Noname manuscript No.
(will be inserted by the editor)

Deep Multiple Classifiers Fusion for Traffic Scene Recognition

Fangyu Wu[1*] · Shiyang Yan[2] · Jeremy S. Smith[3] · Bailing Zhang[2]

Received: date / Accepted: date

Abstract The recognition of traffic scene in still images is an important yet difficult task in an Intelligent Transportation Systems (ITS). The main difficulty lies in how to improve the image processing algorithms against different traffic participants and various layouts of roads while discriminating different traffic scenes. In this paper, we attempt to solve the traffic scene recognition problem by proposing a deep multi-classifier fusion method in the setting of granular computing. Specifically, the deep multi-classifier fusion method which involves local deep-learned feature extraction as one end that is connected to the other end for classification through a multi-classifier fusion manner. At the local deep-learned feature extraction end, the operation of convolution to get feature maps from the local patches of an image is essentially a form of information granulation, whereas fusion of classifiers at the classification end is essentially a form of organization. In addition, we construct a new traffic scene dataset "WZ-traffic", consisting of 6035 labeled images of 20 categories to evaluate the traffic scene recognition performance. Extensive experiments over the benchmark dataset FM2 has also shown that the proposed method significantly outperforms the state-of-the-art approaches for traffic scene recognition.

Keywords Traffic Scene Recognition · Convolutional Neural Networks · Multi-classifier Fusion

✉* Fangyu Wu
E-mail: fangyu.wu@xjtlu.edu.cn
Shiyang Yan
E-mail: shiyang.yan@qub.ac.uk
Jeremy S. Smith
E-mail: J.S.Smith@liverpool.ac.uk
Bailing Zhang
E-mail: bai_ling_zhang@hotmail.com

[1] Department of Computer Science and Software Engineering, Xi'an Jiaotong-liverpool University, SuZhou, JiangSu Province, China
[2] The Institute of Electronics, Communications and Information Technology, Queen's University Belfast, NI Science Park, Queen's Road, Queen's Island Belfast, BT3 9DT
[3] School of Computer and Data Engineering, Ningbo Institute of Technology, Zhejiang University, Ningbo, Zhejiang Province, China
[4] Department of Electrical Engineering and Electronic, University of Liverpool, Liverpool, L69 3BX, UK
* Corresponding Author

1 Introduction

Recognizing the traffic scene in front of a vehicle is an important task for autonomous driving (Huang et al., 1994). Knowledge of the current traffic scene information can have several benefits: e.g., augmenting the driver's situational awareness, reducing driver workload, and automating all/part of the driving process. Despite the progresses in scene recognition (Dixit et al., 2015; Greene et al., 2015; Song et al., 2015), understanding the traffic scene in various environments remains largely unsolved. This is mainly due to the complexity of traffic situations. First, many different traffic participants may be presented and there are a variety of geometric layouts of roads and crossroads. Furthermore, illumination conditions such as cast shadows caused by infrastructure or vegetation add extra complexities.

A traffic scene is generally composed of a collections of entities (e.g. objects) organized in a highly variable layout. This high variability in appearance has made reliable visual representation the primary choice in solving this problem. Among them, an image has

been represented as bags of locally extracted visual features according to bag-of-features (BOF) methods, such as Scale Invariant Feature Transform (SIFT) (Oquab et al., 2014) and Histogram of Oriented Gradient (HoG) (Dalal and Triggs, 2005). For many high level vision tasks, these features can be pooled into an invariant image representation, e.g., Bag of Visual Words (BoVW) (Csurka and Perronnin, 2010), Fisher Vectors (FV) (Dixit et al., 2015), and Vector of Locally Aggregated Descriptors (VLAD) (Jégou et al., 2010).

However, the rich variabilities hidden in the image cannot be reflected by the dominate patch encoding strategies, which are based on hand-crafted features. Recently, Convolutional Neural Networks (CNN) have brought breakthroughs in image representations by emphasizing the significance of learning robust feature representations from raw data (Krizhevsky et al., 2012; Simon et al., 2014). CNN has the ability to detect complex features automatically by training multi-layer of convolutional filters in an end-to-end network, which is a prerequisite for many computer vision tasks, such as action recognition (Yan et al., 2017), vehicle recognition (Wu et al., 2018) and object detection (Girshick, 2015). Despite these achievements, there are still some limitations in deep Convolutional Neural Networks, such as the lack of geometric invariance and the limitations in transferring information about local elements. Besides, a single classifier may have its own advantages and disadvantages in the classification task (Zhou, 2012). For the task of traffic scene recognition, a single classifier may be capable of learning some, but not all, specific characteristics of traffic scene. So it is worth exploring multi-classifier fusion applied to traffic scene recognition to improve the classification performance.

To address the above issues, in this paper, we therefore propose a novel traffic scene recognition methodology in the setting of granular computing, which involves creation of information granulation by extracting the CNN features upon local regions of the image for a compact representation, and design multiple levels of classifiers fusion method through fusing the outputs of the two ensemble classifiers (Random Forests and Gradient Boosted Trees) with the outputs of the selected single classifier. Second, we discuss how to improve the recognition rate by deep multi-classifier fusion method from the perspective of granular computing. With these contributions we are able to create information granulation and diverse classifiers to advance the performance.

The rest of the paper is organized as follows. In section 2, we briefly offer a brief overview of the traffic scene recognition, multi-classifier fusion and granular computing. Section 3 provides a detailed description of the proposed methods. We also present how granular computing concepts are employed to design the framework of deep multi-classifier fusion. In Section 4, we describe the details of the new traffic scene dataset "WZ-traffic" which collected from 20 traffic scenes. Besides, we conduct the experimental study on WZ-traffic and FM2 datasets, and discuss the results in terms of multiple comparison settings. In Section 5, we highlight the contribution of this paper and provide some future directions in this area.

2 Related Work

As an emerging research topic, traffic scene recognition has recently attracted significant interest (Tang and Breckon, 2011; Mioulet et al., 2013; Taylor et al., 2016). In this section, we focus on three relevant research areas: traffic scene recognition, multi-classifier fusion and a review of granular computing concepts.

