Chinese Journal of Electronics
Vol.30, No.6, Nov. 2021

Video Stabilization via Prediction with Time-Series Network and Image Inpainting with Pyramid Fusion

CHENG Keyang[1,4,5], LI Shichao[1], RONG Lan[1], WANG Wenshan[2], SHI Wenxi[3] and ZHAN Yongzhao[1]

(1. *School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China*)

(2. *National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data, Beijing 100846, China*)

(3. *Xinjiang Lianhaichuangzhi Information Technology Co., Ltd., Urumqi 830011, China*)

(4. *Jiangsu Province Big Data Ubiquitous Perception and Intelligent Agricultural Application Engineering Research Center, Zhenjiang 212013, China*)

(5. *Cyber Space Security Academy of Jiangsu University, Zhenjiang 212013, China*)

Abstract — Due to the poor filling effect of the video image defect commonly used in the video stabilization field, the video is seemed still unstable after the image stabilization process, which seriously affects the visual effect. To solve this problem, we improve a video stabilization method based on time-series network prediction and pyramid fusion restoration is proposed to optimize the visual effect after image stabilization. The flow of the proposed method is as follows: First, it is adaptive to determine whether the defect of the corresponding frame at the current time needs padding inpainting. Then, for the frame that needs to be inpainting, the frames generated before the current moment are sent to the model combining the convolutional neural networks and the gate recurrent unit to predict the part to be filled. Next the current defect image and the complete image to be filled are brought into the Laplacian pyramid reconstruction, and the improved weighted optimal suture is introduced for splicing during the fusion. Finally, the video frame is cut after reconstruction. The method is tested on a data set composed of videos commonly used in the field of video stabilization. The experimental results show that the average peak signal to noise ratio of the method is 2 to 5dB higher than that of the comparison algorithm, and the average structural similarity index is improved by about 2% to 7% compared with the contrast algorithm.

Key words — Video stabilization, Video inpainting, Time series network, Multi-resolution fusion, Optimal seam.

I. Introduction

With the widespread use of video images in daily life, electronic image stabilization technology with high flexibility, low cost, and easy maintenance has became a research hotspot. After image stabilization, blank areas of varying sizes will be left on each frame of the image, so that the video that has been initially stabilized will still appear due to incomplete image content "instability" is even worse. Matsushita *et al.*[1] used the method of filling missing frames to improve the focus of the video. Ryu *et al.*[2] used a 2D affine model downsampling bilinear interpolation method to fill and inpainting the image after image stabilization. Yoo *et al.*[3] proposed an image stabilization solution used a mosaic method and motion inpainting method to comprehensively inpainting the image. Fan *et al.*[4] proposed to compensate and fill the current image through adaptive neighboring frame compensation and inpainting methods to solve the problem of error accumulation. However, the algorithm will still perform poorly in the case of severe shaking and foreground occlusion. Patwardhan *et al.*[5] improved the Criminisi image inpainting method and promoted it for video inpainting. Hsu *et al.*[6] used Markov random fields to complete the foreground and background extraction and inpainting. Qiao[7] converted the dynamic background sequence to static based on

Manuscript Received June 15, 2020; Accepted Nov. 30, 2020. This work is supported by the National Natural Science Foundation of China (No.61972183) and the Director Foundation Project of National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC).

 DOI:10.1049/cje.2021.08.006

global motion estimation and motion compensation, and used the accumulation of background information to complete video inpainting. Newson *et al.*[8] layered the spatiotemporal pyramid of the video, maximize the prediction of pixels, and reconstructed the video pyramid to video image inpainting. Luo *et al.*[9] constructed a Gaussian mixture model, and adapted the background reconstruction model to fill and inpainting video, their methods have inspired the construction of the video image stabilization model in this article.

Yu *et al.*[10] introduced a numerical solution method instead of directly using the convolution kernel's inverse approximate deconvolution kernel method. Other researchers[11−13] effectively reduced image edge blurring by extracting video temporal and spatial features in a convolutional neural network and combining adaptive motion compensation. Gate recurrent unit (GRU) is a recurrent neural network model proposed by Cho *et al.*[14] in 2014, which can overcome the timing problems such as poor long-term dependence of RNN processing. The pyramid image fusion model is an image fusion algorithm proposed by Burt *et al.*[15] and Mao *et al.*[16] enriched the background details of the edges of the image by adaptively weighting the high-frequency coefficients and averaging the low-frequency coefficients in the Laplacian pyramid fusion algorithm. Other researchers[17,18] used methods such as PCA image fusion improvement and wavelet transform fusion to improve the edge part inpainting status. The optimal seam algorithm was first proposed by Duplaquet[19]. Gu *et al.*[20] solved the problem of how the stitching lines are delineated by using the difference image weighted optimal seam average partition image fusion method. Qu *et al.*[21] used camera calibration and optimal stitching. In many other research works[22−24], improved methods such as directional gradient histogram, dynamic programming, and weighted average energy equation to eliminate problems.

