Human Action Recognition with Kinect using a Joint Motion Descriptor

ボウボウ, ソマール

（様式３）

氏　　名　：ボウボウ・ソマール

論題名　：Human Action Recognition with Kinect using a Joint Motion Descriptor

（ キネクトと関節運動記述子を用いた人間行動認識 ）

区　　分　：甲

論　文　内　容　の　要　旨

Human action recognition is defined as the visual analysis of human motion that attempts to detect, to track and to interpret human behaviors from an image sequence involving humans. It is viewed as one of the most crucial steps in various applications, such as security surveillance, human-machine interaction and rehabilitation. Recently, Kinect, which is a sensing device capable of simultaneously providing a multi-sensor input, attracts a wide attention by the researchers. Major efforts are focused on developing effective feature descriptors of human body shape and motion acquired by Kinect. The dilemma of the trade-off between the accuracy and the computational cost is considered as a major challenge in case of real-time applications.

Feature descriptors for Kinect, which are typically evaluated by their recognition accuracy and computational time, are mainly divided into two categories: depth-data based descriptors and skeletal-data based descriptors. Descriptors of the first category make use of the richness of the depth-data in order to capture the tiny details of the shape and the motion of the human body and even the surrounding environment. Although this type of descriptors enables high recognition accuracy for complex actions, it is known to be computationally expensive. HON4D, which uses a histogram capturing the distribution of the surface normal orientations in the 4D space of time, depth, and spatial coordinates, is a well-known example for depth-data based descriptors. Although HON4D shows impressive recognition accuracy for complex actions, experiments demonstrate that HON4D consumes a relatively long computation time for feature extraction. On the other hand, skeletal-data based descriptors capture the 3D positions and motions of the skeleton joints representing the human body. Descriptors of this category mostly employ simple features to represent the body posture and motion and achieve competitive computational time, which usually comes with a decline of the recognition accuracy. As an example of skeletal-data based descriptors Chen and Koskela introduced several types of features demonstrating competitive computational time with a drawback regarding its recognition accuracy.

Our aim in this dissertation is to propose a novel efficient approach that can achieve both a faster and more accurate action recognition compared with the-state-of-the-art methods. In this proposed approach, the skeleton sequence is described using a 2D spatial joint histogram capturing the distribution of the orientations of velocity vectors of the skeleton joints in a spherical coordinate system. Since the proposed feature descriptor captures only the orientation of the velocity vector and not the magnitude, it is scale-invariant and speed-invariant. Moreover, due to the fact that the values of the histogram bins are normalized to a close interval [0,1], the descriptor is invariant to the action length. For the same reason, the proposed descriptor is more effective with periodic actions.

In order to decrease the computation time of feature extraction without suffering of the drawback in recognition accuracy, we present two proposals. The first one is based on neglecting the joints with low value data. Low value data is the data resulting from noisy joints or by the joints that show similar trajectories to those of other joints. The second approach is based on grouping joints with similar motion trajectories or similar functionalities into several groups. For example, we considered grouping the neck joints and the head joints into one (similar trajectories) and grouping the joints of a hand or a leg (similar functionalities). Moreover, since we think that actions performed by the right-handed and left-handed people must be considered equivalent, we study the scenario where joints are grouped into three groups (Torso, Upper limbs and Lower limbs).

Through extensive experiments on two collected datasets and one public dataset (MSR3D), the system is tested with different configurations. The proposed descriptor is tested with three classification methods: k-nearest neighbor classifier, Support Vector Machines and Extreme Learning Machines. Compared with the-state-of-the-art methods, our approach that utilizes a simple representation of actions is empirically proved to be more effective. Experimental results demonstrate that the proposed descriptor has much shorter computational time due to the simpler computation needed for feature extraction, e.g., over 600 times faster than HON4D implementation and over 100 times faster than HON4DA implementation. Moreover the proposed descriptor shows higher recognition accuracy. Our proposed descriptor shows 7% improvement in the recognition accuracy compared with HON4D implementation, 12% improvement compared with HON4DA implementation and 10% improvement compared with the descriptor proposed by Chen and Koskela. Finally, the results of the experiments prove the effectiveness of both proposals for decreasing the computational time. The second proposal in particular shows very promising results especially with periodic actions such as wave, hands-clap and boxing.