Applied Soft Computing 46 (2016) 936–952

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Hybrid deep neural network model for human action recognition

Earnest Paul Ijjina*, Chalavadi Krishna Mohan

Visual Learning and Intelligence Group (VIGIL), Department of Computer Science & Engineering, Indian Institute of Technology Hyderabad, Telangana 502285, India

ARTICLE INFO

Article history:
Received 15 June 2015
Received in revised form 10 August 2015
Accepted 13 August 2015
Available online 28 August 2015

Keywords:
Deep neural network
Convolutional neural network (CNN)
Classifier fusion
Action bank features

ABSTRACT

In this paper, we propose a hybrid deep neural network model for recognizing human actions in videos. A hybrid deep neural network model is designed by the fusion of homogeneous convolutional neural network (CNN) classifiers. The ensemble of classifiers is built by diversifying the input features and varying the initialization of the weights of the neural network. The convolutional neural network classifiers are trained to output a value of one, for the predicted class and a zero, for all the other classes. The outputs of the trained classifiers are considered as confidence value for prediction so that the predicted class will have a confidence value of approximately 1 and the rest of the classes will have a confidence value of approximately 0. The fusion function is computed as the maximum value of the outputs across all classifiers, to pick the correct class label during fusion. The effectiveness of the proposed approach is demonstrated on UCF50 dataset resulting in a high recognition accuracy of 99.68%.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

The famous ‘no free lunch’ theorem [1] proposed by Wolpert suggests that there is no single computational view that solves all pattern recognition tasks. This lead to an increased interest in combining several classifier systems that perform information fusion of classification decisions thereby over-coming the limitations of using a single classifier. Several techniques like hybrid intelligent systems, multi-classifier systems, information fusion were proposed in the literature for classification, employing several computational views. While information fusion techniques combine information from different sources to recognize a new view for better classification, multi-classifier systems focus on combining different classifier models for effective classification. Hybrid intelligent systems employs intelligent techniques in various computational phases from data normalization to final decision making to obtain a blend of heterogeneous fundamental views for effective classification.

A true function that cannot be modeled by a single hypothesis can now be modeled as a combination of hypotheses. One of the advantage of using a hybrid intelligent systems is its ability to handle the two extreme cases in availability of training data. The scenario where data samples are scarce can be effectively handled by considering bootstrapping methods like boosting [2] and the scenario with huge number of data samples by combining decisions of classifiers trained on partitions of data [3]. A multi-classifier system can outperform the best individual classifier [4] and this was analytically proved in [5] by considering majority voting on a group of independent classifiers. In case of classifiers using heuristic approaches for optimization, that does not ensure an optimal solution, but a combined approach may increase the probability of finding an optimal model. As stated by Wolpert [1], each classifier has its specific competence domain and choosing an ensemble of heterogeneous classifiers would result in an effective classification model.

The general structure of a multiple classifier system (MCS) consists of a classifier ensemble with a set of diverse classifiers. The most discriminative features are given as input to the classifier ensemble and a fusion method is used to optimally combine the individual classifier outputs for classification. Thus, the main design issues in a MCS are: (1) *system topology*: describing the interconnection between classifiers, (2) *ensemble design*: defining the generation and selection of a pool of classifiers, and (3) *fuser design*: a decision combination function that optimally combines the outputs of classifiers. Some of the multiple classifier system (MCS) proposed in the literature are discussed in the following section. The two popular MCS system topologies are a parallel topology [6] and a serial (or conditional) topology. In a parallel topology, the same input data is fed to all the classifiers and the output generated by these classifiers is used for decision making. As the classification output of one classifier is independent of the output of other

* Corresponding author. Tel.: +91 9494466490.
E-mail addresses: cs12p1002@iith.ac.in (E.P. Ijjina), ckm@iith.ac.in (C. Krishna Mohan).

http://dx.doi.org/10.1016/j.asoc.2015.08.025
1568-4946/© 2015 Elsevier B.V. All rights reserved.

classifiers, this approach is more suited when considering classifiers with low support/confidence in classification. The sequential approach is considered when the cost of classifier exploration is high. The classifiers are arranged in sequence of increasing computation cost i.e., the classifier with the least computation cost will be the first classifier in the pipeline. In [7], each classifier gives an estimate of the certainty of classification and the uncertain data samples are sent to the next classifier in the pipeline. A reject-option [8] can also be used in serial topology and Adaboost [9] is a special case of sequential topology.

Ensemble design in a MCS aims to include mutually complementary classifiers that are characterized by low classifier output correlation [10] and high accuracy [11]. Dietterich in [12] empirically validated that a robust classifier can be built by combining the evidences of complementary classifiers. Brown in [13] suggests that diversity can be achieved using implicit or explicit approaches. Implicit techniques involves use of random techniques to generate individual classifiers while explicit approaches focus on optimizing a diversity metric in an ensemble of classifiers. The wide range of experimental results in [14] suggests that increasing diversity should result in a combined system with better accuracy. According to [6,15], diversity of classifiers can be enforced by manipulation of either individual classifier inputs, outputs, or models. Some of the approaches for diversifying input data are: (1) using different data partitions, (2) using different set of features, and (3) taking local specialization of individual classifiers into consideration. Local specialization is a classifier selection approach that selects the best classifier from a pool of classifiers trained on partitions of the features space. Diversity in MCS can also be achieved by considering classifiers designed to classify only a subset of classes and applying combination technique to restore the whole class label set. Finally, the diversified models in the ensemble should be combined to take advantage of the homogeneous/heterogeneous combination of models. As some classifiers are more efficient for some domains, an ensemble of heterogeneous classifiers would result in an solution well-addressed in multiple domains. As most machine learning algorithms (like neural networks [16]) would try to find an optimal solution from a given initial setup, combining homogeneous (identical) models with various initializations may improve classification performance.

