A New DCT-FFT Fusion based Method for Caption and Scene Text Classification in Action Video Images

[1]Lokesh Nandanwar, [1]Palaiahnakote Shivakumara, [2]Suvojit Manna, [3]Umapada Pal, [4]Tong Lu and [5]Michael Blumenstein

[1]Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. Email: lokeshnandanwar150@gmail.com, shiva@um.edu.my

[2]Dept. of Computer Science & Engineering, Jalpaiguri Govt. Engineering College, Jalpaiguri, India. Email: davsuvo@gmail.com

[3]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India. Email: umapada@isical.ac.in

[4]National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China. Email: lutong@nju.edu.cn

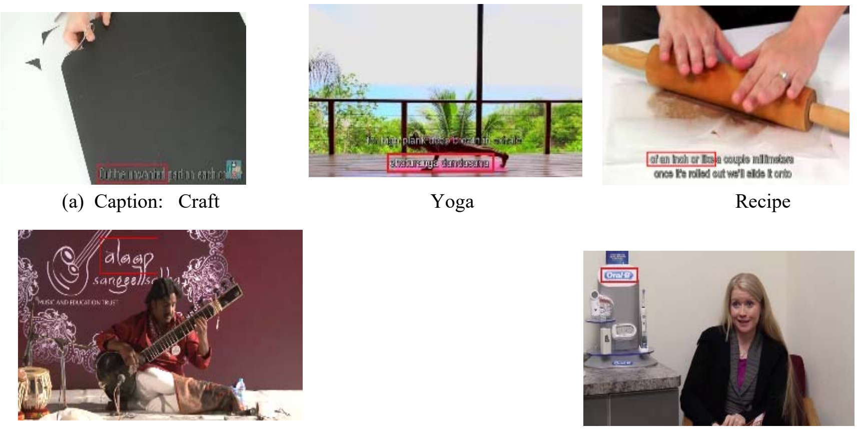
[5]University of Technology Sydney, Australia. Email: michael.blumenstein@uts.edu

Abstract. Achieving better recognition rate for text in video action images is challenging due to multi-type texts with unpredictable backgrounds. We propose a new method for the classification of captions (which is edited text) and scene texts (which is part of an image in video images of Yoga, Concert, Teleshopping, Craft, and Recipe classes). The proposed method introduces a new fusion criterion-based on DCT and Fourier coefficients to extract features that represent good clarity and visibility of captions to separate them from scene texts. The variances for coefficients of corresponding pixels of DCT and Fourier images are computed to derive the respective weights. The weights and coefficients are further used to generate a fused image. Furthermore, the proposed method estimates sparsity in Canny edge image of each fused image to derive rules for classifying caption and scene texts. Lastly, the proposed method is evaluated on images of five above-mentioned action image classes to validate the derived rules. Comparative studies with the state-of-the-art methods on the standard databases show that the proposed method outperforms the existing methods in terms of classification. The recognition experiments before and after classification show that the recognition performance rate improves significantly after classification.

Keywords: Caption text, Scene text, Fusion, Caption and Scene text classification, Action image recognition.

1 Introduction

Understanding text in action video images is a hot topic for researchers in the fields of computer vision and image processing [1, 2, 3]. This is because of its several real world applications such as surveillance, human-computer interaction, robotics and indexing videos [2, 3]. There are powerful methods for recognizing the action in the video proposed in literature [2, 3]. However, due to the difficulty in defining the exact relationship between coherent objects in images with the help of content-based features, achieving better recognition is still considered as elusive goal. The difficulty is further aggravated for still images without temporal frames [4]. Therefore, in this work, we propose to make an attempt to find a solution to the above complex issues based on text information if text is present in the video images. It is true that either caption, which is edited text or scene text, which is part of an image, play a vital role in action recognition. Since text in the action video images is closer and it provides semantic information about the content of the video image, achieving better recognition results improve performance of the action recognition in the video. Sometimes, a few applications may require only caption text as it describes the content of the video while scene text does not. For these applications, caption text separation from scene text is important to obtain accurate results [5].


of an inch or like a couple millimeters
once it's rolled out we'll slide it onto
alaap

(a) Caption: Craft Yoga Recipe

(b) Scene: Concert Teleshopping

Fig.1. Examples of scene and caption texts in action images.

