New Fusion Based Enhancement for Text Detection in Night Video Footage

Chao Zhang[1], Palaiahnakote Shivakumara[2], Minglong Xue[1], Liping Zhu[3], Tong Lu[1(✉)], and Umapada Pal[4]

[1] National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China
zhangchao_nju@126.com, xueml@smail.nju.edu.cn, lutong@nju.edu.cn

[2] Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
shiva@um.edu.my

[3] School of Information Management, Nanjing University, Nanjing, China
chemzlp@163.com

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

Abstract. Text Detection in night video footage is hard due to low contrast and low resolution caused by distance variations between camera and ground under poor light. In this paper, we propose a new fusion based enhancement method for text detection especially in night video footage. The proposed method integrates the merits of color space and frequency based enhanced methods for sharpening low contrast details. Specifically, for each enhanced image, the proposed method derives weighted mean for the pixels values to widen the gap between high contrast (texts) and low contrast (background) pixels. The weighed means are further modified as dynamic weights with respect to enhanced images. These weights are convolved with pixel values of respective enhanced images to produce fused images. The proposed fusion based enhancement method is tested on images collected from night video footage to demonstrate the effectiveness of the method. For the output of each enhancement method including the proposed method, text detection rates are computed to show that the proposed enhancement method outperforms the existing enhancement methods.

Keywords: Night video · Enhancement · Color space · Frequency domain Image fusion · Text detection

1 Introduction

As noticed literature on text detection and recognition in natural scene images and video, researchers are developing new methods to tackle several challenges such as low contrast, uneven illumination effect, multiple scripts or orientations, complex background, and font or font size variations [1, 2]. This shows that text detection and

R. Hong et al. (Eds.): PCM 2018, LNCS 11166, pp. 46–56, 2018.
https://doi.org/10.1007/978-3-030-00764-5_5

recognition is important for real time emerging applications, such as surveillance and navigation apart from image understanding for retrieval [3, 6]. But it is noticed from literature that none of the methods addressed the issue of text detection in night video footage, where texts suffer from poor quality due to poor light conditions including other challenges mentioned above. One such illustration can be seen in Fig. 1, where (a) is an input image captured at night, in which texts are even invisible from our naked eyes, (b) gives the result of the existing enhancement method [7] which explores color space, (c) shows the result of one more state-of-art existing enhancement method [8], which explores frequency domain, and (d) gives the result of the proposed method. Note that to test the effectiveness of text detection performances for enhancement methods, we run the latest method [6] that uses powerful deep learning for text detection in natural scene images. It can be seen that the text detection method fails to detect the texts in the input image, and misses the texts from the results of both the existing enhancement method-1 and method-2. However, the same method [6] detects texts properly for the result of the proposed method. This shows that text detection performance is poor for night images, which can be improved for enhanced images. This is also true that understanding night video footage is essential for surveillance and monitoring applications especially for the purpose of forensic investigations in crime cases. Therefore, we focus on image enhancement for improving text detection performance for night video images in this work.

Fig. 1. Text detection performances prior to enhancement and after enhancement, where (a) is the input image, (b) shows the enhanced image from color space, (c) gives the enhanced image from frequency domain, and (d) shows the enhanced result from the proposed method.

There are methods for general image enhancement and text image enhancement through fusion concept. For example, Sharma et al. [9] proposed contrast enhancement using pixel based image fusion in wavelet domain. The method explores the combination of low and high pass filters. Lee et al. [10] proposed an edge enhancement algorithm for low dose X ray fluoroscopic imaging. This method uses the same filters as mentioned in above in gradient domain. Jiang et al. [7] proposed night video enhancement using improved dark channel priors. The method uses inversion pixel operation in color domain. Rui et al. [8] proposed a medical X-ray image enhancement method based on TV-Homomorphic filter. It has the ability to balance bright and fine pixels in images. Maurya et al. [11] proposed a social spider optimized image fusion approach for contrast enhancement and brightness preservation. The method produces one high sharp image and one more with high peak signals for each input image. Then

it fuses the two images to enhance fine details. From the above reviews, it is noticed that none of the methods considers images with texts for enhancement. In other words, the objective of the above methods is to enhance the content in general images.