We are aiming to solve the problems of inconsistent timing, poor filling image quality, obvious filling boundaries, ghost images, and excessive edge information loss in the above-mentioned video image stabilization technologies commonly are used in the field of image stabilization. We propose a video image stabilization method based on temporal network prediction and pyramid fusion inpainting, in order to improve the inpainting effect of defect images in video image stabilization. The main contributions of this article are: 1) Establish a temporal network model to predict the complete image of the current frame, and improve the quality of the filled image part; 2) We propose a video frame fusion filling scheme to optimize the splicing of the filled part and avoid problems such as ghosting; 3) We propose an optimization strategy to solve the problem of excessive edge information cropping and to improve inpainting efficiency.

II. Method

As mentioned above, there are many problems in current video inpainting technologies like inconsistent timing, poor fill image quality, filling limit is too obvious and ghosting, and excessive edge information loss, *etc.* In order to solve these problems, this paper proposes a video image stabilization method based on temporal neural network and pyramid fusion. The structure flowchart of the method model is shown in Fig.1.

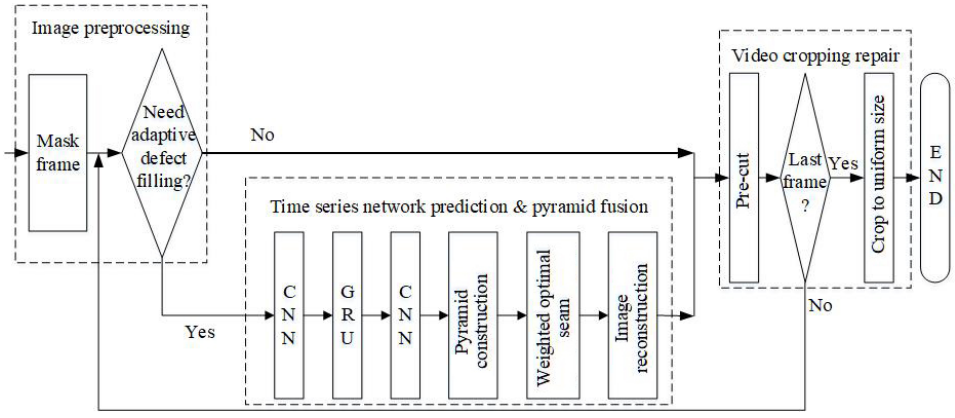

Fig. 1. Flowchart of proposed video stabilization image inpainting method

In the proposed model, we use a CNN and a GRU to construct a prediction network for the complete image of the current frame. We take into account the structural characteristics and timing of the video, and the predicted image frame quality to be filled is close to the original video image frame, which is more suitable for video image filling needs. In pyramid fusion part, we use Gaussian Laplacian pyramid fusion to ensure that the reconstructed frame texture and structural information are reasonable. At the same time, the weighted optimal

seam is introduced in the fusion process to eliminate ghosting, object cutting and other problems to the greatest extent. The video frame inpainting and crop optimization module are composed of three parts: mask frame setting, adaptive defect filling judgment and video frame pre-cropping. This module can reduce the time-consuming and improve the efficiency of the algorithm on the premise of guaranteeing the inpainting effect.

1. Time-series network prediction and pyramid fusion

The convolution pooling part in Fig.2 is mainly composed of a streamlined CNN structure. The purpose of its setting is to extract the characteristics of the target frame well and to simplify the calculations. After a lot of experiments, the convolution pooling part is set to three layers, and the formula can be expressed as:

$$F^1 = \sigma((W_{c(1)}) \otimes D_L + b_1) \tag{1}$$

$$F^i = \sigma((W_{c(i)}) \otimes F^{i-1} + b_i) \tag{2}$$

where F^1 is the first layer output of the convolutional layer, F^i is the output of the i-th convolutional layer, and D_L is the input image of the input layer. $W_{c(1)}$ is the weight of the corresponding i-th convolutional layer. The convolution weights are calculated from a filter bank with number n and size $f \times f$.

The values of n and f need to be adjusted according to the input video quality and the actual situation. In order to take into account the computational efficiency and the effect of generating images, after a large number of experiments, we chose a 3×3 convolution kernel to form a filter bank. $\otimes$ represents convolution and b_i represents the offset to the corresponding layer. σ represents the activation function of the parametric rectified linear unit PReLU. The GRU calculation unit module needs to bring the parameter matrix obtained in the previous section into a time-series neural network to calculate and output a predicted parameter matrix. Its structure is shown in Fig.3. The calculation formula for the status of each door and unit in the GRU is:

$$z_t = \sigma\left(U^z x_t + W^z h_{t-1}\right) \tag{3}$$

$$r_t = \sigma\left(U^r x_t + W^r h_{t-1}\right) \tag{4}$$

$$\hat{h}_t = \tanh\left(U^h x_t + W^h\left(h_{t-1} \times r_t\right)\right) \tag{5}$$

$$h_t = (1 - z_t) \times \hat{h}_t + z_t \times h_{t-1} \tag{6}$$

where z_t represents the update gate, r_t represents the reset gate, $\hat{h}$ represents the candidate activation value set of the current hidden node, h_t represents the output activation value of the current hidden node, σ represents the sigmoid function, x_t represents the input GRU parameters, U and W represents the corresponding weight parameter matrix. After obtaining the current final output state calculated from the current candidate hidden state and the previous final output state by updating the gate weight, it can be brought into the next section.