2.1 Traffic scene recognition

Automatic recognition of visual scenes is an important issue and plays an important role in automatic transpertation and traffic surveillance. A number of studies have been carried out under the daunting challenges of recognizing traffic scene, mostly aimed at automatically analyzing the road environment, or detecting and classifying possible objects in the traffic scene, such as pedestrians and vehicles. For example, Ess et al. (2009) proposed an urban scene understanding method by exploiting a pre-training classifier to label the segmentation regions. Besides, a road classification scheme was introduced by Tang and Breckon (2011), which utilized the color, texture and edge features of the image sub-region. Then they applied a convolutional network for the classification task.

Recently, based on the general data mining process, Taylor et al. (2016) put forward a novel data mining methodology for driving-condition monitoring via CAN-bus data. In (Lu et al., 2016) a generalized Haar filter based deep network was applied for the object detection tasks in traffic scene. A novel concept of the atomic scene has been proposed by Chen et al. (2016), they established a framework for monocular traffic scene recognition by decomposing a traffic scene into atomic scenes.

2.2 Multi-classifier Fusion.

The effectiveness in solving classification tasks has been proven by many machine learning algorithms, such as

support vector machine (SVM) (Cortes and Vapnik, 1995), k-nearest neighbours (KNN) (Altman, 1992), decision tree (DT) (Gondy et al., 1993) and random forest (RF) (Cutler et al., 2004). A simple practice is to retain the best classifier and disregard the others after evaluating their performance. Alternatively, one could fuse the information provided by them to achieve a better recognition rate. Recently, multi-classifier fusion has attracted attention in various computer vision tasks to achieve improved performance. The final result of the classifiers fusion depends on the method of combining the decisions from different classifiers in accordance with the fusion rule.

In (Kuncheva, 2002), six simple classifier fusion methods were theoretically studied, including minimum, maximum, median, average, oracle and majority votes. Due to the simplicity and good performance of these strategies, they may be the most obvious choice when building a multi-classifier system.

To determine the support $S_i(x)$ for class x_i, using the fusion rule R to perform a majority voting on the class-related probability predicted by each classifier, it can be defined as,

$$S_i(x) = R(P_{1,i}(x), , ..., P_{L,i}(x)), i = 1, 2, ..., m. \quad (1)$$

In the majority voting method, the class label of x predicted by each classifier should be computed firstly. Then, the support $S_i(x)$ can be robustly estimated as,

$$S_i(x) = \frac{v + 1}{L + m} \quad (2)$$

where v represents the number of votes received by the class x_i. Compared to frequency-based probability estimation, this probability usually does not affect the final result, while avoiding the problem of certain class labels that do not appear in the basic classifier output (Duin and Tax, 1998).

Fusion of feature sets and classifiers for facial expression recognition has been studied in (Zavaschi et al., 2013). Toufiq and Isalm (2017) developed a dynamic decision selection method for face recognition that uses the least amount of facial information to take correct decision. In (Nanni and Lumini, 2013), a random subspace ensemble of support vector machines (SVM) classifiers has been trained for scene recognition, and then the sum rules were used to combine the classifier results. In this paper, we present a multi-classifier fusion approach by using various classifiers in the setting of ensemble learning which leads to an improvement in the recognition accuracy.

2.3 A review of granular computing concepts

From the aspect of philosophical perspectives, granular computing is a way of structured problem solving at the practical level (Yao, 2005b). There are two commonly concepts in granular computing: granules and granularity (Pedrycz, 2011; Pedrycz and Chen, 2015). In theory, a granule is defined as a collection of smaller units that can form a larger unit.

Various granules involves horizontal relationships and hierarchical relationship. If different granules involves horizontal relationships when if they are located in the same or different levels of granularity. Otherwise, these granules are in hierarchical relationships. For structural information processing, there are different levels of granularity for different sizes of granules. In ensemble learning, an ensemble of classifiers is viewed as a granule. Also, if the combination of classifiers involves different levels, each level represents a level of granularity.

In general, there are two main operations in granular computing including granulation and organization. The granulation operation aims at decomposing larger granule in a higher level of granularity into smaller granules in a lower level of granularity, while organization intends to integrate several parts into one. When designing the top-down and bottom-up approaches from a computer science perspective (Yao, 2005a), the operations of granulation and organization are widely used, respectively (Liu et al., 2018).

In the content of set theory, a set of any formalism is regarded as a granule and each element in a set can be viewed as a particle. There are different formalisms of sets such as probabilistic sets (Liu et al., 2016), fuzzy sets (Zadeh, 1965; Lee and Chen, 2008) and rough sets (Pedrycz, 2011). They belong to information granulation which is one of the fundamental of granular computing. In particular, a probability set can be considered a deterministic set when all elements belong to the set. Probabilistic sets provide a chance space to each set and view it as a granule. The chance space will be divided into subspaces which can be viewed as particles that are considered to be randomly selected to activate the occurrence of an event. Therefore, a whole chance space integrates all these particles.

The fuzzy sets views each set as a granule and gives each element a certain degree of membership in that set (Chen and Wang, 1995; Chen and Tanuwijaya, 2011). In other words, each element belongs to a certain degree of fuzzy set. In the setting of granular computing, a particle represents each part divided from the membership. In the context of rough set context, each set is viewed as a granule. As described in (Liu et al., 2016),

rough set uses a boundary region to recover some elements with insufficient information.

Based on the above description, granular computing is effective in simplifying complex problems by breaking it down into several sub-problems in practice. It can also be used to quantitatively measure qualitative properties in the context of information granulation. In practical applications, the theory of granular computing has been widely used to promote other research fields, such as computational intelligence (Ejegwa, 2018; Khan et al., 2018) and artificial intelligence (Garg and Kaur, 2018; Mandal and Ranadive, 2018).

3 Overview of The Proposed Method

In this section, we describe the details of local deep-learned feature extraction and present the multi-classifier fusion framework. As illustration in Figure 1, the proposed method consists of four steps: 1) generating region proposal, 2) transfer learning, 3) reduction of feature dimension, and 4) classification. The main components in our method will be described in detail. In addition, we will analyze the creativity of the method from the perspective of granular computing.

