Rehabilitation System in 3D Natural Scenes

Amin Safaei, Q. M. Jonathan Wu

Department of Electrical and Computer Engineering
University of Windsor
Windsor, ON, Canada
Email: {safaeia, jwu}@uwindsor.ca

***Abstract*—In this paper, we present a rehabilitation system for patients who suffer wrist injury. The idea to use computer vision for rehabilitation is not new; however, the method proposed in this paper differs significantly from previously proposed methods. We propose a 3D hand model evaluation method that can recognize soft and elaborate representations of hand motions. In practice, hand motion recognition in an unconstrained environment is a difficult task because of intra-class variation. It becomes more challenging when we lose depth data because of projection. However, the emergence of commercial depth sensors, such as Microsoft Kinect and SoftKinect, has overcome this issue. In previous work, we used the data of tip and joints, which was sufficient for simple motion; however, in complex motion, such as grabbing and rotation, it is not possible to track and estimate the depth of tips and joints. In this work, we modify the algorithm that is proposed by Rodriguez *et al* and Hadfield. Instead of using 2D data, we extend the method for 3D data, and for elevation information, Hidden Markov Model (HMM) is used.**

Keywords–Machine Vision; SoftKinetic; Motion Recognition; Video Processing; Rehabilitation.

I. INTRODUCTION

Rehabilitation has been emphasized recently in the field of computer vision. Rehabilitation is defined as a dynamic process that helps patients recover normal functional capability. To reach this milestone, it is necessary to monitor patient activity continuously and correct motions.

In this paper, we proposed a system to evaluate hand motions via depth data (3D) for rehabilitation. Given natural hand features and an uncontrolled environment, the proposed system classifies and differentiates unnatural slowness of motions.

To obtain the data, two main methods are available: sensor based and vision based. Sensor based methods use electromechanical or magnetic sensors to capture activity and then convert the motions to digital signals. The main drawbacks are that this method is expensive and requires calibration and a setup procedure. In contrast, vision based methods require only a camera. The advantages of this method are more natural, unencumbered, non-contact interactions, whereas the disadvantage of this method is requiring environments that are insensitive to lighting. Vision based methods can generally be categorized into appearance hand models (2D Mapping) and 3D hand models. Appearance hand models attempt to learn mapping from feature vectors, such as the positions of tips and joints of the hand. In contrast, 3D hand models rely on a 3D kinematic description of a hand. This method estimates depth data using either a stereo camera or 3D depth cameras (IR cameras). Using a stereo system for hand motion evaluation was tested in [1]; however, because of some issues, such as calibration and complexity, in this work, we used 3D depth cameras [2] [3] for ease of use and higher accuracy than that of the stereo system. To describe hand motions, we use the HMM, which is a state-based model that analyzes data and recognizes patterns. As a special case of human-computer interaction and rehabilitation, several constraints are imposed, which include complexities, background, variable lighting conditions, transforming gesture structures, real time implementation and dependency on user and device characteristics. The remainder of this paper is organized as follows. First, the proposed method for segmentation is discussed and then, the theory behind depth camera is described. Next, feature descriptors and classification method are explained and finally the proposed method for evaluation is discussed.

II. METHOD OF ANALYSIS

The proposed methodology is developed for single-hand motion recognition and evaluation. The procedure consists of segmentation, depth data, UV mapping, classification and recognition. In this section, the theory and implementation of the proposed method are described.

A. Segmentation

Segmentation is defined as the division of an image into different regions, each having different futures. In this work, we need to isolate the hand from the background. Our proposed methods for segmentation are K-means clustering and L*a*b* color space. The image is first transferred from RGB to L*a*b* color space, and then K-means clustering is used to isolate the hand from the background [4][5].

B. Depth Data

A 3D depth camera can provide 3D data using a low-cost CMOS pixel array with an active modulated light source. Its compact construction, ease of use and high accuracy and frame rate make it an attractive solution for a wide range of applications. A 3D depth camera operates by illuminating the scene with a modulated light source and observing the reflected light. In the proposed system, hand images and depth images are captured simultaneously and are then merged via UV mapping[6].

The letters U and V in a UV map denote the axes of the 2D texture because X, Y and Z are used to denote the axes of the 3D object in the model space. For any point P on the sphere, the unit vector from P to the sphere's origin can be calculated. Assuming that the sphere's poles are aligned with

 ISBN: 978-1-61208-422-0

the Y-axis, UV coordinates in the range of [0, 1] can then be calculated using (1).

$$u = 0.5 + \frac{\arctan2(d_z, d_x)}{2\pi};\ v = 0.5 - \frac{\arcsin(d_y)}{\pi} \quad (1)$$

C. 4D Feature Descriptors

In this work [1], we proposed a system that uses the position of joints and tips and then uses classification to recognize hand motions. The primary problem of this method is that it was not reliable for complex hand motions, such as grabbing and rotation. It was based on the assumption that all fingertips are visible and can be detected in the image, so it could not track the position of the all joints and tips when fingers occluded each other. To cope with this problem, two extended future descriptors are used. The first one is the extended method of Laptev et al., which provides a descriptor ρ of the visual appearance and local motion.

$$\rho(\upsilon, \nu, \omega) = (G(I(\upsilon, \nu, \omega)), F(I(\upsilon, \nu, \omega)), D(I(\upsilon, \nu, \omega))) \quad (2)$$

where G is a Histogram Oriented Gradient (HOG), F is Histogram Oriented Flow and D is Histogram Oriented Depth. A bag of words is employed on each ρ. Each ρ represents one type of hand pose. To cluster these poses, K-Means Clustering is performed on all with a Euclidean distance function.