For recognizing text in video and natural scene images, the methods are developed in the past [6, 7] addresses several challenges of natural scene text. However, when the images contain both caption and scene text, the performance of the method degrades. This is due to different nature and characteristics of caption and scene text. Since caption is edited text, one can expect good contrast and visibility while scene is natural text, one cannot predict the nature of text [8]. Sample caption and scene texts of different action images are shown in Fig.1(a) and Fig.1(b), respectively, where texts are marked by red colored rectangles. For the text in the images, the proposed method tested recognition method [7], which is state-of-the-art method before and after classification. Before classification caption and scene texts are considered as input for recognition while after classification individual text is considered as input for recognition. The recognition results of the method reported in Table 1 show that there is incorrect recognition results before classification and correct recognition results after classification. Therefore, it is noted that classification is necessary to improve performance of recognition of the text in video irrespective of applications. However, in this work, we consider action video images as case study for experimentation. In this work, we consider five video images of action classes, namely, Concert which describes music, Teleshopping which describes marketing of goods, Craft which describes art making, Yoga which describes particular exercise, and Recipe which describes food-making as shown in Fig.1, where both captions and scene texts are present in these images. The main reason to choose the above-mentioned action classes is that they are quite popular for indexing and retrieval in the field of computer vision.

Table 1. Recognition results of the method [7] before and after classification of action images

Action Classes	Before Classification	After Classification
Craft	cut the wnwarted	cut the unwanted
Yoga	chatvarga nianiasedu	chaturanga dandasana
Recipe	of an inch orite	of an inch or like
Concert	Aloop	Alaap
Teleshopping	orals	oralb

There are methods developed in the literature [8-12] for the classification of scene and caption texts in video images. However, the methods may not give satisfactory results for texts in video images of action classes because the presence of indoor and outdoor scenes in different action classes makes the problem much more complex. It is evident from the images shown in Fig.1(a) and Fig.1(b), where one can clearly see the effects of indoor and outdoor scenes on text information. Hence, caption and scene texts separation for video images of action classes is challenging and interesting compared to existing methods.

2 Related Work

In this work, we review the methods which focus on the classification of caption and scene texts in video. Raghunandan et al. [8] proposed a method for classifying video, natural scene, mobile and born digital images based on text information. The methods work well when images have different qualities according to classes. This is not necessarily true for texts in video images of actions, where one can expect texts may share the same quality. Shivakumara et al. [9] proposed the separation of graphics and scene texts in video frames based on the combination of Canny and Sobel edge images along with the ring radius transform. Since the method works based on edge quality, it may not perform well for complex images of video. Roy et al. [10] proposed tampered features for the classification of captions and scene text. In this method, DCT coefficients are explored to detect tampered information in caption images such that it can be separated from scene texts. It is noted that DCT coefficients alone may not be sufficient to classify complex caption and scene texts in action images. In a similar way, Bhardwaj et al. [11] explored tampered features which represent superimposed texts in video for text detection by separating scene texts from captioned ones. The method is good for images with high quality and contrast.

Recently, Roy et al. [12] proposed temporal integration for word-wise caption and scene text identification. This method may not work well for still images as it requires temporal frames for achieving accurate results. Later, Roy et al. [13] proposed rough-fuzzy based scene categorization for text detection and recognition in video. The method defines shapes based on edge components for the classification of videos of different text types. However, the scope of this method is to categorize video frames but not caption and scene texts in video. Since use of convolutional network is popular and it solves complex issue, Ghosh et al. [5] proposed identifying the presence of graphical text in scene images using CNN. The method considers text which is edited and text in natural scene images as graphical text for classification. To achieve this, the method explores CNN. However, the method does not consider caption and scene text in video images for classification. From the above discussions, it is noted that none of the methods are perfect for video images of action classes, where one can expect texts with more unpredictable background variations.

