Middle-East Journal of Scientific Research 24 (S2): 103-108, 2016
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2016.24.S2.125

Domain Adaptive Feature Learning for Action Recognition Using Genetic Programming

[1]K. Murugavalli, [1]M. Rajakani and [2]S. Rajesh

[1]Department of CSE, Mepco Schlenk Engineering College (Autonomous), Sivakasi, TamilNadu, India
[2]Department of IT, Mepco Schlenk Engineering College (Autonomous), Sivakasi, TamilNadu, India

Abstract: Action recognition is mainly used for human computer interaction, intelligent surveillance system and video search and retrieval applications. In action recognition there are mainly two stages, first is feature extraction and second is action classification. Extracting features from videos is the most critical step. In some papers, some methods are fixedly followed for feature extractions that are not domain adaptive, all are handcrafted features. To make it domain adaptive, an evolutionary learning technique that automatically learn to extract features for action recognition using genetic programming is developed. In genetic programming the terminal set consists of optical flow information in both vertical and horizontal direction. Function set consists of 3D operators and Max pooling methods. From the available function set and terminal set, the input sequences and filters are randomly chosen to construct individuals. By applying cross over and mutation for some number of generations, the best individual is selected using linear support vector machine as fitness function. The best solution is the optimal feature descriptor. This method is evaluated on KTH, UCF YouTube and HMDB51 datasets. Result shows that, this method outperforms well when compared to hand crafted method.

Key words: Action recognition · Genetic Programming · Support vector machine · Feature extraction · Classification

INTRODUCTION

Activity recognition is an exciting field for the development of robust machine learning techniques. Action recognition is mainly used for human computer interaction, intelligent surveillance system, security-related applications and video search and retrieval applications. In action recognition there are mainly two stages, first is feature extraction and second is action classification. Two methods were mainly followed for the feature extraction stage, local feature extraction method and global feature extraction method. These methods give only handcrafted solution that are not domain adaptive. That is, it gives good results for some domain and poor performance for other domains. To overcome this problem genetic programming approach is used. It is an evolutionary methodology that automatically learns to extract features and solve problems for action recognition without prior knowledge about the dataset and solution. Finally an optimal feature descriptor is obtained through some genetic operators.

The datasets are originally in the form of videos. First videos are converted into frames. Key frame extraction has to be done to get important frames in both space and time domain. Optical flow is done between adjacent frames. Then frames have to be assembled in spatial and temporal alignments. The changes due to motion at some time interval are recorded as optical flow. Velocity field represents the optical flow field, it records the three dimensional motion information of object points across a two dimensional frame. Optical flow is not sensitive to changes in illumination and motion of insignificant objects (e.g., shadows). Genetic programming is a domain-adaptive technique that breeds genetically to solve a problem from a population of computer programs. Genetic programming iteratively generate new generation of programs by applying naturally occurring genetic operators to a population of computer programs. The genetic operations include reproduction, gene deletion, crossover, gene duplication and mutation. Genetic programming follows the ideas of biological evolution to solve the complex problems. Of a number of

Corresponding Author: K. Murugavalli, Department of CSE, Mepco Schlenk Engineering College (Autonomous), Sivakasi, TamilNadu, India.

possible solutions, the effective solution competes and survives in the population by means of cross breeding approach. This will continue until a needed solution is obtained. There are five major preparatory steps in genetic programming. The set of terminals for each branch of the to-be-evolved program. It is the independent variables of the problem. The set of basic functions for each branch of the to-be-evolved program. The fitness function that measure the classification error rate of each individual tree. Some parameters for controlling the run. Finally the termination criterion which stops the execution.

The main objective of this paper is to get best solution selected by means of genetic programming is regarded as optimal feature descriptor. Linear support vector machine classification is used as fitness function to get near-optimal feature descriptor. Ten fold cross validation is performed to estimate the classification error rate of each individual and to improve the recognition rate. By comparing the classification error rate of all individuals, the best individual is selected by means of elitism method. The experimental results of the classification approach are evaluated and a conclusion is made lastly.