3.1 Region proposal and transfer learning

In the setting of granular computing(Liu and Cocea, 2018), a granule generally represents a large particle, which consists of smaller particles that can form a larger unit. Different with most existing methods which use global features extracted from whole images, we consider each image x as a granular and obtain a collection of local features from sub-granule: $x = \{x_1, x_2, ...x_n\}$. So we capture context information from neighboring scenes and objects while preserving key local features. We start our work with a set of region proposals from images to pursue accuracy with affordable computing costs, each region proposal is viewed as a sub-granule of the original image. After observing the experimental results, we find that the top 1000 ranked region proposals are sufficient for the representation of an image.

Once we have the 1000 region proposals which were generated from the original images by EdgeBoxes algorithm, we start the transfer learning in the second stage. We formalize transfer learning as follows: Given a source domain D_S and a target domain D_T, the learning task for D_S and D_T are T_S and T_T, respectively. We aim to use the knowledge from D_S and T_S to boost the learning ability of the target predictive function $f_T(\cdot)$ in T_T, where $D_S \neq D_T$, $T_S \neq T_T$. Transfer learning is particularly relevant when, given labeled source domain data D_S and target domain data D_T, we find that $|D_T| \ll |D_S|$.

In this paper, we transfer knowledge from the ImageNet object recognition task P_1 to the target problem of traffic scene recognition P_2. In P_1, we have the task of object classification with source domain data $D_1 = \{(x_{1_i}, y_{1_i})\}$ from ImageNet that consists of natural images $x_{1_i} \in X_1$ with labels. In P_2, we have a traffic scene prediction task with target domain data $D_2 = \{(x_{2_i}, y_{2_i})\}$ that consists of traffic scene images $x_{2_i} \in X_2$ and image labels. ImageNet is an object classification image dataset which consists of 14 million images belonging to 1000 classes, major breakthroughs have been achieved with the help of sufficient data and CNN models in many computer vision tasks. CNN models trained on the ImageNet dataset are recognized as good generic feature extractors, with low-level and mid-level features such as edges and corners that are able to generalize to many new tasks. We achieve knowledge transfer using the parameters from VGG16 models trained on ImageNet. The VGG16 model has been fine-tuned on the traffic scene dataset using SGD with momentum.

We consider two ways of adapting the original VGG16 network. The first approach is to add a dropout layer before the final convolutional layer to reduce the risk of overfitting. Second, we modify the last fully-connected layer to have K neurons to predict the K- classes, where K is the number of the traffic scene types in the training set. We regard the traffic scene recognition as a multi-class classification problem, and apply the cross-entropy loss to transfer the model outputs to the value of probability for all classes. This corresponds to

$$l = -\sum_{k=1}^{K} log(\sigma p(k) q(k)) \tag{3}$$

where σ denotes the softmax activation function, $p(k) \in [0, 1]$ is the predictive probability of the input image belonging to class K, and $q(k)$ denotes the ground truth distribution. Different with some methods which obtain feature from pooling layer, we extract the 4096-dimensional feature vector from the first full connection layer (FC6) for the region proposals generated from each image. However, it's time-consuming to extract feature of multiple regions (sub-granule) in CNN. To reduce the computational cost and run time, we implement our algorithm on top of a fast R-CNN (Girshick, 2015), in which the RoI projection scheme will complete the feature extraction of an image in only one feed forward process. Fast R-CNN is originally used for object detection and requires object category labels and annotations of bounding boxes. Usually, the annotations

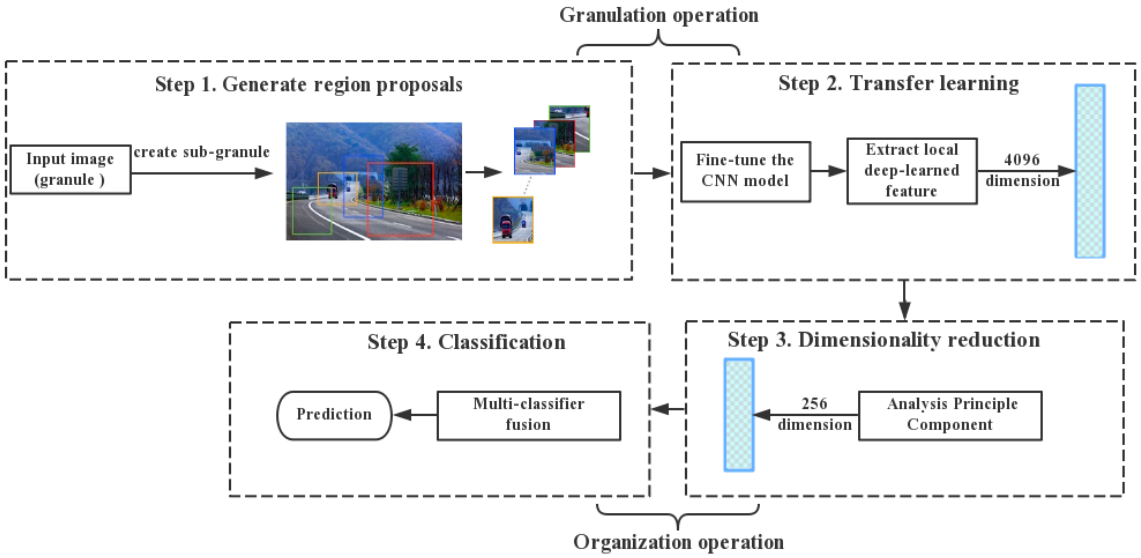

Fig. 1 The workflow of our proposed traffic scene recognition system. The granulation operation includes generate region proposals (sub-granule) for each image (granule) and perform transfer learning to obtain local deep-learned feature. In the organization operation, we analyze the principle component to reorganize the local deep-learned feature and reduce the dimension of it. Besides, the type of traffic scenes can be recognized with multi-classifier fusion that also belongs to the organization operation.

are done manually in general applications. In our work, the parts instances are viewed as objects and annotated automatically. We show the feature extraction process in Figure 2.

3.2 Dimensionality reduction

As has being pointed out in (Jégou et al., 2010), reducing the dimension of the original feature appropriately would further improve the recognition performance. Therefore, after extracting the CNN features from regions, we used principal component analysis (Abdi and Williams, 2010) to reduce feature dimension. However, it is not practical to perform conventional PCA training on all features due to the large number of features. We first select some sample features randomly for training and reduce the CNN features from 4096 to 256 dimensions. Then we perform PCA on all remaining features. In addition, we further investigate the effect of feature dimensions on overall recognition performance by comparing the performance of 512 dimensions.