The deconvolution part implements image reconstruction of the features processed by the time-series model, that is, up-sampling and combining the feature images. The process can be expressed as:

$$F = \sigma(W_d \odot F^i + B) \tag{7}$$

where F is the final output of the deconvolution, W_d is the weight function of the deconvolution layer, $\odot$ is the deconvolution, F^i is the output of the previous part, B is the bias term, and the step size is adaptively adjusted according to the sampling factor.

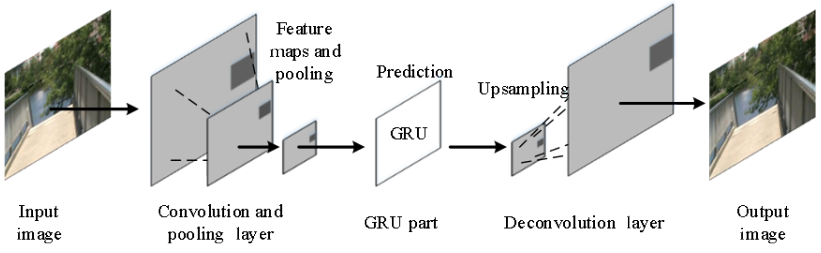

Fig. 2. Construction of frame prediction part

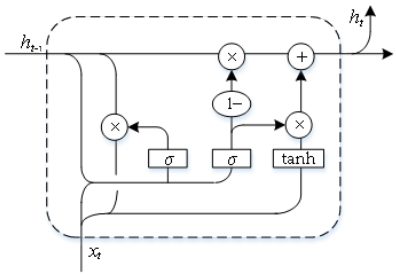

Fig. 3. Structure of GRU

The current frame prediction model of the proposed time-series network requires a certain pre-training process, assuming that the given training data set is $\{X_t^k, Y_t^k\}N_k^N = 1$ where X_t^k represents the image of the "black border" video after motion compensation at time t, Y_t^k represents the image of the complete video without black borders at time t, and N represents the number of training samples. The mean square error Mean square error (MSE) is used as a cost function to continuously adjust the network function $\Theta = \{W, b, B\}$ of each layer. The calculation is shown in formula:

$$L(\Theta) = \frac{1}{N}\sum\nolimits_{k=1}^{N} \|X_t^k - Y_t^k\|^2 \tag{8}$$

The network weight update process uses the stochastic gradient descent method as shown in formula:

$$\Delta_{j+1} = M \times \Delta_j + \eta \times \frac{\partial L}{\partial \theta_j^l}, \theta_{j+1}^l = \theta_j^l + \Delta_{j+1} \quad (9)$$

where j is the number of iterations, M is a constant generally 0.9, η is the learning rate, l is the number of layers, θ_j^l is the weight at the j th iteration of the l th layer, and $\frac{\alpha L}{\alpha \theta_j^l}$ is the partial derivative of the corresponding weight in the cost function.

The pyramid fusion algorithm proposed in our model mainly includes three steps, Gaussian pyramid decomposition, construction of Laplacian pyramid, and image reconstruction. First, Gaussian pyramid decomposition is performed on the stabilized current frame image and the predicted image. Input an image as the first layer in the Gaussian pyramid, then the Gaussian pyramid image is calculated as shown in Eq.(10):

$$G_l(i,j) = \sum_{m=-2}^{2} \sum_{n=-2}^{2} w(m,n) G_{l-1}(2i+m, 2j+n) \quad (10)$$

where $1 \le l \le l, 0 \le i \le R_j, 0 \le j \le C_l$ are the width and height of the l-th layer image in the Gaussian pyramid, and $w(m,n)$ represents a two-dimensional discrete Gaussian convolution kernel function. Suppose that the l-th layer image in the Laplacian pyramid is LP_l, the top-level image is LP_N, G is the corresponding Gaussian pyramid layer image, and G^* represents the upsampled image, then the construction of the Laplacian pyramid is as shown in Eq.(11) shows:

$$\begin{cases} LP_l = G_l - G_{l+1}^*, & 0 \le l < N \\ LP_N = G_N, & l = N \end{cases} \quad (11)$$

