Research Journal of Applied Sciences, Engineering and Technology 5(21): 5128-5132, 2013
DOI:10.19026/rjaset.5.4409
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: November 24, 2012 Accepted: January 05, 2013 Published: May 20, 2013

Research Article

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

[1]Chenglong Yu, [1]Xuan Wang, [2]Muhammad Waqas Anwar and [1]Kai Han
[1]Computer Application Research Center, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
[2]Department 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

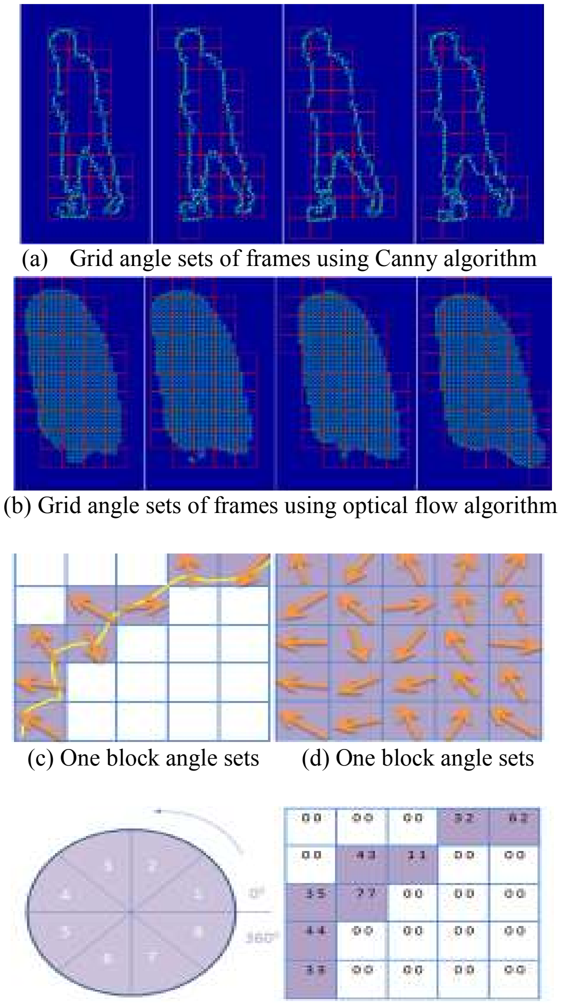
In recent years, many approaches appear with the expansion of human action recognition technology in different application areas (Sharma *et al*., 2012; Khamis *et al*., 2012). However, most of them dependent on not only long time video date, but also a very complicated process of feature extraction. 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. Borzeshi *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 and

Corresponding Author: Chenglong Yu, Computer Application Research Center, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China, Tel.: +86 075526033790

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.



(a) Grid angle sets of frames using Canny algorithm

(b) Grid angle sets of frames using optical flow algorithm

(c) One block angle sets (d) One block angle sets

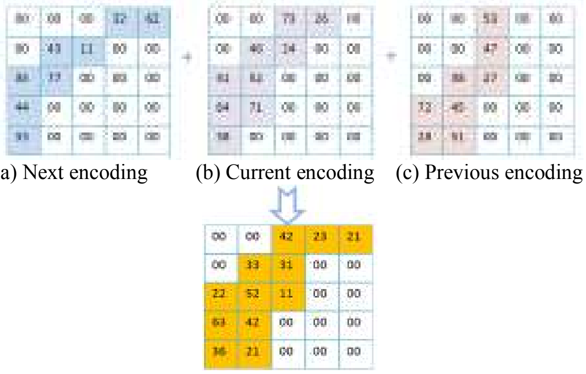
00	00	00	32	62
00	43	11	00	00
35	77	00	00	00
44	00	00	00	00
33	00	00	00	00

(e) Angle encoding disk (f) block encoding values

Fig. 1: Diagram flow of local features encoding method

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



00	00	00	32	62
00	43	11	00	00
88	77	00	00	00
44	00	00	00	00
93	00	00	00	00

(a) Next encoding

00	00	73	26	00
00	40	34	00	00
61	62	00	00	00
64	71	00	00	00
58	00	00	00	00

(b) Current encoding

00	00	53	00	00
00	00	47	00	00
00	86	27	00	00
72	45	00	00	00
28	51	00	00	00

(c) Previous encoding

00	00	42	23	21
00	33	31	00	00
22	52	11	00	00
63	42	00	00	00
36	21	00	00	00

(d) The combinational result

Fig. 2: The combinational encoding process

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:

