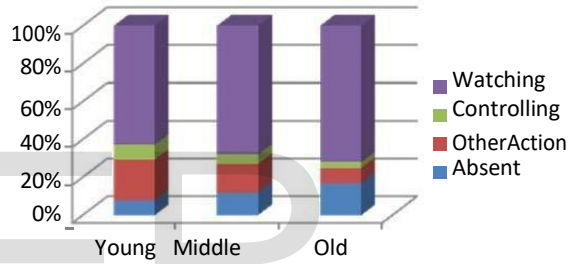


Fig. 11. The user study on practical user states (*Absent, OtherAction, Controlling*, and *Watching*) with respect to different years' old, e.g., young, middle, or old.

C. Automatic User State Recognition

Considering that the action detection algorithm decides the user's action based on the range of motion during certain time, the parameters, P, R, K and Q in (2), (3), (4) and (6), can thus be determined. P is the sampling interval generally no less than 0.1 second which is the delay human eyes can distinguish. K is the length of action sequence, and KxP is the period during which the user's action can be decided. R is the threshold of motion amplitude used to decide the intentional action. Q ($0<Q\leqslant K$) is the threshold of continuous action used to decide that the action sequence is continuous enough for an effective action. T_m=50s is recommended for the inactive time interval of action detection, and T_f=10s for the idle time interval of hand gesture recognition. To get the optimal parameters, some tests are done for normal TV watching. In these tests, more than 100 times of user state transitions during 4 hours are considered. The correct action detection rate (Cadr) is measured with respect to different parameters. Here, Cadr is defined as the f-measure [23] of action detection by considering of both false positive errors and false negative errors. The relation between P, R, K, Q and Cadr are shown in Fig. 12. In the first experiment, the action decision period KxP is tested when R=0.1 and Q=K/2. As can be seen from Fig. 12(a), the correct action detection rate keeps nearly 100% as long as KxP ranges between 1.5 and 4 whenever P=0.05 or

P=0.2. It means that the action decision period is more sensitive than the sampling interval, and the TV user's action can be detected correctly based on 1.5 to 4 seconds' accumulations. In the second experiment, the intentional action threshold R is tested when KxP=2.4, P=0.2 and Q=K/2. As shown in the result in Fig. 12(b), the correct action detection rate keeps nearly 100% when R ranges between 0.07 and 0.17. It means that this group of action amplitude is suitable for deciding user's intentional actions. Taking (P=0.2, K=12, Q=6, and R=0.1) as the optimal parameters, 2 hours of TV user actions including about 50 state transitions are tested, and the correct action detection rate of about 100% is obtained.

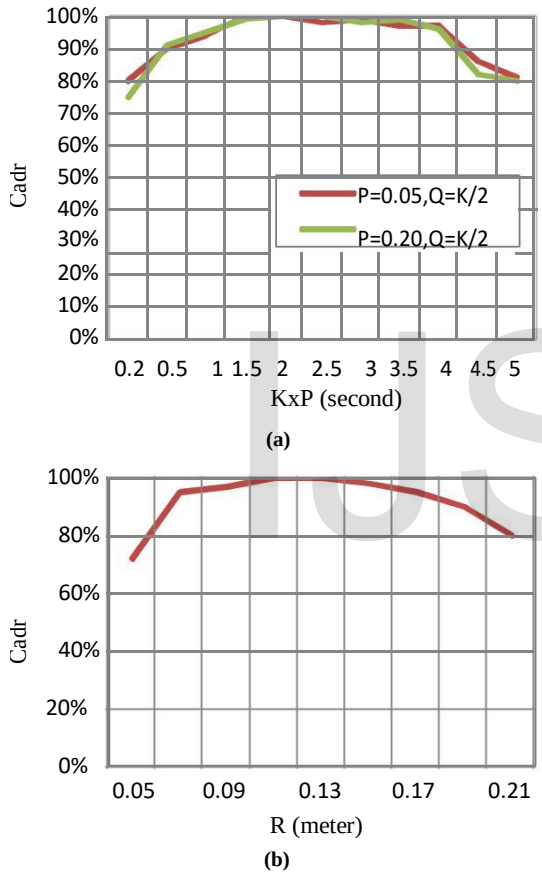


Fig. 12. The relation between the action decision period and the correct action detection rate (a), and the relation between the intentional action threshold and the correct action detection rate (b).