3.3 Design of Multi-classifier Fusion Framework

There are two principles for multi-classifier fusion: a) each individual classifier has its own advantages; b) as indicated in (Zhou, 2012), complementary advantages could to be achieved by encouraging diversity among different classifiers.

Algorithm 1 Proposed traffic scene recognition pipeline

Input: Static traffic scene recognition dataset D including D_{train}, D_{val} and D_{test}
Output: The prediction labels for D_{test}
/*Granulation operation*/
1: Create region proposal (sub-granule) for traffic scene images (granule) in D.
2: Perform transfer learning using D_{train} and D_{val} (see Section 3.1).
3: Extract the local deep-learnt feature matrix H_{train}, H_{val} and H_{test} of selected regions for each images in D_{train}, D_{val} and D_{test}.
/*Organization operation*/
4: Analyze the principal components in H_{train} to obtain the transformation matrix T.
5: **for** $i = 1$ to D_{train} **do**
6: Use the first i transformation vectors of T to compute $H_{train_{transform}}$ by projecting H_{train} to the subspace of principal components.
7: Evaluate the performance of $H_{train_{transform}}$ and save the result as $scores_i$
8: **end for**
9: Obtain the i in which the $H_{train_{transform}}$ achieves the best scores. T_{select} is the first i transformation vectors of T.
10: Compute L_{train}, L_{val} and L_{test} by projecting H_{train}, H_{val} and H_{test} to the principal components subspace using W_{select}
11: Train three basic classifiers KNN, SVM and MLP and two decision tree ensembles RF and GBT using L_{train}, L_{val}.
12: Obtain the posterior probability matrix P_{test} of three basic classifiers and two decision tree ensembles on L_{test}
13: Fuse the multiple P_{test} using algebraic rules.

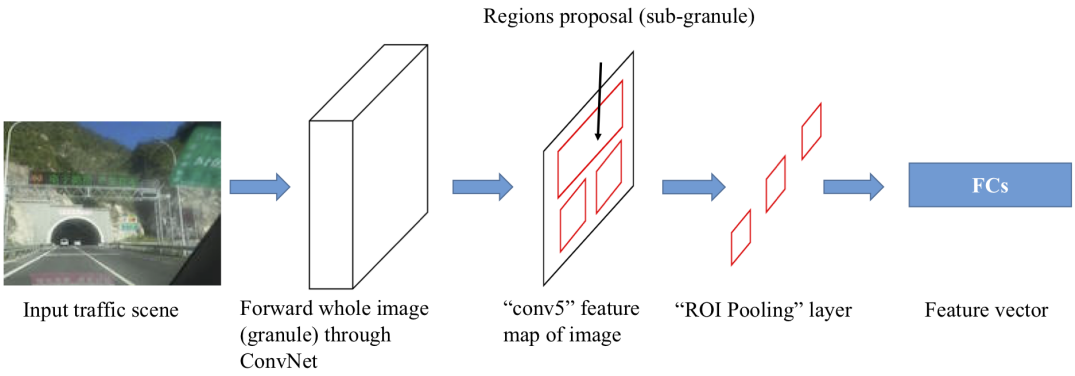

Fig. 2 The process of deep feature extraction. Forward the input traffic scene images (granule) contains a set of region proposals (sub-granule) through CNN model, after generating the conv5 feature map of image, the RoI pooling layer will extract features with one feed forward process.

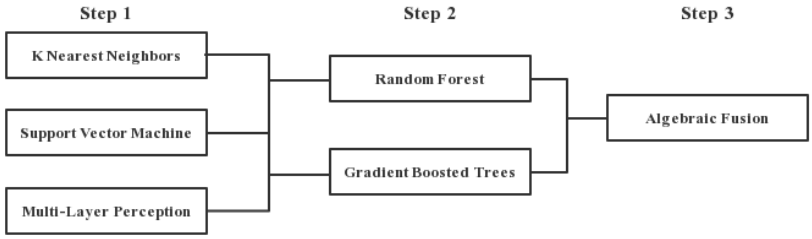

Fig. 3 Pipeline of the multi-classifier fusion. Step 1: train three single classifiers; Step 2: In order to increase the diversity of decision tree classifiers, two decision tree sets are trained by using RF and GBT respectively; Step 3: the trained single classifiers is combined with the decision tree sets through algebraic fusion.

Fig.3 shows the process of multi-classifier fusion. Firstly, we train several single classifiers including the popular SVM, KNN and MLP that have different learning strategies. To boost the recognition performance, in step 2, more diverse decision trees are obtained by training two decision tree ensembles including Random Forests and Gradient Boosted Trees. To reduce the risk of over-fitting and improve the level of generalization, we adopt the 10-fold cross validation to train and validate each classifier.. Finally, we apply the algebraic rule to fuse the results of the two ensemble classifiers with single classifiers for further improving the recognition performance. In particular, the proposed method involves the different levels of granularity. Each ensemble can be viewed as a granule, Random Forests and Gradient Boosted Trees are two independent granules. The final ensemble are organized to include the two ensembles and the single classifiers. Each of the levels of ensembles actually represents a level of granularity.

Voting is the most popular method of classifier combination in the field of classifier fusion. In particular, voting-based set classification can be achieved by selecting the classes provided by most classifiers as their output, i.e. majority voting. In particular, classes provided by most classifiers will be selected as outputs, i.e. majority voting. In this way, voting-based ensemble classification will be realized.

Different with majority voting, weighted voting is another way of voting in which the class output is calculated with the weight of each single classifier. The class that obtains the highest weight will be derived for finally classifying an instance. The overall confidence (accuracy) of a classifier evaluated on a validation set will be used to estimate the weight of this classifier.

The precision or recall for a specific class are also used to measure the confidence in the class level (Liu and Gegov, 2015). Also, due to the high degree of diversity between different instances, the confidence in classifying an instance cannot be represented by the confidence level measured for the classifier or each individual class. In our proposed framework, we use algebraic rules (Zhou, 2012), which are basically based on the median/maximum/average of hidden output (posterior probability of each class) to achieve the fusion of these classifiers trained by using different learning algorithms. Our traffic scene recognition algorithm is summarized in Algorithm 1.

