HUMAN COMPUTER INTERACTION USING VISION TECHNIQUES

SURYA DEVARAJAN
Dept. of Information Technology
IIITM-Kerala
Trivandrum, India
suryamaliyackal@gmail.com

ANNA ROSELU JOSEPH
Dept. of Information Security
IIITM-Kerala
Trivandrum, India
annaroselu91@gmail.com

Abstract—Now a days, the role of computer in daily life is increased, and there by HCI (Human Computer Interaction) is also a growing technology. This paper is a study on various advanced recognition techniques of HCI using computer vision techniques. Computer vision techniques have been widely applied to immersive and perceptual human–computer interaction for applications like computer gaming, education, human action recognition and entertainment. In this paper we discuss computer vision techniques in computer gaming and human action recognition. In vision, relevant techniques are surveyed in terms of image capturing, normalization, motion detection, tracking, feature representation and recognition. In Human action recognition, we discuss the human pose estimation and multiview human action recognition.

Keywords— Normalization, artificial retina chip, tracking and positioning

I INTRODUCTION

Human-computer interaction (HCI) is a fundamental function and problem for efficient communication between users and machines, where various techniques and devices have been developed to fulfill this requirement. In recent years, there has been a trend to combine vision technologies with computer games to develop immersive and perceptual HCI. Through automatic analysis and capturing user's intensions and commands, these applications provide a friendly and natural user interface for intelligent gaming experiences. As a result vision-enabled HCI including object tracking, gesture recognition, face recognition and facial expression recognition have been widely applied in computer gaming. In this paper computer vision enabled HCI techniques in computer gaming applications are reviewed from two viewpoints one is how they are applied in different stages of HCI and other is how the overall game is designed when apply mission techniques.

"Looking at People" is a promising field within computer vision with many applications. Most rely on pose estimation and recognition. It is therefore interesting to get an overview of recent progress in these fields including how the different methods compare. In recent years a wide range of application using 3D human pose estimation and activity recognition has been introduced among those several key applications are, Advanced HCI, Assisted living, Gesture-based interactive games, intelligent driver assistance system, Movies, 3D TV and animations, Physical therapy etc. In this paper Section 2 is dealing with vision HCI in Computer Gaming and Section 3 is dealing with human action recognition.

II HUMAN COMPUTER INTERACTION IN GAMING

In computer gaming the the whole logic of the process can be regarded as a series of interactions between internal objects and external users, where users can control objects' motion and responses . This forms a real word of users and a virtual world in the computer games. MMI techniques can be applied to achieve at least three targets: i) controlling the moving of game objects ii) controlling the actions/responses of the game object iii) combining scenes of the real world with that of the virtual one for immersive gaming experiences.

Figure 1 illustrates a diagram of vision based MMI for HCI in computer gaming applications, where the MMI helps to convert user actions to game controls with the constraints of game logics. Consequently, game state will be updated in response to these commands. Then, the game will wait for user's new commands from MMI, and this process will loop until the end of the game. The Vision MMI contains six main function blocks, i.e. capturing images, normalization, motion detection, feature representation, tracking and positioning, and recognition. Game controls can be results obtained from tracking and positioning, and/or recognized gesture and facial expressions. Technical detail of vision techniques applied in MMI for gaming applications are discussed in accordance with diagram in Fig. 1.

A. Capturing Images and Motion Data

Depending on the requirements of the game, there are several ways to capture the scene images as well as motion data of the real world. One is the use of cameras, where a single camera is useful in 2-D motion detection and positioning. For 3-D positioning, two or more cameras are desired; such as stereo camera pairs used in [5 and 6] and multiple cameras used in [8]. Among these cameras, some of them are web-cams, which reduce the cost of the systems. On the contrary, special cameras are utilized in other applications such as industrial cameras in IEEE 1394 cameras and CCD cameras. Special equipment namely artificial retina chip is even employed in some application for efficiency. This certainly will bring additional cost to users and limit the applications of the associated games.