Recently, we can see a few methods for text image enhancement. Pei et al. [12] proposed multi-orientation scene text detection with multi-information fusion. The method explores convolutional neural networks for detecting low contrast texts in natural scene images. The focus of the method is to enhance text information in natural scene images but not texts in night images. Roy et al. [13] proposed a fractional Poisson enhancement model for text detection and recognition in video frames. However, the target of the work is to enhance images affected by Laplacian operation but not those captured at night. Overall, none of the methods addresses the issue of text enhancement in night images for improving the performance of text detection.

Hence, in this work, we focus on the enhancement of images captured under poor light to improve the performance of text detection. Inspired by the method [11], where two different domains for enhancing low contrast images are used, we exploit the same idea for generating enhanced images using color space and frequency information. When there are texts in images, one can expect high contrast information, and meanwhile the values of such text pixels often have almost the uniform color values. Based on this observation, we propose new criteria for weight derivation for each pixel in enhanced images. Additionally, motivated by the method [14] where fusion is introduced for medical image enhancement, we propose to explore the same fusion concept for combining two generated enhanced images based on weight information. Since our weights are derived from fusion operation based on text properties, the proposed method well enhances text information by suppressing background information. This is our contribution, which is different from the existing methods.

2 The Proposed Method

This work considers images containing texts captured under poor light condition as the input for enhancement. As noticed from the literature, color space and frequency domain are the main concepts for image enhancement. This is because color is sensitive to human perception, while frequency coefficients are sensitive to tiny changes at pixel level. To take the advantage of both the domains, we propose to use color space for generating one enhanced image, and frequency domain for generating another enhanced image for each input image. Then to integrate both the enhanced images, we propose a new fusion method, which derives weights based on text properties to produce the final fused image, resulting in text enhancement. The block diagram of the proposed method can be seen in Fig. 2.

2.1 Enhanced Images Using Color Space and Frequency Domain

As mentioned in the previous section, we consider the idea presented in [7] for obtaining enhanced images using color space. The method treats night video images as foggy or haze removal. As a result, the method proposes a degradation model, which requires estimating the global atmospheric light and medium transmission using an

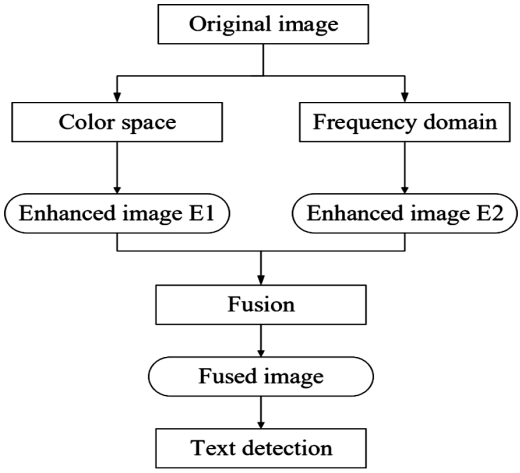

Fig. 2. The block diagram of the propose method.

improving dark channel prior. The institution to estimate the global atmospheric light is that the pixels that correspond to atmospheric light should be located in a large smooth area and during a period of time, the atmospheric light has no changes. In the same way, the main basis for estimating medium transmission using an improving dark channel prior is that the pixels on the same object should have the same or similar depth values. By knowing the transmission map and the global atmospheric light, the method can obtain the inversion of the enhanced image. The steps can be seen in Fig. 3. The effect of the enhancement method can be seen in Fig. 4(a), where we can see image details are enhanced compared to the input image in Fig. 1(a). However, since the method considers image enhancement as degradation, the method alone is insufficient to enhance text details in night video footage.

We thus propose another concept to enhance image details based on TV-Homomorphic filter [8]. It has the ability to reduce low frequency and increase high frequency information simultaneously, which results in reducing illumination changes and sharpening edge pixels. It works based on the incident and reflected light model. For the purpose of filtering, the method uses Fourier transform as the filter, which gives a filtered image for each input image. The total variation model has been used in homomorphic filter as a new transfer function. The total variation model is widely used for image restoration and noise removal. Since homomorphic filter uses the total variation model, it has the ability to adjust brightness and details of enhancement. The effect of the homomorphic filter is shown in Fig. 4(b), where one can see the image is brighter than the result shown in Fig. 4(a). It is also observed from Fig. 4(b) that the method enhances image details but not only text pixels. Overall, we can conclude that method-1 focuses on enhancing foreground information, while method-2 focuses on enhancing background information. Therefore, text pixels that have low contrast and high contrast are enhanced in separate images. To combine both low contrast and high contrast text pixels, we propose a new fusion method in the next section.

