Research Article

A deep action-oriented video image classification system for text detection and recognition

Abhra Chaudhuri[1] · Palaiahnakote Shivakumara[2] · Pinaki Nath Chowdhury[1] · Umapada Pal[1] · Tong Lu[3] · Daniel Lopresti[4] · G. Hemantha Kumar[5]

Received: 10 May 2021 / Accepted: 22 September 2021
Published online: 09 October 2021

Abstract

For the video images with complex actions, achieving accurate text detection and recognition results is very challenging. This paper presents a hybrid model for classification of action-oriented video images which reduces the complexity of the problem to improve text detection and recognition performance. Here, we consider the following five categories of genres, namely concert, cooking, craft, teleshopping and yoga. For classifying action-oriented video images, we explore ResNet50 for learning the general pixel-distribution level information and the VGG16 network is implemented for learning the features of Maximally Stable Extremal Regions and again another VGG16 is used for learning facial components obtained by a multitask cascaded convolutional network. The approach integrates the outputs of the three above-mentioned models using a fully connected neural network for classification of five action-oriented image classes. We demonstrated the efficacy of the proposed method by testing on our dataset and two other standard datasets, namely, Scene Text Dataset dataset which contains 10 classes of scene images with text information, and the Stanford 40 Actions dataset which contains 40 action classes without text information. Our method outperforms the related existing work and enhances the class-specific performance of text detection and recognition, significantly.

Article highlights

1. The method uses pixel, stable-region and face-component information in a noble way for solving complex classification problems.
2. The proposed work fuses different deep learning models for successful classification of action-oriented images.
3. Experiments on our own dataset as well as standard datasets show that the proposed model outperforms related state-of-the-art (SOTA) methods.

Keywords Deep neural networks · Face detection · Action image classification · Text detection · Text recognition

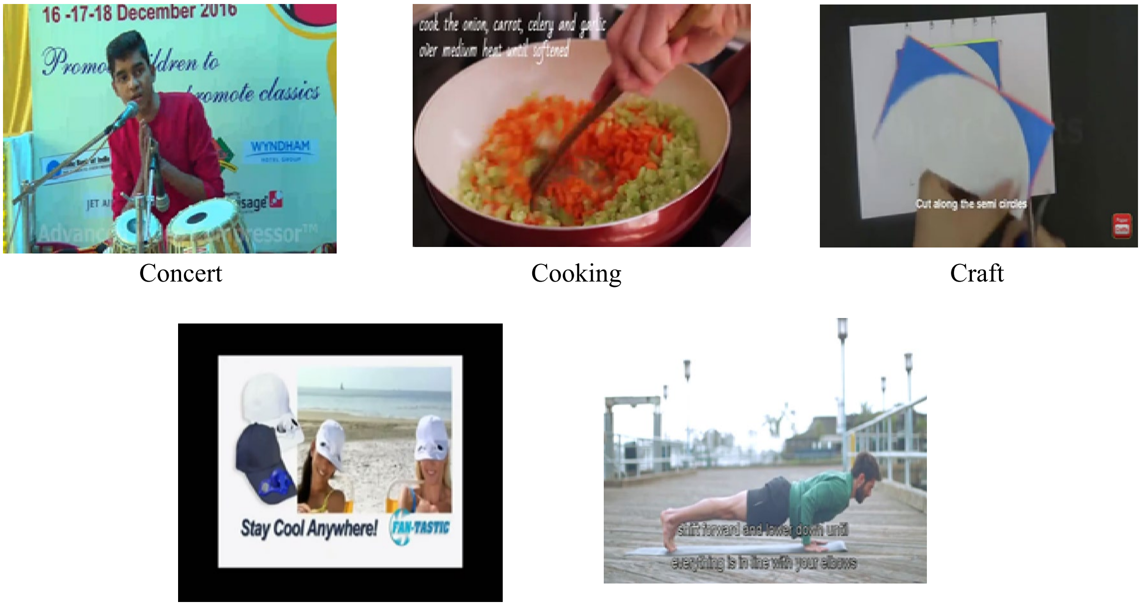
✉ Palaiahnakote Shivakumara, shiva@um.edu.my; Abhra Chaudhuri, abhrachaudhuri.97@gmail.com; Pinaki Nath Chowdhury, pinakinathc@gmail.com; Umapada Pal, umapada@isical.ac.in; Tong Lu, lutong@nju.edu.cn; Daniel Lopresti, lopresti@cse.lehigh.edu; G. Hemantha Kumar, ghk.2007@yahoo.com | [1]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India. [2]Department of Computer System and Technology, Universiti Malaya, Kuala Lumpur, Malaysia. [3]National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China. [4]Computer Science & Engineering, Lehigh University, Bethlehem, PA, USA. [5]Department of Studies in Computer Science, University of Mysore, Mysore, India.

1 Introduction

Understanding action in images and videos is of growing interest for the researchers in the field of artificial intelligence and image understanding related fields. It plays a significant role in real world applications such as activity monitoring, interactions between computers and humans, and video indexing [1–3] etc. The problem remains an ongoing challenge due to factors like uneven illumination effects, partial occlusion, and complex background [4]. However, since temporal information is involved in the process of action detection, the methods consume a greater number of computations [5]. This has motivated many researchers to develop methods which can cope with the challenges of still images for action detection and recognition [6, 7]. Although these methods are more efficient, they are not as accurate as video-based methods. Hence, there is a wide gap between still-image and video-based methods in understanding the content of such images. This issue motivated us for considering text detection and recognition to enhance the performance of action image detection and recognition. The advantage of text detection and recognition is that the process does not require temporal information unlike existing methods that explore temporal frames for action detection [8].

Several methods for detecting and recognizing text in natural scene images can be found in the literature [9–12]. These address many challenges such as arbitrarily oriented text, multi-script text, arbitrarily shaped text etc. However, since most of these methods were developed for natural scene images without much action in them, they may not perform well for our applications of interest because the presence of actions degrades the text information. As a result, text in action images loses quality, contrast, and sharpness and hence hurts the performance of existing methods. As seen from the sample images shown in Fig. 1, which illustrate five classes of interest (Concert, Cooking, Craft, Teleshopping and Yoga), text can be adversely impacted by action in an image, as well as by other confounding factors including poor quality, low contrast, and uneven illumination. This provides motivation for classifying video images based on different actions, allowing the choice of an appropriate method tuned to the complexity of the specific problem to improve text detection and recognition performance. For example, for the images in Fig. 1 for cooking and craft, the background is an indoor scene, while for concert, teleshopping and yoga, the background is an outdoor scene. Likewise, for cooking, craft and yoga actions classes, caption text (which is edited text) is a prominent feature, while for concert and teleshopping, scene text (which is part of the image) is a prominent feature. Based on these observations, we propose a new method for classification of action images drawn from five classes in this work.

Inspired by the tremendous discriminative ability of deep CNNs [13–17] in complex tasks, we explore their use for the classification of action images. The proposed


16 -17-18 December 2016
Promo ldren to
romote classics
WYNDHAM
cook the onion, carrot, celery and garlic
over medium heat until softened
Cut along the semi circles
Stay Cool Anywhere!
FAN-TASTIC
shift forward and lower down until
everything is in line with your elbows

Concert Cooking Craft

Teleshopping Yoga

Fig. 1 Example of five action image classes (Original Source: [42])

method exploits the learning ability of deep neural networks at the pixel and component levels to develop a deep hybrid classification framework. The following are the contributionsof this work:

(i) We demonstrate new classification strategy that adapts existing methods for a novel set of sub-tasks and combines them into an ensemble for solving this complex problem.
(ii) Use of features extracted at the stable-component pixel level and from facial regions for defining unique relationship between foreground and background information in the input image is new.
(iii) The manner in which we combine the individual sub-task specific features obtained from deep neural networks is new for the task of action image classification.
(iv) Lastly, our exploration towards the concept of multi-modality by combining text, face, and general pixel distribution-level information through a set of deep learning models is also new.

The rest of the paper is structured as follows. The critical analysis of existing methods for text detection, recognition and scene image classification is presented in Sect. 2. The proposed hybrid deep learning models for classification of action images are presented in Sect. 3. Section 4 provides variety of experiments to validate the proposed method which includes ablation study, experiments on classification of action images and experiments on text detection and recognition. Section 5 highlights the summary of the proposed work and describes our future work.

2 Related work

Since the objective of our work is to classify action images for enhancing performance of text detection and recognition, here we provide a critical analysis of existing methods on detection, recognition, and scene image classification.

In [9], the authors used a deep neural network for detecting text in natural scene images. It addresses the challenges of arbitrarily oriented texts and complex background images. However, the method is limited to natural scene images but not for video which contain different actions. In [10], a method was presented for detecting scene text using deep reinforcement learning wherein an agent, given a state, learns to estimate future returns. Further, the method makes sequential decisions to find scene texts. The method in [12] used a technique for reading scene texts in the wild based on scene text proposal. The method explores a score function that uses histogram of oriented gradients, and based on their probability of being texts, the method ranks the proposals accordingly. The method in [11] leveraged color prior guided MSER for natural scene text detection. The method extracts stroke features using strokes width distance, which is based on segmented edges. From the above discussions, it is noted that the scope of the above methods is confined to text detection in natural scene images. Therefore, these methods may not be effective for video images because the latter usually contains multi-type texts, namely, graphics (which is superimposed text) and scene text (which is a natural text as in a scene image). To overcome this problem, the method [18] proposed Fourier-Laplacian filtering and Hidden Markov Model for text and non-text classification. It uses hand-crafted features for eliminating false positives to improve performance. In summary, the above text detection methods focus on the challenges such as low resolution, multi-direction, multi-script, and complex background of texts in natural scene images or video frames, but not action images where one can expect unpredictable background complexity (the nature of texts are shown in Fig. 1). For detecting text in natural scene images, the authors in [19] propose a two-staged approach using a quadrilateral scene text detector. However, the structural assumption behind the use of such a structure may not hold for text that is irregular or arbitrarily oriented. The method presented in [20] employs an adaptive Béziercurve-based network for spotting text in scene images. The technique attempts to improve text detection performance by fixing an accurate bounding box for text lines that are arbitrarily oriented. In [21], the method uses similarity estimation between the text components of multiple views of natural scene images for obtaining enhanced performance. However, the requirement of multiple views as input is a constraint. Zhu and Du [22] proposed a scene text detection method based on segmentation approach. The method introduces Text-Mountain architecture, which finds the relationship between center and border of a text to overcome the challenges of text detection.

Similarly, we can find several text recognition methods for natural scene and video images. The method discussed in [23] used an edge descriptor by exploring local binary patterns. However, it is noted that the method is not robust to complex background images. The method [24] explores CNN for multi-lingual text recognition in natural scene images. The scope of the method is limited to natural scene images. As a result, the above methods may not perform well for texts in video images due to the presence of multi-type texts, namely, caption and scene texts, in video, unlike natural scene images which contain only scene texts. To expand the ability of text recognition methods, the method [25] proposes fractals, wavelet transforms and optical flow for tackling the challenge of video and natural scene images. Although, the above methods address the issues of natural scene and video images, it is not sure how the methods behave on action images. The method [26] extracts the dependencies between word tokens in a sentence. This helps to extract 2D spatial dependencies between two characters in a scene text image. The method [27] uses character anchor pooling for scene text recognition. With this step, the method gathers more vital information for recognizing text in the images. The method [28] introduces character awareness network for scene text recognition. The model involves 2D character attention model, which enhances foreground text instances based on character awareness. Lin et al. [29] proposed a sequential transformation attention-based network for text recognition in scene images. The network rectifies irregular text by dividing the task into a series of patch-wise basic transformations. Further, the model uses neighbor information to preserve the shape of characters to achieve recognition results. However, the performance of the method degrades for curved text and arbitrarily shaped text, which are common in the case of action images.