D. Computational Cost and Power Consumption

Compared with traditional TV control systems based on hand gestures [10][11][12], the proposed system introduces the AUSR component composed of such modules as action detection and presentation detection. AUSR determines whether the followed gesture recognition module will be activated or not. The percentages of running time in 6 hours for three computation modules are tested, as shown in Fig. 13. In the proposed scheme, the gesture recognition is replaced by action detection and presentation detection in most of the time (about 90%). Among them, the action detection (about 70% running time) is implemented by low-cost sensors. In the experiments, the whole TV control system's computational cost and power consumption are tested with and without the proposed low-cost scheme respectively. The computational cost is measured by averaged Central Processing Unit (CPU) running percentages, and all the operations related to TV control are counted. The considered power consumption includes the cost of devices and the one of embedded computation modules, while the TV set's cost is skipped. The energy consumption of TV control systems, with or without the proposed AUSR scheme, is tested. As shown in Table 3, by adopting the AUSR scheme, the TV control system saves about 70% of the computational cost and more than 35% of the power consumption. Especially for the systems based on RGB camera [10][11], more than 50% of power consumption is saved. It shows that the proposed AUSR scheme can reduce much computational cost and power consumption.

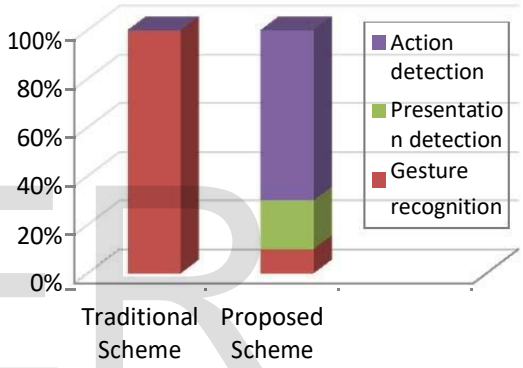


Fig. 13. Compare of the gesture recognition module's running time (measured by percentages) in different schemes. In the proposed scheme, the gesture recognition module keeps slept in most time.

TABLE III
COMPARISON OF COMPUTATIONAL COST AND POWER CONSUMPTION BETWEEN DIFFERENT TV CONTROL SCHEMES

Gesture recognition	Average computational cost (CPU running percentage)		Average power consumption saved (%)
	Traditional	With the proposed scheme	
Icon-based method [10]	41%	11%	56%
Motion-based method [11]	42%	12%	52%
Depth camera method [12]	58%	14%	35%

V. CONCLUSIONS

In this paper, an automatic user state recognition scheme is proposed to reduce free-hand TV control's computational cost and power consumption. In this scheme, the low-cost ultrasonic distance sensor array is constructed for detecting the user's intentional actions, the face detection method based on restricted face library is proposed to decide whether the user is present and actively watching TV, and the TV user's states are defined according to the finite-state machine (FSM). Based on the results of action detection and presentation detection, the user state is initialized and updated automatically. The camera

device and gesture recognition module is activated or closed with respect to the recognized user state. For example, only when the user state is *Controlling*, the hand gesture recognition module is activated, otherwise not. The implemented prototype and experimental results show that the proposed scheme reduces existing gesture control systems' computational cost and power consumption greatly. This will improve the practical usability of the hand gesture based TV control. In the future, this scheme can be extended to some other touchless TV control systems, such as the voice-based control system and so on.

REFERENCES

[1] I. S. Mackenzie and W. Buxton, "Extending fitts' law to two-dimensional tasks," in *Proceedings of CHI'92*, pp. 219-226, May 1992.

[2] S. K. Kim, G. H. Park, S. H. Yim, S. M. Choi, and S. J. Choi, "Gesture recognizing hand-held interface with vibrotactile feedback for 3D interaction," *IEEE Trans. Consumer Electronics*, vol. 55, no. 3, pp. 1169-1177, Aug. 2009.

[3] H. Heo, E. C. Lee, K. R. Park, C. J. Kim, and M. C. Whang, "A realistic game system using multi-modal user interfaces," *IEEE Trans. Consumer Electronics*, vol. 56, no. 3, pp. 1364-1372, Aug. 2010.

[4] D. W. Lee, J. M. Lim, S. W. John, I. Y. Cho, and C. H. Lee, "Actual remote control: a universal remote control using hand motions on a virtual menu," *IEEE Trans. Consumer Electronics*, vol. 55, no. 3, pp. 1439-1446, Aug. 2009.

[5] R. Aoki, M. Ihara, A. Maeda, M. Kobayashi, and S. Kagami, "Expanding kinds of gestures for hierarchical menu selection by unicursal gesture interface," *IEEE Trans. Consumer Electronics*, vol. 57, no. 2, pp. 731-737, May 2011.

