스마트 홈에서의 TV 제어 시스템을 위한 손 제스처 인식 방법

김대환° 조상호 천영재 김대진
포항공과대학교 컴퓨터공학과 지능형미디어 연구실
{msoul98°, scho, yjcheon, dkim}@postech.ac.kr

A hand gesture recognition method for an intelligent smart home TV remote control system

Daehwan Kim° Sang-ho Cho, Young-jae Cheon, Daijin Kim
IM Lab, Dept. of Computer Science & Engineering, POSTECH

Abstract

This paper presents a intuitive, simple and easy smart home TV remote control system using the hand gesture recognition. Hand candidate regions are detected by cascading policy of the part of human anatomy on the disparity map image. Exact hand region is extracted by the graph-cuts algorithm using the skin color information. Hand postures are represented by shape features which are extracted by a simple shape extraction method. We use the forward spotting accumulative HMMs for a smart home TV remote control system. Experimental results show that the proposed system has a good recognition rate of 97.33 % for TV remote control in real-time.

1. Introduction

Recently, it has researched a part of human-computer interface (HCI), specific machinery [1] [2]. For convenient environment, it is researching about smart home to be automatic. The present situation is manual, physical control using remote controller and etc. But, we want to have a simple manner for controlling equipments. That characteristic must be indirect, simple. HCI such as face, fingerprint, speech, gesture recognition can be a kind of that. Among those things, for using in smart home, gesture is the most intuitive, convenient. We focus on using the hand artificial gesture recognition in this work.

A hand gesture recognition in real time has been studied more studied than 10 years ago. Various technologies have been developed to recognize the hand gesture all the while. The concerns of the hand gesture recognition are to detect where the hand is and to recognize what the gesture is.

Shin, Lee, and et al. [3] presents the method that can adaptively obtain the hand region in the change of lighting or individual's difference. It is obtained by measuring the entropy from the color and motion information between continuous frames. Tanibata, Shimada and Shirai [4] proposed a method to get hand features from input images. This method use the color information and template matching to extract the hand and face. To recognize the gesture, many researchers have used the HMM because it can model the spatial and temporal characteristics of gestures effectively. Lee and Kim [5] proposed an HMM based threshold model that computed the likelihood threshold of an input gesture pattern and could spot the start and end points by comparing the threshold model with the predefined gesture models. Deng and Tsui [6] proposed an evaluation method based on HMM for gesture patterns that accumulated the evaluation scores along the input gesture pattern. Song and Kim [7] proposed a forward spotting scheme that performs gesture segmentation and recognition at the same time.

There are several applications that applies the hand gesture recognition to the TV remote control. Freeman and Weissman [8] developed the television control system by the gesture recognition of the open hands. They use the normalized correlation of templates to analyze the hand. Bretzener and et al. [9] presents algorithms and a prototype system for hand tracking and hand posture recognition. They used hierarchies of multi-scale color image features at different scales.

We propose a intuitive, simple and easy TV remote control system using the hand gesture recognition. Our system consists of two steps. The fist is to detect the hand using cascading policy, color

information and graph-cuts algorithm. The second is to recognize the sequence of hand gesture using a forward spotting scheme. Figure. 1 shows the flow chart of our system.

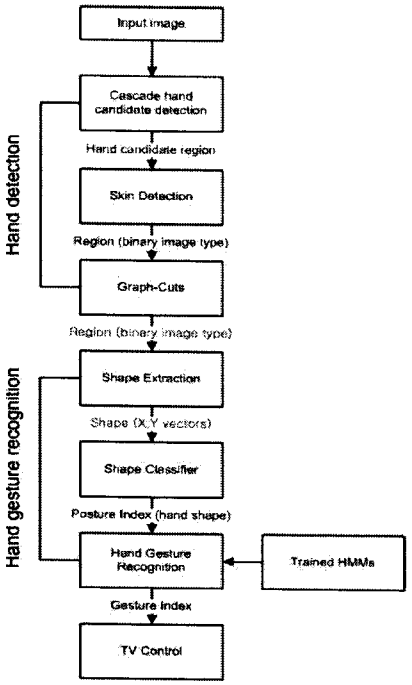


Fig. 1. the TV remote control system

2. Hand detection

The human hand is a non-rigid object which has five fingers. It has a various form and shape with different views and is very difficult to be found. So other hand detections except using multi-cameras are almost view-dependent. Our system is also view-dependent because of using single stereo camera. However this problem should be overcame according to the objective of each application.