Related Works: Most of the previous work has used handcrafted feature extraction technique that is not domain independent. Scovanner *et al*. presented a new concept using Histogram of 3D oriented gradients and is called *HOG3D*. It is a novel spatio-temporal descriptor for action recognition. It is very similar to the popular SIFT descriptors. Spherical coordinates and magnitude of the gradient has to be calculated, then it is binned in 3D spherical coordinate space in case of HOG. In HOG3D, codebooks are generated using bag-of-features. Then obtained features in histogram representation are fed into SVM classifier for action recognition. Main drawback here is, local based feature extraction technique is used. The obtained features are not invariant and distinctive. 3D feature extractors are derived from 2D, that is not fully exploit the essential difference between dynamic videos and static images. It is good for static images, not for dynamic video sequences. Another disadvantage is, during codebook generation there is structural configuration loss and quantization error.

Laptev *et al*. detected interest points in the spatio-temporal space. By measuring optical flow in spatio-temporal volume, motion descriptors are obtained to recognize human actions. Histograms of optical flow in their neighborhood and Histograms of gradient are calculated, that are local descriptors for video sequences. Then feature vectors are formed by concatenating normalized histograms. Here only spatial gradients are used and temporal information is obtained from histogram based optical flow. It is good for static images, not for dynamic video sequences. The disadvantage here is optical flow computation is rather expensive and the result is mainly depends on the choice of regularization method.

Han *et al*. explained object oriented saliency map detection method for action recognition. It is holistic based feature extraction method that is actions are represented using whole sequence visual information. Here saliency map is computed by using both pixel rarity and objectness into consideration simultaneously. To measure the rarity global contrast is used. Objectness is a property of group of pixels. Using contextual information, three attributes compactness, continuity and center bias are measured that are reflecting salient objectness of pixel. The drawback here is, it achieve good performance for particular given domain and result in poor performance on other applications.

Tao *et al*. detected gabor features for gait recognition. Here for gait recognition, sum of gabor filter responses over directions, scales, scales and directions are used for image representation. Using general tensor discriminant analysis (GDTA) approach features are extracted and classification is done using LDA. The disadvantage here is it requires lot of engineering work to tune and design. It is not adaptive for different datasets.

Bobick *et al*. proposed motion templates, namely motion energy images and motion history images (MHI). Motion history image is developed to represent the motion movement. The motion history at a location is detected using pixel intensity, where brighter pixel corresponds to more recent actions. MHI is robust to silhouette noises, missing parts and shadows. Using MHI the motion patterns in video sequences can be identified. MHI approach is computationally inexpensive and gives good performance in case of static background. For dynamically changing background, it results in poor performance.

Hinton *et al*. used a methodology called deep belief network (DBN) that automatically learn to recognize actions. It consists of multiple layers of hidden units. DBN contains large number of parameters to be learned. Sometimes the parameters are too large relative to available training samples, which needs proper tune and it restricts the applicability.

In local based feature extraction method, performance is good for static images and not for dynamic videos. In global feature extraction method, preprocessing steps are needed and it is sensitive to shifting, photometric and geometric distortions. The methods discussed above are all used handcrafted feature extraction technique that is not domain adaptive. To overcome these problems, genetic programming approach is used for action recognition that automatically learns to extract features without proper knowledge about the datasets and solutions. Ultimately an optimal feature descriptor is obtained using genetic programming.

System Design:

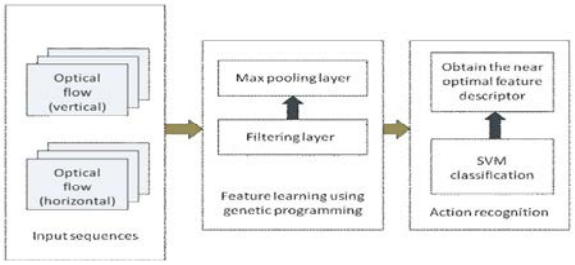

Fig. 1: System Design

In this work, the optical flow information in both vertical and horizontal direction form the terminal set. The filtering layer consists of lot of 3D filters and max-pooling layer consists of two pooling techniques, both form the function set. From the available terminal set and function set, the input sequences, 3D filters and max poolers are randomly chosen to construct individual trees using genetic programming. These individuals represent the population. Genetic programming iteratively generate new generation of programs by applying naturally occurring genetic operators to population. The genetic operations include crossover and mutation. Fitness is calculated for each individual to estimate the classification error rate. Classification error rate is calculated from average accuracy that is obtained through ten fold cross validation method using linear support vector machine. Finally optimal feature descriptor is obtained by using elitism selection method.