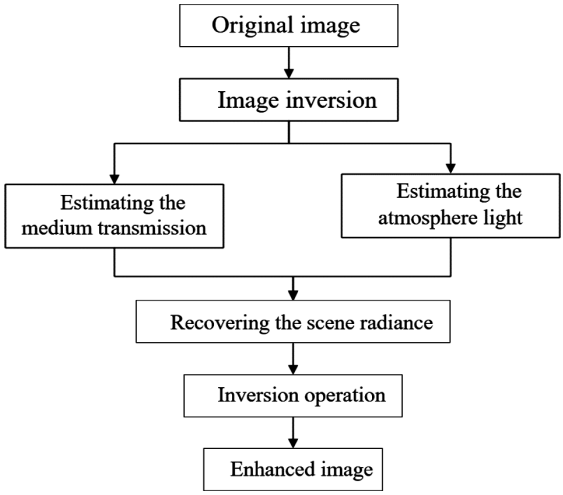

Fig. 3. The method for obtaining enhanced image-1 using color space.

(a) (b) (c)

Fig. 4. Fusing two enhanced images for generating a better enhanced image, where (a) is the enhanced image from color space, (b) is the enhanced image from frequency domain and (c) is the enhanced image from the proposed method.

2.2 Fusion Criteria for Text Enhancement

As mentioned in the Proposed Method Section, for each pixel in the two enhanced images given by the methods presented in the previous section, the proposed method computes mean as defined in Eq. (1). It is true that when there is an edge that represents text pixels, there will be high contrast values compared to its background. Therefore, the mean of such pixels gives high values. If there are no text pixels, the mean gives low values because of low contrast background information. To extract such difference, the proposed method multiplies the mean values with the current pixel values as defined in Eq. (2), which we call weights for each pixel in the two enhanced images. This operation increases text pixel values and decreases non-text values simultaneously. However, sometimes, arbitrary contrast variations in an image cause problem for identifying the gap between text and non-text pixels. To alleviate this problem, we recalculate weights based on the fact that an image can contain three type of values,

namely, high values which represent edges (text), low values which represent background, and middle values which usually represent text pixels affected by arbitrary contrast variations. Therefore, we recalculate weights as if a weight is greater than 0.6 in the values (text pixels) of the range, 0 to 1 must be multiplied by 0.8. If the weight is less than 0.6 and greater than 0.2 (affected values), the values are multiplied by 0.5 in the range of 0 and 1 values. The weight recalculation is defined in Eq. (3). The main objective of recalculating weights is to classify the pixels which represent texts as white and the pixels which represent non-texts as black. The values are determined empirically. The recalculated weights for each pixel in the enhanced images are then used for obtaining the fused image as defined in Eq. (4). The effect of the proposed fusion can be seen in Fig. 4(c), where one can see text pixels are enhanced compared to the images in Fig. 4(a) and (b). This is the advantage of the proposed work. The above steps are formulated mathematically as follows.

We define the grayscale image of the enhanced image from color space as $f^1(x,y)$ and enhanced image from frequency domain as $f^2(x,y)$. For two enhanced images, we calculate the local mean value of the $(i,j)^{th}$ pixel over an $n \times n$ window as $m^1_{(i,j)}$, $m^2_{(i,j)}$ using the Eq. (1):

$$m^k_{(i,j)} = \frac{1}{n*n}\sum_{x=0}^{n-1}\sum_{y=0}^{n-1} f^k(x,y) \tag{1}$$

where k denotes the index of the enhanced images and $k \in \{1,2\}$, while n denotes the size of the window and is set as 7 according to experiments. We compute the corresponding weight using Eq. (2):

$$w^k_{(i,j)} = \frac{m^k_{(i,j)}}{\sum_{l=1}^{l=p} m^l_{(i,j)}} \tag{2}$$

where $w^k_{(i,j)}$ denotes the weight at position (i,j) of the image, p is the number of the enhanced images and is set as 2, $k \in \{1,2\}$. Then we apply the following Eq. (3) to change weight values:

$$w^k_{(i,j)} = \begin{cases} w^k_{(i,j)} \times \mu_1 & t_1 \le w^k_{(i,j)} < 1 \\ w^k_{(i,j)} \times \mu_2 & t_2 \le w^k_{(i,j)} < t_1 \\ w^k_{(i,j)} & 0 \le w^k_{(i,j)} < t_2 \end{cases} \tag{3}$$

where $k \in \{1,2\}$, $w^k_{(i,j)}$ is the weight at position (i,j) of the enhanced image. $\mu_1 = 0.8$, $\mu_2 = 0.5$, $t_1 = 0.6$, $t_2 = 0.2$ and these values are determined empirically.