Hence, this work focuses on the classification of caption and scene texts in video images of different action classes. As mentioned in the Introduction Section, caption text exhibits good quality, contrast and uniform color, while the nature of scene text is unpredictable. To extract such an observation, inspired by the method in [10] where it is shown that DCT coefficients help in extracting distinct features for the differentiation of caption texts from scene ones via classification, we explore the DCT domain in a different way in this work. Since the problem being considered is a complex one, we are inspired by the method in [14] where Fourier coefficients are used for separating coefficients that represent actual edge pixels from those that represent noise ones for identifying printers. We explore the Fourier coefficients in a different way for the classification of scene and caption texts in this work. The main contributions of the work are as follows: (1) the way in which the advantages of both DCT and Fourier coefficients are integrated and utilized to obtain fused images is novel, and (2) exploring sparsity at the component level for Canny edge images of fused ones for classifying caption and scene texts is new.

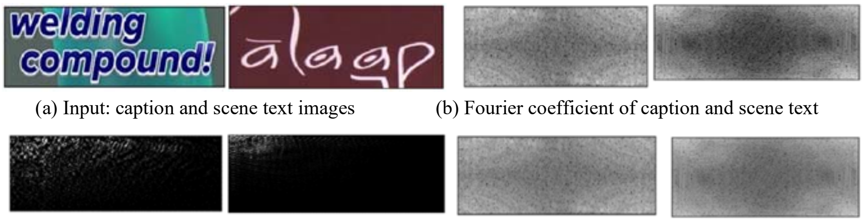
3 Proposed Method

This work considers detected texts from video images of action classes as the input for the classification of scene and caption texts. This is because there are methods which detect texts well irrespective of text types in videos [15]. As mentioned in the previous section, DCT and Fourier coefficients are useful in extracting distinct properties of caption and scene texts. We exploit the above coefficients as motivated by the method in [16], where the fusion concept is introduced in the wavelet domain to enhance degraded medical images. In other words, the method in [16] fuses coefficients that represent edge pixels in different wavelet sub-bands. This observation motivated us to propose a new fusion approach by combining coefficients in the DCT and Fourier domains. This is justifiable because we believe that if caption text has good quality and contrast, the same thing should be reflected in the fusion image. When the nature of scene text is unpredictable, one cannot expect regular patterns in the fused image represented by coefficients. As a result, the fused image of a scene text image results in sparsity, which is opposite to caption text.

Therefore, we can conclude that a fusion image of caption text provides edge information, while a fusion image of scene text provides sparse information. To extract such an observation for classifying caption and scene texts, we perform the inverse Fourier transform on fused images of both scene and caption texts. If the reconstructed image of scene text provides sparse information, the Canny edge operator gives nothing; while for caption text, the Canny edge operator gives edges of text information. In this way, the proposed method explores sparsity for classifying caption and scene texts in this work.

3.1 DCT and FFT Coefficients for Fusion

For the input caption and scene text line images shown in Fig.2(a), the proposed method obtains Fourier and DCT coefficients, respectively using Equation (1) and Equation (2) as shown in Fig.1(b) and Fig.1(c). It is observed from Fig.1(b) and Fig.1(c) that the caption text of Fourier and DCT coefficients is brighter than the scene text of Fourier and DCT coefficients. This is understandable because caption text has good clarity, contrast and clear differences between the text and its background, while scene text does not. As a result, the Fourier and DCT coefficients of scene text images appear darker compared to caption text images as shown in Fig.2(b) and Fig.2(c). With this cue, we propose to compute the variance of each pixel in the input image by defining a local window as defined in Equation (3) for the respective Fourier and DCT coefficients images. For variance computations, the proposed method considers the respective coefficients corresponding to pixels of the local window defined over the input image. In other words, the variance is computed using frequency coefficients. However, the local window is moved pixel by pixel by referring to the input image. The variances of the respective Fourier and DCT coefficients are used to derive weights as defined in Equation (4) and Equation (5). Finally, the derived weights are combined with the respective Fourier and DCT coefficients as defined in Equation (6), which results in a fused image. The effect of fusion can be seen in Fig.2(d) for both caption and scene texts, where we can see clearly that the fused image of caption text appears brighter than that of the scene text. This is true because the variance of coefficients of Fourier and DCT matches for caption text, while it mismatches for scene text. As a result, one can expect the fused image of caption text must contain high frequency coefficients that represent edge pixels, while that of scene text contains more low frequency coefficients that represent zero.