There are methods that use text recognition for different applications. For example, [30] proposed a method for defect extraction in sewers of CCTV inspection videos utilizing text recognition. The method detects and recognizes texts in video images captured by CCTV camera for the purpose of location identification and severity of cracks. The paper [31] proposed a flash flood categorization system using scene text recognition. The approach detects and recognizes texts in bridge images, which works well for different situations such as fog, sunny afternoon and during dusk time. The method [32] uses text detection and recognition for person re-identification. The approach uses deep learning using CNN and LSTM. The discussions on the above-mentioned methods show that for tackling different situations, text detection and recognition has been explored, which includes natural scene, video images, etc. However, one can note that the purposes of the methods are specific and work only with specific situations. None of the methods focuses on the situation like action images.

If we consider the posed problem as general scene categorization, we can find several methods like the ones above for classification of different scenes and images. Bosch [33] used a different color space and leverages SIFT features and analysis of probabilistic latent semantics, which results in a hybrid generative/discriminative approach. Dunlop [34] proposed a method based on the semantic segmentation of images and videos of indoor and outdoor scenes for the purpose of classification. A public API made available by Google, known as the Google Vision API [35], can be used for annotating scene images. The underlying system uses deep-learning-based features for labeling query images. The method [36] uses combination of CNN and LSTM to classify scene images of different categories based on multiple views and levels. The method considers each image with multiple views to deal with the variations of the images. The above methods work well for general scene images, where we can see multiple objects with clear shapes in images. But this restriction may not be true in case of action images considered in this work. In addition, the objective of the above method is to classify scene images but not enhancing performance of text detection and recognition methods. The method [37] explores color spaces, gradient distribution and Gabor wavelet binary pattern for classifying water images of different types. Xie et al. [38] proposed a knowledge distillation strategy based on EfficientNet for ImageNet classification. The model achieves high accuracies by introducing noise to the training procedure and making the student network equal to or even more powerful (relative to its size) than the teacher. Dosovitskiy et al. [39] proposed a model that overcomes some of the drawbacks of conventional convolutional neural networks by entirely replacing convolutions with attention blocks, i.e., by employing a pure transformer directly to the sequence of patches of an input image for classification. The methods are applicable to general images but ignore text information.

Several methods proposed for improving text detection and recognition performance in literature. For example, the method [40] uses a combination of local and global features for video image categorization. However, the success of methods depends on the success of text detection. In the same way, the method [41] employs the combination of face, torso detection and text detection methods for enhancing the performance of bib number detection and recognition. The method is developed specifically for Marathon images, which depicts a very specific scene setting. The methods [8] explores the combination of rough set and fuzzy for classifying scene images based on text and background information. The method extracts feature for each classified edge component on the classification of images. However, the methods are not tested on action images without text information. Recently, the method [42] proposes the combination of Discrete Cosine Transform and Fast Fourier Transform for classifying caption and scene texts in action images to improve text recognition results. The method generates a fused image for the input and then the average of sparsity and non-sparsity counts in terms pixel values of zero or non-zeros is computed for classification. However, the method is limited to text line image classification and not for action images.

It is noted from the above review on text detection, recognition, scene image classification and the classification methods for enhancing text detection and recognition performance, none of the methods aims at classifying action images for amplifying text detection and recognition performance. In addition, there are methods that focus on image types or situations for classification but without targeting text detection and recognition performance. As a result, one can conclude that there is a need for developing a method that can work for action images. Thus, we propose a new hybrid deep learning model for classifying action-oriented images to improve performance of text detection and recognition.

3 The proposed methodology

Since the classification of action images is a complex problem, our idea is to extract information at different levels, like pixels, components, and facial information if available, to ease the problem. It is observed that the relationship between foreground and background information can be represented in a unique way. For example, in the case of Yoga, Concert, Teleshopping images, it can be expected outdoor scenes represent background while actions represent foreground. Similarly, for Cooking and Craft class images, indoor scene represent background and actions represents foreground. To extract such observations, the proposed work detects region of interests, which includes text information as well as dominant region of background and facial information because these features represent foreground while the pixel distribution and values represent both foreground and background of the images. This is the main institution to combine region of interest (text components) given by Maximally Stable Extremal Regions (MSER) [43], Face information given by Multitask Cascaded Convolutional Neural Network (MTCCN) [44] and pixel information by VGG16 model. Overall, the way the proposed method combines features and deep learning models is able to tackle the challenges of action images classification.

The use of the Residual Network as a general classifier was motivated by the method [17], where it is mentioned that the deep residual learning technique converges much faster than standard solvers that are unaware of the residual nature of the solutions. In addition, ResNet50, being a neural network of a very deep architecture, has a high entropic capacity compared to other relatively shallower networks and, hence, can better cope with large intra-class variations and random noisy pixels in the images. These factors motivated us to explore ResNet50 for learning parameters at the pixel level.

This step outputs a vector of five elements indicating the probabilities of the membership in the input image of respective classes, which is denoted as P1. According to our experiments, the P1 architecture achieves more than 99% classification rate for Yoga action class. As a result, the proposed method considers other four classes, namely, concert, cooking, craft, teleshopping for training at components given by MSER. The distribution of stable extremal regions tends to be relatively similar in the images of a particular class but differ largely between images of different classes. Due to its computational simplicity and robustness, the proposed work considers the detection of Maximally Stable Extremal Regions (MSERs). The outputs of the MSER detection are fed to an MSER-based classifier for training the VGG16 net [16]. It is noted that MSER usually extracts components that have the uniform color [43]. In general, character and object in images are formed by the uniform color. Besides, those two play a prominent role in representing the content of images. To extract such observations, we train VGG16 at component level. This step outputs a vector of 4 elements, indicating the probabilities of the membership of the input image of respective classes, which is denoted as P2. In the same way, to exploit face information in action images, we propose to use Multitask Cascaded Convolutional Network [44], which employs deep learning for face detection. The output of face detection is fed to VGG16 [13] to train at face component level. Since the Craft and Cooking classes do not provide faces and for Yoga class, P1 gives more than 99% classification rate, we train VGG16 with face components of other two classes, namely, Concert and Teleshopping, which gives a vector of 2 elements denoted as P3.

The three obtained vectors, P1, P2 and P3, are supplied to a Fully Connected Neural Network (FCNN) that aims at learning an effective combination of the probability vectors for classifying the original images. The FCNN has one hidden layer of 16 units with ReLU activations, and one output layer of 5 units with Sigmoid activations that denote the final class confidences. The number of output units in the final layer of the fully connected network can be varied based on the number of classes in dataset. Furthermore, the proposed method finds correct architecture which fits for the proposed classification according to predefined experimental results. Increasing the number of hidden layers or the number of units in the hidden layer would always lead to overfitting, and even regularization or dropout could not be compensated for. The optimal performance achieved from such a simple architecture is because the probabilities given by deep neural networks are mostly in agreement with each other regarding the class of input image and hence, it is relatively easy for the final network to learn a combination of those probabilities. The major role of the model combination step is to determine the weights that are to be assigned to the outputs of the classifiers for the input probability distributions given by the three classifiers with the actual ground truth labels of the image. The framework of our approach is shown in Fig. 2.

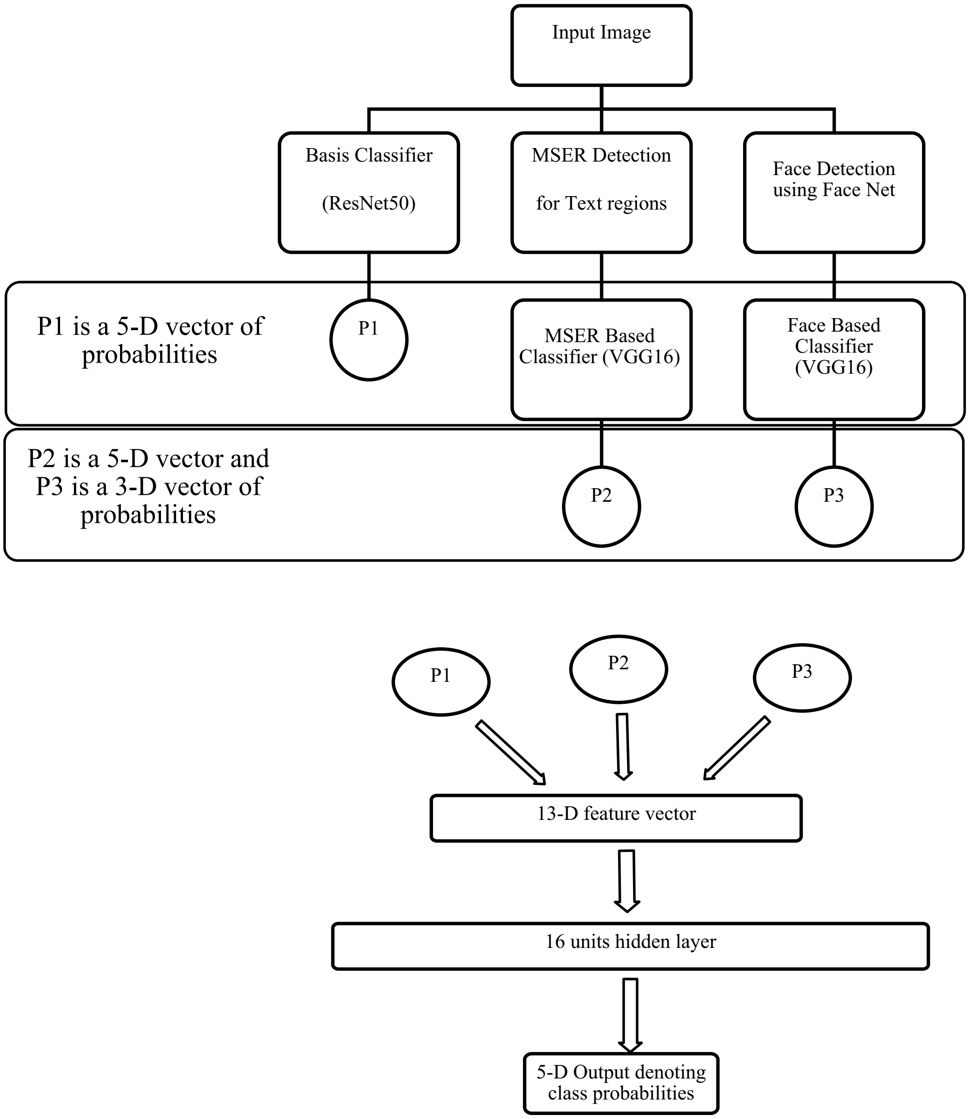

Input Image
Basis Classifier
(ResNet50)
MSER Detection
for Text regions
Face Detection using Face Net
P1 is a 5-D vector of probabilities
P1
MSER Based Classifier (VGG16)
Face Based Classifier (VGG16)
P2 is a 5-D vector and P3 is a 3-D vector of probabilities
P2
P3
P1
P2
P3
13-D feature vector
16 units hidden layer
5-D Output denoting class probabilities

Fig. 2 The framework of our model. (P1 indicates features vectors extracted from the pixel in formation, P2 indicates features extracted from MSER regions and P3 indicates, the features extracted from only Face region

The proposed method explores ResNet50 and VGG16 models for learning parameters both at pixels and component levels to extract distinct features for action images of different classes, which results in feature vectors containing probabilities. The proposed method consists of four sub-sections. First, Subsect. 3.1 presents ResNet50 for studying the overall pixel distribution to predict classes. VGG16 is then explored for learning at component level given by MSER for getting probability score for the classes in Subsect. 3.2. Since many classes considered in this work contain face information, VGG16 is also explored for learning facial information at face level given by MTCNN and it is discussed in Subsect. 3.3. Finally, the unification of the three deep neural networks is presented in Subsect. 3.4 for action image classification.

