Multi-Features Encoding and Selecting Based on Genetic Algorithm for Human Action Recognition from Video

1Chenglong Yu, 1Xuan Wang, 2Muhammad Waqas Anwar and 1Kai Han
1Computer Application Research Center, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
2Department of Computer Science, COMSATS Institute of Information Technology, Abbottabad, Pakistan

Abstract: In this study, we proposed multiple local features encoded for recognizing the human actions. The multiple local features were obtained from the simple feature description of human actions in video. The simple features are two kinds of important features, optical flow and edge, to represent the human perception for the video behavior. As the video information descriptors, optical flow and edge, which their computing speeds are very fast and their requirement of memory consumption is very low, can represent respectively the motion information and shape information. Furthermore, key local multi-features are extracted and encoded by GA in order to reduce the computational complexity of the algorithm. After then, the Multi-SVM classifier is applied to discriminate the human actions.

Keywords: Feature encoding, feature selecting, genetic algorithm, human action recognition, multi-features

INTRODUCTION

In recent years, many approaches appear with the expansion of human action recognition technology in different application areas (Sharma et al., 2012; Khamis et al., 2012). However, most of them dependent on not only long time video date, but also a very complicated process of feature extraction. Therefore, human action recognition can be done by using these methods after actions carry out some time later or a few cycles in order to extract and calculate effectively the human action features. Different from these existing machine recognition methods, humanity can instantaneously discriminate and analyze complex human behaviors. To solve the above problem, in the spatial domain, some features such as contour, texture, shape, edge and so on are used to analyze human actions. And then, in the temporal, some features are employed. Also, in the space-time domain, the central cuboids or nearby cuboids may be utilized as effective features in human action recognition.

In this study, we proposed multiple local features with statistical ability that hold the high ability of expression in the classification of human actions from video and have good robustness. The multiple local features were obtained from the simple feature description of human actions in video. The simple features are two kinds of important features, optical flow and edge, to represent the human perception for the video behavior. As the video information descriptors, optical flow and edge, which their computing speeds are very fast and their requirement of memory consumption is very low, can represent respectively the motion information and shape information. Furthermore, key multi-features are extracted and encoded by GA in order to reduce the computational complexity of the algorithm. After then, the Multi-SVM classifier is applied to discriminate the human actions.

Related work: Many approaches have been proposed for human action recognition; however, the problem of a suitable feature set which can well classify the human actions in a swift manner is still partially unresolved. Borzeszki et al. (2011) use graphs model, which are converted into a suitable feature vector, to represent the shape of human actions. The experimental result shows that the embedded graphs can effectively describe the deformable human action shape and its evolution along the time. As one of the most useful and important features, high-quality edges can admirably characterize boundaries of objects in computer vision or image processing, whatever, how to obtain them is a problem of fundamental importance. There are many edge detecting algorithms to be proposed and used and as one of these algorithms, canny edge detection algorithm, which is proposed by Canny (1986), is well known as the optimal edge detector. Nieblesand
Fei-Fei (2007) propose a hierarchical model of shape and appearance for human action recognition and human actions in a frame by frame basis can be classified through using this model. The experiments show that the classification performance is improved by the proposed mixture of hierarchical models. Wang and Suter (2006) used a sequence of human silhouettes from videos that are converted into representations: average motion energy and mean motion shape to characterize actions. Wang and Suter (2007) used Locality Preserving Projections (LPP) to project continuous moving silhouettes into a low-dimensional space for characterizing the spatiotemporal property of actions. Abdelkader et al. (2007) focused on the use of shape of the object contour for recognizing human actions. They used Dynamic Time Warping (DTW) to align trajectories of silhouettes using elastic geodesic distances. And also they used a graphical model method to cluster the gesture shapes on the shape space. Their proposed approaches successfully represent shape of different human actions for recognition. Eweiwi et al. (2011) propose the approach that combines the methodologies of the key pose and motion: Motion History Images (MHI) and Motion Energy Images (MEI) for human action recognition. The experiment achieves high recognition rates over Weizmann data sets and the MuHAVi data sets. As another feature, optical flow is also utilized to describe the dynamic information of human actions in this study. The concept of optical flow was brought out firstly by Gibson (1950). It shows the instantaneous velocities of the spatial motion object's pixels in the imaging plane. And it uses changes in image sequences' pixels and relationships between adjacent frames to calculate this motion information. The classical and common methods of optical flow field calculation are L-K (Lucas and Kanada) method and H- S (Hom and Schunck) method. Many researchers proposed other methods. Ramadass et al. (2010) presented an extended Optical Flow algorithm for human action recognition and used Frame Jump restricts to detect useful features from video. Senst et al. (2011) present a novel speed and directional independent motion descriptor to detect people carrying objects. Wu et al. (2011) used based on Lagrangian particle trajectories, which a set of dense trajectories obtained by Optical Flow are used, to capture the motions of the scene and the approach obtained promising experimental results. Kovashka and Grauman (2010) proposed to obtain the shapes based on space-time feature neighborhoods and encode them into the visual vocabulary for human action recognition. The experiment results show that the approach has the high classification performance on the UCF Sports and KTH datasets.

**Multi-features selection:** Firstly, canny and optical flow features of the whole frames from the training human action dataset given are extracted and calculated. Secondly, local features in the fixed sub spaces are gathered from the global image optical flow and edge spaces and the features of the high discriminatory power are learned from the above features. Finally, the frequency distributions of different types of local features are used to assist calculating feature values and then the classifiers are trained by these features in order to sort human action categories.