An effective fuser is a crucial requirement for an efficient classifier. A fuser combines the outputs of the selected classifiers from the ensemble to give a final decision of the MCS system. The outputs of the classifier could be the class label associated with the test instance or the support (confidence value) for test instance to belong to a class. Early implementations of fusion models considered majority voting [6] that determined the final class label by (1) *unanimous voting* where the decision is unanimous, or (2) *simple majority* decision made if majority is more than half of the selected classifiers, or (3) *majority voting* where decision is to select the class with highest number of votes. Later, alternate voting methods [6,17] were proposed that assign different weights to the outputs of the classifiers. Fusion models based on support, use a support function to compute the confidence of a classifier in its decision. Some of the well know approaches are the ranking based approach of Borda count [18], the posterior probability approaches [19–21] and combination of accuracy of neural networks [22]. Trainable fusers were proposed by considering the weights used to combine classifier outputs as a learning process [23,24]. Perception learning with evolutionary approaches were used by Wozniak in [25] to train a fuser and Zheng used data envelopment analysis in [26]. A experimental comparison of various fusion functions along with their sensitivity analysis was done in [27].

Among the high-dimension data, human action recognition in videos poses a unique challenge due to the existence of temporal dimension whose length varies with each instance and subject executing the action. The inconsistencies in the execution of actions, the environment and capturing conditions further complicates the observed data, that is in-turn used as input for recognition algorithms. Some of the most commonly used features for human action recognition are histogram of oriented gradients (HOG) [28], histogram of optical flow (HOF) [28], motion boundary histograms (MBH) [29] and motion interchange patterns (MIP) [30]. These features are used with some classical approaches like support vector machine (SVM), neural networks and k-nearest neighbor to compute the base results for most of the action recognition datasets. Nazli Ikizler-Cinbis et al. used different features with multiple instance learning (MIL) framework to utilize the entities related to an action like the scene, objects and people for action recognition in [31]. In [32], Fabian Caba Heilbron et al. used dense point trajectories to extract context from foreground motion to recognize actions in videos with camera motion. In [33], Salah Althloothi et al. also used multiple features for human action recognition in RGB-D videos by using multiple kernel learning. An ensemble of homogeneous models are used by Karen Simonyan et al. [34] for object recognition in ImageNet Large Scale Visual Recognition Competition (ILSVRC) [35]. Samira Ebrahimi Kahou et al. [36] used fusion of models trained on different modalities to improve the efficiency of their model in Emotion Recognition In The Wild (EmotiW) [37] challenge. Mengyi Liu et al. in [38] combined multiple kernel methods trained for different modalities using a trained fusion function. Multi-resolution CNN architecture with time information fusion is used by Andrej Karpathy et al. in [39] for human action recognition. These approaches assert the need for using multiple features and classifiers to design an effective classification model.

In this paper, we propose a hybrid classifier for action recognition by fusion of evidences generated by homogeneous models arranged in a parallel topology. A convolutional neural network classifier designed to recognize human actions from action bank features is used to build the ensemble of classifiers. The novelty of the proposed approach lies in achieving the diversity of models by manipulation of the input data using complementary features and by varying the initialization of neural network weights. Also, we use a fusion function that exploits the high confidence value of classifiers for correct prediction, to pick the correct class label across outputs. The reminder of this paper is organized as follows: Section 1 gives an introduction to multi-classifier systems and the various approaches in the literature for classifier fusion. Section 2 introduces the proposed hybrid system along with the approaches used to diversify the models and the fusion function. Section 3 covers the experiment setup and results followed by analysis of results. Section 4 gives the conclusions and future directions of this work.

2. Human action recognition using fusion of CNN classifiers

This section presents the CNN classifier architecture used to generate the ensemble of classifiers in the proposed fusion model. Different initial weights are used to generate multiple CNN classifiers. These weights are determined by a random number generator that is initialized by a seed value. By using n unique seed values, n different weight initializations of CNN classifier are constructed. Corresponding to each weight initialization, we train one CNN classifier on action bank features and another CNN classifier on complementary action bank features. The complementary action bank features are computed by taking the complement of action bank features interpreted as an image. As a result, $2n$ number of CNN classifiers (that are assumed to be implicitly diverse) are constructed in the ensemble. The block diagram of the proposed model to recognize 'c' classes using fusion of $2n$ models is shown in Fig. 1. The CNN classifier initialized with seed value i and using action

Download English Version:

https://daneshyari.com/en/article/494716

Download Persian Version:

https://daneshyari.com/article/494716

Daneshyari.com