Our system assumes that the TV watching environment looks at each other and the mechanism for controlling the TV like remote control gesture toward it. We think that this assumption is very natural for our application. Figure. 2 shows the structure of the TV watching environment.

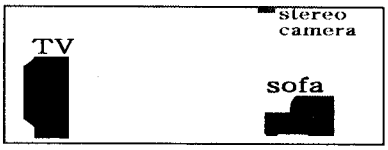


Fig. 2. the TV watching environment

2.1 Cascade hand candidate region detection

We find the hand candidate region as cascading the part of human anatomy. The first is to observe the head of the human. The second is to detect whether there is the arm or not. The last is to designate a rough hand region. This cascade procedure is accomplished at the disparity map.

First, we project the disparity image to the depth axis. Second, We apply the depth window mask to the projected disparity map. Because of being the camera at the ceiling, the maximal disparity regions are candidates of the human head. The method of observing the head is to measure how to be similar with circle. For verifying whether each maximal disparity region is the head or not, we analyze the rate of the first eigenvalue and the second eigenvalue. If the rate is close to 1, this shape is similar with circle, If not, it will be similar with a stick shape. In like previous manner, we find an arm region in each disparity regions smaller than the head disparity region. The arm region should be similar with a stick shape. Finally, we regard a hand candidate region as the end of the arm region. Figure. 3 shows the cascade hand candidate region detection process.

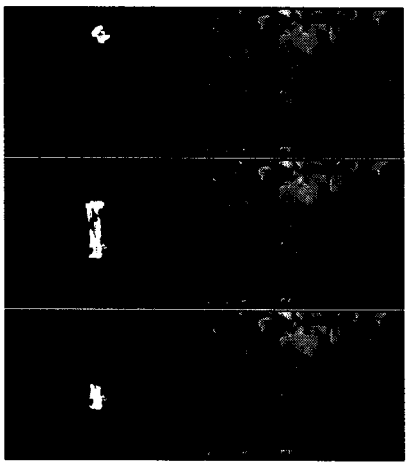


Fig. 3. the cascade hand candidate region detection, 1st row : head region, 2st row : arm region, 3rd row : hand candidate region

2.2 Hand detection using skin color and graph-cut

For exactly detecting the hand, we use the skin color and graph-cut technique. Because the region detection using just skin color has a illuminant variation problem, it is incongruent to extract a

shape of the region. We use the YCbCr color space for detecting the skin color region. The YCbCr color space is less sensitive to the illumination than the RGB space. Because the luma component is also dependent on the illumination, we just use Cb, Cr chroma components among Y, Cb and Cr. The skin color classifier is modelled by a unimodal Gaussian [10] per each Cb, Cr component. In the training step, the hand skin pixel values of each component are obtained manually.

$$p(X_{cb}|skin_{cb}) = g(X_{cb};m_{cb}, C_{cb})$$
$$p(X_{cr}|skin_{cr}) = g(X_{cr};m_{cr}, C_{cr})$$

where $g(X_{cb};m_{cb}, C_{cb})$, $g(X_{cr};m_{cr}, C_{cr})$ is the Gaussian distribution of pixel values of color component Cb, Cr with m_{cb}, m_{cr} mean of the distribution, and C_{cb}, C_{cr} covariance of the distribution.

We apply a simple threshold to the multiplication value of the each skin component's conditional pdf. Intuitively, the multiplication of two pdfs will be the mean-centered pdf. So such a distribution has more separable capabilities.

$$P(X_{cb}|skin_{cb}) \times P(X_{cr}|skin_{cr}) \geq \tau$$

where τ is threshold.

The detection result using just color model is almost on the illumination limitation. However it is good to use by means of the subsidiary information. So we need another method for segmenting the more accurate hand region as utilizing color information.

The graph-cuts algorithm [11] [12] is used to fid the globally optimal segmentation of the image. The obtained solution gives the best balance of boundary and region properties among all segmentations *satisfying the color constraints.*

We use the graph cuts algorithm suggested by Boykov and Funka-Lea [11]. We apply this algorithm to each component of YCbCr. We consider a set of pixels P in the input image, and all unordered pairs of neighboring pixels of that pixels defie N . We use a standard 8-neighbor system. A vector $L = (L_1, \cdots, L_p, \cdots, L_{|P|})$ describes labels to pixels p in $|P|$. The labels identify pixels to the hand and background. Each region of the hand and background is obtained by the hand skin classifier at the previous step. The color values of pixel p specify a vector I_p. The vector $I_p=(Y_p,B_p,R_p)$ defines each color components -Y, Cb, Cr - of the pixels.