Terminal Set Formation: The datasets are all in the form of videos. First key frame extraction has to be done for all videos. In a video all frames are not equally important, few informative frames only needed. There are number of ways to get key frames from a video. Here entropy difference method is chosen to extract key frames. In this method, entropy is calculated for each frame. Using entropy, difference between two consecutive frames is measured. If the difference is greater than threshold, consider that frame as key. This process should be continued for all frames in the video. The names of the frames are not consecutive after key frame extraction, so it has to be renamed for proper processing. For basic action recognition, very short snippets (1-7 frames) are sufficient, so 21 frames are selected.

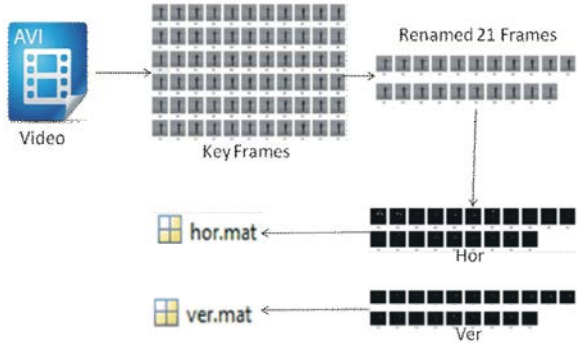

Fig. 2: Terminal Set Formation

The optical flow between adjacent frames has to be calculated to get motion information for action recognition. Majorly two methods are followed for optical flow calculation. They are Lucas Kanade method and Horn Schunck method. Horn Schunck method is more sensitive to noise and it needs high computational time, so Lucas Kanade method is chosen for optical flow calculation. Motion information in both vertical and horizontal directions are obtained through optical flow calculation. Then horizontal and vertical optical flow frames are separately assembled to form two 3D matrices. These horizontal and vertical 3D matrices form the terminal set.

Feature Extraction Using Genetic Programming: Feature extraction is the critical step in action recognition. In previous work, handcrafted feature extraction method is used for recognizing action that is not domain adaptive. To make it domain adaptive, genetic programming has to be done. Feature extraction is done using 3D operators that are assembled to get domain-specific feature descriptor. Tree structure is used to represent genetic programs that are evolved using genetic operators like crossover and mutation in every generation. Finally optimal feature descriptor is selected as solution to the particular domain. To construct this model, three important concepts should be defined first that are terminal set, function set and fitness function.

Terminal Set: For each video sequence V_i, terminal set T_i is defined as $T_i = \{V_{Fx}, V_{Fy}\}$.

Table 1

Terminal	Type	Description
V_{Fx}	Sequence	Optical flow sequence along horizontal direction.
V_{Fy}	Sequence	Optical flow sequence along vertical direction.

Function Set: Function set consists of two layers, filtering layer and max-pooling layer. Filtering layer is the bottom layer that consists of unary and binary operators. Max-pooling layer is the top layer that consists of unary operators. The order of these layers are always fixed in the genetic program tree structure. In filtering layer, meaningful features are extracted using 3D filters. 3D gaussian filter is used to smoothen the image and to remove detail and noise. This filter is somewhat similar to mean filter, but it uses different kernel. Kernel represents the shape of gaussian hump. Gaussian standard deviation is used to determine the degree of smoothing. The output of gaussian is weighted average of each pixel's neighborhood. Gentler smoothing is provided by gaussian filter that preserves edge detail better than mean filter.

Laplacian filter is used to sharpen the image for edge detection. It is second order derivative filter. 3D laplacian filter is used to find the fine details in the 3D image. Laplacian operator is used to enhance any feature with sharp discontinuity. Laplacian kernel can be constructed in various ways. In 3D LOG operator, first 3D gaussian filter is applied to smoothen the image in order to reduce the noise. Then laplacian operator is used to highlight regions of rapid intensity change for edge detection by using zero crossings. 3D LOG is obtained by combining 3D Gaussian and 3D laplacian operator that is to create a single LOG filter 3D gaussian and laplacian are convoluted together. If the image is basically uniform, LOG will give zero. If there is change, LOG give positive response on darker side and negative response on lighter side. The 3-D DWT is implemented along x, y and z axes. The x and y directions denote the spatial coordinates of an image and z is the spectral axis. It results in eight sub bands after one level decomposition. After wavelet transform, a sub band is chosen it consists of 3-D wavelet coefficients. Approximation LLL sub band is selected here that contains maximum quantity of information when compared to other sub bands. Downsampling technique is avoided here to get output of same size as input. Mean filter, or average filter is windowed filter of linear class, that smoothes image. The basic idea behind this filter is for any element of the image take an average across its neighborhood.