(a) Input: caption and scene text images (b) Fourier coefficient of caption and scene text

(c) DCT coefficient images of caption and scene (d) Fusion of DCT and Fourier coefficient

Fig.2. Examples of obtaining a fused image for caption and scene text images.

$$I_{FFT}(\mathrm{u},\mathrm{v}) = \frac{1}{MN}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} I(x,y)e^{-2\pi i\left(\frac{ux}{M}+\frac{vy}{N}\right)} \tag{1}$$

$$I_{DCT}(\mathrm{u},\mathrm{v}) = \frac{2}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} \lambda(x).\lambda(y).\cos\left[\frac{u\pi}{2M}(2x+1)\right].\cos\left[\frac{v\pi}{2N}(2y+1)\right].I(x,y) \tag{2}$$

$$Var(\mathrm{x},\mathrm{y}) = \frac{1}{M \times M}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}[abs(p(i,j) - \mu)]^2 \tag{3}$$

where Var(x,y) denotes the variance of a local window of size 3×3, $p(i, j)$ is the coefficient value at position (i, j), and μ is the mean of the coefficients of the local window.

$$\omega_{I_{FFT}} = \frac{Var_{I_{FFT}}}{Var_{I_{FFT}} + Var_{I_{DCT}}} \tag{4}$$

$$\omega_{I_{DCT}} = \frac{Var_{I_{DCT}}}{Var_{I_{FFT}} + Var_{I_{DCT}}} \tag{5}$$

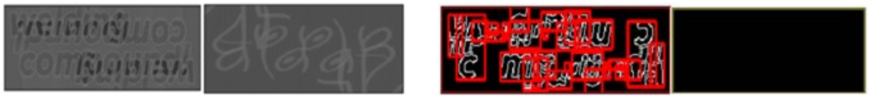
Where ω_{IFFT} denotes the weight with respect to Fourier coefficients, ω_{IDCT} denotes the weight with respect to DCT coefficients, and I denotes an image.

$$F_{coeff} = I_{FFT}\ o\ \omega_{I_{FFT}} + I_{DCT}\ o\ \omega_{I_{DCT}} \tag{6}$$

where ◦ is the hadamard product between two given matrices. F_{coeff} denotes the fused image, I_{IFFT} denotes the Inverse Fourier transform, and I_{DCT} denotes the Inverse Discrete Cosine transform.

3.2 Classification of Caption and Scene Text

It is noted from the fused image obtained from the previous step that a caption text image contains vital information, while a scene text image results in sparsity. To extract such an observation, the proposed method applies the Inverse Fourier transform for both the fused images of caption and scene texts using Equation (7). This step outputs reconstructed images of caption and scene texts as shown in Fig.3(a). When we look at the reconstructed images of caption and scene texts, it is non-trivial to notice sparsity and non-sparsity. Therefore, the proposed method applies the Canny edge operator on the fused images as shown in Fig.3(b), where one can see clearly that caption text provides significant edge information, while scene text provides nothing. This is the advantage of the fusion of Fourier and DCT coefficients in the frequency domain.



(a).