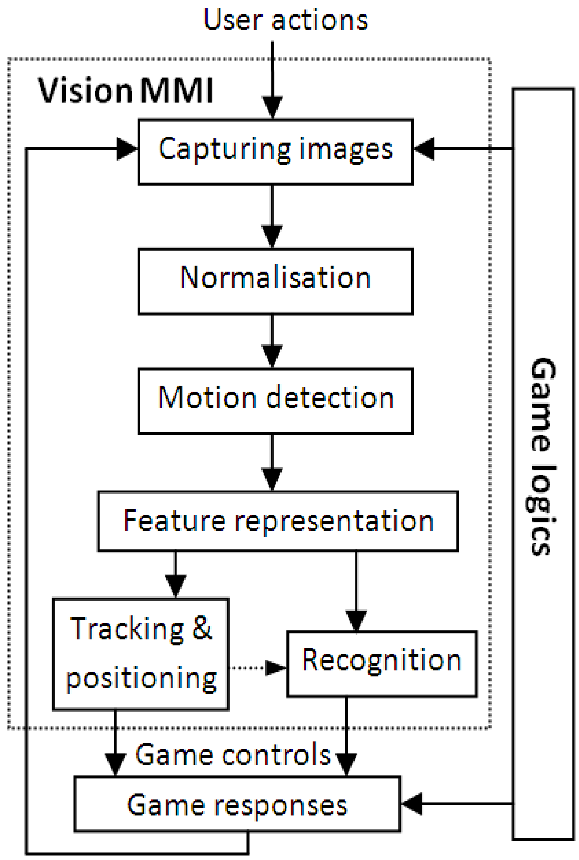

Figure 1: Diagram of vision based MMI interface in computer gaming applications.

B. Normalization

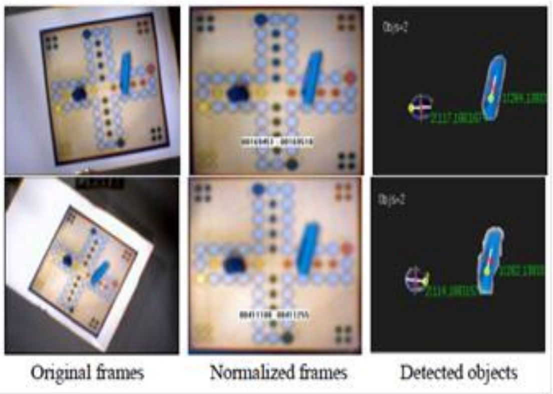

Figure 2: Examples of spatial and illumination normalization in board games for robust moving object detection.

With captured images, pre-processing like normalization is necessary for consistent measurement to deal with the changes in illumination and spatial coordinates. For spatial normalization, calibration is commonly used, especially for 3-D gaming with multiple cameras. For applications with a single camera, geometric warping is often adopted following detected corner points, such as the bilinear transform used in vision-based board-games where changes in terms of camera position and board location are involved.

C. Motion Detection

Motion detection also refers to motion segmentation, which aims at extraction and segmentation of moving or changing objects in the scene. Vision based automatic motion detection is still desirable as it has no constraints for general applications. There are three main techniques used for vision-based motion detection, which include background subtraction, image differencing and optical flow based approaches.

Background subtraction often applies to the cases when the camera is fixed. Firstly, an object-free background image is obtained. Then, scene changes can be determined as the difference between scene image and the background. For its simplicity, image differencing is widely utilized in motion detection. Firstly, the difference of two frames is obtained as a difference image. Then, pixels whose values exceed a predefined threshold are labelled as foreground ones. Optical flow is a 2-D motion field, which is estimated via optimal determination of pixel shifts between two images.

D. Feature Representation

When blobs of moving objects regions are determined via motion detection, several features can be extracted from each blob for further tracking and recognition. Three main categories of features that can be derived are color, shape and motion relevant measurements.

Regarding color features, color histogram and dominant color are usually utilized. Color histogram has been applied in detection of hand and face and recordings of game play in HSV and RGB spaces, respectively. Shape is another very popular feature for blobs, which can be represented by the orientation of main axis, location of centriod, moments, size (2-D area and 3-D volume) and bounding box. More importantly, specific shape modeling is employed for the determination of hands, fingers, face, nose tip and shape from silhouette.

Velocity and orientation is often used for motion features, which respectively measure the magnitude and phase of the corresponding motion vector.

E. Tracking and Positioning

Tracking in vision games usually refer to continuous positioning of human body parts, which can be obtained via analysis of motion detection results. When human body parts or other content of interest are detected, their spatial locations are determined and used for positioning. The body parts and objects used include feet, hands, fingers, face/head, wrists, and external objects like game pieces and markers. Accordingly, appearance-based modeling of these specific body parts and objects are needed for their accurate detection. In addition, facial component like eyes, nose, even lips, thumb, and chin, can be used for tracking, and a comprehensive comparison suggests that tracking of nose tip seems more reliable against lighting changes. To simplify the difficulty in accurate location of human body parts, controlled environment is used.

F. Recognition and Interpretation

Although heuristic approaches can be applied for gesture recognition using thresholding and rule- based reasoning, Hidden Markov Model (HMM) is widely employed for this purpose. The reason behind is that HMM is capable of statistically modeling multi-state temporal sequence, and a recognition rate of 98% for 40 sign language gestures can be achieved in a lab environment. Other approaches used for gesture recognition include artificial neural network, moment or optical-flow based shape recognition and example-based clustering.

