A Vision Approach to Game Interface using Object History Images

Hyun Kang [1] and Keechul Jung [2]

[1] Electronics and Telecommunications Research Institute, Daejeon, South Korea
hkang@etri.re.kr

[2] School of Media, College of Information Science, Soongsil University, Seoul, South Korea
kcjung@ssu.ac.kr

Abstract. For a vision-based game interface, we introduce a simple and powerful representation of human gestures. The representation is an image that tells us where objects have occurred and how the objects have moved in an image sequences. As the objects are user's body parts (a head and both hands), this objects' information allow us to know what movements a user acts.

1 Introduction

A vision-based game interface interprets user's gestures as commands of a game using computer vision. In Sony's EyeToy^{TM}, while stands/sits in front of TV and camera, a user interact with the game using his/her gestures instead of mouse/joysticks [1].

Such video games, a gesture-based interface must be required to have very fast and reliable algorithms and also to meet economic constraints [2]. In this paper, gestures are generated by only a head and both hands in upper-body of a user. While the user of a video game gesticulates intuitively and clearly, well-defined gestures are used as inputs of the game.

Motion-Energy Images (MEI) and Motion-History Images (MHI) [3] are well known to recognize human movements. Bobick and Davis presented these view-specific representations of movements. The view-specific representation is useful to represent where body parts appear and how body parts move in given images.

Upgrading the works of Bobick and Davis, we represent gestures of upper-body using Object History Images (OHI). Information of objects such as a head and both hands more exactly describes what intention the user has that one of his/her motion in a image sequence. The objects are easily extracted by skin-color model [4].

As similar in MHI and MEI, we represent movements as OHI, which tell us where body parts have occurred and how the body parts have moved in an image sequence. Using Tangent Distance, clustering is performed in training data to obtain prototypes (or templates). Given a new image sequence, we recognize a movement as a gesture from prototypes.

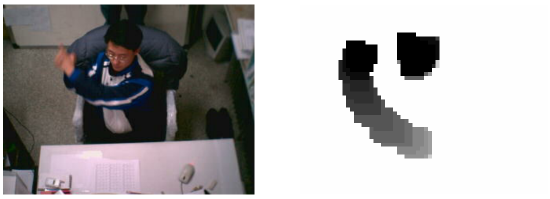
2 Object History Images

We use OHI to obtain pattern that represent where body parts appear in and how the body parts move in given image. Our goal is to construct a view-specific representation of gestures, which is a pattern that represents where body parts appear in and how the body parts moves.

OHI is a static vector-image where the vector value at each point is a function of the object properties at the corresponding spatial location in an image sequence.

$$O_\tau(x,y,t)=\begin{cases}\tau & if \quad S(x,y,t)=1\\ \max(0,\ O_\tau(x,y,t-1)-1) & otherwise.\end{cases}$$

We use $S(x,y,t)$ as an image has skin color regions so that make $H_\tau(x,y,t)$ as shown in Fig. 2 where τ = 15.



(a) the last frame of given an image sequence

(b) OHI of image (a)

Fig. 2. OHI

To detect regions from a given image, Park et al. [4] use skin color model in indoor scene. Fig. 3 shows the color distribution of human skins, obtained from 200 test images, in chromatic color space. The color distribution of human skins is clustered in a small area of chromatic color space and can be approximated by a 2D-Gaussian distribution. Therefore, the skin-color model is approximated by a 2D-Gaussian model $N(m, \Sigma^2)$, where the mean and variance are as follows.

$$m = (\bar{r}, \bar{g}) \text{ where } \bar{r} = \frac{1}{N}\sum_{i=1}^{N} r_i \text{, and}$$

$$\bar{g} = \frac{1}{N}\sum_{i=1}^{N} g_i \cdot \Sigma = \begin{bmatrix} \sigma_{rr} & \sigma_{rg} \\ \sigma_{gr} & \sigma_{gg} \end{bmatrix}$$