3.1 ResNet50 for overall pixel-distribution based class prediction

We take the following approaches for training deep net classifiers at different levels. For every class, 90% of samples used for training and 10% of samples are used for testing. Inspired by the method [45] where transfer learning has been introduced for deep neural networks, we explore the same for the purpose of classifying action images. In this work, we use neural networks that are pre-trained on ImageNet [46] and fine-tuned [47] with the experimental datasets corresponding to this work. The process involves the removal of the top fully connected layers and the introduction of a new randomly initialized fully connected layer of size 1024 units with ReLU activations. The weights of all the previous residual layers are fixed. The output layer is a new fully connected layer with 5 outputs having sigmoid activations. Having trained these two newly added fully connected layers, the weights of the top-most (last) residual block were made trainable and subsequently fine-tuned (trained with a very low learning rate) along with the two final fully-connected layers. The reason behind this approach is that the convolutional blocks towards the beginning learn to detect very general structures like edges, curves, shapes, etc., while the layers towards the end detect more class-specific features. So, keeping the earlier layers fixed, the layers towards the end were trained in order to make the network learn the parameters corresponding to those features which are relevant to the classes in our classification problem. The training was done in two phases in order to prevent large and noisy gradient updates from severely disturbing the previously learned weights of the model. More details about the architecture of ResNet50 can be found in [17].

The loss function used in all the classifiers presented in this work is categorical cross-entropy as defined in Eqs. (1) and (2).

The categorical cross-entropy loss for a single training example is given by:

$$L_i = -\sum_{c=1}^{M} y_c log\left(p_c\right) \tag{1}$$

and for the overall training set is given by:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M} y_c^{(i)} log\left(p_c^{(i)}\right) \tag{2}$$

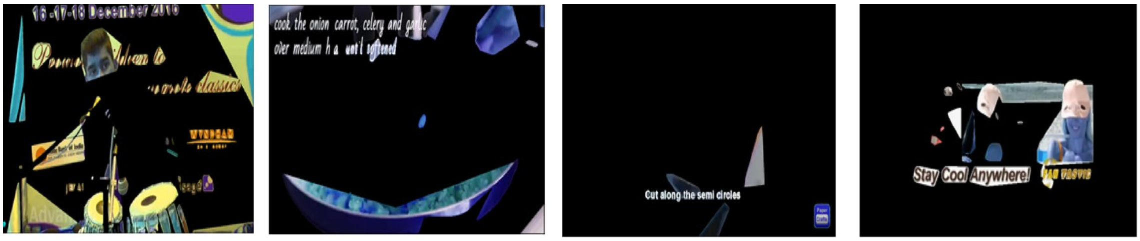
where N is the number of samples in the training set, M is the number of classes, $y_c^{(i)}$ is the binary indicator (0 or 1) if class label *c* is the correct classification for observation *i*, $p_c^{(i)}$ is the predicted probability observation *i* of class *c*.

3.2 MSER-based classifier for learning dominant stable-region based features

It is noted that Maximally Stable Extremal Region (MSER) works well for grouping the pixels that share similar values or properties in the images, which results in connected components [43]. In other words, MSER detection aims at extracting regions that stay nearly the same through a wide range of thresholds. The sample results of MSER for input images are shown in Fig. 3, where one can see, the components are formed based on similarity of intensity values. The training set for the MSER based classifier is prepared by running the MSER detection algorithm on all the images of four class datasets. After MSERs are detected, all the pixels that do not belong to the MSER are masked out and thus, the training set consists of images with only their MSERs. A transfer-learning based procedure, similar to the one that was adopted for training the ResNet50, is used for fine-tuning and training VGG16 (pre-trained on ImageNet) at component level by feeding the output of the MSER extraction step to VGG16 and optimizing the categorical cross-entropy loss. This results in a vector containing five probabilities representing the confidence of the MSER-based classifier, regarding the membership of the image to each of the four classes.

3.3 Face-based classifier for learning facial image features

It has been observed that facial features (orientation, expression, degree of occlusion, etc.) of action images have much less variations within a class compared to the variations between images from different classes. Face detection is carried out over the images of Concert and Teleshopping classes. This is because in general, Cooking and Craft may not provide face information and for Yoga class, P1 achieves the best results. Therefore, the following steps are used for detecting faces in images of the above-mentioned two classes. Face detection uses Multitask Cascaded Convolutional Network (MTCNN) as in [44]. The candidate of facial windows and their bounding



Concert Cooking Craft Teleshopping

Fig. 3 The result of MSER extraction

box regression vectors using fully convolutional Proposal Network (P-Net) are obtained. The candidates are passed to next level of CNN called Refine Network (R-Net), which conducts bounding box regression and Non-Maximum Suppression (NMS) to filter out false candidates. The sample face detection results for the images of Concert and Teleshopping are shown in Fig. 4. The reason to choose the above MTCNN for face detection is that the method is robust for the proposed work according to our preliminary experiments. The faces extracted from images of respective classes are used as a training set for the face classifier. For this, we explore VGG16 as discussed above for an MSER based classifier. This results in a vector containing three probabilities representing the confidence of the face-based classifier, regarding the membership of the image to the two respective classes.

In this work, both RestNet50 and the face-based classifier were trained and fine-tuned for 50 epochs with learning rates of 0.001 during training and 0.0001 during fine-tuning. The Adam optimizer was used to update the weights during training and the Stochastic Gradient Descent optimizer with a momentum of 0.9 was used for fine-tuning the networks.

3.4 Unification of the 3 deep convolutional neural networks

The proposed method explores ResNet50 for training at overall pixel-distribution level and VGG16 at component levels, which results in three deep neural networks. The outputs obtained from the three networks are concatenated to obtain a 11-dimensional vector of probabilities (5 from pixels, 4 from MSER and 2 from face components). Next, the vector with 11-dimensions is supplied to a FCNN that learns a mapping between the output of the three classifiers and the actual class of the image. The final fully-connected neural network comprises 1 hidden layer consisting of 16 hidden units with Leaky ReLU activation, and an output layer of 5 units with Sigmoid activation, indicating the probabilities of the membership of the input image to the 5 respective classes. The proposed model considers the categorical cross-entropy loss as the objective function for training the network. The final architecture of the proposed work is illustrated in Fig. 5.

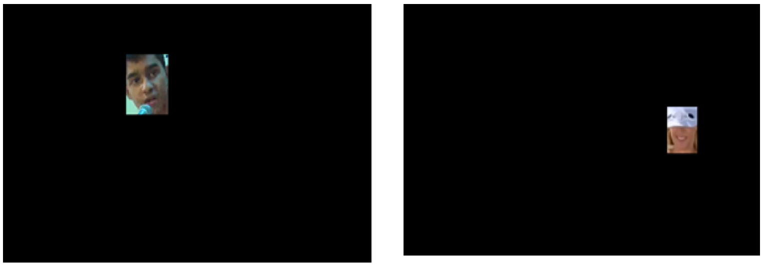
4 Experiments

Our experimental analysis is presented in four sub-sections. Description of the datasets, different performance measures and different existing methods used for comparison purpose are given in Sect. 4.1. To show the effectiveness of the key steps we are using, Sect. 4.2 presents an ablation study. Section 4.3 provides experimental results for the proposed method and existing techniques on classification. To validate the proposed classification is useful, Sect. 4.4 provides experimental analysis of the text detection and recognition.

4.1 Dataset, performance measure and work for comparison

Dataset To evaluate the proposed method for classifying action images, we create our own dataset which includes five classes, namely, Concert (Ct), Cooking (Ck), Craft (Cr), Teleshopping (Te) and Yoga (Yg). We use different internet sources, such as YouTube and other social media for dataset collection. Our dataset includes 5078 images of five classes. For training and testing of each class, the proposed method considers 90% and 10%, respectively. The same criterion is followed for all the classification experiments in this work. To test the scalability and effectiveness of the proposed method on classification, we consider the dataset used for video image type categorization in [8], where the method considers 10 classes, namely, Defense (D), Economics (Ec), Sports (S), Medical (M), Weather (W), Animation (A), e-learning (e-L), Technology (T), Outlet (O) and Animal Planet (AP). Since each class comprises 3,000 frames, for the 10 classes, it provides 30,000 frames for validating the proposed method. Note that this dataset includes natural scene images having multi type texts of video, while our dataset includes scene images with different actions and multi-type texts. In addition, both the datasets pose challenges like poor resolution, contrast,

Fig. 4 The results of face detection method for Concert and Teleshopping images shown in Fig. 1



Concert

Teleshopping

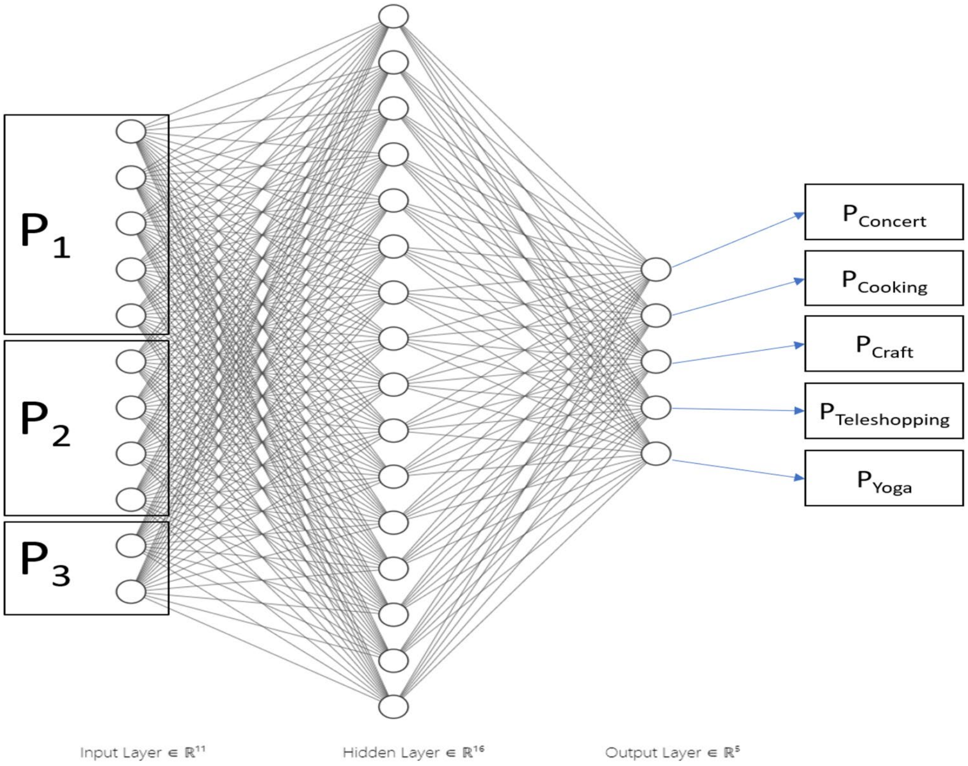

P_1
P_2
P_3
$P_{Concert}$
$P_{Cooking}$
P_{Craft}
$P_{Teleshopping}$
P_{Yoga}
Input Layer $\in \mathbb{R}^{11}$
Hidden Layer $\in \mathbb{R}^{16}$
Output Layer $\in \mathbb{R}^{5}$

Fig. 5 The final combination phase of the proposed Hybrid Deep Net for action image classification. (P1, P2 and P3 are discussed in Fig. 2)

font and font size, background variations, arbitrarily oriented texts, etc., due to different nature and characteristics. In summary, our and the standard datasets provide, in total, 35,078 frames for conducting experiments.

To test the scalability and generality, we also consider the benchmark dataset called Stanford40 Actions [47], which contains 40 classes of general actions, and each class contains 1,80,300 images. Therefore, it gives total 9532 images for all the 40 classes. The focus of this dataset is to capture general actions of persons. Another standard dataset, called Scene Text Dataset (STD), which provides 10 classes of scene images containing text information [8]. However, our dataset is to capture person actions having multiple type text information. In addition, our dataset may not include images of person faces compared to Stanford40 Actions dataset. The experiment involves total 44,610 images from three datasets. We believe that the dataset collection with different focuses and nature ensure the fair evaluations of the proposed approach on classifying action images.

Performance Measure To evaluate the performance of the proposed approach, we consider Average Classification Rate (ACR) of confusion matrix as measures for evaluating the methods. We estimate standard measures, namely, Recall, Precision and F-Score for text detection experiment, and Recognition Rate (RR) for text recognition to validate the usefulness of the proposed method. The recognition rate is estimated using the edit distances based on insertion, substitution, deletion operations for all the experiments. In case of text detection and recognition experiments, we estimate the measures for before and after classification to show the advantage of the proposed classification. It is expected that the text detection and recognition methods should report better results after classification compared to a prior classification. It is not necessarily true for all the classes due to the limitations of the text detection and recognition methods. Note that for text detection and recognition experiments, we consider only our dataset because Stanford40 Actions dataset does not provide text information.