3.4 Application of granular computing concepts

We design the deep multi-classifier fusion method in the setting of granular computing, which is a paradigm of information processing. In the local deep-learned feature extraction part, granulation is operated through decomposing the information of original images into multiple region proposals which involves local information. Organization is operated through analyzing the principal components to reduce the feature dimension. Different with general feature selection, we reorganize the various feature into a low-dimension feature with no information loss. A principal component is a feature that is regarded as a large information particle, which contains a plurality of features called small information particles. The whole process of dimensionality reduction belongs to information fusion, which utilizes organizational operations in granular computing.

On the other hand, the framework of multi-classifiers fusion involves multiple levels of classifiers fusion, and we view each of the levels as a specific level of granularity. In this setting, a primary ensemble containing three base classifiers is viewed as a granule at the basic level of granularity, whereas the final ensembles that may involves both base classifiers and lower level ensembles is viewed as a granule at the top level of granularity.

Multi-classifier Fusion vs. Deep multi-classifier fusion. Multi-classifier fusion and the proposed deep multi-classifier fusion have the same objective of outputting the prediction labels for testing data. Multi-classifier fusion focuses on classification task and leverages different classifiers to improve the performance. Deep multi-classifier fusion seamlessly integrates the two components including local deep-learned features extraction framework (step 1 to step 3) and multi-classifier fusion into a unified system. In principle, the two components should collaborate with each other effectively: the former operation of granulation is essentially decomposition of a whole into multiple parts in a top-down information processing manner through extraction features for local patches through the FC6 layer of CNN, whereas the latter organization operation is essentially integration of multiple parts into a whole in a bottom-up information processing manner through achieving the complementary advantages of different classifiers.

4 Experiments and Results

We will first describe the implementation details, and then briefly outline the experimental set up and performance comparison on the WZ-traffic dataset and FM2 dataset.

4.1 Implementation Details

Deep Feature Extraction. Our experiments were conducted under the Linux operating system. The implementation of the deep feature extraction was undertaken on the Caffe deep learning framework (Jia et al., 2014). We employed the VGG16, VGG1024 and Cafenet models which were pre-trained on ImageNet, and then fine-tuned on specific datasets. We set the maximum number of training iterations and the learning rate to 10000 and 0.0001, respectively. Other parameters are the same as fast R-CNN (Girshick, 2015).

Setting of Multi-classifier Fusion. The multi-classifier fusion experiment was built on the KNIME Analysis Platform, which has abundant nodes for applying machine learning algorithms. All experiments were conducted with 10-fold cross-validation. We divided each dataset into 10 parts including 7 parts for training and 1 parts for validation and the rest for testing. The performance of three popular standard learning methods, SVM, KNN and MLP, were first evaluated. We used the RBF kernel in the SVM learner and set the values of the sigma and overlap penalty to 13 and 1, respectively.

For K nearest neighbor, we set the value of K equal to 7. In addition, we trained the MLP classifier through 150 iterations with 2 hidden layers and 10 units in each layer. Secondly, we used the random forest learner (RF) and Gradient boosted trees learner (GBT) to improve the performance of decision tree learning. As for random forest learner, the information gain ratio was used for for split criterion in tree ensemble learner, we set the ensemble size which means the number of decision trees that make up a random forest to 150. In addition, for gradient boosted trees learner, the tree depth, number of models and learning rate were set as 10, 20 and 0.1, respectively. In multi-classifier fusion stage, mean, median, maximum rule of algebraic fusion were used to boost the prediction accuracy.

4.2 WZ-traffic dataset

To facilitate the research on traffic scene recognition, we created a new dataset of labeled traffic scenes, called the WZ-traffic dataset (Wu, 2019). It contains 6035 labeled images of 20 categories: highway, country road, gas station, indoor parking, outdoor parking, crossing, city stress, scenic gate, bridge, car wash, train station, autodrome, traffic circle, tunnel, tunnel entrance, bus station, booth, bus parking and traffic jams. The images were collected by us from both the image search engine as well as personal photographs, and took into account sufficient variations in the background and viewpoints.

Table 1 VGG16: Mean AP result on the WZ-traffic dataset (Wu, 2019) with different methods.

Method	Mean AP(%)
FC6 features (pre-trained model) (Simonyan and Zisserman, 2014)	83.12
FC6 features (fine-tuned model)	85.71
1000 regions+FC6 features+PCA256	**87.43**
1000 regions+FC6 features+PCA512	87.10
2000 regions+FC6 features+PCA256	87.30
3000 regions+FC6 features+PCA256	87.12

Fig.4 presents sample examples from the corresponding traffic scene categories in this dataset.

Fig. 4 Some examples of the WZ-traffic dataset.

We followed the step of deep feature extraction as previously explained, and applied multiple classifiers fusion for the final prediction. To compare and evaluate the performance from different models, we selected the pre-trained CNN models VGG16, VGG-M-1024 and CaffeNet for following fine-tuning. We implemented the training process in the fast R-CNN framework (Girshick, 2015). After applying the region proposal algorithm EdgeBoxes (Zitnick and Dollár, 2014) on each image, and we extracted FC6 features from each region. Multi-classifier fusion was accomplished after PCA dimensionality reduction and feature clustering. More details about the experiment procedure are described as follows:

(1) Result from VGG16.

First, the fine-tuned CNN model and the pre-trained CNN model were applied for extracting FC6 features. As shown in Table 1, with the same experimental setting, the fine-tuned model obtains about a 3% improvement (from 83.12% to 85.71%) in recognition performance. This result indicates that the fine tuning of the CNN model can significantly boost the feature representation ability. Then, we provided recognition results for 2000 and 3000 boxes per image to verify that 1000 regions per image are sufficient for deep feature representation. From Table 1, we can clearly observed that 1000 boxes yields the best performance. To reduce the feature dimension, the PCA was used to reduce the CNN features from 4096 dimensions to 256. We repeated the same experimental process and reduced the CNN features to 512 dimensions for comparison. It can be clearly seen from Table 1, the mAP results of 512 dimensions are slightly worse. Hence, the CNN features of 1000 regions with 256 dimensions will be the focus for most of the experiments.

Table 2 VGG16: Mean AP result on the WZ-traffic dataset (Wu, 2019) with individual and fusion classifiers.