(b)

Fig.3. Sparse estimation at the component level for classification of caption and scene texts. (a) The result of the inverse Fourier transform on the fused image of caption and scene text images. (b) Canny edge images of a fused image of caption and scene text images

To extract the sparsity property in edge images of caption and scene text images, the proposed method considers edge components as defined by the bounding box shown in Fig.3(c). For each pixel of each edge component, we define the window of size 3×3 to check whether it contains more than one white pixel. If the window satisfies the above condition, the proposed method counts it as non-sparsity, else it is counted as sparsity. In this way, the proposed method obtains the number of sparsity and non-sparsity counts for each component in the image. Then it computes the average for sparsity and non-sparsity counts of all the edge components separately in the image. If the average of the non-sparsity count is larger than that of the sparsity count, the image is considered as caption text, else it is scene text. Since most of the time, for scene text images, the edge operator gives nothing. Therefore, scene text is classified with high sparsity.

$$I_{IFFT}(\mathrm{u},\mathrm{v}) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} F_{coeff}(x,y)e^{2\pi i\left(\frac{ux}{M}+\frac{vy}{N}\right)} \quad (7)$$

4 Experimental Results

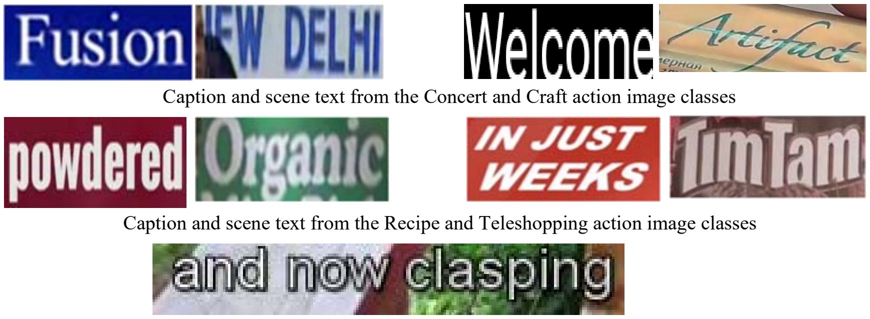
Since our target is to classify caption and scene text in action video images, we create our own dataset by collecting from different sources, such as YouTube, Instagram, the Internet and from our own camera, which comprises 429 video images for Concert, 509 for Recipe, 517 for Craft, 507 for Teleshopping and 524 for Yoga action classes. In total, 2486 video images for all the five action classes are considered. The dataset includes video images of different resolutions ranging from 426×240 to 1980×1080, different contrasts, foregrounds and background complexities. We detect caption and scene text images from the above dataset, which consists of 2814 text images including caption and scene texts. Since the dataset includes text with different varieties, the dataset is considered to be complex and fair for evaluating the performance of the proposed and existing methods. To test the objectiveness of the proposed method, we also conducted experiments on a standard dataset, which is used for caption and scene text classification in video images [10]. This dataset consists of 900 caption and 650 scene texts, which gives a total of 1550 text images for experimentation.

To illustrate the effectiveness of the proposed method, we implement two state-of-the-art methods for comparative studies. Roy et al. [10] proposed new tampered features for scene and caption text classification in video, which explores DCT coefficients for extracting tampered features for classification. Ghosh et al. [5] proposed a method for identifying the presence of graphical text in natural scene images by exploring CNN. The reason to choose the above methods is that the method [10] focuses on caption and scene text in video as the proposed method while the method [5] focuses on use of CNN for detecting the presence of graphical text in the natural scene images. In this work, we modify the CNN for classification of caption and scene text rather than identifying the presence of graphical text. We consider these two methods for comparative study in this work to show that only DCT used in method [10] is not adequate to achieve better results. The method [5] is considered to show that CNN approach does not perform well compared to the proposed rule based method. This can be justifiable because CNN works well when we have a large number of samples for training and learning while the rule based methods does not. In the same way, to validate the effectiveness of the proposed classification, we implement the following text recognition methods. Luo et al. [7] which proposes Multi-Object Rectified Attention Network (MORAN) for scene text recognition and Shi et al. [6] which proposes an Attention Scene Text Recognizer with Flexible Rectification (ASTER). The reason to choose the above two methods is that the above methods explored deep learning models for addressing several challenge of scene text recognition. In addition, since the codes of the above two methods are available, we prefer to use the same for recognition experiments in this work.