$$F(i,j) = \sum_{q=1}^{n} w^q_{(i,j)} \times f^q(i,j) \tag{4}$$

where F denotes the fused image, $f(i,j)$ is the gray pixel value at position (i,j) in the enhanced image, while n denotes the number of enhance images, which is set as 2.

3 Experimental Results

As there is no standard dataset available for evaluating the proposed method on night video footage, we collect our own data by capturing images at night, which includes 500 images under different poor light conditions. The dataset comprises images of Chinese and English captured by mobile phones in dark scenes. Besides, it also includes images of book covers, daily necessities, and boxes containing texts. To test the effectiveness of the proposed method, we collect low contrast images from standard ICDAR 2015 video [15] and YVT video, which provide 500 images. In total, we consider 1000 images for experimentation and evaluation of the proposed and existing methods.

To show the superiority to existing methods, we implement three state of the art existing enhancement methods, namely, Jiang et al.'s method [7] which uses color space and a degradation model for enhancing general images, Rui et al.'s method [8] which uses a TV-Homomorphic filter for enhancing details in images, and Roy et al.'s method [13] which is developed for enhancing text information in images affected by Laplacian operation. The reason to consider the above three methods is that Jiang et al.'s method is the state-of-the-art method for enhancing details in night video images as the proposed work, Rui et al.'s method is the state-of-the-art method that enhances image details in low contrast images, while Roy et al.'s method [13] is the state-of-the-art method that focusses on enhancement of text pixels affected by video noises as the proposed work.

To validate the result of enhancement given by the proposed and existing methods, we implement two well-known text detection methods to run on input images and the result of enhancement images, that is, Zhou et al.'s method [5] which proposes an efficient and accurate scene text detector based on deep learning and a large number of features, and Shi et al.'s method [4] which detects oriented texts in natural scene images based on deep learning and linking segments. The reason to consider these two methods is that the former one is good for low contrast and low resolution images, while the latter is good for images affected by blur and uneven illuminations. In addition, both the methods used deep learning tools for achieving their results. The results are obtained for prior to enhancement and after enhancement to show that the proposed enhancement method helps in improving text detection performance compared to prior to enhancement.

To measure the performance of the proposed method, we use standard measures, namely, Recall (R), Precision (P) and F-Measure (F) as defined in Eqs. (5)–(7), where True Positive (TP) means the number of items labeled correctly and belong to the positive class, True Negative (TN) means the number of items labeled correctly and belong to the negative class, False Positive (FP) and False Negative (FN) means the number of items labeled incorrectly in positive class and negative class, respectively.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{F-Measure} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7}$$

Qualitative results of the proposed and existing enhancement methods are shown in Fig. 5, where we can see that text detection methods detect texts properly for the output of the proposed enhancement method compared to the results of the other three existing enhancement methods. It is also observed from Fig. 5 that proposed fusion is better than the existing methods in terms of fine details of texts. This shows that the proposed enhancement is better than the existing enhancement methods. The reason of the existing methods for poor results is that the existing methods are developed with specific objectives but not text enhancement in night images. In addition, the way the proposed method integrates the advantage of enhancement results is something new and contributes for enhancing text pixels in night images.