The second descriptor is defined based on the algorithm that was developed by Oshin *et al*[7] and Hadfield[8], which is called the Relative Motion Descriptor (RMD). We extended the algorithm for pose estimation in 3D data. For each frame, we consider volume i_{xyz} with 3D dimensions. The sum of interest point s is defined based on the interest point detection and their strengths.

$$s(x, y, z, t, \tfrac{X}{\sigma}, \tfrac{Y}{\sigma}, \tfrac{Z}{\sigma}) = \sum_{x'=x}^{x+\frac{X}{\sigma}} \sum_{y'=y}^{y+\frac{Y}{\sigma}} \sum_{z'=z}^{z+\frac{Z}{\sigma}} \sum_{t'=t}^{t+\frac{T}{\sigma}} \iota(x', y', z', t') \quad (3)$$

where ι is the representation of the frame of the specific pose.

D. Classification

To classify each of the motions, the extended method of MACH filter proposed by [9] is used and is given by:

$$F(u, \upsilon, \omega, q) = \sum_{t=0}^{T-1} \sum_{z=0}^{N-1} \sum_{y=0}^{M-1} \sum_{x=0}^{L-1} f(x, y, z, t) \quad (4)$$

where f(x,y,z,t) are the 3D data corresponding to the temporal derivative and depth of the input sequence, and F(u,v,w,q) is the result in the frequency domain. To detect similar action in a testing video sequence, Inverse Fourier is applied to the filter and then to the video sequence.

$$F(u, \upsilon, \omega, q) = \sum_{t=0}^{T-1} \sum_{z=0}^{N-1} \sum_{y=0}^{M-1} \sum_{x=0}^{L-1} s(x, y, z, t) H(x, y, z, t) \quad (5)$$

where H is the filter in the time domain and s is the test video of hand motion.

E. Evaluation

To evaluate and differentiate any unnatural slowness of motions, after we classify the motion, we employ a HMM, which is widely used with time series data, such as speech and gesture recognition. We consider that hand motion consists of discrete hand poses and that each hand pose can be represented as a state. We define a bounded left-right model with the well-known transition model[10][11].

III. Conclusion

We have proposed a framework for 3D automatic hand motion evaluation with a SoftKinetic camera that solves some of the drawbacks of the existing methods. It can evaluate complex motions, whereas previous models are suitable for only simple motions. The proposed system captures images and depth data and then implements segmentation to extract the object of interest. UV mapping was used to merge RGB data with depth data. The classifier learned the characteristics of the points of interest based on the extracted features and then classified the hand postures. Finally, an HMM was used to evaluate and recognize movements based on the rate of the evolving motions.

Acknowledgment

The work is supported in part by the Canada Research Chair program, AUTO21 Networks of Centers of Excellence, the Natural Sciences and Engineering Research Council of Canada.

References

[1] A. Safaei and M. Jahed, "3D Hand Motion Evaluation Using HMM," Journal of Electrical and Computer Engineering Innovations (JECEI), vol. 1, 2013, pp. 11–18, ISSN: 2322-3952.

[2] "Kinect," 2015, URL: https://www.microsoft.com/en-us/kinectforwindows/ [accessed: April, 2015].

[3] "SoftKinetic," 2015, URL: http://www.softkinetic.com/ [accessed: April, 2015].

[4] "CIELab," 2015, URL: http://www.optelvision.com/documents /optelvision-s-explanation-on-cielab-color-space.pdf [accessed: April, 2015].

[5] T.-W. Chen, Y.-L. Chen, and S.-Y. Chien, "Fast image segmentation based on k-means clustering with histograms in hsv color space," in Multimedia Signal Processing, 2008 IEEE 10th Workshop on, Oct 2008, pp. 322–325.

[6] "UVMapping," 2015, URL: http://wiki.blender.org/index.php/Doc:2.4 /Manual/Textures/Mapping/UV [accessed: April, 2015].

[7] O. Oshin, A. Gilbert, and R. Bowden, "Capturing the relative distribution of features for action recognition," in Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, March 2011, pp. 111–116.

[8] S. Hadfield and R. Bowden, "Hollywood 3d: Recognizing actions in 3d natural scenes," in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, June 2013, pp. 3398–3405.

[9] M. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, June 2008, pp. 1–8.

[10] M. Elmezain, A. Al-Hamadi, J. Appenrodt, and B. Michaelis, "A hidden markov model-based continuous gesture recognition system for hand motion trajectory," in Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, Dec 2008, pp. 1–4.

[11] G. A. Fink, Markov models for pattern recognition: from theory to applications. Springer Science & Business Media, 2014.

 ISBN: 978-1-61208-422-0