In order to reduce the computational time and the feature dimensionality, local features in the continuous frame from video are encoded into simple and typical vector sets, which also are convenient for computing and training. Figure 1 shows that both canny features and optical flow ones are encoded through an average calculating way in the full range space into a new vector sets. Each frame block is cut equally into several small grids; furthermore, canny and optical flow features in every grid are quantified as the fixed value according to their directions, which are consistent with the direction disk. In Fig. 1a and b demonstrate

Fig. 1: Diagram flow of local features encoding method
respectively grid angle sets of frames using canny algorithm and grid angle sets of frames using optical flow algorithm from video.

For the canny space, these edges, to be single, can form a curve, so we represent the angle between the curve radian direction and horizontal coordinate direction as the canny edge angle and then these features in the curve are encoded into the fixed value according to the above way. From Fig. 1c, we can judge that the curve angle value in the grid 4 is the same to 3 fan-shaped area direction of the encoding disk, so these features are encoded into value 3.

In addition, for the optical flow features, they discrete points are messy and irregularly. They do not use the edge encoding method to finish this step. We combine and encode these feature angles into a new angle with the vector synthesis rule. In Fig. 1d, these discrete features’ angles in the grid 4 are toward different directions and then, they encoded into a new value 2. After we get both canny edge encoding value and optical flow encoding value, they can be combined into value 32. According to the encoding result (f), the strategy taken in the study is that grids no having values in the same positions are encoded as 00. Region area having optical flow values is greater than the area having the contour values in our implementation system. So we use the edge grid as the reference, the optical flow grid is chosen if the former has the value.

In this study, we select the next frame and previous frame of current frame to build the combinational encoding features for preserving the context information of the image correlations and also reducing feature numbers. Figure 2 shows the process of combinational encoding vectors. The encoding values of three frames are added together and then their average values are used as new feature values.

**Features extracting based on genetic algorithm:** Compared with traditional search methods, Genetic algorithm has some advantages. The operation method of Genetic algorithm has selection, crossover and mutation. After the encoding process is completed, we use GA to form models of different human action categories. In Fig. 2, we obtain the encoding value sets of the block for n classes of human actions from videos. Sub grid sets in blocks are selected by GA as the encoding model sets in Fig. 3 column three (Selection i). GA implementation procedure is as follows:

**Step 1:** Input control parameters
**Step 2:** Randomly generate initial group G and calculate the individual fitness, selection probability and group average fitness
**Step 3:** Use genetic operator to produce a new generation of group G
**Step 4:** Evaluate a new generation of groups and genetic cycle times increase 1
**Step 5:** Judge whether breeding cycle times is greater than the provisions cycle times. If so, turn step 6; otherwise, turn step 3
**Step 6:** Generation result is sorted and choose the highest fitness value as the most optimal solution according to the fitness value

**EXPERIMENTAL RESULTS**

In this study, Multi-SVM classifier is used to learn and build the classification model for human action recognition from video after we consider the identification accuracy and calculation simplicity. WEIZMANN dataset (Blank et al., 2005) and KTH dataset (Schuldt et al., 2004) are used to test our approach. Eighty-one video sequences are contained in the Weizmann dataset and are divided into ten types (bend, jack, jump, run, side, skip, walk, wave1, wave 2) (Fig. 4). And Five hundred and ninety-nine video sequences are contained in KTH dataset and are divided into six types (walk, run, jog, box, hand wave, hand clap) (Fig. 5). For training, 5 sequences and 50 ones are selected in each class of the former dataset and the latter dataset. Each video size of WEIZMANN dataset is 180*144 and of KTH dataset is 160*120. The ROI image size of datasets is designated as 125*125 and each block sizes 25*25. And accordingly, each grid sizes 5*5. Every frame has 25*25 features used as...
multi-feature vectors. The key frames are sampled by step 2 and the whole number is 25, so each video has 25*25*25 feature vectors. Genetic algorithm is used to do the optimization and the control parameters selected in the experiment: Breeding group scale $M = 100$; Crossover probability $p = 0.8$; Mutation probability $Q = 0.001$; Genetic cycle times $T = 1000$.

In the first experiment, our approach is implemented in two datasets for testing the performance of human actions recognition. Table 1 and 2 show recognition rates of our proposed approach and the experimental results demonstrate that our proposed method yields state-of-the-art performance on the KTH and the WEIZMANN datasets.

In the second experiment, Multi-feature selection based on GA is tested on the KTH and the WEIZMANN datasets. Frame selection of video and GA parameters are fixed, furthermore, feature numbers of each video both the test dataset and the train dataset are set for 15625 and feature number is selected for 10000. In Table 3, feature number decrement rate reduces 36%, but recognition rate reduces 0.9% approximately in the two datasets, the conclusion is drawn that our proposed method can get high-accuracy action recognition and is feasible.

**CONCLUSION**

In this study, we presented multiple local features using optical flow and shape for human action categorization. Firstly, Canny and optical flow features of the whole frames from human actions given are extracted and calculated. Blocks are fixed in the whole frame in video and then grids are segmented in each block. Secondly, Sub grid sets in blocks are selected by GA as the encoding model set. Finally, Multi-SVM classifier is used to learn and build the classification model for human action recognition. The experiment results show that our approach has the high classification performance on the KTH dataset and the WEIZMANN dataset.

Future works include adopting robust features and key frames selections for improving implementation time and unconstrained environments. We believe this approach has the potential to be able to characterize more complex human activities.

**ACKNOWLEDGMENT**

We would like to thank laboratory colleagues for giving us insightful discussions and comments. This study was supported in part by the Basic Research Program of Shenzhen City (No. JC201005260112A). Any opinions and conclusions expressed in this material are those of the author(s) and do not necessarily reflect those of the BRP of Shenzhen City. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.
REFERENCES