III HUMAN ACTION RECOGNITION

In this section we discussed about human action recognition, this is explained under two topic, i) 3D Human Pose Estimation and ii) Multi-view Human Action Recognition. While 2D human action recognition has received high interest during the last decade, 3D human action recognition is still a very unexplored field. Human actions performed in real 3D environments; however, traditional cameras only capture the 2D projection of the scene Vision-based analysis of 2D activities carried out in the image plane will therefore only be a projection of the actual actions. As a result, the projection of the actions will depend on the viewpoint, and not contain full information about the performed activities. To overcome this shortcoming, the use of 3D representations of reconstructed 3D data has been introduced through the use of two or more cameras. The use of 3D data allow for efficient analysis of 3D human activities.

A. 3D Human Pose Estimation

As mentioned earlier in this section we focus on model-based approaches using multi-view video input and aim to extract real 3D posture. Fig. 3 shows common steps in model-based approaches for human pose estimation using multi-view input including: camera calibration/data capture, voxel reconstruction, initialization/segmentation (segment voxel data into different body parts), modeling/estimation (estimating pose using the current frame only), and tracking (use temporal information from the previous frames in estimating body pose in the current frame). In each step, different methods may have different choices of approaches: There are methods using 3D features (e.g., voxel data) reconstructed from multiple views while others may still use 2D features (e.g., silhouette, contour) extracted from each view. They may have manual or automatic initialization step. Some methods may not have tracking step. Some methods are for a generic purpose while others are application specific for efficiency. In the following section we will discuss the factors mentioned above.

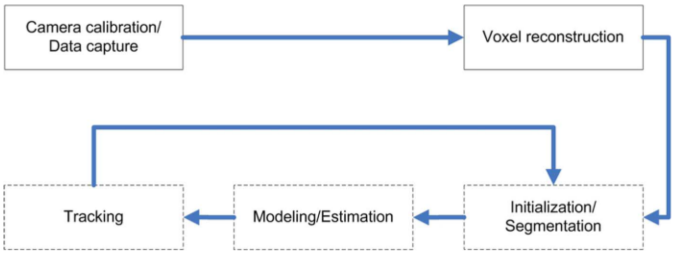

Figure.3. Common steps in model-based methods for articulated human body pose estimation using multi-view input. Dashed boxes mean that some methods may or may not have all of these steps.

B. Using 2D Versus 3D Features From Multi-view

Since the real body pose is in 3D, using voxel data can help avoiding the repeated projection of 3D body model onto the image planes to compare against the extracted 2D features. Furthermore, reconstructed voxel data help to avoid the image scale issue. These advantages of using voxel data allow the design of simple algorithms and we can make use of our knowledge about shapes and sizes of body parts.

C. Tracking-Based Versus Single Frame-Based Approach

The modelling and tracking steps can be considered as a mapping from input space of voxel data Y and information in the predefined model (e.g. kinematic constraints) C to the body model configuration space Ө:

$$M : (Y , C) \Rightarrow \Theta \quad (1)$$

The body model configuration contains both static parameters (i.e., shape and size of each body component) and dynamic parameters (i.e., mean and orientation of each body component), in which the static parameters are estimated in the initialization step. Methods are different in the way they use and implement the mapping procedure M.

D. Manual Versus Automatic Initialization

Some methods have automatic initialization step like while others require *a priori* known or manually initialized static parameters. In some methods, the specific shape and size of the head was used to design a hierarchical growing procedure for initialization. In another method, a database of human body shapes was used for initial pose-shape registration. In , the user was asked to start at a specific pose (e.g., stretch pose) to aid the automatic initialization.

E. Generic Purpose Versus Application Specific approaches for efficiency

Depending on applications, human pose tracking may focus on different body parts including full body pose; upper body pose, hand pose, and head pose. Due to the complexity of human body pose estimation task, there are tradeoffs between developing a generic approach versus an approach integrated to some specific cases for efficiency.

IV MULTI-VIEW HUMAN ACTION RECOGNITION

In this section we give an outline of approaches which solely apply 2D multi-view image data, and 3D-based techniques.

A. 2D Approaches

Some work concentrates solely on the 2D image data acquired by multiple cameras. Action recognition can range from pointing gesture to complex multi-signal actions, including both coarse level of body movement and fine level of hand gesture. Some authors perform action recognition using motion features or a combination of static shape and motion features from image sequences in different viewing angles. Some method apply principal component analysis (PCA) of optical flow velocity and human body shape information, and then represent each action using a set of multi-dimensional discrete hidden Markov models (HMMs) for each action and viewpoint. Some researcher proposed a method for multi-user, prop-free pointing detection using two camera views. The observed motion are analyzed and used to refer the candidates of pointing rotation centres and then estimate the 2D pointer configurations in each image. Based on the extrinsic camera parameters, these 2D pointer configurations are merged across views to obtain 3D pointing vectors.

