Recent advancements in commodity depth sensors, such as the Kinect and Leap Motion, have made quality RGBD data accessible at an affordable cost. The new data modality these sensors provide has driven the creation of many new methods for human activity recognition in 3D environments. Most of these methods focus on leveraging the 3D representation of sequential data; consequently, the internal temporal dynamics of the human activity sequence are not sufficiently exploited. In this dissertation, we propose to address the challenge of 3D human activity recognition from three different perspectives; namely, 3D spatial relationship modeling, temporal dynamic quantization, and temporal order modeling.

We propose a novel octree-based algorithm for computing 3D spatial relationships between objects from a 3D point cloud captured by a Kinect sensor. A set of 26 3D spatial directions are defined to describe the spatial relationship of an object with respect to a reference object. These 3D spatial directions are implemented as a set of 3D spatial operators, such as "AboveSouthEast" and "BelowNorthWest," of an event query language to query human activities in an indoor environment; for example, "A person walks in the hallway from north to south." The performance is quantitatively evaluated in a public RGBD object dataset and qualitatively investigated in a live video computing platform.

In order to address the challenge of temporal modeling in human action recognition, we introduce dynamic quantization, a clustering-like algorithm to quantize human action sequences of varied lengths into fixed-size quantized vectors. A two-step optimization algorithm is proposed to jointly optimize the quantization of the original sequence. In the aggregation step, frames falling into the sample segment are aggregated by max-polling and produce the quantized representation of the segment. During the assignment step, frame-segment assignment is updated according to dynamic time warping, while the temporal order of the entire sequence is preserved. The proposed technique is evaluated on three public 3D human action datasets and has achieved state-of-the-art performance.

Finally, the above model-driven approaches to human activity recognition are not scalable to a large-scale scenario in the real world. To this end, we propose a novel temporal order encoding algorithm that models the temporal dynamics of the sequential data for human activity recognition. The algorithm encodes the temporal order of the latent patterns extracted by the subspace projection and generates a highly compact First-Take-All (FTA) feature vector representing the entire sequential data. A stochastic gradient descent-based optimization algorithm is further introduced to learn the optimized projections in order to increase the discriminative power of the FTA feature. The compactness of the FTA feature makes it extremely efficient for human activity recognition using nearest neighbor search based on Hamming distance. Experimental results on three public human activity datasets have demonstrated the advantages of the FTA feature over state-of-the-art methods in both accuracy and efficiency.