[6] Y. M. Han, "A low-cost visual motion data glove as an input device to interpret human hand gestures," *IEEE Trans. Consumer Electronics*, vol. 56, no. 2, pp. 501-509, May 2010.

[7] L. C. Miranda, H. H. Hornung, M. C. .Baranauskas, "Adjustable interactive rings for iDTV," *IEEE Trans. on Consumer Electronics*, Vol. 56, No. 3, pp. 1988-1996, August 2010.

[8] J.-S. Park, G.-J. Jang, J.-H. Kim, S.-H. Kim, "Acoustic interference cancellation for a voice-driven interface in smart TVs," *IEEE Trans. on Consumer Electronics*, Vol. 59, No. 1, pp. 244-249, February 2013.

[9] I. Papp, Z. Saric, N. Teslic, "Hands-free voice communication with TV," *IEEE Trans. on Consumer Electronics*, Vol. 57, No. 1, pp. 606-614, February 2011.

[10] W. T. Freeman and C. D. Weissman, "Television control by hand gestures," in *Proceeding of IEEE International Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, pp. 179-183, June 1995.

[11] S. Jeong, J. Jin, T. Song, K. Kwon, and J. W. Jeon, "Single-Camera Dedicated Television Control System using Gesture Drawing," *IEEE Trans. on Consumer Electronics*, Vol. 58, No. 4, pp. 1129-1137, November 2012.

[12] M. Takahashi, M. Fujii, M. Naemura, and S. Satoh, "Human gesture recognition using 3.5-dimensional trajectory features for hands-free user interface," in *Proceeding of ARTEMIS'10*, pp. 3-8, Oct. 2010.

[13] X. Liu, and K. Fujimura, "Hand gsture recognition using depth data," in *Proceedings of 6th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 529-534, May 2004.

[14] D. Ionescu, B. Ionescu, C. Gadea, and S. Islam, "An intelligent gesture interface for controlling TV sets and set-top boxes," in *Proceeding of 6th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pp. 159-164, May 2011.

[15] D. Lee, and Y. Park, "Vision-based remote control system by motion detection and open finger counting," *IEEE Trans. Consumer Electronics*, vol. 55, no. 4, pp. 2308-2313, Nov. 2009.

[16] S. Lenman, L. Bretzner, and B. Thuresson, "Using marking menus to develop command sets for computer vision based hand gesture interfaces," in *Proceeding of NordiCHI'02*, pp. 239-242, Oct. 2002.

[17] P. Premaratne, and Q. Nguyen, "Conumer electronics control system based on hand gesture moment invariants," *IET Computer Vision*, vol. 1, no. 1, pp. 35-41, March 2007.

[18] N. Henze, A. Locken, S. Boll, T. Hesselmann, and M. Pielot, "Free-hand gestures for music playback: deriving gestures with a user-centered process," *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, no. 16, Dec. 2010.

[19] M. Chen, L. Mummert, P. Pillai, A. Hauptmann, and R. Sukthankar, "Controlling your TV with gestures," In *Proceeding of International Conference on Multimedia Information Retrieval (MIR)*, pp. 405-408, March 2010.

[20] M. V. Bergh, and L. V. Gool, "Combining RGB and ToF cameras for real-time 3D hand gesture interaction," in *Proceeding of IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 66-72, Jan. 2011.

[21] A. Wilson, and N. Oliver, "GWindows: robust stereo vision for gesture based control of windows," in *Proceeding of ICMI'03*, pp. 211-218, Nov. 2003.

[22] Y.-W. Bai, L.-S. Shen, Z.-H. Li, "Design and implementation of an embedded home surveillance system by use of multiple ultrasonic sensors," *IEEE Trans. on Consumer Electronics*, Vol. 56, No. 1, pp. 119-124, February 2010.

[23] S. Lian, W. Hu, X. Song, and Z. Liu, "Smart Privacy-Preserving Screen Based on Multiple Sensor Fusion," *IEEE Trans. on Consumer Electronics*, Vol. 59, No. 1, pp. 136-143, February 2013.

[24] H. C. Lee, D. T. Luong, C. W. Cho, E. C. Lee, K. R. Park, "Gaze tracking system at a distance for controlling IPTV," *IEEE Trans. on Consumer Electronics*, Vol. 56, No. 4, pp. 2577-2583, November 2010.

[25] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, vol. 1, pp. 511-518, 2001.

[26] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," In *Computational Learning Theory: Eurocolt '95*, pages 23-37. Springer-Verlag, 1995.