The third step includes image restoration and reconstruction. The process of restoration and reconstruction is the reverse of the construction process. Let us assume that the image of the first layer of the Laplacian pyramid is LP_l, the top image is LP_N, and G_l is the image of the first layer in the Gaussian pyramid. The formula for image reconstruction is as follows:

$$\begin{cases} G_l = LP_l + G_{l+1}^*, & 0 \le l < N \\ LP_N = G_N, & l = N \end{cases} \quad (12)$$

In the proposed model, we add image stitching with weighted optimal seam between Laplacian pyramid construction and image reconstruction. For each sub-image on the Laplace pyramid, weighted optimal seam is used. The stitching method is used for stitching, and then the image reconstruction is performed. The lowest layer image obtained is the final fusion effect image frame. The calculation criterion of the optimal seam line is shown in Eq.(13):

$$E(x,y) = E_c^2(x,y) + E_g(x,y) \quad (13)$$

In the formula, $E_c(x,y)$ represents the color difference intensity value of the image, and $E_g(x,y)$ represents the image structure difference intensity value. Bringing the image gradient calculation into $E_g(x,y)$, it can be rewritten as:

$$\begin{aligned} E_g(x,y) =& (S_x(x,y) - S_x(x+1,y))^2 + (S_y(x,y) \\ & - S_y(x,y+1))^2 \end{aligned} \quad (14)$$

where S_x and S_y represent a 3×3 Sobel operator template. Considering that the suture may have cut objects when stitching images, we choose to introduce Canny operator to optimize and improve the optimal seam delineation formula in the proposed model, which makes the delineation of the suture more reasonable. The introduced Canny operator is used to calculate the edge gradient information in all directions, and the direction of the suture is bypassed to avoid the situation where the object is cut. Eq.(13) is weighted to ensure that the edges of the object are complete. The weighted optimal seam rule is shown in Eq.(15):

$$E_t(x,y) = w_e[E_c^2(x,y) + E_g(x,y)] \quad (15)$$

where E_t represents the optimal seam line of the image at the current moment, and the weight of w_e is calculated from the 2×2 Gaussian convolution template and the gradient is used as the weight to be brought into the formula. The formula for calculating weights is shown in Eq.(16):

$$w_e = \frac{w_x + \varepsilon}{w_y + \varepsilon} = \frac{G_x(x,y) + G_x(x+1,y) + \varepsilon}{G_y(x,y) + G_y(x,y+1) + \varepsilon} \quad (16)$$

where G_x and G_y represent the gradient information in the x and y directions detected by the Canny operator, respectively. The constant coefficient ε is added in the calculation of the weight w_e to avoid the situation where the coefficient is zero. When weighting the edges, the points in the overlapping area are divided into two types, one is the edge node connected to the edge, and the other is the node in the common area. In order to ensure that the suture is delineated in the coincident area, the weight of the edge nodes is taken as ∞. The weight of the nodes in the common area can be calculated by Eq.(15). The delineation of the suture is shown in Fig.4.The largest area is delimited in the incomplete area of the frame, and the suture is determined.

2. Video frame repair cropping optimization strategy

The video frame repair and cropping optimization strategy proposed in this paper include image preprocessing and video cropping and repair. Fig.5 is a flow chart of the video cropping and repairing part. The specific process is as follows:

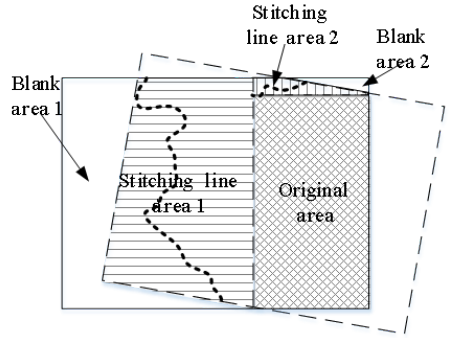

Fig. 4. Cutting position of optimal seam in video frames

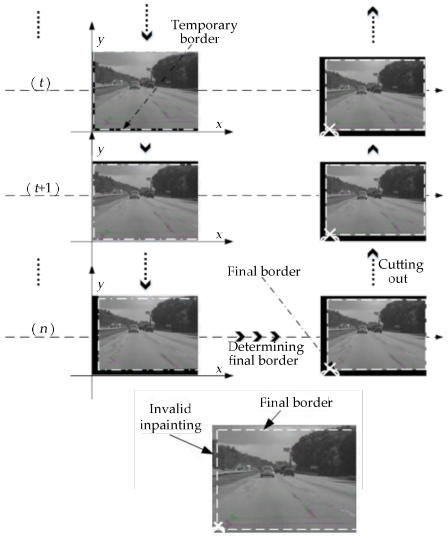

Fig. 5. Schematic diagram of the proposed adaptive optimized cutting process

1) Initially, all pixels of the mask frame are valid. Let E_0 be its initial state area. The first frame of the video is p_1, that is, $E_0 = Area(p_1)$. After obtaining the final effect frame at the end of the first frame image reconstruction process, $Area(p_1) = \Omega_1 + D_1$ is obtained, where Ω represents the effective area of the frame and D represents a small amount of blank area. The effective area E_1 of the mask frame is updated to $E_1 = Area(p_1) - D_1 = \Omega_1$, and the points outside the E_1 area are set as invalid points and cannot be reversed.

