
Mahsa T. Pourazad, Student Member, IEEE, Ali Bashashati, and Panos Nasiopoulos, Member, IEEE

Abstract—An efficient Random Forests (RF)-based method for estimating the depth of human-body gestures in real-time using a single video camera is proposed. The potential application of the proposed method is in computer video games that include a 3D gesture interaction system, where human-body movements are translated to computer commands and interact directly with virtual content. The main advantage of our approach is its simplicity and cost effectiveness, since it eliminates the need for designing and implementing depth sensing cameras or 3D video cameras on computers. Performance evaluations show that our approach results in estimating very realistic depth of human-body gestures on real time basis.

I. INTRODUCTION

Body gestures provide natural and device-free means for human-computer interactions. Many applications can be controlled more intuitively with body gestures rather than traditional controllers like a joystick, mouse or keyboard. For instance, in computer gaming, the gesture of the player’s body can be translated to computer commands and the content in virtual space can be maneuvered according to the user’s movements (Microsoft Natal project) [1]. This provides the player with a much more realistic gaming experience. In order to detect the human body gesture, the depth information of the scene is required. Solutions include the use of sophisticated and very expensive depth sensing cameras or highly demanding stereo cameras (in terms of cost, calibration and setup) [1,2]. Both approaches require complex implementation of new hardware, which in addition to design issues, will inevitably increase the cost of a computer unit (especially in the case of laptops). A much more attractive and cost effective approach would be estimating the depth information from existing single cameras, a cost-effective approach that will require the development of appropriate algorithms.

In this paper, we present an effective real-time technique to estimate the depth of human gestures from regular existing 2D video cameras. To estimate a very realistic depth map, our proposed scheme is designed based on the human visual depth perception mechanism, in a sense that it utilizes several monocular depth cues such as motion parallax, occlusion, texture, and edge information (perspective). In the proposed approach, these monocular depth cues are extracted from 2D video and are used by a Random Forests (RF) machine-learning algorithm, which once appropriately trained, estimates extremely accurate depth model of human body gestures in real-time.

II. PROPOSED SCHEME

In recent years, machine learning has received increasing attention as a tool for estimating depth maps [3, 4]. To this end, in our proposed study the Random Forest machine learning algorithm is utilized to construct a model using multiple monocular depth cues as input features for estimating the depth information of human body gestures (for details on the RF algorithm see [5]). In our study, RF is chosen, since as we show it is a powerful tool in integrating depth cues as decision trees (DTs) and finding a generalized depth model for unseen data. While individual monocular depth cues are not quite successful in predicting the depth of a scene, integration of these cues provides a much more accurate depth-map estimate.

In our proposed scheme, we train the RF model using some video sequences whose depth information is also available, and which are very carefully chosen to effectively represent the case under study, i.e., video games and body gestures. The following subsections elaborate on the different steps involved in our proposed scheme.

A. Extracting features representing depth cues

In our approach we estimate the depth value using cues such as motion parallax, texture variation, perspective, sharpness and occlusion. The rest of this subsection explains the procedure of extracting features that represent these cues.

1) Motion parallax: motion may be seen as a form of “disparity over time”, represented by the concept of motion field. In our study, in order to extract features representing motion depth-cue we implement a matching-block method for each 4x4 block between consecutive frames.

2) Texture variation: to capture texture information, we used 9 Laws’ texture energy masks [6]. Texture information is mostly contained within the frame’s luma information. Thus, to extract features representing this depth-cue we apply Laws’ texture energy masks to the luma information of each 4x4 block \(I(x, y)\) as:

\[
E_t(n) = \sum_{(x,y)\in Block} |F(x,y) * F(x,y)|^k \quad k \in \{1,2\}
\]

where \(F\) refers to each of the Laws’ texture energy masks. By applying the 9 masks to the luma component of each block using equation (1), we obtain a feature set that includes 18 features for each block within a frame (\(k\) can be 1 or 2).

3) Perspective (edge information): edge information of each frame is derived by applying the Radon Transform in 6 orientations with 30° intervals. Then, the amplitude and phase of the most dominant edge within a block are selected as features representing the block’s edge information.

