Genetic Programming-Evolved Spatio-Temporal Descriptor for Human Action Recognition

Li Liu

Ling Shao

Peter Rockett

Department of Electronic and Electrical Engineering,
University of Sheffield

Human action recognition has attracted a great deal of attention due to its potential usage in areas such as: video search and retrieval, intelligent surveillance systems and human-computer interaction.

In this paper, we propose a novel method by using Genetic Programming (GP) [2] to automatically generate a highly-performing low-level spatio-temporal descriptor for high-level human action recognition tasks. Our method is outlined in Fig. 1. For a given group of 3D sequence processing operators, GP first randomly assembles them into a variety of descriptors as the initialized population. The population is then continually evolved/evaluated by calculating the recognition error rate to evolve, hopefully, better-performing individuals into the next generation. Finally, one best-so-far individual can be selected as the final spatio-temporal descriptor. Genetic Programming (GP), as an evolutionary computation methodology, allows the computer to solve pre-defined tasks without requiring users to specify the form or structure of the solution in advance. GP can also escape local optima which may trap deterministic methods. Because of this, the use of GP is not limited to any research domain and can create relatively generalized solutions for target tasks. The basic Genetic Programming flow is shown in Algorithm. 1.

Algorithm 1 Genetic Programming

Start

Initialization Randomly create an initial population of computer programs from the available primitives (terminal set & function set).

Repeat

(1) Execute each program and evaluate its fitness.

(2) Choose programs from the population with a particular probability based on the fitness to involve genetic operations

(3) Create new generation programs by applying genetic operations.

If An acceptable solution is found or reach the maximum number of generations defined by user.

Stop

Return The best-so-far solution selected by Genetic Programming.

End

In our GP architecture, we have pre-defined three significant components as follows:

Terminal set: We flatten the 3D action sequences into 1D vectors as the programs' external inputs for GP.

Function set: We apply 12 unary 3D operators (*i.e.* Gaussian pyramid filters, Laplacian pyramid filters, Wavelet pyramid filers, *etc.*) and 4 basic binary arithmetic functions (*i.e.* Add, Subtraction, Multiply, Absolute subtraction) as our function set.

Fitness function: We use the classification error E_r for evaluating the performance of candidate descriptors. A support-vector machine (SVM) is adopted as the classifier to compute corresponding error rate. To achieve a fairer and more accurate result, ten-fold cross-validation is simultaneously employed on our dataset. We define the fitness function as follows:

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

where $SVM[acu_i]$ denotes the classification accuracy of the fold i by the SVM, n indicates the total number of cross-validation folds, here, $n = 10$.

We test our GP architecture on a mixed dataset combining the KTH dataset[1] [3] with the Weizmann dataset[2] [1] to obtain a promising spatio-temporal descriptor. The parameter settings for our GP running are listed in Table 1.

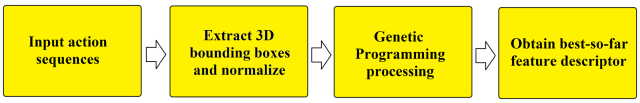
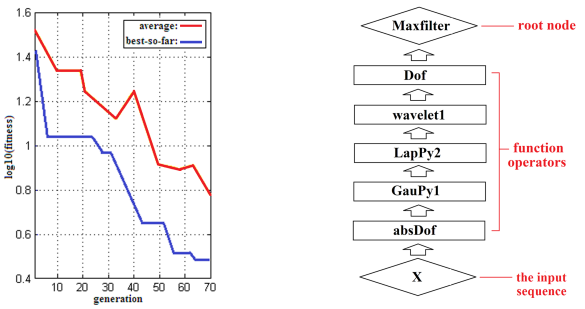


Figure 1: The outline of our proposed method



(a) (b)

Figure 2: (a) Evolutional average and best-so-far values of fitness (b) Tree-based genomic structure for the best-so-far program

We calculate the average error rate as the fitness value and obtain an accuracy of 96.9% for the GP-generated spatio-temporal descriptor. Fig. 2(a) shows the evolution of the average and best-so-far fitnesses. The final descriptor is illustrated in Fig. 2(b).

To demonstrate the generalizability of our method, the descriptor has further been tested on the more challenging IXMAS dataset [4] (composed of eleven human daily actions performed by ten actors and recorded from five different viewpoints.) The accuracy is 93.6% for multi-view fusion. We observe that our method achieves improvements and significantly outperforms previous work. The details can be seen in our paper.

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *International Conference on Computer Vision*, volume 2, pages 1395–1402, Beijing, China, 2005.

[2] R. Poli, W.B. Langdon, and N.F. McPhee. *A field guide to genetic programming*. Published via `http://lulu.com` and freely available at `http://www.gp-field-guide.org.uk`, 2008.

[3] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *International Conference on Pattern Recognition*, volume 3, pages 32–36, Cambridge, UK, 2004.

[4] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3D exemplars. In *IEEE 11th International Conference on Computer Vision*, pages 1–7, Rio de Janeiro, Brazil, 2007.

[1]The KTH dataset contains six types of human action examples (boxing, handwaving, handclapping, jogging, running and walking) performed by 25 different subjects with four scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. From http://www.nada.kth.se/cvap/actions/.

[2]The Weizmann dataset contains ten actions types (bend, jack, jump, pjump, run, side, skip, walk, wave1, wave2) performed by nine different subjects

Table 1: The parameter settings for GP

Population Size: 100	Generation: 70	Crossover Rate: 90% Mutation Rate: 10%
Selection for Reproduction: 'Lexictour'	Survival Method: 'Keepbest'	Stopping Conditions: Equal or lower than 2% of error rate