Accelerated Learning of Discriminative Spatio-temporal Features for Action Recognition

Munender Varshney
School of Computing and Electrical Engineering
Indian Institute of Technology
Mandi, Himachal Pradesh, India
Email: munender_kumar@students.iitmandi.ac.in

Renu Rameshan
School of Computing and Electrical Engineering
Indian Institute of Technology
Mandi, Himachal Pradesh, India
Email: renumr@iitmandi.ac.in

***Abstract*—Recently, paradigm has shifted from hand-designed local feature learning to unsupervised learning in order to extract features from raw data. In action recognition, good results are achieved using deep learning techniques such as stacking and convolution to extend the idea of independent subspace analysis (ISA). Albeit performance is good, it takes significant amount of time on big datasets due to high computational complexity and sequential implementation. We propose two methods for speeding up feature learning using ISA. We also propose input data modification which increases the classification performance. One method of faster feature learning is parallelization - we use the scalable programming model, MapReduce to parametrize ISA algorithm by distributing datasets into equal disjoint sets. The second method for increasing speed is by using spatio-temporal interest point detectors to extract "important" blocks from video. The latter not only enhances the speed but also improves the classification accuracy. We modified input as the gradient of video and achieved a better classification accuracy on all the datasets that were tested. We also created a dataset of water activities and used the ISA network for feature extraction. We achieved speed up by a factor of 4 and 2.4 in first and second method respectively.**

I. Introduction

Human action recognition is an area of research which has a variety of unsolved problems and has applications in surveillance, shopping behaviour analysis and robotics *etc.* These systems are challenging to build because recognition is inherently a difficult task [1]. Recognition is done in two phases, first is feature extraction to discriminate the activities, followed by classification based on these extracted features. Feature extraction can be either supervised or unsupervised. Supervised feature extraction is problem-dependent and requires domain knowledge while in unsupervised methods feature is extracted directly from the raw data.

In a recent work by Wang *et al.* [2] it is concluded that there is no best hand designed feature that can work for all datasets. This necessitates the learning of feature from data itself. Inspired by the success of deep neural networks, researchers have used convolutional neural nets(CNN) [3], [4], [5], deep belief nets [6] and sparse learning algorithms which are based on biological systems [7], [8] for learning hierarchical representation of local features. Quoc *et al.* [9] used independent subspace analysis (ISA), an extension of independent component analysis (ICA) to learn filters that resembles the receptive field of simple cells in primary visual cortex [7], [10]. They extended the idea of ISA to video domain for learning hierarchical representation of spatio-temporal features using convolution and stacking to make the algorithm fast and scalable. This methodology works well and leads to good results in classification but has a drawback of expensive computation in terms of time *i.e.,* $3\ \mathrm{h}\ 43\ \mathrm{min}$ for a dataset of size 3.4 GB on a machine with 16 cores Intel(R) Xeon(R) @2.3 GHz processor and 16 GB of RAM, for training the network. This is due to the high computational complexity of matrix operations involved.

With increasing data volume, there is a need to speed up the system by parallel processing. We propose to use MapReduce [11] by Google which works similar to single instruction multiple data (SIMD) architecture of a processor. In MapReduce [11], data is divided equally into disjoint sets and then processed independently by workers in the cluster, which are controlled by a master. Each worker executes a task, which is a higher level map or reduce function [11]. Map processes key/value pairs and generates a set of intermediate key/value pairs which are then shuffled and sorted according to these intermediate keys. The reduce function merges the intermediate values according to these intermediate keys. This model can express many real world tasks in a simple way.

The contributions of this paper are summarized below :

- *used distributed algorithm for training the ISA layers in the form of higher level MapReduce functions using RDD [12], an abstraction (provided by spark) responsible for efficient processing of an iterative algorithm and achieved a speed up by a factor of 4.*
- *used spatio-temporal interest point detectors [13] to extract the discriminative cuboids around the interest point which are used to train ISA and achieved a speed up of 2.4. We also note that in addition to enhanced speed, this gives the best classification performance.*
- *the classification performance is improved with modified input gradient while it is reduced with edges information.*
- *we have also built a dataset for human water activities for surveillance purposes which can be useful in real time decision making to avoid accidents.*

In Section II we give an overview of related work and section III describes the ISA and it's multilayer architecture in detail. Next section discuss different method of speed enhancement for ISA. Section V introduce various methods

978-1-5090-1746-1/16/$31.00