Method for Comparison The following state-of-the-art methods are used for comparative study in this work. Roy et al. [8], proposes a method for classification of scene images containing text information to improve text detection and recognition performance. Qin et al.'s method [40] which uses statistical, structural, spatial features and color spaces with SVM for classification of video text frames. The method [38] uses self-supervised approach for scene image classification (Noisy Student) and transformer-based (Vision Transformer) method [39] for scene image classification. Both the methods exploit the deep-learning literature for achieving the results for classification of scene images.

Apart from the above, we also evaluate the performance of the Google Vision API on our task, which is a standard deep learning based public API for performing various tasks on scene images [35]. Given an input image, the API returns confidence scores for a set of pre-defined classes. As an example, if an input image from the 'Animal Planet' class is given to the API, it can return high confidence scores for classes like trees, nature, animals, birds, agriculture, etc. For an input image belonging to the class 'Animation', labels like cartoon, amusement park, amusement ride, etc., could be returned. The returned labels are native to the Google Vision API system. Using the above confidence scores, a feature vector for each class is created for the training samples. A cut-off threshold for confidence score is set at 85%. This value is determined empirically based on the results of the training set. As observed from our experiments, increasing this threshold incorporates irrelevant labels, and decreasing it discards useful labels, which leads to an eventual loss of information in the representation.The above-mentioned methods represent state-of-the-art for comparison purpose.

In addition, to demonstrate the impact of our approach on downstream processing stages, we present experimental results for detecting and recognizing text in an input image before and after our classification step. For that purpose, we implement the well-known method called EAST, which is developed for accurate scene text detection [9]. The method uses deep convolutional neural networks for tackling challenges of arbitrarily oriented text detection in natural scene images. Zhu et al. [22] describes a method which employs a TextMountain network for scene text detection. For recognition, we implement the E2E-MLT [24], which proposes a model for multi-language scene text recognition. The approach uses deep learning models for recognizing texts in scene images irrespective of scripts. A method by Lin et al. [29] uses an attention-based network for recognizing text in natural scene images. The motivation to choose the above four methods for text detection and recognition experiment is that the approaches employ deep learning, which are capable of handling complex situations of text detection and recognition.

4.2 Ablation study

There are three steps of the proposed method, namely, training the deep learning model, ResNet50 at pixels level for classifying action mages, and VGG16 model for MSER and Face components. We carried out experiments for calculating the classification rate (mean of the confusion matrix diagonal) separately for each of the components of our framework, in order to understand their individual contributions. The Average Classification Rates (ACR) for ResNet50, VGG16 of MSER and VGG16 for face components are 87.6%, 87.7% and 76.5%, respectively. It is noted that ResNet50 and VGG16 of MSER components score almost the same results, while VGG16 of face component repots a bit low score compared to ResNet50 and VGG16 of MSER. This is due to the absence of faces in Craft and Cooking classes. For these two classes, the face detection method outputs either nothing or false results. When it gives nothing, the proposed method adds zeros to feature vectors for the purpose of implementation. On the other hand, ResNet50 and VGG16 for MSER give good results compared to VGG16 of face components because those models get sufficient information from each image class. However, when we combine all the three deep nets, the proposed method shows 90.0% ACR. Therefore, we conclude that all the three mentioned deep nets are effective to achieve better results.

As discussed in the above sections, the proposed method combines three classifiers to achieve better results for action image classification. In order to assess the effectiveness of each classifier, we conduct experiments on each classifier and the combined classifier for our dataset and the results are recorded in Table 1. It is noted from Table 1 that each classifier is effective and contributes equally for classifying action images of respective classes. Hence, the combined components provide a higher classification rate than any of the individual components. In Table 1, '-' indicates that the classifier does not consider the class for training and testing.

Table 1 Confusion matrices of the proposed individual models and combined model on our dataset

Models	Basic Classifier (ResNet50)					MSER based Classifier (VGG16)					Face based Classifier (VGG16)					The Proposed Combined				
Class	Ct	Ck	Cr	Te	Yg	Ct	Ck	Cr	Te	Yg	Ct	Ck	Cr	Te	Yg	Ct	Ck	Cr	Te	Yg
Ct	**93.0**	3.0	0.0	1.0	3.0	**91.0**	0.0	2.0	7.0	–	**83.0**	–	–	17.0	–	**93.0**	1.0	0.0	5.0	1.0
Ck	2.0	**89.0**	4.0	3.0	2.0	0.0	**92.0**	5.0	2.0	–	–	–	–	–	–	3.0	**90.0**	5.0	1.0	1.0
Cr	7.0	7.0	**80.0**	7.0	7.0	1.0	3.0	**94.0**	2.0	–	–	–	–	–	–	1.0	3.0	**90.0**	2.0	4.0
Te	6.0	8.0	6.0	**77.0**	3.0	4.0	11.0	11.0	**74.0**	–	30.0	–	–	**70.0**	–	5.0	5.0	9.0	**78.0**	3.0
Yg	0.0	1.0	0.0	0.0	**99.0**	–	–	–	–	–	–	–	–	–	–	0	1.0	0.0	0.0	**99.0**
ACR	87.6					87.75					76.5					90.0				

Bold indicates the best results

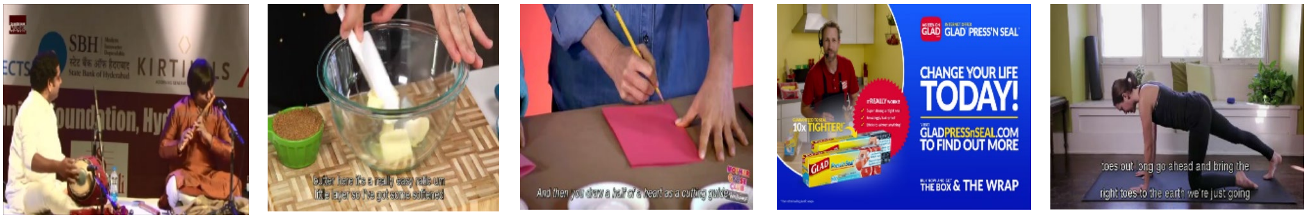
4.3 Evaluating the proposed classification approach

Qualitative results of the proposed method on our dataset are shown in Fig. 6, where we can see the images with clutter background and multi–type texts are classified successfully. Quantitative results for the proposed and existing approaches on our dataset are compared in Table 2 by the means of confusion matrix and classifications rate. It can be clearly seen from Table 2 that the proposed technique outperforms all of the other methods we studied in terms of classification rate. The main reason for the poor results of the existing methods, namely GOOGLE API [35], Qin et al. [40], Xie et al. [38] and Dosovitskiy et al. [39] is that they only work well when multiple objects in images preserve their shapes. However, Roy et al.'s method [8] defines limited shapes for edge components in images for classification. Limited shapes are the main cause for poor results in case of the action dataset because of unpredictable shapes. It is noted from Table 2 that the method [39] gives the second highest accuracy compared to the other existing approaches. This is because the model replaces the inductive bias introduced by the convolution operation with the general idea of attention, which can be used to directly model all possible relationships among the set of image patches. Since the model considers high-level semantic information for classification, it has better generalization ability compared to the other methods. However, it still falls short of the technique we have presented here. The proposed method combines the strength of pixels, dominant components given by MSER, and face components, to achieve superior classification performance.

The proposed and existing approaches are tested on Roy et al.'s dataset (STD) [8] which provides 10 classes of different video types, and the Standford40 dataset [47] which provides 40 general action image classes through confusion matrix and classification rate. These two datasets are considered as the standard ones. Qualitative results of the proposed technique on STD are shown in Fig. 7, where all the images are classified successfully. It is noted from Fig. 7 that sample images of all the classes contain texts of different types. Quantitative results of the proposed and different existing approaches on Roy et al.'s dataset are reported in Tables 3, 4, 5, 6, 7, 8. It is observed from Tables 3, 4, 5, 6, 7, 8 that the proposed method is the best at classification rate compared to the existing methods. When we compare the results of this dataset with five classes of the action dataset, the proposed method scores low. This makes sense because the proposed method is developed for action images of five classes but not scene images of 10 classes. Interestingly, the results of our dataset and the STD dataset show that one can see the high variations in the classification rate for our dataset, while for STD dataset, the classification rates of the other methods are more similar. It can be inferred that variations in action images are less predictable compared to scene images. However, in terms of overall classification rate, the proposed method outperforms all of the other methods.

To test the proposed approach does not depend on text in the images for classification of actions images, we conduct experiments on the benchmark dataset called Stanford40 Action dataset [47], which does not contain text information. Qualitative results of the proposed approach for Stanford 40 Actions dataset are shown in Fig. 8, where one can see that the proposed method classifies different action images successfully. Quantitative results of the proposed and existing methods are reported in Table 9, where we list classification rates for each class. Even though the number of classes increases to 40 from 5 and images of different classes do not contain text, the proposed method achieves the best classification rate, that is, AoA (Average of Average), compared to the existing methods.

This is due to the way the proposed method trains the three different deep nets at different levels. Therefore, we can infer that the proposed approach is generic and capable of classifying different classes irrespective of the number of classes and content. However, the existing methods perform poorly because of the inherent limitation of the methods. It is noted from the experiments on the 5-class, 10-class and 40-class datasets that the proposed method scores the highest for the 5-class dataset, slightly lower for the 10-class dataset, and still lower for the 40-class dataset. This is valid because as the number



Concert Cooking Craft Teleshopping Yoga

Fig. 6 Sample images of successful classification of the proposed modelon our dataset. Original Source: [42]

of classes increases, the complexity of the problem also increases. But if we consider the overall performance in terms of classification rate, the proposed method outperforms the others.

4.4 Validating the proposed classification using text detection and recognition experiments

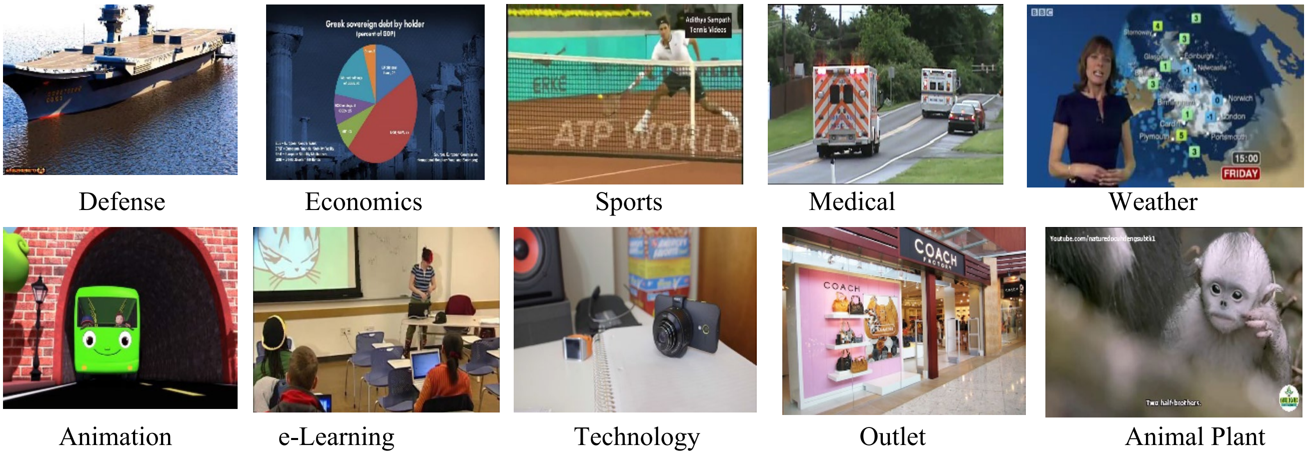
The experiments in the previous sub-sections show that the proposed classification method is effective, scalable and generic. In order to show the advantage of the proposed classification, we conduct experiments for text detection and recognition on dataset containing 5 classes (our dataset) and another dataset containing 10 (STD dataset). However, there is no text detection and recognition experiment on the Stanford40 Actions dataset because it does not provide text information. Experiments on before and after classification are done to see the advantage of our proposed classification method. In case of before classification experiments, we gave all the images without separating classes as the input for text detection and recognition. The number of training and testing samples is determined based on the total number of images of all the classes. In case of after classification, the text detection and recognition method considers each class as the input for calculating measures, because with the proposed classification technique, one can do the same. Therefore, we can treat each class as one problem rather than all classes to enhance the performance of text detection and recognition after classification. For the text detection experiments we use EAST method [9], Zhu et al. [22], and for the recognition experiments we use E2E-MLT method [24], Lin et al. [29].