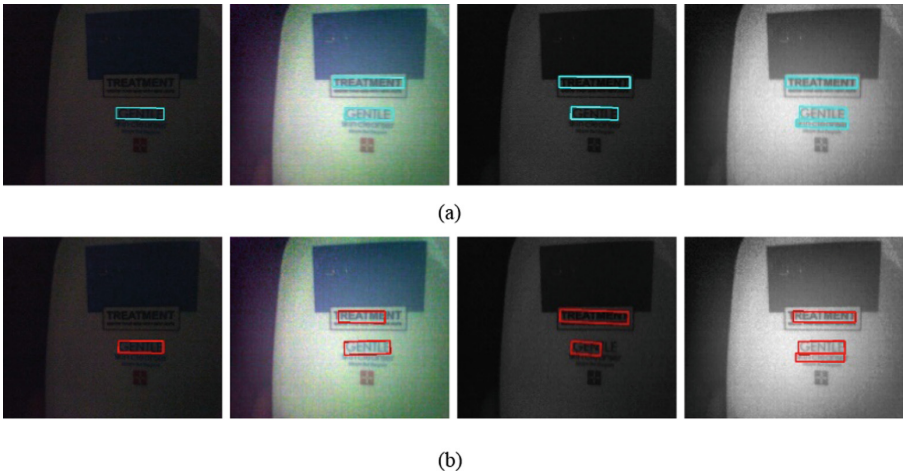

Fig. 5. Text detection performances for the enhanced results of proposed and existing methods where (a) is text detection results of EAST [5] and (b) is text detection results of SegLink [4]. The enhancement images in (a) and (b) are obtained by color space method, frequency domain method, Roy et al. [13] and the proposed method, respectively.

Quantitative results of the text detection methods for the output of the proposed and existing enhancement methods on our dataset are reported in Table 1. Table 1 shows that the text detection performance for input images is lower than that of enhanced images. When we compare the text detection performance on the output of the proposed method with the text detection performance on input images, the text detection

performance for the proposed enhanced images is improved significantly especially on F-measure. Similarly, the text detection results for the output of the proposed and existing enhancement methods on low contrast dataset are reported in Table 2, where the same conclusions can be drawn as Table 1. It is observed from Tables 1 and 2 that when we compare the text detection performances of the existing enhancement methods with the text detection performance of the proposed method, text detection performance of proposed enhancement is better than those of the existing enhancement methods. This shows that the proposed enhancement is better than the existing enhancement, and hence we can conclude that text detection performance improves significantly for the output of the proposed enhancement method.

Table 1. Text detection performance for the output of proposed and existing enhancement methods on our dataset.

Methods	SegLink [4]			EAST [5]		
	P	R	F	P	R	F
Input (prior to enhancement)	43.6	16.0	23.4	73.4	18.6	29.7
Enhancement-1 [7]	41.5	21.2	28.1	69.8	38.2	49.4
Enhancement-2 [8]	43.2	33.0	37.4	69.6	54.8	61.3
Roy's et al. [13]	41.4	31.4	35.7	70.8	50.7	59.1
Proposed method	**46.5**	**36.6**	**41.0**	**73.7**	**57.1**	**64.3**

Sometimes, for the images captured under full dark as shown in Fig. 6, the proposed enhancement method does not work well due to the limitation of automatic weight calculation. As a result, text detection method [6] does not detect texts properly for both input and enhanced images. This is valid because our naked eyes fail to notice texts in input images. Though the proposed method enhances text details compared to input images, it is insufficient to improve text detection performance. This shows that there is a scope for improvement in future. To overcome this issue, it is necessary to consider context information for text detection in such full dark images.

Table 2. Text detection performance for the output of the proposed and existing enhancement methods on low contrast dataset collected from standard video dataset.

Methods	SegLink [4]			EAST [5]		
	P	R	F	P	R	F
Input (prior to enhancement)	70.3	71.9	71.1	74.5	75.4	74.9
Enhancement-1 [7]	69.4	**76.5**	72.8	77.2	74.2	75.7
Enhancement-2 [8]	74.5	68.4	71.3	78.0	72.7	75.3
Roy's et al. [13]	69.5	73.8	71.6	73.2	**77.0**	75.0
Proposed method	**79.1**	71.2	**74.9**	**80.1**	74.5	**77.2**

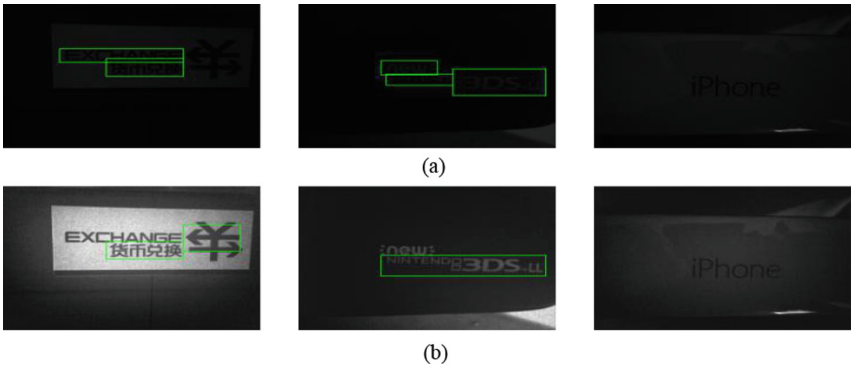

Fig. 6. Limitation of the proposed method. Text detection results are given by the CTPN [6]. (a) Gives input images of different situations, and (b) shows fused images.

