We propose a depth architecture networks which trained from semantic action features for real-time action recognition.

- The action features are extracted based on the velocity feature of body joints, relational feature concept and statistic method.
- We convert the extracted features into sequences of images which are feed into multilevel networks in depth architecture.
- These networks not only can learn the spatial and temporal from extracted features, but also can focus on objects more detail to recognize actions in real-time.

2. EXTRACT ACTIVE PARTS AND ACTIVE FRAMES

We define the body parts from skeleton joints and calculate the vector velocity of body parts to extract the active parts for each action category and active frames for each sample.

3. EXTRACTION OF SEMANTIC ACTION FEATURES

- We use Relational Features convert raw skeleton sequences into a sequence of binary codes for representing the occurrence of predefined unit actions.
- The set of Relational Features (RFs), which is created by a small number of Generic Relational Features (GRFs), is created by applying the set of GRFs to special poses and joints.
- We use the improved TF-IDF to calculate the words weighting then select the set of top-N words which are considered as action features.

4. DEPTH ARCHITECTURE NETWORK

- The Action Features are converted to Sequences of Images
- Trained by N categories, the depthNet_L1 gives top \( k_1 \), \( (k_1 < N) \) candidates from the input image.
- The depthNet_L2, which is trained from sequences of images containing \( k_1 \) categories as a time, gives top \( k_2 \), \( (k_2 < k_1) \) candidates to the higher networks.
- The top network gets information from its lower networks and the input image to recognize the final winner category.

5. ACTION RECOGNITION

- At each time slot we get an image \( \text{Img}_t \) by projecting the set of \( NJ \) body joints into the set of all RFs A.
- This image is an input to the depthNet_L1. From the output, we select set of \( N \), highest values candidates.
- We create the an image \( \text{Img}_t \) by projecting the set of \( NJ \) body joints into the set of RFs \( \{ \alpha^{k_2}_{A} \} \). This image is an input to the higher network depthNet_L2 to get \( k_2 \) candidates.
- This process is continuous until it reaches the top network. The final output is the recognition result.

6. EXPERIMENTS

We use the 3D motion capture data from HDM05 motion capture database for performance validation. To extract the action features, we collect twenty different actions. Each action has from eight to thirty-five samples and performed by five different actors.