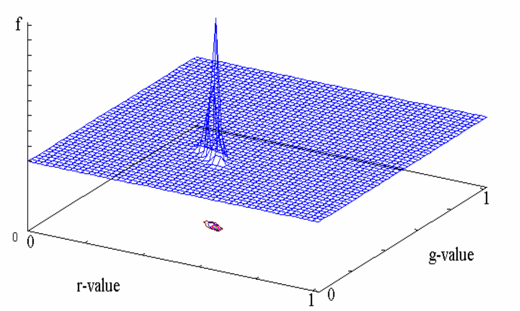


Fig. 3. Color distribution of human skins

Table 3.2 shows the mean and covariance matrix of the skin color model obtained from 200 sample images.

Table 1. Actual 2D-Gaussian parameters.

	$\bar{r}$	$\bar{g}$	σ_{rr}	σ_{rg}	σ_{gr}	σ_{gg}
Values	117.588	79.064	24.132	-10.085	-10.085	8.748

3 Gesture Recognition

To construct a recognition system, we need to define a matching algorithm for discriminations of OHIs. Because we are using an appearance-based approach, we must first define the desired invariants for the matching techniques.

Given an image, we classify a spatio-temporal pattern of the image into one of predefined gestures. To determine gestures, clustering method is used from training data that are collected as defined gestures.

We use *k*-means as clustering method because the *k*-means algorithm is very simple and works well in practice. It requires one to specify the number of classes *k*, where each class corresponds to a gesture of the user's upper body. It automatically generates clusters by minimizing the sum of squared errors from all patterns in a cluster to the center of the cluster, while initial cluster centers are chosen randomly from training points.

In our interface, clustering has two different stages: a training phase and a lookup phase. During the training phase, users generate different gestures in several times while capturing the training image sequence with a single camera. It constructs clusters that correspond to gestures of users. During the lookup phase, the interface estimates a user's poses by comparing the best matches of spatio-temporal patterns in given frame and clusters of gestures. In Fig. 4, gestures are shown as results of the k-means clustering.

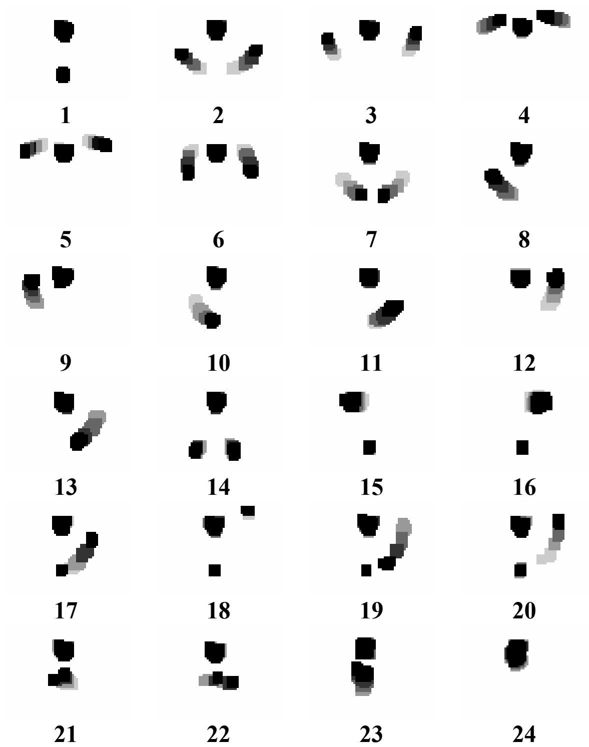


Fig. 4. Prototypes from training data.

For good results, choice of distance measure between two patterns is important. We use tangent distance [5] to guarantee local transformation invariance. Because we are using an appearance-based approach, we must first define the desired invariants for the matching techniques. Bobick and Davis use Hu moments to match instances of their representations. Hu moments is a shape descriptor which is known to yield reasonable shape description in a translation- and scale- invariant manner.

We do not look at shape information, but we are interested in relationship of objects appeared in an image. So, we use Tangent Distance which is known to recognize optical digit character recognition. Tangent Distance yields local transformation-invariant distance between two image patterns [5].