4 Conclusion and Future Work

In this work, we have proposed a new method for enhancing texts in night video footage to improve text detection performance. The proposed method generates enhanced images based on color and frequency domains respectively to take the advantage of respective domains. We propose a new fusion criterion for integrating the advantages of color domain and frequency domain, which results in text enhancement in night images. Experimental results on the proposed and existing enhancement methods show that the proposed method is better than the existing enhancement methods. Additionally, text detection performance of after enhancement is better than prior to enhancement. However, when an image contains non-uniform blur or variations in degree of illumination effect, the performance is still poor. This would be our near future work.

Acknowledgment. The work described in this paper was supported by the Natural Science Foundation of China under Grant No. 61672273 and No. 61272218, the Science Foundation for Distinguished Young Scholars of Jiangsu under Grant No. BK20160021, and Scientific Foundation of State Grid Corporation of China (Research on Ice-wind Disaster Feature Recognition and Prediction by Few-shot Machine Learning in Transmission Lines).

References

1. Ye, Q., Doermann, D.: Text detection and recognition in imagery: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **37**(7), 1480–1500 (2015)
2. Yin, X.C., Zuo, Z.Y., Tian, S., Liu, C.L.: Text detection, tracking and recognition in video: a comprehensive survey. IEEE Trans. Image Process. **25**(6), 2752–2773 (2016)
3. Tian, S., Yin, X.C., Su, Y., Hao, H.W.: A unified framework for tracking based text detection and recognition from web videos. IEEE Trans. Pattern Anal. Mach. Intell. **40**(3), 542–554 (2018)

4. Shi, B., Bai, X., Belongie, S.: Detecting oriented text in natural images by linking segments. In: Proceedings CVPR, vol. 3 (2017)
5. Zhou, X., et al.: East: an efficient and accurate scene text detector. arXiv preprint arXiv:1704.03155 (2017)
6. Tian, Z., Huang, W., He, T., He, P., Qiao, Y.: Detecting text in natural image with connectionist text proposal network. In: ECCV, pp. 56–72 (2016)
7. Jiang, X., Yao, H., Zhang, S., Lu, X., Zeng, W.: Night video enhancement using improved dark channel prior. In: ICIP, pp. 553–557. IEEE (2013)
8. Rui, W., Guoyu, W.: Medical X-ray image enhancement method based on tvhomomorphic filter. In: 2017 2nd International Conference on Image, Vision and Computing (ICIVC), pp. 315–318. IEEE (2017)
9. Sharma, S., Zou, J.J., Fang, G.: Contrast enhancement using pixel based image fusion in wavelet domain. In: 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), pp. 285–290. IEEE (2016)
10. Lee, M.S., Park, C.H., Kang, M.G.: Edge enhancement algorithm for low-dose X-ray fluoroscopic imaging. Comput. Methods Programs Biomed. **152**, 45–52 (2017)
11. Maurya, L., Mahapatra, P.K., Kumar, A.: A social spider optimized image fusion approach for contrast enhancement and brightness preservation. Appl. Soft Comput. **52**, 575–592 (2017)
12. Pei, W.Y., Yang, C., Kau, L.J., Yin, X.C.: Multi-orientation scene text detection with multi-information fusion. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 657–662. IEEE (2016)
13. Roy, S., Shivakumara, P., Jalab, H.A., Ibrahim, R.W., Pal, U., Lu, T.: Fractional poisson enhancement model for text detection and recognition in video frames. Pattern Recogn. **52**, 433–447 (2016)
14. Xu, X., Wang, Y., Chen, S.: Medical image fusion using discrete fractional wavelet transform. Biomed. Signal Process. Control **27**, 103–111 (2016)
15. Karatzas, D., et al.: ICDAR 2015 competition on robust reading. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1156–1160. IEEE (2015)