Method (VGG16)	Mean AP(%)
MLP (Gardner and Dorling, 1998)	87.43
SVM (Cortes and Vapnik, 1995)	88.26
KNN (Altman, 1992)	86.57
RF (Cutler et al., 2004)	86.43
GBT (Friedman, 2002)	88.06
Median-based fusion	89.90
Maxmium-based fusion	90.15
Mean-based fusion	**90.30**

Table 2 shows the results for different single classifiers and multi-classifier fusion. The outputs of the two ensemble classifiers (Random Forests and Gradient Boosted Trees) were fused with the outputs of the 3 single classifiers for advancing the performance further. The three multi-classifier fusion methods prove their capabilities of improving the recognition performance comparing with the single classifiers. The highest recognition rates was obtained from the mean-based fusion method. Overall, the results show very supportive evidence for multi-classifier fusion towards advancing the overall classification performance.

(2) Results from VGG-M-1024 and CaffeNet.

To compare the performance with other CNN models, we select the middle scale CNN model VGG-M-1024 and small scale CNN model CaffeNet. The experiments were undertaken using the same conditions as VGG 16. Comparing the deep multi-classifier fusion methods result with VGG-M-1024 and CaffeNet which are 88.90% and 88.11%, respectively, in terms of the recognition rate, VGG16 performs better than VGG-M-1024 and CaffeNet models. Fig.5 shows the confusion matrix of our best recognition results on the WZ-traffic dataset. From the confusion matrix, we can observe that the proposed method performed well in recognizing tunnel, traffic circle and car wash. For the other types of traffic scene, our method also performed reasonably well.

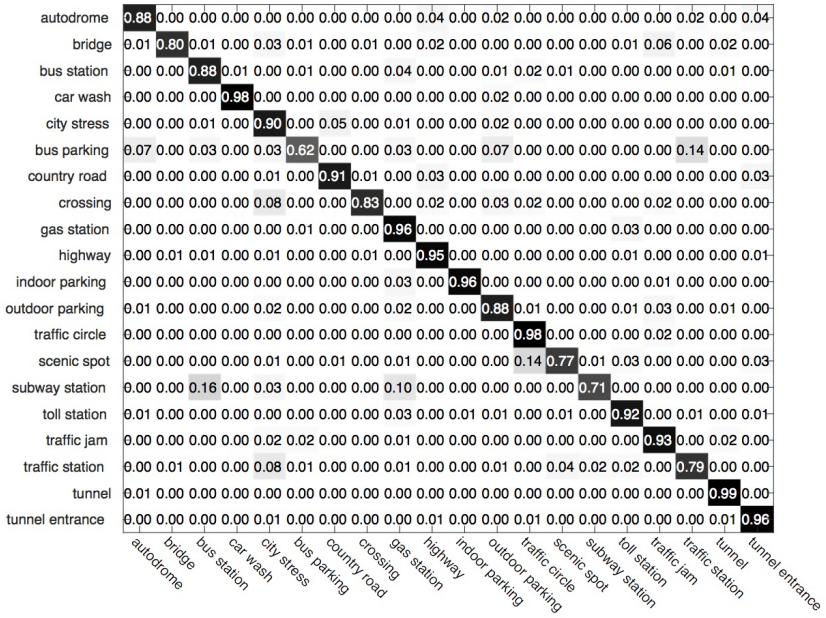

Fig. 5 Confusion matrix of the best recognition results on the WZ-traffic dataset (Wu, 2019) (mean AP is 90.30%). The labels in the leftmost column and on the bottom represent the ground truth, the number in each row represents the corresponding prediction results.

4.3 FM2 dataset

The FM2 dataset has been introduced by Sikiric et al. (Sikiric et al., 2014) and contains 6237 traffic scene images from the perspective of the driver. The images were extracted from videos of several drives on European roads, obtained using a camera installed in a vehicle. The traffic scene consists of dense traffic,highway, overpass, road, tunnel, exit, toll booth and settlement. Fig.6 provides some examples of the traffic dataset FM2.

Fig. 6 Some examples of the FM2 Dataset (Sikiric et al., 2014).

There is no ground-truth region provided in FM2 dataset, therefore, we fine-tuned the pre-trained VGG16 model which achieved the best performance on the WZ-traffic dataset compared with VGG-M-1024 and CaffeNet. When the training process of CNN model was completed, we extracted the CNN features for the top 1000 regions that produced from Edgeboxes. Multi-classifier fusion was accomplished after PCA dimensionality reduction. We can observe the following results from Table 3 and Table 4: On this dataset, satisfactory results are obtained when only the image-level CNN features are considered. Besides, the performance increased 0.87% (from 96.25% to 97.12%) when we implement the multi-classifier fusion on CNN feature. This improvement proves the complementarity of multi-classifier fusion and CNN

Table 3 VGG16: Mean AP result on the FM2 dataset (Sikiric et al., 2014) with individual and fusion classifiers.

Method (VGG16)	Mean AP(%)
FC6 features(pre-trained model) (Simonyan and Zisserman, 2014)	93.41
FC6 features(fine-tuned model)	95.65
PCA256+FC6 features	**96.25**

Table 4 VGG16: Mean AP result on the FM2 dataset (Sikiric et al., 2014) with individual and fusion classifiers.

Method (VGG16)	Mean AP(%)
MLP (Gardner and Dorling, 1998)	96.25
SVM (Cortes and Vapnik, 1995)	96.46
KNN (Altman, 1992)	95.87
RF (Cutler et al., 2004)	96.13
GBT (Friedman, 2002)	95.70
Median-based fusion	96.82
Maxmium-based fusion	96.95
Mean-based fusion	**97.12**

features. Compared with the other methods shown in Table 5, we also obtained the most state-of-the-art results on FM2 dataset.

We describe the details about the experimental process and the three comparison settings as follows:

(1) CNN features.

We directly extracted the CNN features from the first fully connected layers of the fine-tuned VGG16 model for each images without applying the region proposal algorithm to generate candidate objects. As shown in Table 3, the mAP accuracy is 95.65%. We evaluate the stand-alone performance of the fine-tuned VGG16 model by comparing the results of pre-trained VGG16 model in Table 3. The fine-tuned model leading to over 2.24% improvement (from 93.41% to 95.65%) over the pre-trained model.

(2) Dimension reduction.