For measuring the performance of the proposed and existing methods for classification and recognition, we calculate classification rate through confusion matrix of caption and scene text classification. For recognition, we calculate character recognition rate before and after classification. As mentioned earlier, before classification both caption and scene text are considered as input for recognition while after classification individual classes are considered as input for recognition. It is expected the recognition methods should report better results for after classification compared to before classification. This is valid because the recognizer is trained with caption and scene text together for before classification. On the other hand, the recognizer is trained separately for each class after classification.

4.1 Evaluating the Proposed Classification Method

Sample qualitative results of the proposed method for each action class are shown in Fig.4, where we can see that the proposed method classifies caption and scene texts successfully regardless of background complexities. Quantitative results of the proposed and existing methods are reported in Table 2 for our dataset. Table 2 shows that the proposed method is the best at the classification rate for all the five action classes compared to existing methods. It is noted from Table 2 that the proposed method scores the highest average classification rate which is mean of diagonal elements of confusion matrix for the Craft class and the lowest for the Concert class. This is due to texts in the Craft class that are not exposed much to open environments, while texts in the Concert class are exposed to open and closed environments.



Caption and scene text from the Concert and Craft action image classes

Caption and scene text from the Recipe and Teleshopping action image classes

Caption text from the Yoga action image class

Fig.4. Sample images where successful classification is performed by the proposed method.

When images are exposed to both open and closed environments, they get affected by multiple factors such as uneven illumination, dim lighting, different writing styles, and the effect of rough surfaces, perspective distortion, and occlusion. The main reason for the poor results of the existing methods is that they are developed for classifying texts with different objectives but not video images of action classes. When we compare the results of two existing methods, the method [5] scores better results than the method [10]. This shows the method which explores CNN is better than the method which uses only DCT based features for classification. On the other hand, the proposed method is better than the existing methods because of the fusion criteria that have been introduced, which consider the advantages of the Fourier and DCT coefficients in the frequency domain for extracting sparsity. This demonstrates a difference in comparison to existing methods to achieve the best results. Note that the symbol '-'indicates that no classification results reported in Table 2 and this is because the Yoga action class does not provide any scene text. It is reported classification rate of only caption text. It is noted from Table 2 that the methods including proposed one score the best results at classification rate for the caption text compared to scene text. This justifies that caption text provide good clarity and visibility compared to scene text.

Table 2. Confusion matrix of the proposed and existing methods on action image classes (in %).

Methods	Actions	Recipe		Concert		Crafts		Teleshopping		Yoga	
	Classes	Scene	Caption	Scene	Caption	Scene	Caption	Scene	Caption	Scene	Caption
Proposed	Scene	**71.4**	28.5	**74.3**	25.6	**97**	3	**71.8**	28.1	**-**	-
	Caption	16.2	**83.8**	27.0	**72.9**	0	**100**	20.4	**79.5**	3.42	**96.5**
	Average	**77.6**		**73.6**		**98.5**		**75.6**		**96.5**	
Roy et al. [10]	Scene	64.2	35.7	62.4	37.5	66.6	33.3	63.8	36.1	-	-

	Caption	32.5	67.4	35.2	64.7	33.5	66.4	34.5	65.4	33.2	66.7
	Average	65.8		63.55		66.5		64.6		66.7	
Ghosh et [5]	Scene	60.15	39.85	61.61	38.39	64.01	35.99	61.03	38.97	-	-
	Caption	18.7	81.3	8.1	91.9	7.3	92.7	13.4	86.6	6.98	93.02
	Average	70.7		76.7		78.3		73.8		93.0	

To test objectiveness of the proposed method, we conduct experiments on a standard dataset for classification. Quantitative results of the proposed and existing methods are reported in Table 3. It is observed from Table 3 that the proposed method is better than the existing methods in terms of average classification rate. The main drawback of existing methods is that the existing features do not have sufficient discriminative power when compared to the proposed method. When we compare the results on our dataset and the standard dataset, the proposed method gives almost consistent results if we consider the average classification rate, while the existing methods report inconsistent results. This is mainly because of the differences in complexity of the two datasets.