Then, we try to minimize the Potts energy.

$$E(L) = \lambda \cdot D(L) + V(L)$$

$D(L)$ is a regional term, and $V(L)$ is a boundary term. This two terms define that

$$D(L) = \sum_{p \in P} D_p(L_p)$$
$$V(L) = \sum_{(p,q) \in N} K_{(p,q)} \cdot T(L_p \neq L_q)$$

The coefficient λ controls the importance of the regional term $D(L)$ and the boundary term $V(L)$ in the energy function.

The regional term $D(L)$ notifies the assignment penalties for respective pixels p in the set. We regard the hand and background as 'h' and 'b'. The individual penalty of pixel p about the object is $D_p('h')$ and calculated as :

$$\begin{aligned} D_p('h') &= -\ln \Pr(I_p|'h') \\ &= -\ln \Pr(Y_p|'h')\Pr(B_p|'h')\Pr(R_p|'h') \\ &= -\ln \Pr(Y_p|'h') - \ln Pr(B_p|'h') - \ln Pr(R_p|'h') \end{aligned}$$

The pixel about the background is $D_p('b')$ and also calculated as the same way above.

$$D_p('b') = -\ln Pr(Y_p|'b') - \ln Pr(B_p|'b') - \ln Pr(R_p|'b')$$

$\Pr(I_p| \cdot)$ could be known from the previous Gaussian modeling. The boundary term $V(L)$ specifies the boundary penalty of labeling L. $V_{p,q}$ means the penalty for the discontinuity between the point p and q. $V_{p,q}$ is bigger, the value of two points is more similar.

$$V_{p,q} \propto \exp(-\frac{(I_p - I_q)^2}{2\sigma_{p,q}^2}) \cdot \frac{1}{dist(p,q)}$$

The result of calculation of (Ip− Iq)2 is obtained as the calculation of a square of the Euclidean distance between the two vectors, Ip and Iq. σ is estimated as camera noise.

(a) original image (b) Extracted area by a hand skin color (c) Apply the graph cut algorithm

Fig. 4. Result of hand detection

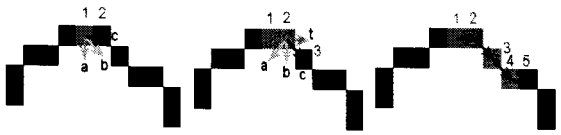
3. HAND POSTURE CLASSIFICATION AND GESTURE RECOGNITION

3.1 Hand shape extraction

We need to extract the hand shape from a segmented hand region in the previous section. First, we apply the canny edge detection technique to the hand region. Second, we obtain the sequence of pixels on the edge.

We now present a simple shape extraction method how to arrange that pixels. First of all, the centroid of the edge should be calculated, and the topmost point, p1, which has same horizontal position with the centroid. The point is regarded as a reference point.

Then we find 8 nearest neighbors from p1and should decide a next pixel p2 on the clockwise manner. 1) The fist search direction is bottom and search the next point on the anticlockwise direction until a point is found. 2) Search a nearest neighbor until a point is found from one more rotated direction than the opposite direction from the point found in the previous step to the current point. 3) Repeat step 2 until all pixels on the hand edge is ordered. 4) For matching between each shape, normalize it into the number of uniform points Figure. 5 shows the shape extraction process.



(a) First step (b) Second step (c) Iterated 5 times

Fig. 5. Result of hand detection

The extracted hand shapes have various scale, translation and rotation factors which are removed. The shape alignment process can eliminate the above factors. This process [13] is a general pre-processing step for matching between two shapes.

3.2 Hand posture classification and gesture recognition

We should classify a input hand shape into predefined postures. We select the nearest posture as measuring the distance between the input hand shape and each predefined hand posture. Each predefined hand shape is the mean of the same hand class shapes.

Main concern of gesture recognition is how to segment some meaningful gestures from a continuous sequence. Existing method use generally the backward spotting scheme that fist detects the end point, then do back-tracing to the start point. Song and Kim [7] introduced the forward spotting accumulative HMMs for solving the problem about the time delay between the gesture segmentation and recognition. This method is suitable for our real-time application system.

4. EXPERIMENTS

4.1 Experiment setup

We apply the TV control system to on/off power, up/down volumes and up/down channels. This system is composed of one TV at the front of an user and a Bumblebee stereo camera which attached to the ceiling.

In this work, fie gestures were defied for use as control commands. Figure. 6 displays the five gestures which are composed of nine postures.