4 Experimental Results

The proposed gesture-based interface is applied to a first-person action game, Quake II. Overall system consists of one projector/monitor, one standard personal computer and one web camera with respects to layout A and B in described in chapter 3. To implemented software of the interface, we use OpenCV that means Intel® Open Source Computer Vision Library. It is a collection of C functions and few C++ classes that implement some popular algorithms of Image Processing and Computer Vision.

To applied to game, we use Quake II DLL (http://www.quake2.com/dll/) that is presented Dynamic Linked Library by Id Software® for the purpose to enhance ability of Quake II's game components. In the applied system, the interface has processing ability of 15 frames per seconds (fps). Although the definition of 'real-time' is not clear [6], we call our interface as the real-time interface since our interface can process present frame until arriving next frame. Fig. 5 shows the applied system and player's playing in Quake II.

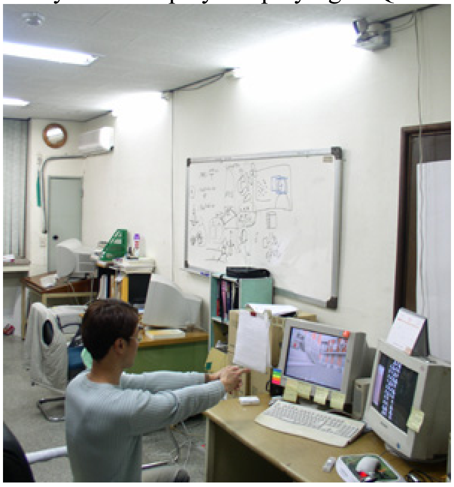


Fig. 5. Implemented Interface.

Table 2. Experimental results

Total	Correct	Miss	Results
518	493	36	95.17%

5 Conclusion

In this paper, we presented the gesture-based interface for video games. Real-time processing, accuracy, reliability, and chief equipments are needed for video games. For real-time performance of gesture-based interface for video games, we introduce a simple and powerful pattern that represents where objects appear and how the objects moves as an image – Object History Image.

Former researches in vision-based game interfaces have simple algorithm to recognize simple interactions due to limitation of real-time performance and chief equipments. We propose new representation of gestures that is simple and powerful, can make real-time interface using chief web camera, and apply to Quake II, famous first-person action games, with 10 gestures.

The proposed OHI is another representation of appearance-based approaches. With Tangent Distance the representation works well in recognition of manipulative gestures that is known as hard to be recognized using appearance based approaches. This may be updated one of appearance-based approaches.

We have plans that the interface has more powerful performances for more reliable interface for a game as followings: In various illuminations, adaptive skin color model will be researched.

References

[1] www,eyetoy.com

[2] William T. Freeman, David B. Anderson, Paul A. Beardsley, Chris N. Dodge, Michal Roth, Craig D. Weissman, and William S. Yerazunis, "Computer Vision for Interactive Computer Graphics," *IEEE Computer Graphic and Application,* vol. 18, pp. 42-53, 1998.

[3] A.F. Bobick and J.W Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, no. 3, pp. 257-267, March 2001.

[4] S. H. Park, E. Y. Kim, S. W. Hwang, Y. C. Lee, and H. J. Kim, "Face Detection for Security System on the Internet," *in Proceedings of IEEE International Conference on Consumer Electronics,* pp. 276~277, June, 2001.

[5] P. Y. Simard, Y. A. L. Cun, J. S. Denker, and B. Victorri, "Transformation Invariance in Pattern Recognition - Tangent Distance and Tangent Propagation," In: Orr, G. B. Muller, K-R(eds.): *Neural Networks: Tricks of the Trade*, Chapter 12. Springer, (1998)

[6] Thomas B. Moeslund and Erik Granum, "A Survey of Computer Vision-Based Human Motion Capture," *Computer Vision and Image Understanding,* vol. 81, pp. 231-268, 2001.