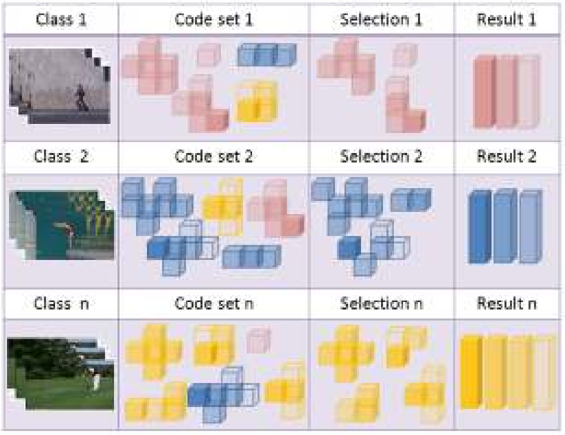


Fig. 3: Model selecting and forming using GA

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

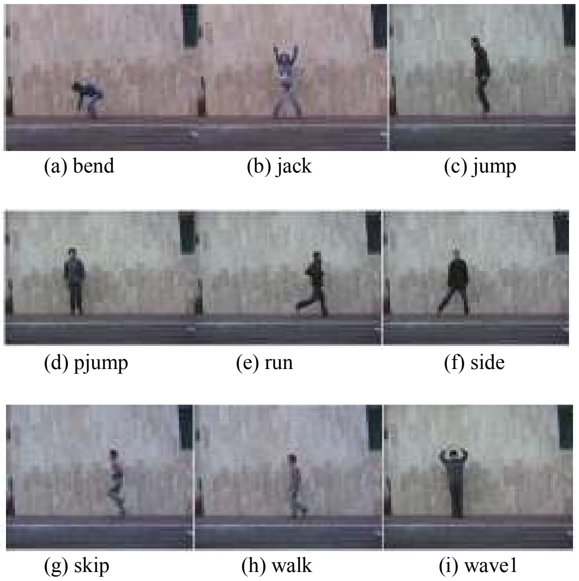
Step 4: Evaluate a new generation of groups and genetic cycle times increase 1

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, pjump, 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

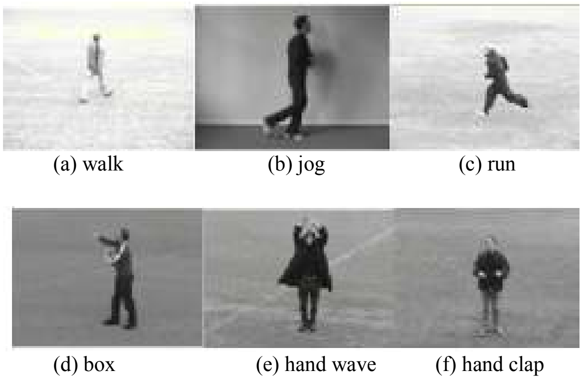


(a) bend (b) jack (c) jump

(d) pjump (e) run (f) side

(g) skip (h) walk (i) wave1

Fig. 4: Weizmann human action dataset



(a) walk (b) jog (c) run

(d) box (e) hand wave (f) hand clap

Fig. 5: KTH human action dataset

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.

Table 1: WEIZMANN dataset recognition accuracy results

Methods	Recognition rate
Our method	99.3%
Wang and Suter (2006)	97.8%
Bregonzio *et al*. (2009)	96.6%
Ali *et al*. (2007)	92.6%
Jia and Yeung (2008)	90.9%
Scovanner *et al*. (2007)	84.2%

Table 2: KTH dataset recognition accuracy results

Methods	Recognition rate
Our method	96.5%
Bregonzio *et al*. (2009)	93.2%
Savarese *et al*. (2008)	86.8%
Nieblesand and Fei-Fei (2007)	81.7%
Dollar *et al*. (2005)	81.5%

Table 3: Recognition result comparisons

Dataset	Feature selection using GA	Future number	Recognition rate
WEIZMANN	Done	10000	98.1%
	No	15625	99.2%
KTH	Done	10000	95.6%
	No	15625	97.8%

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 highhigh-accuracyaction 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

Abdelkader, M.F., W. Abd-Almageed, A. Srivastava and R. Chellappa, 2007. Silhouette-based gesture and action recognition via modeling trajectories on riemannian shape manifolds. Comp. Vision Image Understanding, 115(3): 439-455.