Fig. 6. The hand gestures. 1st col : Power On/Of 2nd col : Channel Up, 3rd col : Channel Down, 4t col : Volume Up, 5th col : Volume Down

4.2 Hand gesture recognition

To perform the experiments, some modules had to be trained. First, we should obtain each mean of predefined hand shape postures. Second, we trained the six HMMs for fie gesture models and one non-gesture model using a set of training posture sequences.

The extracted hand shape data has a size of 160 dimensions, we apply PCA to the shape data to represent it in a reduced from using the basis vectors that were obtained from the aligned training sample set. The size of reduced dimension is 80. Then we are able to obtain the means of each hand data shape. The number of the mean is 9.

We tested total 75 gesture sequences where each 5 gesture consists of 15 gesture sequences. Table 1 summarizes the hand gesture recognition results. This table shows that the recognition accuracy of the proposed system is enough to use as the TV

remote control.

Table. 1. Gesture recognition accuracy of the TV remote control

Gestures	The number of test gesture sequences	Recognition rate (%)
1	15	15 (100.00%)
2	15	14 (93.33%)
3	15	15 (100.00%)
4	15	14 (93.33%)
5	15	15 (100.00%)
Total	75	73 (97.33%)

5. CONCLUSION

We present a intuitive, simple and easy smart home TV remote control system using the hand gesture recognition. For exactly detecting the hand, fist, we fid the hand candidate region by cascading policy of human anatomy and use the graph-cut algorithm to search the exact hand region using the skin color as reference information. We also applied the forward spotting accumulative HMMs to the hand gesture recognition for a smart home TV remote control system. The accuracy of our proposed hand gesture recognition system is good to use as the TV remote control.

Acknowledgments

This work was partially supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center (BERC) at Yonsei University. Also, It was financially supported by the Ministry of Education and Human Resources Development(MOE), the Ministry of Commerce, Industry and Energy(MOCIE) and the Ministry of Labor(MOLAB) through the fostering project of the Lab of Excellency.

REFERENCES

[1] T. Starner, J. Weaver, A. Pentland, Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video, IEEE Transaction on Pattern Analysis and Machine Intelligence, 20(12), 1371-1375, (1998).

[2] F. Quek, Toward a Vision-Based Hand Gesture Interface, Proc. of Virtual Reality System Technology Conf., 17-29, (1994).

[3] J. Shin, J. Lee, S. Kil, D. Shen, J. Ryu, E. Lee, H. Min, and S. Hong, Hand Region Extraction and Gesture Recognition using entropy analysis, International Journal of Computer Science and Network Security, 6(2), 216-222, (2006).

[4] N. Tanibata, N. Shimada, and Y. Shirai, Extraction of Hand Features for Recognition of Sign Language Words, Proceedings of the International Conference on Vision Interface, 15, 391-398, (2002).

[5] H. Lee and J. Kim, An HMM-based threshold model approach for gesture recognition, IEEE Transaction on Pattern Analysis and Machine Intelligence, 21(10), 961-973, (1999).

[6] J. Deng and H. Tsui, An HMM-based approach for gesture segmentation and recognition, Proceedings of the 15th International Conference on Pattern Recognition, 679-682, (2000).

[7] J. Song, and D. Kim, Simultaneous Gesture Segmentation and Recognition based on Forward Spotting Accumulative HMMs, Proceedings of the International Conference on Pattern Recognition, 1, 1231-1235, (2006).

[8] W. T. Freeman, and C. D. Weissman, Television control by hand gestures, IEEE International Workshop on Automatic Face and Gesture Recognition, (1995).

[9] L. Bretzner, I. Laptev, T. Lindeberg, Hand gesture recognition using multi-scale color features, hierarchical models and particle fitering, Proceedings of the fith IEEE International Conference on Automatic Face and Gesture Recognition 2002, (2002).

[10] S. Phung, A. Bouzerdoum, and D. Chai, Skin Segmentation Using Color Pixel Classification: Analysis and Comparison, IEEE Transaction on Pattern Analysis and Machine Intelligence, 27(1), 148-154, (2005).

[11] Y. Boykov, and G. Funka-Lea, Graph Cuts and Effiient N-D Image Segmentation, International Journal of Computer Vision, 70(2), 109-131, (2006).

[12] Y. Boykov, and V. Kolmogorov, An Experimental Comparison of Min-Cut/Max-Flow Algorithm for Energy Minimization in Vision, IEEE Transaction on Pattern Analysis and Machine Intelligence, 26(9), 1124-1137, (2004).

[13] T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, Active shape models their training and application, Computer Vision and Image Understanding, 61(1), 38-59, (1995).