2) After initially stabilizing the t-th frame, $Area(p_1) = \Omega_t + D'_t$ is obtained, where D'_t represents a temporary blank area in the current state. If Ω_t contains all the pixels in $E_t - 1$ that is, $\Omega_t \supseteq E_{t-1}$, then turn $E_t = E_t-$ and go to 4), otherwise continue to execute.

3) Before the t-th frame image reconstruction step is completed, a template frame is first used to perform certain content processing on the frame to be output. Because $\Omega_t \subset E_{t-1}$, the pixels outside the $E_{t-1} - \Omega_t$ part are set to blank, so as to generate the final image of the t-th frame.

4) Update E_t with: $E_t = Area(p_t) - (D_{t-1} + \triangle D) = Area(p_t) - D_t$ where $\triangle D$ represents the increment of the blank area, $\triangle D \geq 0$.

5) Repeat 2) to 4) until the last frame is reached to determine the area of E_n.

6) A rectangle E_R i-th the largest area is determined in the finally updated mask frame effective area E_n, and the size from the first cropping frame to the n-th frame is cropped with the size of the final cropping boundary.

III. Experimental Results and Analysis

The experimental environment is Intel (R) Pentium (R) CPU, 3.60GHz, 8GB memory, 500G hard disk, a host of Windows 10 operating system and 110G memory, 7T disk, tesla P100 graphics card 16G×2, Linux operating system One server, the experimental platform is PyCharm2017 + Python3.6 + OpenCV3.The self-built video image stabilization dataset[25,26] is used in the experiment, which contains video segments commonly used in the field of video image stabilization[27,28].

1. Ablation study

In order to verify the effectiveness of each part of the proposed method, the ablation study was used to prove[29]. Because the model structure of the existing method is not completely the same as that of the proposed method, and some of the methods proposed in this paper can not show the effect independently. Therefore, we chose to replace some of the modules in the repair model in the ablation experiment to form a new temporary model to compare and verify the effectiveness of the proposed method. In the experiment, we replaced the time-series network prediction and pyramid fusion repair with reference frame filling and multi-frame stitching methods that commonly used in video stabilization image repair for comparison.

The selection of video segment is to randomly select 00016.avi from the data set to show the video processing effect. In Fig.6, (a) represents the 504th frame in the original video, (b) represents the image that has not been repaired and filled with black borders after video stabilization, (c) represents the result image using the combination of reference frame filling and pyramid fusion, (d) represents the result of time-series network prediction and Speeded up robust features (SURF) image stitching, (e) represents the experimental effect of our algorithm.

From the highlighted part in the result pictures, we can see that the part of the filled image in Fig.6(c) is similar to the edge of the effective part in Fig.6(b), and the part of the filled image is blurred. This is because

the corresponding part in the reference frame image is relatively blurred and does not fully meet the needs of the image structure.

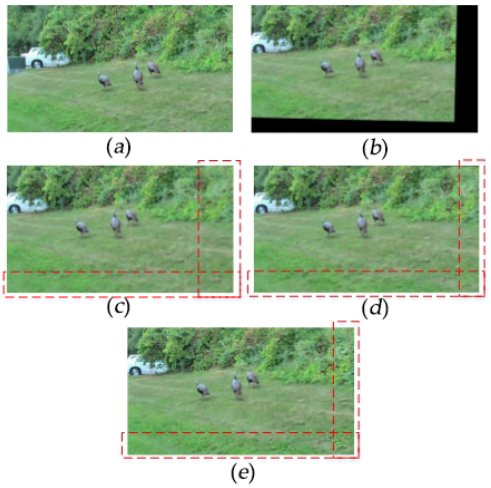

Fig. 6. Comparison of repair performance in ablation study experiment

The filled part in Fig.6(d) conforms to the structure but the boundary is obvious at the image splice. This is because the image quality predicted by the time-series network is high, the stitching adopts the SURF image stitching method to match the stitches according to the feature points but cannot deal with the boundary problem at the stitching point. Fig.6(e) performs well in image structure, image quality and splicing, and has the best visual effect. The PSNR and SSIM in Table 1 also verified the effectiveness of each part of our method. Compared with the methods combined in the ablation study, our method is more delicate in image restoration and better in detail processing.