Qualitative results of the text detection technique [9] before (images of all the classes are input) and after classification (images of individual classes are input) for our dataset and STD are shown in Figs. 9 and 10, respectively, where it can be seen that the method provides better results after classification compared to before classification for almost all the classes. It is expected because the classifier used in the text detection method [9] is trained and parameters are tuned according to complexity of the classes. However, sometimes, the text detection method may not achieve better results after classification compared to before classification. This is due to the limitation of the text detection method and the lack of relevant samples to be trained. The same conclusion can be drawn from the qualitative results of the recognition method [24] on our dataset and STD dataset as shown in Figs. 11 and 12, respectively, where it can be noted that recognition results are better for after classification compared to before classification for almost all the classes. In this way, the proposed classification helps to improve text detection and

Table 2 Confusion matrices of the proposed and existing approaches on the proposed dataset

Met	Proposed					Roy et al. (2018) [8]					Google API [35]					Noisy Student (2020) [38]					ViT (2021) [39]					Qin (2016) [40]				
Class	Ct	Ck	Cr	Te	Yg	Ct	Ck	Cr	Te	Yg	Ct	Ck	Cr	Te	Yg	Ct	Ck	Cr	Te	Yg	Ct	Ck	Cr	Te	Yg	Ct	Ck	Cr	Te	Yg
Ct	**93.0**	1.0	0.0	5.0	1.0	**87.8**	10.2	0.0	0.8	1.2	**84.0**	6.0	2.2	1.1	6.7	**74.5**	5.3	7.5	9.0	3.7	**94.1**	0	1.9	1.2	2.8	**58**	10.7	10.9	12.3	8.1
Ck	3.0	**90.0**	5.0	1.0	1.0	12.4	**85**	0.0	0.4	2.2	2.8	**68.8**	3.9	12.0	12.5	5.2	**82.5**	6.3	4.8	1.2	4.4	**85.2**	6.8	1.6	3.0	10.2	**52.0**	13.4	6.6	17.8
Cr	1.0	3.0	**90.0**	2.0	4.0	37.6	28.2	**25**	3.6	5.6	12.0	0.0	**80**	2.3	5.7	3.8	2.3	**90.2**	1.7	3.0	2.5	11.5	**79**	4.5	2.5	7.9	11.0	**48.7**	22.9	9.5
Te	5.0	5.0	9.0	**78.0**	3.0	35.2	19.6	0.6	**41.8**	2.8	11.6	8.2	3.6	**74**	2.6	7.3	6.0	3.5	**75.7**	7.5	9.0	3.2	2.8	**83.0**	2.0	10.4	18.8	3.7	**60.9**	6.2
Yg	0.0	1.0	0.0	0.0	**99.0**	11.4	0.2	0.0	0.0	**88.4**	1.8	3.6	5.0	2.7	**86.9**	2.0	4.9	2.7	5.3	**85.1**	2.3	0.0	0.0	1.0	**96.7**	1.2	13.1	15.0	16.7	**54.0**
ACR	**90.0**					**65.6**					**78.7**					**81.6**					**87.7**					**54.7**				

Bold indicates the best results


Defense Economics Sports Medical Weather

Animation e-Learning Technology Outlet Animal Plant

Fig. 7 Examples of successful classification of the proposed approach on STD dataset. Original Source: [8]

Table 3 Confusion matrix of the proposed approach on STD dataset (Average Classification Rate: 77.89%)

Classes	Defense	Ecnmcs	Sports	Medical	Wthr	Animation	E-Lrg	Tech	Outlet	Anl.Plt
Defense	**89**	1.18	0	0	0	0	0	9.58	0	0
Ecnmcs	0	**100**	0	0	0	0	0	0	0	0
Sports	0	0.5	**99**	0.17	0	0.25	0	0	0.08	0
Medical	11.98	0.34	0.08	**80**	0.17	1.01	0	0.08	0.17	5.7
Wthr	0.33	1.5	1.5	0.5	**73**	1.92	0.17	0.17	12.25	8.67
Animation	0.8	0.98	0.23	0	0	**97**	0.06	0.69	0.17	0
E-Lrg	3.37	6.25	0.64	1.76	5.05	39.58	**19**	6.73	14.9	2.56
Tech	0.72	1.18	0	2.26	0	10.23	4.71	**78**	2.44	0.45
Outlet	3.35	6.2	2.35	1.73	2.23	17.6	16.11	3.84	**46**	0.62
Anl.Plt	0	0	0	0	0	2.5	0	0	0	**98.5**

Bold indicates the best results

Table 4 Confusion matrix of the Roy et al. method [8] on STD dataset (ACR: 76.0%)

Classes	Defense	Ecnmcs	Sports	Medical	Wthr	Animation	E-Lrg	Tech	Outlet	Anl.Plt
Defense	**83.9**	1.03	2.7	1.7	2.23	2.3	1.07	1.6	2.4	1.04
Ecnmcs	1.3	**75.5**	2.04	1.04	1.7	2.67	4.22	1.53	5.8	4.2
Sports	2.21	6.47	**72.7**	2.8	4.7	1.05	3.9	2.4	1.05	2.69
Medical	5.8	1.36	2.8	**71.5**	4.06	6.53	1.23	2.12	3.07	1.49
Wthr	1.8	2.3	1.32	1.17	**80.6**	3.5	1.47	2.53	3.71	1.56
Animation	1.4	3.26	2.51	2.4	1.1	**77.7**	2.78	4.9	2.41	1.54
E-Lrg	1.4	3.26	2.51	2.4	1.1	1.78	**78.7**	4.9	2.41	1.54
Tech	1.83	3.28	6.9	1.07	1.82	1.32	2	**75.8**	1.78	4.2
Outlet	12.9	1.82	2.96	1.06	1.93	2.45	1.48	3.73	**69.5**	2.17
Anl.Plt	1.37	2.78	1.05	2.67	6.58	4.72	2.47	6.48	2.38	**69.5**

Bold indicates the best results

recognition performance for classes of different complexities. In this study, we use existing text detection method [9] and recognition method [29] to show that the result after classification improves compared to the results of before classification. Based on our experiments, the same inferences can be drawn if we use other existing methods for comparison.

Quantitative results of the text detection methods [9, 22] before and after classification on our dataset and STD dataset are reported in Table 10, where it can be

Table 5 Confusion matrix of the GOOGLE API method [35] on STD dataset (ACR: 71.7%)

Classes	Defense	Ecnmcs	Sports	Medical	Wthr	Animation	E-Lrg	Tech	Outlet	Anl.Plt
Defense	**72.8**	3.1	2.4	3.1	2.9	3.2	2.9	2.4	4.7	2.4
Ecnmcs	3.0	**73.1**	2.4	2.5	2,5	2.7	6.9	2.1	2.1	2.6
Sports	2.5	2.4	**78.6**	2.3	2.4	2.2	2.5	2.1	2.9	2.1
Medical	2.7	2.0	2.7	**77.4**	2.9	2.6	2.6	2.6	2.1	2.5
Wthr	2.5	2.5	2.2	4.2	**72.8**	3.6	3.2	2.4	3.8	2.7
Animation	2.5	2.8	2.1	2.5	3.3	**68.7**	3.9	3.5	5.7	5.0
E-Lrg	2.5	3.7	2.0	2.5	2.1	2.6	**75.6**	3.2	3.0	2.7
Tech	3.1	2.2	2.1	2.9	2.9	4.4	3.4	**72.4**	4.3	2.3
Outlet	5.5	3.9	4.5	3.7	6.6	5.8	8.2	9.5	**48.9**	3.4
Anl.Plt	2.6	2.7	2.5	2.3	2.1	3.2	2.3	2.7	2.8	**76.8**

Bold indicates the best results

Table 6 Confusion matrix of Noisy Student [38] on STD dataset (ACR: 72.6)

Classes	Defense	Ecnmcs	Sports	Medical	Wthr	Animation	E-Lrg	Tech	Outlet	Anl.Plt
Defense	**79.3**	1.6	3.1	2.7	1.9	1.2	1.3	4.4	1.0	2.5
Ecnmcs	0.0	**82.1**	1.1	1.9	4.8	2.2	3.9	2	1.5	1.5
Sports	2.0	0.0	**95.2**	0.0	0.0	0.0	0.0	0.0	2.8	0.0
Medical	1.2	2.9	3.8	**63.9**	3.1	1.0	2.0	5.5	5.5	11.0
Wthr	9.8	4.2	1.7	3.8	**68.3**	1.0	5.0	2.2	2.0	2.0
Animation	1.0	0.0	0.0	2.2	0.0	**93.8**	0.0	0.0	0.0	3.0
E-Lrg	1.0	3.4	2.0	1.1	13.9	14.0	**51.6**	11.5	2.0	8.5
Tech	13.5	10.0	1.0	1.0	8.0	1.4	6.6	**55.5**	1.0	2.0
Outlet	11.5	3.2	18.8	3.0	7.0	2.3	1.5	1.7	**49.2**	1.8
Anl.Plt	1.2	1.0	3.8	0.0	1.0	5.9	0.0	0.0	0.0	**87.1**

Bold indicates the best results

Table 7 Confusion matrix of Vision Transformer (ViT) [39] on STD dataset (ACR: 73.4)

Classes	Defense	Ecnmcs	Sports	Medical	Wthr	Animation	E-Lrg	Tech	Outlet	Anl.Plt
Defense	**82.7**	2.0	2.9	1.0	4.0	0.0	2.1	5.3	0.0	0.0
Ecnmcs	0.0	**96.5**	0.0	0.0	2.5	0.0	1.0	0.0	0.0	0.0
Sports	4.2	1.0	**80.2**	0.0	2.8	0.0	0.0	3.6	5.8	2.4
Medical	0.0	1.3	0.0	**88.4**	3.6	0.0	2.0	4.7	0.0	0.0
Wthr	1.7	13.5	0.0	0.0	**68.1**	1.1	9.2	6.3	0.0	0.0
Animation	0.0	0.0	0.0	0.0	2.0	**94.5**	3.5	0.0	0.0	0.0
E-Lrg	0.0	14.2	0.0	0.0	17.5	9.2	**51.8**	4.0	0.0	3.3
Tech	13.0	12.5	2.2	0.0	8.3	1.0	1.0	**56.0**	6.0	0.0
Outlet	5.1	3.1	15.1	12.0	1.2	2.0	0.0	4.3	**51.0**	6.2
Anl.Plt	1.8	1.0	6.2	1.0	2.0	9.3	9.2	2.1	3.1	**64.3**

Bold indicates the best results

noted that there is a significant improvement for Yoga class of our dataset after classification at Recall, Precision and F-score compared to before classification. For other classes, we cannot see much improvement compared to before classification. This is because the number of training samples used may not be represented by the variations of classes, especially for Concert class. With these experiments, one can understand that the text detection

Table 8 Confusion matrix of the Qin et al. technique [40] on STD dataset (ACR: 70.2)

Classes	Defense	Ecnmcs	Sports	Medical	Wthr	Animation	E-Lrg	Tech	Outlet	Anl.Plt
Defense	**53.1**	15.3	1.8	7.4	8.8	1.1	2.7	0.5	0.3	9.0
Ecnmcs	0.0	**97.9**	0.2	0.0	0.1	1.3	0.0	0.0	0.0	0.5
Sports	1.5	0.3	**85.6**	0.5	4.3	2.3	1.9	2.7	0.4	0.4
Medical	2.3	1.7	1.3	**80.1**	3.3	2.7	4.2	0.1	0.1	4.2
Wthr	0.9	0.8	0.8	2.1	**93.6**	0.0	0.0	0.1	1.1	0.6
Animation	0.6	13.8	2.8	0.5	2.2	**74.7**	2.2	0.6	0.5	1.9
E-Lrg	2.8	0.5	3.0	2.9	2.9	1.5	**80.5**	1.8	0.7	3.4
Tech	6.8	10.1	13.1	14.9	4.4	2.2	2.8	**37.5**	2.6	5.6
Outlet	5.6	6.7	13.7	11.8	7.2	14.7	13.7	7.4	**7.7**	11.6
Anl.Plt	1.1	1.0	1.7	1.7	1.8	0.7	0.1	0.1	0.0	**91.7**

Bold indicates the best results

method is good for Yoga class but not for the other classes. The reason is that usually the images of Yoga class contain caption text (which is edited text) but not scene text (natural text) like other classes. Since caption text is edited text, it has good clarity and visibility while scene text does not. Therefore, it leads one more option to choose an appropriate text detection method or develop a new method for achieving the best results for such deprived classes. This is the advantage of the proposed classification. For STD dataset, the text detection method gives better results at Recall for almost all the classes after classification compared to before classification. So, there is a significant difference between before classification and after classification at Recall.