In the test phase, we used the Edgeboxes algorithm to generate 1000 region proposal for each image, which are represented by 4096-dimensional CNN features. To reduce the amount of feature computation and improve the performance, we perform dimension reduction on CNN feature through PCA algorithm and reduce the CNN features to 256 dimensions. Table 3 shows the mAP accuracy increases 0.6% (from 95.65% to 96.25%). In this setting, the multi-classifier fusion has not been taken into account.

(3) Deep multi-classifier fusion.

Finally, we fuse the hidden outputs (probability for each class) of SVM, KNN, MLP, RF and GBT classifiers through the mean, median, maximum rule of algebraic fusion. Table 4 shows the detailed comparison results between our methods and five single classifiers baseline methods. Experimental results indicate that adding the multi-classifier fusion does improve the overall performance and the best performance in terms of mAP accuracy of mean-based fusion is 97.12%, Fig.7 presents the confusion matrices of the best recognition results on the FM2 database. From the confusion matrix, we can see that the proposed method recognizes most of the traffic scene well, such as highway, tunnel and settlement. Fig.8 shows some correctly recognized examples in this dataset. For example, our method recognized Fig.8(a) as booth with a 99.99% (0.9999) probability. We also compared our method with the state-of-the-art method in Table 5, and the comparisons indicate the competitiveness of the proposed method on the FM2 dataset.

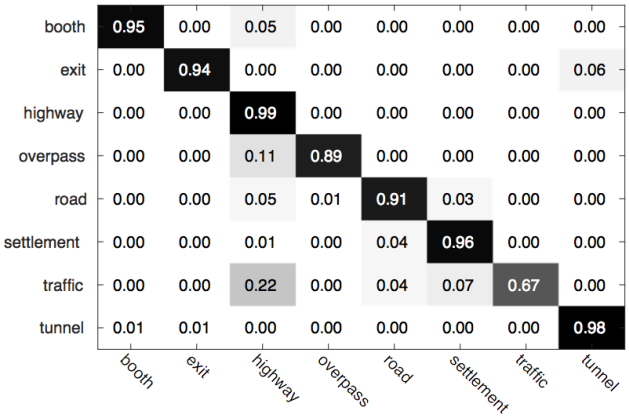

Fig. 7 Confusion matrix of the best recognition result (mean AP is 97.12%) on the FM2 database (Sikiric et al., 2014). The labels in the leftmost column and on the bottom represent the ground truth, the number in each row represents the corresponding prediction results.

Fig. 8 Some examples of correct recognition in the FM2 dataset (Sikiric et al., 2014), the predicted label and corresponding probability are provided for each image.

Table 5 Mean AP result on the traffic scene dataset FM2 with previous results in (Sikiric et al., 2014).

Method	Mean AP(%)
BoW (Csurka et al., 2004)	93.55
LLC (Wang et al., 2010)	92.68
SFV (Krapac et al., 2011)	95.09
GIST (Oliva and Torralba, 2001)	93.30
Ours (Deep Multiple Classifiers Fusion)	**97.12**

5 Conclusion

In this paper, we have proposed a traffic scene recognition system using local deep-learned features and multi-classifiers fusion in the setting of granular computing. We have designed to create information granulation through extracting CNN features of region proposal generated for each image. In addition, organization is operated by analyzing the principal components to reduce the feature dimension. The multi-classifier fusion method which involves multiple levels of granularity to improve the performance. In practice, we use the local deep-learned features to train three basic classifiers, namely KNN, SVM and MLP. Furthermore, RF and GBT are used to train two decision tree ensembles. Finally, we apply the three algebraic rules, mean, median and maximum to fuse the above classifiers. We conduct

experiments on two different traffic scene datasets, including public dataset and our own dataset. The experimental results show that the information of the local patches and the global background are significant to improve the performance of traffic scene recognition, while the deep multi-classifiers fusion method brings performance improvement to traffic scene recognition. In the future, the deep multi-classifier fusion will be further improved to study the relationship between classes in a granular computing setup. Specifically, we will identify the relationships between information granules where each class is viewed as a granule. Besides, it is worth of future research to use fuzzy sets (Chen and Chang, 2011; Cheng et al., 2016; Chen and Huang, 2003) to deal with the deep multiple classifiers fusion.

References

Abdi H, Williams LJ (2010) Principal component analysis. Wiley interdisciplinary reviews: computational statistics 2(4):433–459

Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. American Statistician 46(3):175–185

Chen CY, Choi W, Chandraker M (2016) Atomic scenes for scalable traffic scene recognition in monocular videos. In: Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, IEEE, pp 1–9

Chen SM, Chang YC (2011) Weighted fuzzy rule interpolation based on ga-based weight-learning techniques. IEEE Transactions on Fuzzy Systems 19(4):729–744

Chen SM, Huang CM (2003) Generating weighted fuzzy rules from relational database systems for estimating null values using genetic algorithms. IEEE Transactions on Fuzzy Systems 11(4):495–506

Chen SM, Tanuwijaya K (2011) Fuzzy forecasting based on high-order fuzzy logical relationships and automatic clustering techniques. Expert Systems with Applications 38(12):15425–15437

Chen SM, Wang JY (1995) Document retrieval using knowledge-based fuzzy information retrieval techniques. IEEE Transactions on Systems, Man, and Cybernetics 25(5):793–803

Cheng SH, Chen SM, Jian WS (2016) Fuzzy time series forecasting based on fuzzy logical relationships and similarity measures. Information Sciences 327:272–287

Cortes C, Vapnik V (1995) Support-Vector Networks. Kluwer Academic Publishers

Csurka G, Perronnin F (2010) Fisher vectors: Beyond bag-of-visual-words image representations. In: International Conference on Computer Vision, Imaging and Computer Graphics, Springer, pp 28–42

Csurka G, Dance C, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV, Prague, vol 1, pp 1–2

Cutler A, Cutler DR, Stevens JR (2004) Random forests. Machine Learning 45(1):157–176

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, IEEE, vol 1, pp 886–893

Dixit M, Chen S, Gao D, Rasiwasia N, Vasconcelos N (2015) Scene classification with semantic fisher vectors. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2974–2983

Duin RP, Tax DM (1998) Classifier conditional posterior probabilities. In: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Springer, pp 611–619

Ejegwa PA (2018) Distance and similarity measures for pythagorean fuzzy sets. Granular Computing pp 1–14