Table 2

No	Operator	Input	Type
1	3D Gaussian	1sequence	filter
2	3D Median	1sequence	filter
3	3D Laplacian	1sequence	filter
4	3D LOG	1sequence	filter
5	3D Mean	1sequence	filter
6	Add	2sequence	Arithmetic
7	Sub	2sequence	Arithmetic
8	3D Wavelet	1sequence	filter

In Max-pooling layer, the output of filtering layer is given as input to this layer. Two types of max-pooling techniques are used here, max-pooling10 (size 10×10×10) and max-pooling (size 20×20×20). The input image is divided into either 10×10×10 or 20×20×20 non-overlapping sub blocks. Maximum value of each sub block is concatenated into vectors that form the input for SVM classification. In case of action recognition, max-pooling is the key mechanism for robust response. The output of max-pooling layer is having same size as input.

Table 3:

Max-pooling tier	Input	Window size	Type
Max-pooling10	1 sequence	10×10×10	Filter
Max-pooling20	1 sequence	20×20×20	Filter

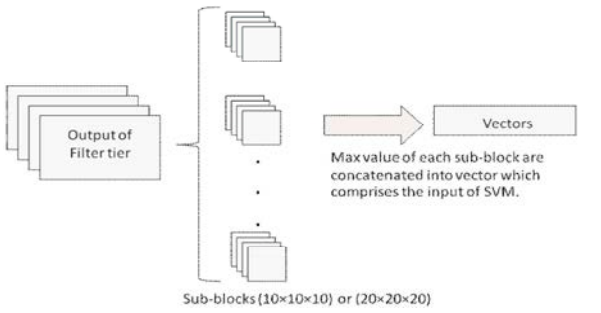

Fig. 2: Process of Max-pooling layer

Fitness Function: Fitness function is used to measure the performance of individual tree by using linear support vector machine. SVM classification is used for multiclass classification. The classification method used here is ten fold cross validation. The whole feature set is divided into 10 equal parts. 1 part is used for testing. The remaining 9 parts are used for training. In every fold testing and training set will change. In each set, predicted labels and accuracy is obtained as output. In ten fold cross

validation 10 accuracy values are obtained as output. By doing average of these 10 accuracy values, classification error rate is obtained for each individual.

$$E_r = \left(1 - \left(\sum_{i=1}^{n} \left(SVM\left[acu_i\right]\right)/(n \times 100)\right)\right) \times 100\% \quad (1)$$

SVM [*acu*] - recognition accuracy of fold 'i' by svm.

n - total no of folds.

E- Average of 'n' SVM test fold error rates, this value is less for best individual feature descriptor.

From the available terminal set and function set, the input sequences, 3D filters and max poolers are randomly chosen to construct individual trees using genetic programming. These individuals represent the population. Genetic programming iteratively generate new generation of programs by applying naturally occurring genetic operators to population. The genetic operations include crossover and mutation. Fitness is calculated for each individual to estimate the classification error rate. Classification error rate is calculated from average accuracy that is obtained through ten fold cross validation method using linear support vector machine Finally optimal feature descriptor is obtained by using elitism selection method.

GP Implementation: Action recognition system is implemented using genetic programming.

A) Population size: The initial population is generated using 10 individuals.

B) Generation: Number of generation is 10.

C) Genetic operators: Mutation and crossover are used.

D) Selection method: Total elitism method chosen for selecting the individual for reproduction. In every generation, top 2 individual is restored and if strong individual is found in that generation,weak individual is replaced by the strong individual in the population.

E) Stopping condition: Upto 10 generation and classification error rate less than 5%.

Experiments and Results: The Proposed method is tested using KTH, YouTube and HMDB51 action datasets. The KTH dataset is a benchmark action dataset. It includes 6 action classes, hand clapping, hand waving, jogging, running, walking and boxing. Each class contains 100 videos performed by 25 subjects in 4 different scenarios outdoors, outdoor with different clothes, outdoor with scale variation and indoor with lighting variation. Using the outdoor scenario, proposed method is tested using 25 videos from each class of total 150 features are given as input to SVM ten fold cross validation. From this, one optimal feature descriptor is obtained that gives 92.834% average accuracy and 7.16% classification error rate. The HMDB51 dataset consists of 51 classes. In this case 6 classes were taken for research that includes hand shake, pullup, pushup, throw, brush hair and sword. Using genetic programming, the average accuracy of 80% is obtained. The YouTube dataset consists of 11 classes. 5 classes were taken into account for research that includes basketball, diving, golf, juggling and tennis. Using genetic programming, the average accuracy of 87% is obtained. The numbers in the tree represents corresponding filter in the Table 2.