Ali, S., A. Basharat and M. Shah, 2007. Chaotic invariants for human action recognition. Proceeding of the IEEE 11th International Conference on Computer Vision.

Blank, M., L. Gorelick, E. Shechtman, M. Irani and R. Basri, 2005. Actions asspace-time shapes. Proceeding of Internatinal Conference on Computer Vision. Beijing, China, pp: 1395-1402.

Borzeshi, E.Z., R. Xu and M. Piccardi, 2011. Automatic human action recognition in videos by graph embedding. Lect. Notes Comput. Sc., 6979(2): 19-28.

Bregonzio, M., S.G. Gong and T. Xiang, 2009. Recognizing action as clouds of space-time interest points. Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp: 1948-1955.

Canny, J., 1986. A computational approach to edge detection. IEEE T. Pattern Anal., 8(6): 679-698.

Dollar, P., V. Rabaud, G. Cottrell and S. Belongie, 2005. Behavior recognition via sparse spatio-temporal features. Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp: 65-72.

Eweiwi, A., S. Cheema, C. Thurau and C. Bauckhage, 2011. Temporal key poses for human action recognition. Proceeding of IEEE International Conference on Computer Vision Workshops. Barcelona, Spain, pp: 1310-1317.

Gibson, J.J., 1950. The Perception of the Visual World. Houghton Mifflin, Boston.

Jia, K. and D.Y. Yeung, 2008. Human action recognition using local spatio-temporal discriminant embedding. Proceeding of IEEE Conference on Computer Vision and Pattern Recognition.

Khamis, S., V.I. Morariu and L.S. Davis, 2012. A flow model for joint action recognition and identity maintenance. Proceeding of IEEE Conference on Computer Vision and Pattern Recognition. Providence, Rhode Island, pp: 1218-1225.

Kovashka, A. and K. Grauman, 2010. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. Proceeding of IEEE International Conference on Computer Vision and Pattern Recognition. San Francisco, CA, pp: 2046 -2053.

Nieblesand, J.C. and L. Fei-Fei, 2007. A hierarchical model of shape and appearance for human action classification. Proceeding of IEEE International Conference in Computer Vision and Pattern Recognition (CVPR). Minnesota, USA, pp: 1-8.

Ramadass, A., M. Suk and B. Prabhakaran, 2010. Feature extraction method for video based Human action recognitions: Extended optical flow algorithm. Proceeding of IEEE International Conference on Acoustics Speech and Signal Processing. Texas, USA. Mar. 14-19, pp: 1106 -1109.

Savarese, S., A. DelPozo, J.C. Niebles and F.F. Li, 2008. Spatial-temporal correlatons for unsupervised action classification. Proceeding of IEEE Workshop on Motion and Video Computing, 2008.

Schuldt, C., I. Laptev and B. Caputo, 2004. Recognizing human actions: Alocal SVM approach. Proceedings of the International Conference on Pattern Recognition. Cambridge, UK, pp: 32-36.

Scovanner, P., S. Ali and M. Shah, 2007. A 3-dimensional sift descriptor and its application to action recognition. Proceedings of the 15th International Conference on Multimedia, pp: 357-360.

Senst, T., R.H. Evangelio and T. Sikora, 2011. Detecting people carrying objects based on an optical flow motion model. Proceeding of IEEE Workshop on Applications of Computer Vision. Kona, Hawaii, pp: 301-306.

Sharma, G., F. Jurie and C. Schmid, 2012. Discriminative spatial saliency for image classification. Proceeding of IEEE Conference on Computer Vision and Pattern Recognition. Providence, Rhode Island, pp: 3506-3513.

Wang, L. and D. Suter, 2006. Informative shape representations for human action recognition. Proceeding of the International Conference on Pattern Recognition. Kowloon Tong, Hong Kong, pp: 1266-1269.

Wang, L. and D. Suter, 2007. Learning and matching of dynamic shape manifolds for human action recognition. IEEE T. Image Process., 16(6): 1646-1661.

Wu, S.D., O. Oreifej and M. Shah, 2011. Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. Proceeding of IEEE International Conference on Computer Vision. Barcelona, Spain, pp: 1419-1426.