Motion characteristics of human actions can be represented by the position variation of skeleton joints. Traditional approaches generally extract the spatial-temporal representation of the skeleton sequences with well-designed hand-crafted features. In this paper, in order to recognize actions according to the relative motion between the limbs and the trunk, we propose an end-to-end hierarchical RNN for skeleton-based action recognition. We divide human skeleton into five main parts in terms of the human physical structure, and then feed them to five independent subnets for local feature extraction. After the following hierarchical feature fusion and extraction from local to global, dimensions of the final temporal dynamics representations are reduced to the same number of action categories in the corresponding data set through a single-layer perceptron. In addition, the output of the perceptron is temporally accumulated as the input of a softmax layer for classification. Random scale and rotation transformations are employed to improve the robustness during training. We compare with five other deep RNN variants derived from our model in order to verify the effectiveness of the proposed network. In addition, we compare with several other methods on motion capture and Kinect data sets. Furthermore, we evaluate the robustness of our model trained with random scale and rotation transformations for a multiview problem. Experimental results demonstrate that our model achieves the state-of-the-art performance with high computational efficiency.