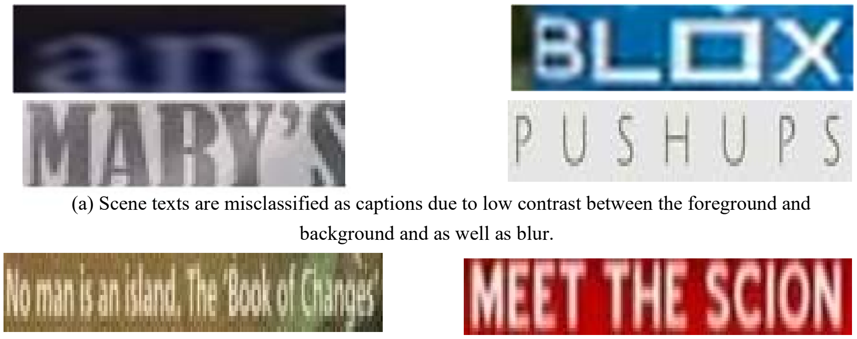
Table 3. Confusion matrix of the proposed and existing methods on Dataset [10] (in %)

Types	Proposed Method		Ghosh et al.[5]		Roy et al. [10]	
	Scene	Caption	Scene	Caption	Scene	Caption
Scene	**79.11**	20.88	69.53	15.9	65.69	34.31
Caption	21.99	**78.01**	26.01	79.67	32.62	67.38
Average	**78.56**		74.6		66.53	

In order to show the usefulness of the proposed classification method, we conduct experiments for text recognition methods using the ASTER and MORAN methods [6, 7] respectively. The recognition results of both the methods for all the five action video image classes before and after classification are reported in Table 4. For before classification, we use pre-trained model with same parameters and values while for after classification, we tune the parameters, namely, "Epochs and "Batch size" according to data of individual classes. This is the advantage of the classification. For training and testing of the recognition methods, we use 75% data for training and 25% data for testing. Quantitative results of the two recognition methods before and after classification for all the five action video image classes are reported in Table 4, where it is noted that both the recognition methods achieve better results after classification compared to before classification for all the five classes on both caption and scene text. It can also be seen in Table 4 that the MORAN is better than ASTER before and after classification. This is due to the MORAN focusses on irregular shaped text along with distortion but the ASTER focusses on distorted images. However, in case of the proposed work, one can expect both irregular shaped text and text affected by distortion. Therefore, we can conclude that the classification is useful and effective to improve recognition performance for video irrespective of applications.

Table 4. Recognition performance of the methods before and after classification (in %)

Methods	ASTER [6]			MORAN [7]		
	Before Classification	After Classification		Before Classification	After Classification	
Classes	Caption + Scene	Caption	Scene	Caption +Scene	Caption	Scene
Concert	71.2	73.6	77.8	74.4	73.3	77.2
Recipe	72.4	89.0	63.0	82.2	90.2	76.2
Crafts	74.2	80.8	70.0	75.0	81.9	70.5
Teleshopping	66.8	71.8	67.2	72.9	76.8	72.3
Yoga	79.0	80.4	-	81.2	82.0	-
Dataset [10]	61.8	69.6	66.0	63.0	63.2	69.3



(a) Scene texts are misclassified as captions due to low contrast between the foreground and background and as well as blur.

(b) Caption texts are misclassified as scene text due to background and foreground complexity variations.

Fig.5. Limitations of the proposed method.

Though the proposed classification method works well for several cases, sometimes, if the images suffer from very low contrast where we cannot differentiate between the foreground and background, the proposed method misclassifies them as per the samples of misclassifications shown between scene and caption text images in Fig.5. Therefore, there is scope for undertaking improvements in the future.