Table 1. Comparison of the average PSNR and average SSIM in ablation study

Method	PSNR(dB)	SSIM
Original	22.58	0.7154
Without filling repair	21.72	0.7531
Reference frame filling + Pyramid fusion	25.09	0.8155
Time-series network prediction+SURF image stitching	26.32	0.8373
Ours	28.45	0.8554

2. Repair performance comparison

In order to verify the effectiveness of the proposed video image stabilization method, we select three videos to display restoration effects, there respectively are 7_6_input.mp4, 2WL_input.mp4 and 18AF.avi. The image size of 7_6_input.mp4 and 2WL_input.mp4 are 1280×720, images in 18AF.avi with size of 640×360. In Fig.7, Fig.8 and Fig.9, (a) is original image, (b) shows stabilized without filling repair (c) represents Fast matching method (FMM)[27] algorithm, (d) represents improved Criminisi algorithm[5], (e) represents mixed filling method[3], (f) represents video spatio-temporal pyramid layering method[8] and (g) represents our method. The structure of the video named 7_6_input.mp4 shows in Fig.7 is more complicated, and the defective part of the video picture is located at the boundary of various objects. Analysis of PSNR and SSIM data in Table 2 can show that our algorithm greatly improves the visual effect and is more obvious. However, in the video named 2WL_input.mp4 shows in Fig.8, the structure is relatively stable and the defective part is mostly located in the video structure such as roads and sky.

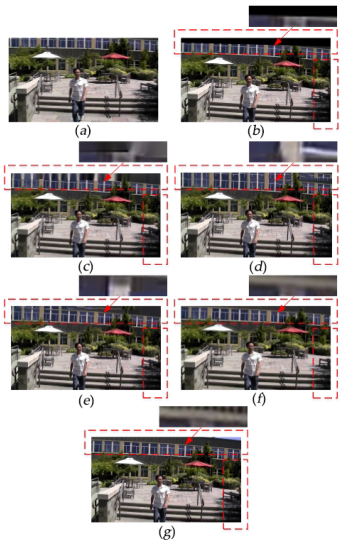

Fig. 7. Comparison of the repair of the 131st frame of video 7_6_input.mp4 under different algorithms

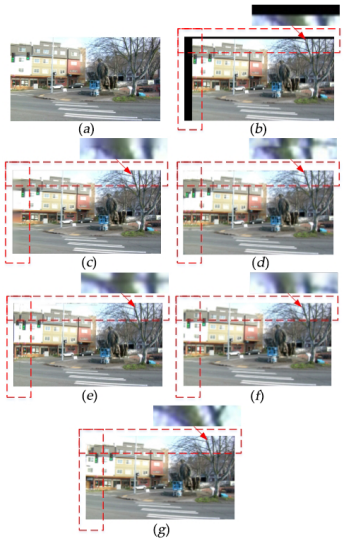

Fig. 8. Comparison of the repair of the 221st frame of video 2WL_input.mp4 under different algorithms

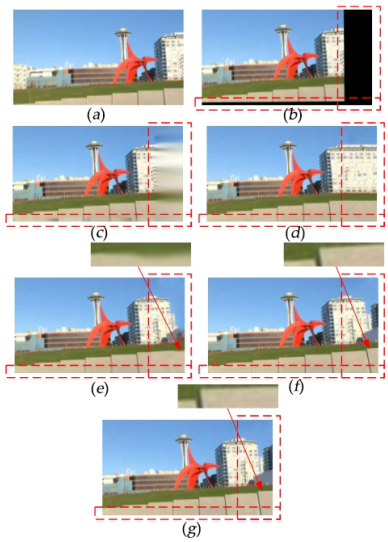

Fig. 9. Comparison of the repair of the 98th frame of video 18AF.avi under different algorithms

The improvement of the video effect of proposed method is equivalent to the comparison algorithm. The video named 18AF.avi shows in Fig.9 has a large motion range, violent shaking and a certain angle transformation. Through comprehensive analysis, the proposed algorithm can greatly improve the repair effect of video images when a small amount of time is added, and it can obtain better results when repairing defects requires a larger filling area or a more complex video structure. Table 3 shows the comparison of the average PSNR and average SSIM index of the five algorithms including the proposed algorithm on the 40-segment video data set we constructed. Experiments show that the average PSNR of this algorithm is improved by about 2–5dB and the average SSIM is increased by about 2%–7% compared with the comparison algorithm. In the case of little time-consuming increase, the proposed algorithm has obvious video repair effect. Especially for large video defects and complex video structures, the algorithm in this paper is significantly improved compared to the comparison algorithm, which can better improve the image quality in the video and facilitate the post-processing of the video.