Quantitative results of the recognition method [24, 29] on our dataset and STD dataset reported in Table 11 show that there are huge improvements for the results after classification compared to before classification at recognition rate for almost all the classes of both the datasets. However, for E-Learning and Economics classes, the recognition methods show poor results compared to before classification. This is justifiable because texts in those classes suffer from too small font and low contrast, which demands for more and relevant samples for training parameters of the classifier. Therefore, to achieve better results for those two classes, we can choose different methods which can cope with the challenges of E-Learning and Economics classes.

Sometimes, when images share the similar background or foreground as shown in Fig. 13, the proposed approach misclassifies those. This is the limitation of the proposed model. Therefore, there is a demand for improvement in future. In this situation, rather than extracting using a few modalities, such as text and face, it is necessary to extract the relationship between foreground and background by adding more modalities, such as video information, which can be considered in near future.

5 Conclusion and remark

We have proposed a new idea for the classification of action images to support the detection and recognition of embedded text. The proposed work combines three deep nets, namely, ResNet50 for overall pixel-distribution based classification, VGG16 for MSER components, and VGG16 for Face components for achieving enhanced classification performance. The proposed method exploits content of images at both pixel and component levels for tackling the complex classification problems at hand. Experimental results on our dataset of action images and two standard datasets show that the proposed method is effective and scalable. It is also noted from the experimental results that the proposed method outperforms existing methods in terms of classification rate for all three datasets. To show the significance of the proposed classification technique, text detection and recognition experiments before and after classification are conducted. Our future work would be dedicated towards solving such other challenges by combining features of text, image content and video.

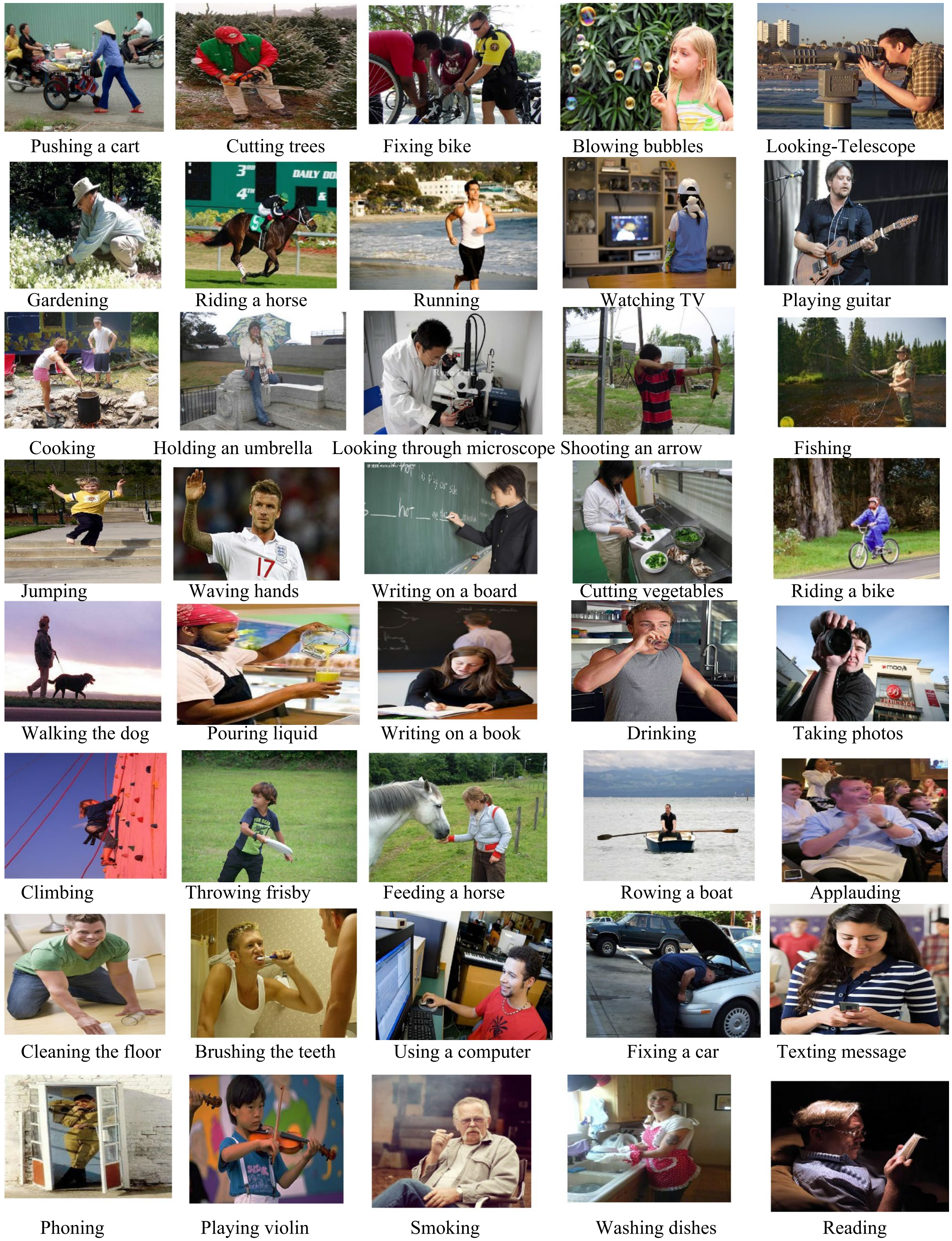


Fig. 8 Examples of successful classification of the proposed approach on Stanford 40 Actions dataset. Original Source: [47]

Table 9 Average Classification Rate of the Proposed and Existing Approach on Stanford 40 Actions Dataset

No	Classes	Proposed Method	Roy et al. [8]	GOOGLE API [35]	Noisy Student 2020 [38]	ViT 2021 [39]	Qin et al. [40]
1	Pushing_a_cart	69.9	89.9	80.5	71.0	72.5	30.4
2	Cutting_trees	50.0	73.0	73.00	61.2	87.0	39.1
3	Fixing_a_bike	80.0	68.9	40.56	64.3	65.7	42.4
4	Blowing_bubbles	64.9	67.0	69.56	51.0	41.1	53.5
5	Looking_through_a_telescope	75.0	67.0	63.2	89.2	68.2	50
6	Gardening	75.0	64.9	61.8	75.0	55.0	58.2
7	Riding_a_horse	60.0	64.9	72.8	42.0	45.1	49.3
8	Running	80.0	64.9	69	35.0	61.0	57.2
9	Watching_TV	69.9	64.9	52.1	54.5	67.2	3.5
10	Playing_guitar	85.0	62.2	60.4	58.0	51.0	48.1
11	Cooking	75.0	62.0	61.8	60.0	38.0	42
12	Holding_an_umbrella	100.0	62.0	58.4	73.5	51.5	43.5
13	Looking_through_a_microscope	89.9	61.0	54	51.5	47.1	43.5
14	Shooting_an_arrow	85.0	61.0	63.7	57.6	52.3	51
15	Fishing	80.0	60.0	67.2	61.9	45.0	58.4
16	Jumping	55.0	58.9	60.4	35.2	62.0	59.3
17	Waving_hands	34.9	58.9	55.7	72.0	67.5	42.1
18	Writing_on_a_board	80.0	57.9	59.1	91.0	75.4	61.7
19	Cutting_vegetables	55.0	56.9	47.3	65.3	92.0	42.5
20	Riding_a_bike	100.0	56.0	60.5	55.8	54.0	49.1
21	Walking_the_dog	89.9	55.0	61.5	42.3	66	51
22	Pouring_liquid	50.0	54.0	50.08	57.8	72.3	31.5
23	Writing_on_a_book	69.9	54.0	59.03	66.1	87.6	51.5
24	Drinking	34.9	52.9	56.6	43.1	75.1	40.1
25	Taking_photos	34.9	52.9	52.9	55.0	53.0	38.6
26	Climbing	85.0	51.9	48.3	53.9	46.0	39.6
27	Throwing_frisby	80.0	51.9	52.6	78.0	63.0	20.5
28	Feeding_a_horse	69.9	50.9	43.7	84.2	78.1	29.4
29	Rowing_a_boat	89.9	50.9	53.2	56.1	71.1	38.5
30	Applauding	40.0	50.0	43.5	67.0	66.0	41.9
31	Cleaning_the_floor	80.0	50.0	46.6	76.3	49.4	33.8
32	Brushing_teeth	44.9	47.9	52.4	48.0	64.5	40.4
33	Using_a_computer	64.9	47.9	49.1	85.0	86.8	40.3
34	Fixing_a_car	75.0	44.9	60.9	35.0	75.0	58.2
35	Texting_message	20.0	44.9	49	68.0	73.2	31.1
36	Phoning `	40.0	43.9	45.4	47.1	65.1	18.5
37	Playing_violin	64.9	43.9	46.4	63.7	61.0	28.5
38	Smoking	34.9	40.0	44.2	74.0	51.0	33.9
39	Washing_dishes	40.0	37.0	39.3	41.8	59.0	25.7
40	Reading	44.9	36.0	20.2	37.2	42.5	40.8
AoA		**65.4**	56.1	55.14	60.20	62.7	45.8