Ess A, Müller T, Grabner H, Van Gool LJ (2009) Segmentation-based urban traffic scene understanding. In: BMVC, vol 1, p 2

Friedman JH (2002) Stochastic gradient boosting. Computational statistics and data analysis 38(4):367–378

Gardner MW, Dorling S (1998) Artificial neural networks (the multilayer perceptron)?a review of applications in the atmospheric sciences. Atmospheric environment 32(14-15):2627–2636

Garg H, Kaur G (2018) Novel distance measures for cubic intuitionistic fuzzy sets and their applications to pattern recognitions and medical diagnosis. Granular Computing pp 1–16

Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448

Gondy LA, Thomas CRB, Bayes N (1993) Programs for machine learning. Advances in Neural Information Processing Systems 79(2):937–944

Greene MR, Botros AP, Beck DM, Fei-Fei L (2015) What you see is what you expect: rapid scene understanding benefits from prior experience. Attention, Perception, & Psychophysics 77(4):1239–1251

Huang T, Koller D, Malik J, Ogasawara G, Rao B, Russell SJ, Weber J (1994) Automatic symbolic traffic scene analysis using belief networks. In: AAAI, vol 94, pp 966–972

Jégou H, Douze M, Schmid C, Pérez P (2010) Aggregating local descriptors into a compact image repre-

sentation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, pp 3304–3311

Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia, ACM, pp 675–678

Khan MSA, Abdullah S, Ali A, Amin F, Rahman K (2018) Hybrid aggregation operators based on pythagorean hesitant fuzzy sets and their application to group decision making. Granular Computing pp 1–14

Krapac J, Verbeek J, Jurie F (2011) Modeling spatial layout with fisher vectors for image categorization. In: 2011 International Conference on Computer Vision, IEEE, pp 1487–1494

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105

Kuncheva LI (2002) A theoretical study on six classifier fusion strategies. IEEE Transactions on pattern analysis and machine intelligence 24(2):281–286

Lee LW, Chen SM (2008) Fuzzy risk analysis based on fuzzy numbers with different shapes and different deviations. Expert Systems with Applications 34(4):2763–2771

Liu H, Cocea M (2018) Granular computing-based approach of rule learning for binary classification. Granular Computing pp 1–9

Liu H, Gegov A (2015) Collaborative Decision Making by Ensemble Rule Based Classification Systems. Springer International Publishing

Liu H, Gegov A, Cocea M (2016) Rule-based systems: a granular computing perspective. Granular Computing 1(4):259–274

Liu H, Cocea M, Ding W (2018) Multi-task learning for intelligent data processing in granular computing context. Granular Computing 3(3):257–273

Lu K, Li J, An X, He H (2016) Generalized haar filter based deep networks for real-time object detection in traffic scene. arXiv preprint arXiv:161009609

Mandal P, Ranadive A (2018) Hesitant bipolar-valued fuzzy sets and bipolar-valued hesitant fuzzy sets and their applications in multi-attribute group decision making. Granular Computing pp 1–25

Mioulet L, Breckon TP, Mouton A, Liang H, Morie T (2013) Gabor features for real-time road environment classification. In: Industrial Technology (ICIT), 2013 IEEE International Conference on, IEEE, pp 1117–1121

Nanni L, Lumini A (2013) Heterogeneous bag-of-features for object/scene recognition. Applied Soft Computing 13(4):2171–2178

Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision 42(3):145–175

Oquab M, Bottou L, Laptev I, Sivic J (2014) Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1717–1724

Pedrycz W (2011) Information granules and their use in schemes of knowledge management. Scientia Iranica 18(3):602–610

Pedrycz W, Chen SM (2015) Granular computing and decision-making: interactive and iterative approaches, vol 10. Springer

Sikiric I, Brkic K, Krapac J, Segvic S (2014) Image representations on a budget: Traffic scene classification in a restricted bandwidth scenario. In: Intelligent Vehicles Symposium Proceedings, 2014 IEEE, IEEE, pp 845–852

Simon M, Rodner E, Denzler J (2014) Part detector discovery in deep convolutional neural networks. In: Asian Conference on Computer Vision, Springer, pp 162–177

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556

Song X, Jiang S, Herranz L (2015) Joint multi-feature spatial context for scene recognition on the semantic manifold. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1312–1320

Tang I, Breckon TP (2011) Automatic road environment classification. IEEE Transactions on Intelligent Transportation Systems 12(2):476–484

Taylor P, Griffiths N, Bhalerao A, Anand S, Popham T, Xu Z, Gelencser A (2016) Data mining for vehicle telemetry. Applied Artificial Intelligence 30(3):233–256

Toufiq R, Isalm MR (2017) Face recognition system using soft-output classifier fusion method. In: International Conference on Electrical, Computer and Telecommunication Engineering, pp 1–4

Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y (2010) Locality-constrained linear coding for image classification. In: 2010 IEEE computer society conference on computer vision and pattern recognition, Citeseer, pp 3360–3367

Wu F (2019) WZ-traffic dataset. `https://github.com/Fangyu0505/traffic-scene-recognition`,

[Online; accessed 10-March-2019]

Wu F, Yan S, Smith JS, Zhang B (2018) Joint semi-supervised learning and re-ranking for vehicle re-identification. In: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, pp 278–283

Yan S, Smith JS, Zhang B (2017) Action recognition from still images based on deep vlad spatial pyramids. Signal Processing: Image Communication 54:118–129

Yao J (2005a) Information granulation and granular relationships. In: 2005 IEEE International Conference on Granular Computing, IEEE, vol 1, pp 326–329

Yao Y (2005b) Perspectives of granular computing. In: 2005 IEEE international conference on granular computing, IEEE, vol 1, pp 85–90

Zadeh LA (1965) Fuzzy sets. Information and control 8(3):338–353

Zavaschi THH, Jr ASB, Oliveira LES, Koerich AL (2013) Fusion of feature sets and classifiers for facial expression recognition. Expert Systems with Applications 40(2):646–655

Zhou ZH (2012) Ensemble Methods: Foundations and Algorithms. Taylor and Francis

Zitnick CL, Dollár P (2014) Edge boxes: Locating object proposals from edges. In: European Conference on Computer Vision, Springer, pp 391–405