Table 2. Comparison of the average PSNR, average SSIM, and average time per frame of five algorithms in three videos

		FMM	Improved criminisi algorithm	Mixed filling method	Video spatio-temporal pyramid layering method	Ours
7_6_input.mp4	PSNR(dB)	27.02	29.45	27.16	31.89	**33.26**
	SSIM	0.8092	0.8233	0.8366	0.8438	**0.8654**
	Time(ms)	223	**195**	204	543	663
2WL_input.mp4	PSNR(dB)	35.04	34.64	33.53	36.34	**37.89**
	SSIM	0.9165	0.9095	0.9111	0.9276	**0.9343**
	Time(ms)	364	**291**	353	791	1094
18AF.avi	PSNR(dB)	22.61	25.43	29.51	31.29	**33.52**
	SSIM	0.7835	0.8381	0.8530	0.8997	**0.9088**
	Time(ms)	**90**	93	92	176	205

Table 3. Comparison of the average PSNR and average SSIM of all videos of the five algorithms in the dataset

Methods	40 videos	
	PSNR(dB)	SSIM
FMM	29.76	0.8167
Improved criminisi algorithm	31.33	0.8470
Mixed filling method	30.67	0.8316
Video spatio-temporal pyramid layering method	33.24	0.8671
Ours	34.59	0.8893

Applicable scenarios include videos taken by mobile phone cameras, law enforcement recorders, drones and other devices, which are suitable for the needs of video stabilization algorithms for mobile shooting platforms[30–32].

IV. Conclusions

In this paper, we propose a video image stabilization method based on temporal neural network and pyramid fusion. Aiming at the problem of poor image quality and obvious boundaries in the video image stabilization repair part, a repair model composed of CNN, GRU, pyramid fusion and optimal joint weighting scheme is designed. It effectively improves the image quality of the filled part and removes the existing stitching borders, ghosting and other problems. Aiming at the problem of low efficiency and time-consuming single-frame repair algorithm, an optimized cropping and repairing strategy is proposed, which effectively reduces the time-consuming algorithm while ensuring the integrity of the video edge information. Experiments show that the performance indicators such as PSNR and SSIM of the algorithm proposed in this paper are significantly improved, and can obtain better image stabilization visual effects.

References

[1] Y. Matsushita, E. Ofek, W. Ge, *et al.*, "Full-frame video stabilization with motion inpainting", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.28, No.7, pp.1150–1163, 2006.

[2] Y. G. Ryu and M. J. Chung, "Robust online digital image stabilization based on point-feature trajectory withoutaccumulative global motion estimation ", *IEEE Signal Processing Letters*, Vol.19, No.4, pp.223–226, 2012.

[3] S. Yoo, A. K. Katsaggelos, G. Jo, *et al.*, "Video completion using block matching for video stabilization", *Proc. of The 18th IEEE International Symposium on Consumer Electronics*, JeJu Island, South Korea, pp.1–2, 2014.

[4] Fan Yeping, Guo Zheng and Zhang Rui, "Research on electronic image stabilization algorithm based on subsample gray-scale projection", *Industry and Mine Automation*, No.4, pp.22–27, 2017. (in Chinese)

[5] K. A. Patwardhan, G. Sapiro and M. Bertalmío, "Video inpainting under constrained camera motion", *IEEE Transactions on Image Processing*, Vol.16, No.2, pp.545–553, 2007.

[6] H. A. Hsu, C. K. Chiang and S. H. Lai, "Spatio-temporally consistent view synthesis from video-plus-depth data with global optimization", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.24, No.1, pp.74–84, 2014.

[7] Qiao Xiaotian, "3D Reconstruction and display of multi-view video", *Ph. D. Thesis*, University of Zhejiang, China, 2016.

(in Chinese)

[8] A. Newson, A. Almansa, M. Fradet, *et al.*, "Video inpainting of complex scenes", *SIAM Journal on Imaging Sciences*, Vol.7, No.4, pp.1993–2019, 2014.

[9] G. B. Luo, Y. S. Zhu, Z. T. Li, *et al.*, "A hole filling approach based on background reconstruction for view synthesis in 3D video", *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp.1781–1789, 2016.

[10] Yu Haibao, Shen Qi and Feng Guocan, "Introduce numerical solution to visualize convolutional neuron networks based on numerical solution", *Computer Science*, No.S1, pp.146–150, 2017. (in Chinese)

[11] Liu Cun, Li Yuanxiang, Zhou Yongjun, *et al.*, "Video image super-resolution reconstruction method based on convolutional neural network", *Application Research of Computers*, Vol.36, No.4, pp.1256–1260, 2019. (in Chinese)

[12] L. He, J. Tan, C. Xie, *et al.*, "A novel two-step approach for the super-resolution reconstruction of video sequences", *Proc. of International Conference on Digital Home*, Guangzhou, China, pp.85–90, 2014.

[13] Li Sumei, Lei Guoqing and Fan Ru, "Depth map super-resolution reconstruction based on convolutional neural networks", *Acta Optica Sinica*, Vol.37, No.12, pp.124–132, 2017. (in Chinese)

[14] K. Cho, van B. Merrienboer, D. Bahdanau, *et al.*, "On the properties of neural machine translation: Encoder-decoder approaches", *Proc. of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation*, Doha, Qatar, pp.103–111, 2014.

[15] P. J. Burt and E. H. Adelson, "A multiresolution spline with application to image mosaics", *ACM Transactions on Graphics*, Vol.2, No.4, pp.217–236, 1983.