5 Conclusion and Future Work

In this paper, we have proposed a novel method for the classification of caption and scene texts in action video images of different classes. The proposed method introduces a new fusion concept to integrate the advantages of Fourier and DCT coefficients to extract sparsity for the classification of scene and caption texts in action images. The variances are computed in the frequency domain for both Fourier and DCT coefficients by referring to pixel positions in the input images. The variances are used to derive weights with respect to the Fourier and DCT coefficients. Furthermore, the proposed method uses weights and coefficients to generate fused images. Finally, sparsity is estimated for reconstructed images of the fused ones for classification. Experimental results on our own dataset and a standard dataset show that the proposed method outperforms the existing methods in terms of average classification rate. Experimental results on recognition show that the proposed classification improves recognition performance significantly after classification compared to before classification. However, when an image suffers from very low resolution and low contrast, the proposed method does not perform well. We will explore such situations in our future work.

References

1. A. Ullah, J. Ahmad, K. Muhamad, M. Sajiad and S. W. Baik, “Action recognition in video sequences using deep Bi-directional LSTM with CNN features”, IEEE Access, pp 1155-1166, 2018.
2. T. Qi, Y. Xu, Y. Quain, Y. Wang and H. Ling, “Image based action recognition using hint enhanced deep neural networks”, Neurocomputing, 267, pp 475-488, 2017.
3. G. Taniski, C. Zalluhoglu and N. L. Cinbis, “Facial descriptors for human interaction recognition in still images”, Pattern Recognition Letters, 73, pp 44-51, 2016.
4. S. Yuan, J. S. Smith and B. Zhang, “Action recognition from still images based on deep VLAD spatial pyramids”, Signal Processing: Image Communication, 54, pp 118-129, 2017.
5. M. Ghosh, H. Mukherjee, S. M. Obaidullah, K. C. Santosh, N. Das and K. Roy, "Identifying the Presence of Graphical Texts in Scene Images using CNN," In Proc. ICDARW, pp. 86-91, 2019.
6. B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao and X. Bai, “ASTER: An attentional scene text recognizer with flexible rectification”, IEEE Trans. PAMI, pp 2035-2048, 2019.
7. C. Luo, L. Jin and Z. Sun, “MORAN: A multi-object rectified attention network for scene text recognition”, Pattern Recognition, Vo. 90, pp 109–118, 2019.
8. K. S. Raghunandan, P. Shivakumara, G. H. Kumar, U. Pal and T. Lu, “New Sharpness Features for Image Type Classification based on Textual Information”, In Proc. DAS, pp 204-209, 2016.
9. P. Shivakumara, N. V. Kumar, D. S. Guru and C. L Tan, “Separation of graphics (superimposed) and scene text in videos”, In Proc. DAS, pp 344-348, 2014.
10. S. Roy, P. Shivakumara, U. Pal, T. Lu and C. L. Tan, “New Tampered Features for Scene and Caption Text Classification in Video Frame”, In Proc. ICFHR, pp 36-41, 2016.
11. D. Bhardwaj, V. Pankajakshan, “Image Overlay Text Detection Based on JPEG Truncation Error Analysis”, IEEE Signal Process. Lett. pp 1027-1031, 2016.
12. S. Roy, P. Shivakumara, U. Pal, T. Lu and A. W. B. A. Wahab, “Temporal integration of word-wise caption and scene text identification”, In Proc. ICDAR, pp 350-355, 2017.
13. S. Roy, P. Shivakumara, N. Jain, V. Khare, A. Dutta, U. Pal and T. Lu, “Rough-fuzzy scene categorization for text detection and recognition in video”, Pattern Recognition, pp 64-82, 2018.
14. Z. Wang, P. Shivakumara, T. Lu, M. Basavanna, U. Pal and M. Blumenstein, “Fourier-residual for printer identification”, In Proc. ICDAR, pp 1114-1119, 2017.
15. G. Liang, P. Shivakumara, T. Lu and C. L. Tan, “Multi-Spectral Fusion Based Approach for Arbitrarily-Oriented Scene Text Detection in Video Image”, IEEE Trans. Image Processing, pp 4488-4500, 2015.
16. X. Xu, Y. Wang and S. Chen, “Medical image fusion using discrete fractional wavelet transform”, Biomedical Signal Processing and Control, 27, pp 103-111, 2016.