4) Sharpness: we measure the sharpness of each 4x4 block based on the diagonal Laplacian method [7].

5) Occlusion: to capture the occlusion depth cue, multi-resolution hierarchical approach is implemented, i.e., we extract all the above mentioned feature sets for each 4x4 patch at three different image-resolution levels (1, 1/2, and 1/4). Extraction of feature sets at different resolution levels not only gives occlusion information but also guarantees that the extracted features are globally and locally accountable.

Taking into account all the above-mentioned extracted feature sets for each 4x4 block at each image-resolution level we can form a 22-dimensional feature vector representing...
local depth cues. Adding the occlusion depth cue, which involves 3 image-resolution levels, results in a 66-dimensional feature vector for each 4x4 block.

B. RF model estimation

In our proposed scheme, we manually identify key frames (shot transitions) from the training sequences. We select key-frames of these sequences and divide each frame into 4x4 blocks. Then a set of global and local features, which represents specific depth cues, is extracted at three image scales. In general each block within a picture contains pixels of one or more objects. To ensure the training set represents the depth cues of different objects more accurately, the chosen blocks include mostly the pixels of a single object (more than 8 pixels) rather than multiple objects (mean-shift image segmentation applied to separate objects [8]).

The RF-based depth model is estimated using the feature vectors of the selected blocks as input and their average ground-truth depth-value as output. Once the RF model is trained, the depth information of unseen video sequences - not used for training our model - is predicted based on their blocks’ feature vectors (monocular depth cues). The resulting depth information is block-based. To find the object-based depth map, we perform image segmentation on each video frame and then calculate the average depth value of all the pixels within each segmented object and assign that to all the pixels. We base this on the assumption that all the pixels of a segmented object have one depth value.

III. PERFORMANCE EVALUATION

A. Experimental data preparation

In our experiment we require true depth map of the scene for training the RF model and also for comparing the results. To remedy this problem we capture the test data with three full HD parallel cameras with baseline distance of 9cm. The middle view is used as our main experimental video sequence and the two other views are used only for estimating the true depth map of the scene by applying a block matching technique to the streams [9]. Please note the true depth map is required for training the RF model, and that is the reason we need to generate true depth maps at this point. Once this model is trained we are not required to have the depth map, except for comparing our results. We captured five video sets of different body gestures of five different people in different indoor environments (about 20s each). In these videos, the subjects were asked to pretend that they were playing different sports which involved a variety of gestures. The captured videos were rectified to ensure the alignment of the views. The middle view of these captured video sets and their estimated depth maps are used to train the RF model. The key frames for training are chosen such that the training set includes different human body gestures.

For evaluating the performance of our method, we captured two different video sets of two other people. The trained model is used to estimate the depth of the human body gesture of the middle views of these two multiview video streams. Once more, the left and right view were used for finding the true depth in these two cases for the purpose of comparing it with the depth map obtained from our model.

B. Experimental results

Fig. 1 shows a snapshot of the original 2D streams, the original (ground-truth) depth maps, and the estimated depth maps by our approach. As it can be observed, our method yields highly realistic depth information of human body gestures, which can be used as a controlling tool in 3D video games.

We also investigated the importance of different depth-cues in estimating the depth map while training our RF model. We found that in our experimental set, motion cue with the average measure of importance of 44.96% is the most important monocular depth cue for estimating the depth map model, human body gesture compare to texture (32.59%), perspective (13.49%) and sharpness (8.96%). Please note since the occlusion features also represent the global version of other features, their importance factor is embedded in all other depth cues. Considering that 3D computer video-gaming involves lots of motion from the players, the recommended technique will successfully find the depth information of the player’s body gestures.

IV. CONCLUSION

We present an efficient Random Forests (RF)-based method for estimating the depth of human-body gestures in real-time using a single video camera. Our method incorporates several monocular depth cues to approximate a very accurate depth estimate, thus eliminating the need for expensive and sophisticated depth or stereo video camera setups for computer-based 3D video game applications.

REFERENCES