[16] R. Mao, X. Fu, P. Niu, *et al.*, "Multi-directional Laplacian pyramid image fusion algorithm", *Proc. of International Conference on Mechanical, Control and Computer Engineering*, Huhhot Inner Mongolia, China, pp.568–572, 2018.

[17] S. K. Verma, M. Kaur and R. Kumar, "Hybrid image fusion algorithm using Laplacian pyramid and PCA method", *Proc. of the Second International Conference on Information and Communication Technology for Competitive Strategies*, The Papandayan Hotel, Bandung, Indonesia, pp.68–72, 2016.

[18] M. J. Li, Y. B. Dong and X. L. Wang, "Image fusion algorithm based on wavelet transform and Laplacian pyramid", *Advanced Materials Research*, Vol.860, No.3, pp.2846–2849, 2014.

[19] M. L. Duplaquet, "Building large image mosaics with invisible seam lines", *Proc. of Visual Information Processing VII*, Tokyo, Japan, pp.369–377, 1998.

[20] Gu Yu, Zhou Yang, Ren Gang, *et al.*, "Image stitching by combining optimal seam and multi-resolution fusion", *Journal of Image and Graphics*, Vol.22, No.6, pp.0842–0851, 2017. (in Chinese)

[21] Z. Qu, T. Wang, S. An, *et al.*, "Image seamless stitching and straightening based on the image block", *IET Image Processing*, Vol.12, No.8, pp.1361–1369, 2018.

[22] D. Lee and S. Lee, "Seamless image stitching by homography refinement and structure deformation using optimal seam pair detection", *Journal of Electronic Imaging*, Vol.26, No.6, pp.1–6, 2017.

[23] W. X. Yan, C. C. Liu and J. Hu, "Optimal seam line detection in laplacian pyramid domain for image stitching", *Journal of Computers*, Vol.29, No.1, pp.209–219, 2018.

[24] J. Xue, S. Chen, X. Cheng, *et al.*, "A new optimal seam method for seamless image stitching", *Proc. of Ninth International Conference on Digital Image Processing*, Hong Kong, China, pp.1–6, 2017.

[25] F. Liu, M. Gleicher, H. Jin, *et al.*, "Content-preserving warps for 3D video stabilization", *ACM Transactions on Graphics*, San Francisco, USA, pp.1–9, 2009.

[26] W. C. Hu, C. H. Chen, Y. J. Su, *et al.*, "Feature-based real-time video stabilization for vehicle video recorder system", *Multimedia Tools and Applications*, Vol.77, No.5, pp.5107–5127, 2018.

[27] Liu Guanglong, "Research on electronic image stabilization based on feature optical flow", *Ph.D. Thesis*, University of Harbin Institute of Technology, China, 2015. (in Chinese)

[28] M. Wang, G. Yang, J. Lin, *et al.*, "Deep Online Video Stabilization With Multi-Grid Warping Transformation Learning", *IEEE Transactions on Image Processing*, Vol.28, No.5, pp.2283–2292, 2019.

[29] BAI Dongdong, WANG Chaoqun, ZHANG Bo, *et al.*, "CNN feature boosted SeqSLAM for real-time loop closure detection", *Chinese Journal of Electronics*, Vol.27, No.3, pp.488–499, 2018.

[30] QU Hua, ZHANG Yanpeng, LIU Wei, *et al.*, "A robust fuzzy time series forecasting method based on multi-partition and outlier detection", *Chinese Journal of Electronics*, Vol.28, No.5, pp.899–905, 2019.

[31] WANG Haitao, HE Jie, ZHANG Xiaohong, *et al.*, "A short text classification method based on N-gram and CNN", *Chinese Journal of Electronics*, Vol.29, No.2, pp.248–254, 2020.

[32] G. Fei, W. Meng, W. Jun, *et al.*, "A novel separability objective function in CNN for feature extraction of SAR images", *Chinese Journal of Electronics*, Vol.28, No.2, pp.634–640, 2019.

CHENG Keyang (corresponding author) was born in 1982.He is professor and Member of CCF. His research interests include computer vision and machine learning. (Email: kycheng@ujs.edu.cn)

LI Shichao was born in 1993. M.S. candidate at Jiangsu University. His research interests include computer vision and video image processing. (Email: tayloraustin@foxmail.com)

RONG Lan was born in 1996. She is an M.S. candidate at Jiangsu University. Her research interests include computer vision and machine learning. (Email: 1748675160@qq.com)

WANG Wenshan is an M.S. candidate at Columbia University. Her research interests include statistical analysis and machine learning. (Email: wenshanwang2468@163.com)

SHI Wenxi was born in 1987. She is a senior engineer. Her research interests include statistical analysis and machine learning. (Email: swxcetc@163.com)

ZHAN Yongzhao was born in 1962. He is a professor and member of CCF. His research interests include computer vision and machine learning. (Email: yzzhan@ujs.edu.cn)