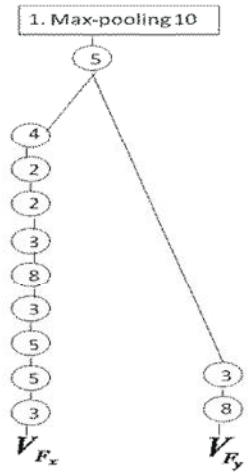

Fig. 3: Optimal feature descriptor for KTH dataset

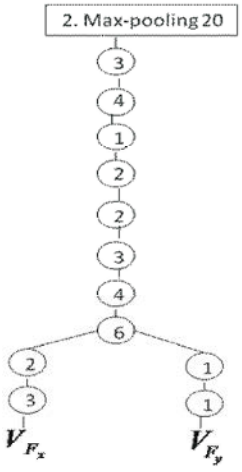

Fig. 4: Optimal feature descriptor for YouTube dataset

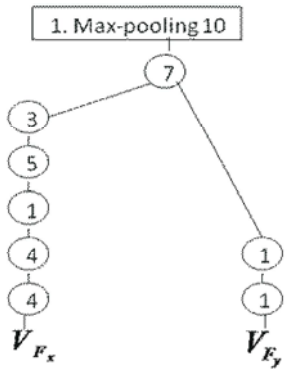

Fig. 5: Optimal feature descriptor for HMDB51 dataset

CONCLUSION

By using Genetic programming approach it has been proved that classification accuracy is improved and it is domain adaptive when compared with traditional handcrafted feature extraction method. Handcrafted feature gives good results for some domain and poor performance for other domains. To overcome this problem genetic programming approach is used. It is an evolutionary methodology that automatically learns to extract features and solve problems for action recognition without prior knowledge about the dataset and solution. The proposed system is evaluated using KTH, Hmdb51 and YouTube datasets. From the results obtained, proposed system outperforms well when compared with traditional handcrafted feature extraction method. It achieves significantly higher recognition rate in recognizing action. In future work, the number of population and number of generations will be increased in order to get best feature descriptor for particular domain.

REFERENCES

1. Li Liu, Ling Shao, Xuelong Li, and Ke Lu, "Learning Spatio-Temporal Representations for Action Recognition: A Genetic Programming Approach," IEEE Transactions on Cybernetics.
2. Scovanner, P., S. Ali and M. Shah, 2007. "A 3-dimensional SIFT descriptor and its application to action recognition," in Proc. 15th Int. Conf. Multimedia, Augsburg, Germany, 2007, pp. 357-360. R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," IEEE Trans. Syst. Man Cybern., vol. SMC-3, 6: 610-621.
3. Laptev, I., 2005. "On space-time interest points," Int. J. Comput. Vis., 64(2): 107-123.
4. Han, J., *et al*., 2013. "An object oriented visual saliency detection framework based on sparse coding representations," IEEE Trans. Circuit Syst. Video Technol., 23(12): 2009-2021.
5. Tao, D., X. Li, X. Wu and S.J. Maybank, 2007. "General tensor discriminant analysis and Gabor features for Gait recognition," IEEE Trans. Pattern Anal. Mach. Intell., 29(10): 1700-1715.
6. Klaser, A. and M. Marszalek, 2008. "A spatio-temporal descriptor based on 3D-gradients," in Proc. 19th Brit. Mach. Vis. Conf., Leeds, U.K., pp: 995-1004.
7. Bobick, A.F. and J.W. Davis, 2001. "The recognition of human movement using temporal templates," IEEE Trans. Pattern Anal. Mach. Intell., 23(3): 257-267.
8. Hinton, G., S. Osindero and Y. The, 2006. "A fast learning algorithm for deep belief nets," Neural Comput., 18(7): 1527-1554.
9. Schindler, K. and L. Van Gool, 2008. "Action snippets: How many frames does human action recognition require?" in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Anchorage, AK, USA, pp: 1-8.
10. Poli, R., W. Langdon and N. McPhee, 2008. A Field Guide to Genetic Programming. Morrisville, NC, USA: Lulu Press.
11. Lukas, B. and T. Kanade, 1981. "An iterative image registration technique with an application to stereo vision," in Proc. DARPA Image Und. Workshop, Vancouver, BC, Canada, pp: 674-679.