論文の内容の要旨

A Hierarchical Action Recognition System Based on Spatio-Temporal Local Motion Feature Descriptors
(ローカルな動きの時空間特徴表現を用いた階層的行動認識システム)

Name : 包 蕊寒（Ruihan BAO）

Gesture perception or action recognition is receiving growing attentions due to its applications in smart surveillance, sign language interpretation, advanced user interface and intelligent robot control. As compared to static image recognition, action recognition usually requires handling overwhelmingly large amount of data because a whole set of video sequences must be analyzed. Moreover, if action recognition is subject to cluttered background, results are often degraded significantly. Therefore, it sometimes requires taking additional measures, i.e., tracking windows or background estimation on the frame-basis. In some cases, it is desirable to build the recognition system directly in the VLSI hardware such as ASICs (application specific integrated circuits) or FPGAs in order to achieve real-time performance. Therefore several constraints need to be further imposed on the algorithms. One important requirement is that the background elimination should be incorporated to the system so that video sequences can be taken as direct input. Another constraint is that computation in the system should be simple enough to be implemented on VLSI circuits either by analog or digital technology.

Among various algorithms, recognition based on local features is receiving great attention due to its robustness to space and time variation. In order to apply local features for recognition, interest points containing essential information of the movement are usually detected by spatio-temporal detectors, which are inspired by object tracking and object recognition. Once interest points are detected, motion descriptors are extracted from the video and used for building models representing certain motions. Beside the algorithms that apply machine-learning methods for motion analysis, bio-inspired hierarchical models based on both local features and associate memory principles show promising advantages for recognition tasks. One of the latest models for action recognition extended the similar structures for established object recognition system, and good results are reported. Nevertheless, the system applied only spatial patches in the processing for lower layers, therefore may not fully take advantages of the temporal relationship in the video sequences.

Inspired by this work as well as our previous efforts on VLSI based recognition system applying associate memory principles, a VLSI-hardware-friendly action recognition algorithm using spatio-temporal motion-field patches has been introduced in this thesis. The system employs a hierarchical two-layer structure so that the robust recognition can be achieved gradually. At the lower level, primary features called motion field maps that represent local features such as speed and direction are calculated from video sequences, further blurred by max filters. At the higher level, a collection of so-called template/prototype patches are used to recognize query actions by comparing local features in the query videos with those prototypes. In addition, in order to design a system for real-time performance, we intentionally simplify all the calculations into summation operations or Boolean operations so that the algorithm can be directly implemented on ultra high-speed VLSI chips without much effort. Our proposed system is at first developed for the application of gesture perception and promising results were reported compared to our previous researches based on global fea-

tures.

As an improvement, we have further proposed an enhanced processing to estimate motion field maps based on so-called essential directional edge displacement map. As the results, clean motion field maps can be calculated at the lower level. In addition, we simply the computations for feature vectors generation, by introducing more hardware friendly updating scheme. The proposed method not only reduce the computational cost significantly compared to our previous system but also fully take advantages of parallel processing for hardware implementation.

For most researches applying local features, descriptors play an important role in achieving high accuracy recognition. However, for most of researches, pixels within the descriptors are directly concatenated and Principle component analysis (PCA) is usually carried out for dimension reduction, which are not only unsuitable for hardware implementation but also inefficient to some extent. Based on our previous systems developed for face recognition. We have extended the face descriptor to the spatio-temporal form by coding temporal information along with spatial information. In addition, another effective descriptor coded for local maximum was also proposed. We show that the choices of the descriptors have significant influence to the recognition result.

Finally, once feature vectors for a given video is extracted, classifiers will be applied so that query videos can be labeled from the information of learning samples (already labeled samples). Among various classification models, sparse representation classification (SPC) becomes popular due to its powerfulness in face recognition. Recently, an extension of SPC called Fisher discriminant dictionary learning (FDDL) has been also proposed in which a structured dictionary is learned instead of raw vectors from learning samples. Because the success of the algorithm for face recognition, we employed it in our action recognition system and compared the recognition results with kNN. We show that FDDL is an effective classification method for our local feature based recognition system.