Bold indicates the best results

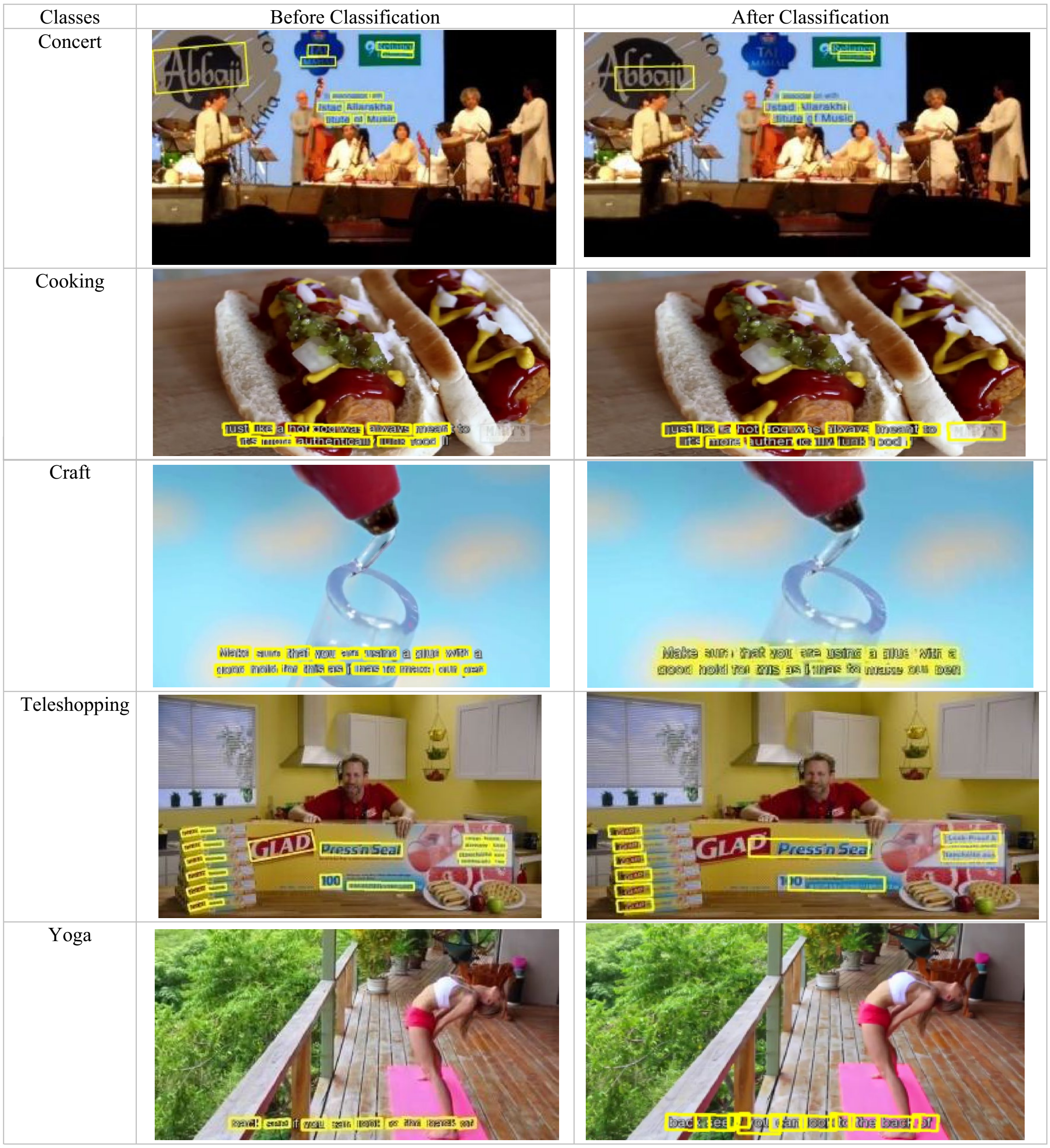
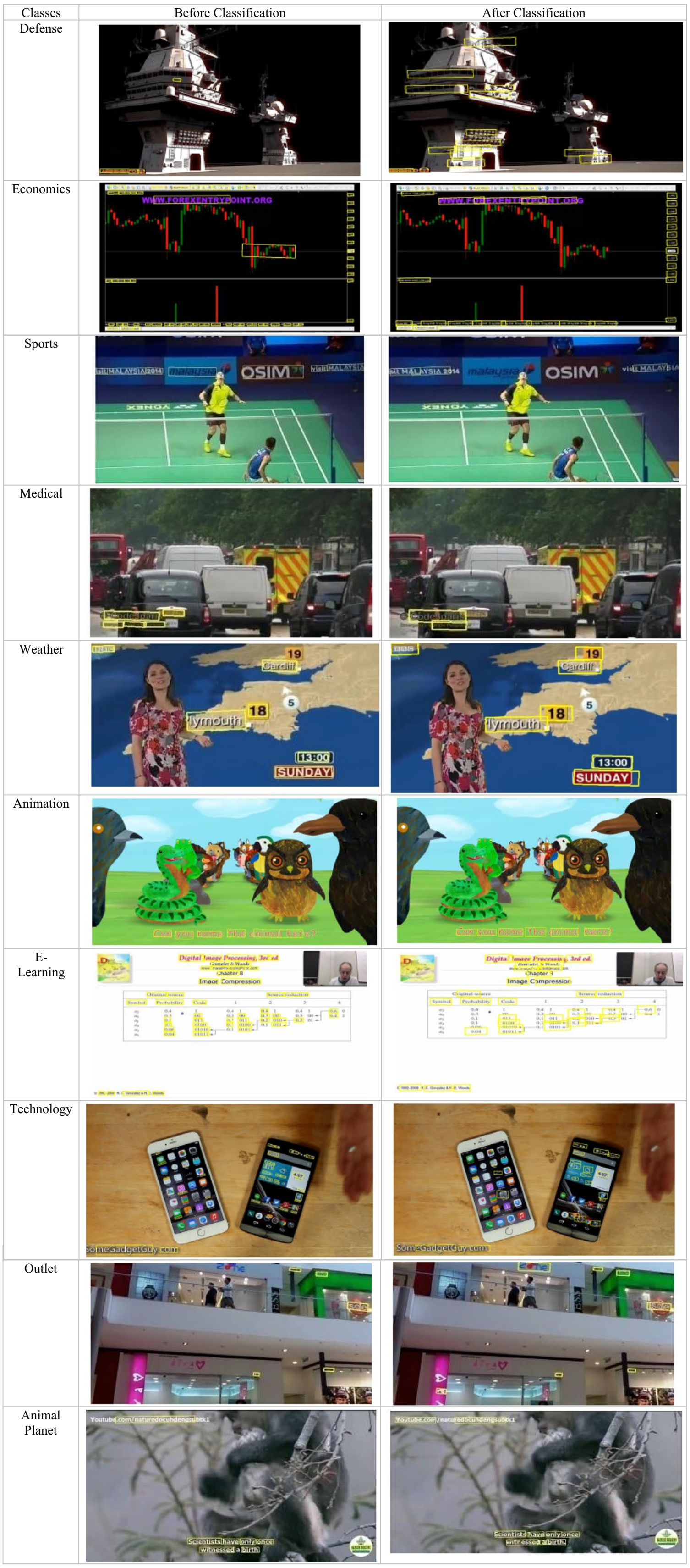


Fig. 9 Sample results of the text detection [9] on our dataset before and after classification. Original Source: [8, 42]

Fig. 10 Qualitative results of the text detection [9] on STD dataset before and after classification. Original Source: [8, 42]



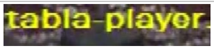
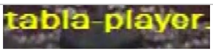
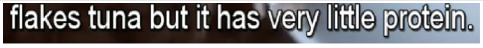
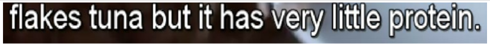
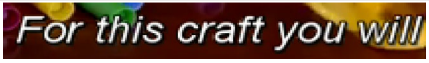
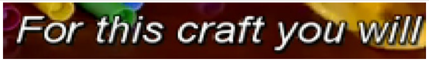
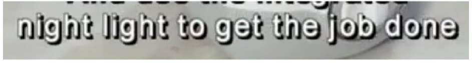
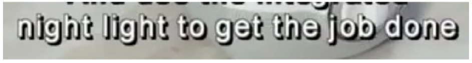
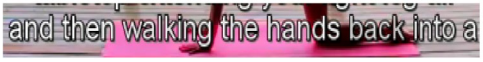
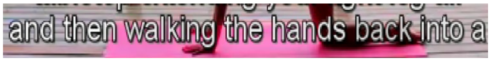
Classes	Before Classification	After Classification
Concert	 “$$$$$” player GT: tabla player	 tabla player GT: tabla player
Cooking	 flakes tuna but it has very little protein. GT: flakes tuna but it has very little protein.	 flakes tuna “$$$” “$$” has very little protein GT: flakes tuna but it has very little protein.
Craft	 For this craft you will GT: For this craft you will	 For “$$$$” craft you will GT: For this craft you will
Teleshopping	 night light “$$” get the job done GT: night light to get the job done	 night light t“$” get the job done GT: night light to get the job done
Yoga	 “$$$” “$$$$” walking the hands back into a GT: and then walking the hands back into a	 and “$$$$” walking the hands back into a GT: and then walking the hands back into a

Fig. 11 Qualitative results of the recognition method on our dataset before and after classification. GT denotes ground truth, “$” denotes recognition method gives nothing

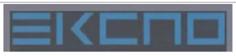
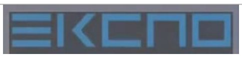
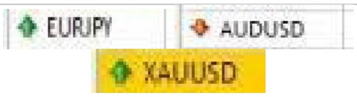
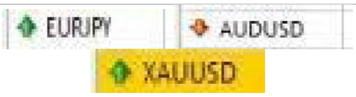
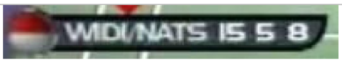
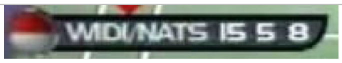
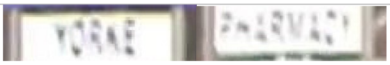
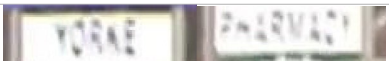
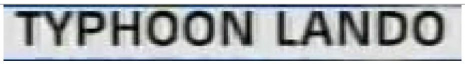
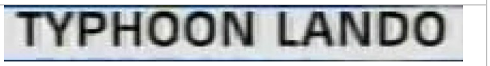
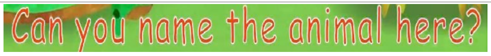
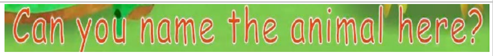
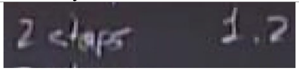
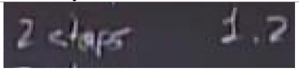
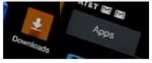
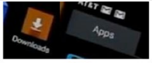
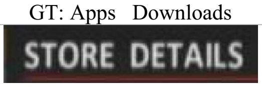
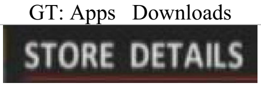
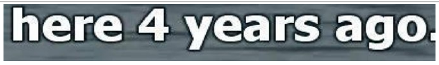
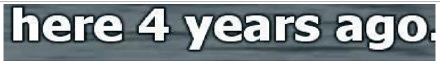
Classes	Before Classification	After Classification
Defense	 EK”$$$” GT: EKCNO	 EKC”$”O GT: EKCNO
Economics	 AUDUSD EURIPY XAUUSD GT: AUDUSD EURJPY XAUUSD	 AUDUSD “$$$$$$” “$$$$$$” GT: AUDUSD EURJPY XAUUSD
Sports	 WIDI NATS 15 5 8 GT: WIDI/NATS 15 5 8	 WIDI NATS “$$” “$” “$” GT: WIDI/NATS 15 5 8
Medical	 Y”$”64E P”$$”R”$”A”$$” GT: YORKE PHARMACY	 Y”$$”SE P”$”AR”M”A”$$” GT: YORKE PHARMACY
Weather	 TYPHOON LA GT: TYPHOON LANDO	 TYPHOON LANDO GT: TYPHOON LANDO
Animation	 Can you name the animal here”$” GT: Can you name the animal here?	 Can you name the animal here? GT: Can you name the animal here?
E-learning	 2 s”$”eps 1.2 GT: 2 steps 1.2	 2 s”$$$$” “$$$” GT: 2 steps 1.2
Technology	 Apps Downloads GT: Apps Downloads	 Apps Downloads GT: Apps Downloads
Outlet	 “$$$$$” DETAILS GT: STORE DETAILS	 STORE DETAILS GT: STORE DETAILS
Animal Planet	 here “$” years ago GT: here 4 years ago	 here 4 years ago GT: here 4 years ago