B. 3D Approaches

Another line of work utilizes the full reconstructed 3D data for feature extraction and description. A number of other recent 3D approaches are there. Approaches which use 3D shape and pose features are, Johnson and Hebert proposed the spin image, and Osada *et al.* the shape distribution. Ankerst *et al.* introduced the shape histogram, and Kazhdan *et al.* applied spherical harmonics to represent the shape histogram in a view-invariant manner. Later Huang *et al.* extended the shape histogram with color information.

V FUTURE DIRECTIONS AND RECENT WORKS

Computer vision techniques have been successfully applied in computing gaming in many applications. According to their characteristics, these games can be classified into several categories such as, vision enabled pointing and positioning, vision for manipulating object. For efficiency and correctness, Internet-based network gaming and error anti-cheating scheme are emphasized.

Although vision-based interface facilitates more natural and friendly HCI while controlling the game, some issues need to be fully addressed before migrating the relevant systems from lab to real applications. This is mainly due to the limitations of vision techniques used, where immature approaches may constrain such migration especially in robust and efficient detection, tracking and recognition of user's motion and intension under an unconstrained environment. Consequently, to solve these challenging problems will no doubt be of interest as future directions for further investigations.

A multi-level human body pose estimation system, such
as: combined information from different level of details is more useful (e.g., in intelligent environment, the combination of body pose, hand pose, head pose would give better interpretation of human status/intention); Information from different levels can complement each other and help to improve the estimation performance. However, typical approaches in this area deal with each of these tasks (body pose estimation, hand pose estimation, head pose estimation, etc.), separately. Therefore, it is useful to conduct further studies to analyze the reasons to why typical approaches only deal with one task at a time, and find a way to achieve the goal of a full body model (e.g., including body, head, and hand). Another related open-ended research area that is important put more emphasis on, is the issue of pose estimation and tracking of multiple objects simultaneously.

VI CONCLUSION

Computer vision techniques have been successfully applied in computing gaming in many applications. Automatic detection and tracking of human body is a difficult task in real world even though it's easy in controlled environment. Therefore robustness is one of the first priorities in developing such systems. Usability and speed also plays key roles in for successful computer gaming. Hardware support including graphics processors (GPU) and special image processing devices like artificial retina chip as well as fast algorithms are necessary to fulfill these requirements.

Human pose recognition is also another HCI method. Using multiple cameras human pose is captured and it is used to further calculation to interact with computer. In recent years other prominent vision-based sensors for acquiring 3D data have been developed. ToF range cameras, which are single sensors capable of measuring depth information, have become popular in the computer vision community.

REFERENCE

[1] Jinchang Ren, Theodore Vlachos, and Vasileios Argyriou EImmersive and Perceptual Human-Computer Interaction Using Computer Vision Techniques, 2010 IEEE

[2] Michael B. Holte, Cuong Tran, and Mohan M. Trivedi. Human Pose Estimation and Activity Recognition From Multi-view vedios,2012 IEEE

[3] A. Jaimes and N. Sebe. Multimodal human-computer interaction: a survey. Computer Vision and Image Understanding. 108(1-2): 116-134, 2007.

[4] L. Szirmay-Kalos. Machine vision methods in computer games, 2009. KEPAF Conf. Image Analysis and Pattern Recognition.

[5] D. Brehme, F. Graf, F. Jochum, et al. A virtual dance floor game using computer vision. 2006. Proc. 3rd European Conf. Visual Media Production (CVMP), 71-78, London.

[6] F. Sparacino, C. Wren, A. Azarbayejani, and A. Pentland. 2002. Browsing 3-D spaces with 3-D vision: body-driven navigation through Internet city. Proc. 1st Int. Symposium on 3D Data Processing Visualisation and Transmission (3DPVT), Padova, Italy.

[7] M. Betke, J. Gips, and P. Fleming. The camera mouse: visual tracking of body features to provide computer access for people with severe disabilities. IEEE Trans. Neural Systems and Rehabilitation Engineering. 10(1): 1-10, 2002.

[8] C. Tran and M. Trivedi, “Introducing XMOB: Extremity movement observation framework for upper body pose tracking in 3D,” in *Proc.IEEE Int. Symp. Multimedia*, 2009, pp. 446–447.

[9] T. Moeslund, A. Hilton, and V. Krüger, “A survey of advances in visionbased human motion capture and analysis,” *Comput. Vis. Image Understand.*, vol. 104, no. 2–3, pp. 90–126, 2006.

[10] D. Weinland, R. Ronfard, and E. Boyer, “Free viewpoint action recognition using motion history volumes,” *Comput. Vis. Image Understand.*, vol. 104, no. 2, pp. 249–257, 2006.