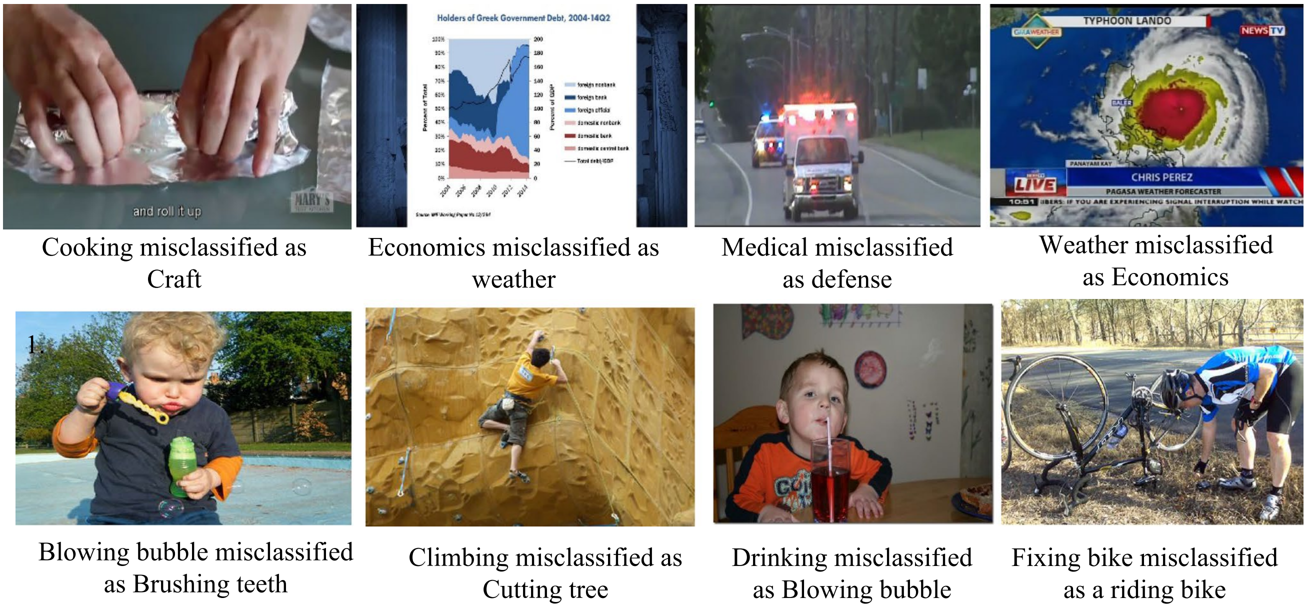
Fig. 12 Qualitative results of the recognition method on STD dataset before and after classification. GT denotes ground truth, “$” denotes recognition method gives nothing

Table 10 Performance of text detection methods (EAST and TextMountain) before and after classification on our and STD datasets

Datasets	Class	Before Classification			After Classification (EAST)			After Classification (Text Mountain [22])		
		Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
Our Dataset	Concert	0.99	0.91	0.95	0.93	0.73	0.82	0.95	0.80	0.86
	Cooking				0.99	0.93	0.96	0.97	0.93	0.94
	Craft				1	0.94	0.96	1	0.93	0.96
	Teleshopping				0.95	0.88	0.91	0.96	0.91	0.93
	Yoga				1	1	1	1	1	1
STD Dataset	Defense	0.95	0.76	0.84	0.58	0.9	0.72	0.62	0.93	0.74
	Economics				0.98	0.84	0.99	0.85	0.91	0.87
	Sports				0.93	0.75	0.83	0.89	0.86	0.87
	Medical				0.93	0.7	0.79	0.91	0.82	0.86
	Weather				1	0.82	0.90	0.98	0.80	0.88
	Animation				0.95	0.95	0.95	1	1	1
	E-Learning				0.98	0.59	0.74	0.99	0.77	0.86
	Technology				0.84	0.68	0.75	0.86	0.81	0.83
	Outlet				1	0.81	0.89	0.95	0.92	0.93
	Animal Plant				1	0.98	0.99	1	1	1

Table 11 Performance of Text Recognition methods E2E-MLT [24] and STAN[29] on our and STD datasets before and after classification

Datasets	Class	Before Classification	After Classification (E2E-MLT)	After Classification (STAN)
		Recognition Rate	Recognition Rate	Recognition Rate
Our Dataset	Concert	0.85	0.87	0.89
	Cooking		0.73	0.72
	Craft		0.81	0.80
	Teleshopping		0.96	0.97
	Yoga		0.88	0.88
STD Dataset	Defense	0.73	0.93	0.96
	Economics		0.55	0.63
	Sports		0.62	0.71
	Medical		0.8	0.75
	Weather		0.82	0.87
	Animation		0.82	0.85
	E-Learning		0.55	0.52
	Technology		0.73	0.71
	Outlet		0.79	0.82
	Animal Plant		0.99	0.96


and roll it up
Holders of Greek Government Debt, 2004-14Q2
TYPHOON LANDO
NEWS TV
CHRIS PEREZ
PAGASA WEATHER FORECASTER
LIVE
1.

Cooking misclassified as Craft

Economics misclassified as weather

Medical misclassified as defense

Weather misclassified as Economics

Blowing bubble misclassified as Brushing teeth

Climbing misclassified as Cutting tree

Drinking misclassified as Blowing bubble

Fixing bike misclassified as a riding bike

Fig. 13 Samples of unsuccessful classification results of the proposed approach on different datasets. Original Source: [42] and [8]

Acknowledgements Tong Lu received support for this work from the Natural Science Foundation of China under Grant 61672273, and Palaiahnakote Shivakumara received support for this work from the Faculty Grant: FP104-2020, FRGS, University of Malaya, Malaysia. This work is also partly supported by Technology Innovation Hub of Indian Statistical Institute, Kolkata.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Informed consent The images used in the proposed work are originally taken from [8, 42]. The authors of this paper and the authors of [8, 42] have given consent to use the data for publication.

References

1. Sharif A, Khan MA, Javed K, Umer HG (2019) "Intelligent human action recognition: A framework of optimal features selection based on Euclidean distance and strong correlation. Control Eng Appl Inf 21:3–11
2. Khan MA, Javed K, Khan SA, Saba T (2020) Human recognition using fusion of Multiview and deep features: an application to video surveillance. Multimed Tools Appl
3. Khan MA, Zhang YD, Attique M, Rehaman A, Seo S (2020) A resource conscious human action recognition framework using 26-layered deep convolutional neural network. Multimed Tools Appl
4. Sahoo SP, Ari S (2019) On an algorithm for human action recognition. Expert Syst Appl 115:524–534
5. Hernández-García R et al (2018) Improving bag-of-visual-words model using visual n-grams for human action classification. Expert Syst Appl 92:182–191
6. Nweke HF et al (2018) Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: state of the art and research challenges. Expert Syst Appl 105:233–261
7. Sreela S, Idicula SM (2018) Action recognition in still images using residual neural network features. Procedia Comput Sci 143:563–569
8. Roy S et al (2018) Rough-fuzzy based scene categorization for text detection and recognition in video. Pattern Recogn 80:64–82
9. Xinyu Z, Yao C, Wen H, Wang Y, Zhou S, He W, Liang J (2017) EAST: an efficient and accurate scene text detector. CVPR, pp. 2642–2651
10. Wang H, Huang S, Jin L (2018) Focus on scene text using deep reinforcement learning. In: Proc. ICP, pp. 3759–3765
11 Zhang X, Gho X, Tian C (2018) Text detection in natural scene images based on color prior guided MSER. Neurocomputing 307:61–71
12 Van Nguyen D, Lu S, Tian S, Ouarti N, Mokhtari M (2019) A pooling-based scene text proposal technique for scene text reading in the wild". Patten Recognit 87:118–129
13. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Proc. NIPS, pp. 1–9

14. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proc. CVPRW, pp. 1–9
15. Schroff F, Kalenichenko D, Philbin J (2015) FaceNet: a unified embedding for face recognition and clustering. In: Proc. CVPRW
16. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large scale image recognition. In: Proc. ICLR, pp. 1–14
17. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proc. CVPRW
18. Sain A, Bhunia AK, Roy PP, Pal U (2018) Multi-oriented text detection and verification in video frames and scene images. Neurocomputing 1549(275):1531
19. Wang S, Liu Y, He Z, Wang Y, Tang Z (2020) A quadrilateral scene text detector with two-stage network architecture. Pattern Recognit 102:107230
20. Liu Y, Chen H, Shen C, He T, Jin L, Wang L (2020) ABCNet: real-time scene text spotting with adaptive Bezier curve network. In: Proc. CVPR
21 Wang C, Fu H, Yang L, Cao X (2020) Text co-detection in multi-view scene. IEEE Trans IP 29:4627–4642
22. Zhu Y, Du J (2021) TextMountain: accurate scene text detection via instance segmentation. Pattern Recognit 110:107336
23. Yang CSY, Yang YH (2017) Improved local binary pattern for real scene optical character recognition. Pattern Recognit Lett 100:14–21
24. Bušta M, Patel Y, Matas J (2018) E2E-MLT-an unconstrained end-to-end method for multi-language scene text. In: Springer Asian Conference on Computer Vision, pp. 127–143
25. Shivakumara P, Wu L, Lu T, Tan CL, Blumenstein M, Anami BS (2017) Fractals based multi-oriented text detection system for recognition in mobile video images. Pattern Recognit 68:158–174
26. Lee J, Park S, Baek J, Oh SJ, Kim S, Lee H (2020) On recognizing text of arbitrary shapes with 2D self-attention. In: Proc. CVPRW, pp. 2326–2335
27. Long S, Guan Y, Bian K, Yao C (2020) A new perspective for flexible feature gathering in scene text recognition via character pooling. In: Proc. ICASSP, pp. 2458–2462
28. Shang M, Gao J, Sun J (2020) Character region awareness network for scene text recognition. In: Proc. ICME
29 Lin Q, Luo C, Jin L, Liu S, Lai S (2021) STAN: A sequential transformation attention-based network for scene text recognition. Pattern Recognit 111:107692
30 Dang LM, Hassan SI, Im S, Mehmood I, Moon H (2018) Utilizing text recognition for the defects extraction in sewers CCTV inspection videos. Comput Ind 99:96–109
31. Basnyat B, Roy N, Gangopadhyay A (2018) A flash flood categorization system using scene text recognition. In: Proc. ICSC, pp. 147–154
32 Xu P, Yang Y, Xu Y (2017) Person re-identification with end-to-end scene text recognition. Springer, New York, pp 363–374
33. Bosch A, Zisserman A, Munoz X (2008) Scene classification using hybrid generative/discriminative approach. IEEE Trans PAMI 30:712–727
34. Dunlop H (2010) Scene classification and video via semantic segmentation. In: Proc. CVPRW, pp. 72–79
35. Google Vision API, "https://cloud.google.com/vision/ ".
36. Bai S, Tang H, An S (2019) Coordinate CNNs and LSTMs to categorize scene images with multi-views and multi-levels of abstraction. Expert Syst Appl 120:298–309
37. Xue M, Shivakumara P, Wu X, Lu T, Pal U, Blumenstein M, Lopresti D (2020) Deep invariant texture features for water image classification. SN Appl Sci 2:1–19
38. Xie Q, Luong M-T, Hovy E, Le QV (2020) Self-training with noisy student improves ImageNet classification. In: Proc. CVPR, pp. 10684–10695
39. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16×16 words: transformers for image recognition at scale. In: Proc. ICLP
40. Qin L, Shivakumara P, Lu T, Pal U, Tan CL (2016) Video scene text frames categorization for text detection and recognition. In: Proc. ICPR, pp. 3875–3880
41. Shivakumara P, Raghavendra R, Qin L, Raja KB, Lu T, Pal U (2017) A new multi-modal approach to bib number/text detection and recognition in Marathon images. Patten Recognit 61:479–491
42. Nandanwar L, Shivakumara P, Manna S, Pal U, Lu T, Blumenstein M (2020) A new DCT-FFT fusion based method for caption and scene text classification in action video images. In: Proc. ICPRAI, pp. 80–92
43. Matas J, Chum O, Urban M, Pajdla T (2002) Robust wide baseline stereo from maximally stable extremal regions. In: Proc. BMVC, pp. 384–396
44. Xiang J, Zhu G (2017) Joint face detection and facial expression recognition with MTCNN. In: Proc. ICISCE, pp. 424–427
45. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks. In: Proc. NIPS'14, pp. 3320–3328
46. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: Proc. CVPRW, pp. 248–255
47. Yao B, Jiang X, Khosla A, Lin AL, Guibas L, Fei-Fei L (2011) Human action recognition by learning bases of actions attributes and parts. In: Proc. ICCV, pp. 1331–1338

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.