



# Video Inpainting Using Advanced Homography-based Registration Method

B. Janardhana Rao<sup>1</sup> · Y. Chakrapani<sup>2</sup> · S. Srinivas Kumar<sup>3</sup>

Received: 2 March 2020 / Accepted: 1 July 2022 / Published online: 25 July 2022  
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

## Abstract

A new video inpainting technique for videos taken from free moving cameras is suggested in this research paper. The effective results of video inpainting can be achieved by maintaining spatiotemporal coherence while filling the holes in the target frames. This is possible only with the proper registration of source frames to the target frame. Image registration plays a vital role in the process of video inpainting to obtain effective results. An advanced homography-based image registration method is introduced, based on HALF-SIFT: high accurate localization feature for SIFT to extract feature points without localization error. The covariance matrix has been used to estimate the localization error. Further, new inlier selection method using CW MLESAC and refining is carried out for homography matrix with CW L-M. This iteration process can improve the accuracy of image registration. After registering frames to the target frame, the hole is inpainted by globally minimizing the energy cost function. The proposed video inpainting is applied to several complex video sequences. Experimental results are outperformed in visual quality when compared with the state-of-the-art methods. The performance metrics like peak signal-to-noise ratio and Structural Similarity Index are determined and compared with existing methods for different video sequences.

**Keywords** Video inpainting · Image registration · Homography · HALF-SIFT · Covariance matrix · PSNR · SSIM

## 1 Introduction

Video inpainting is a process of removing undesired objects in the video frames and filling them with sophisticated algorithms without any artifacts. It is also a technique used to reconstruct the damaged parts of the video sequence. The frames in the video sequence contain undesired objects are called target frames, the hole created after removing the object is called the target region. The frames which are having similar pixel information to fill the target region are called source frames. Video inpainting finds a lot more applications in video processing such as video restoration, video stabilization, film post-processing, etc. Image inpainting is the technique of restoring the damaged images and removing unwanted information in the image and filling the hole with appropriate data from the remaining region of the image.

### 1.1 Video Inpainting: Related Work

The total variation (TV) and diffusion-based methods [1, 2] are used for inpainting small regions and removal of text and scratches. Large region removal and inpainting are obtained with exemplar-based inpainting methods [3]. The basic exemplar-based image inpainting was proposed by Criminisi et.al [4], which is further modified as robust exemplar-based inpainting using region segmentation [5]. Video inpainting is obtained by extending the Criminisi method for images with taking care of spatiotemporal coherence between the frames of the video [6–9]. These methods achieve good results with compromise in camera movement. Granados et al. [10] introduced graph-cuts optimization to implement video inpainting. Homography-based image registration was used in this to align the input video frames to the target frame. The missing pixels in the video frames are filled with the information taken from registered frames. The cost function minimization is utilized to find the best pixels values to fill in the target region [11, 12]. This entire process takes more time even for low-resolution videos and long-duration sequences.

✉ B. Janardhana Rao  
janardhan.bitra@gmail.com

<sup>1</sup> CVR College of Engineering - Ibrahimpatan Campus, Hyderabad, India

<sup>2</sup> ACE Engineering College, Hyderabad, Telangana, India

<sup>3</sup> UCEK, JNTUK, Kakinada, Andhra Pradesh, India

Newson et al. [13] improved the Granados inpainting by using the patch match algorithm [14] to the spatiotemporal domain with pyramids of frames. In this method, approximate nearest neighbor (ANN) was efficiently calculated after aligning all the frames to the middle frame using affine transformation. This method produces poor results because of aligning all the frames to the middle frame, the side part regions which does not intersect with all the frames were not aligned properly.

Ebdelli et al. [15] achieved a well-reconstructed video by aligning more frames to the target frame using region-based homography transformation. For the proper alignment, it produces excellent results and adverse results with incorrect frame alignment. Huang et al. [16] introduced a video completion technique by including both optical flow and color information of pixels in the target region of the frames in a video. In this temporal consistency is maintained from both pixel-wise flow field and patch-based optimization. The reconstruction of the frames was obtained through the iterative computation of forward and backward flow fields. This increases the computational complexity of the algorithm.

Recently, the novel video inpainting technique is implemented with a hybridization of the cuckoo search algorithm and multi-verse optimization (CS-MVO) [17] for optimizing the patch matching and recurrent neural network (RNN) for categorizing the patch as smooth or structured. This method produced the optimal video inpainting results compared to available methods. There is another video inpainting technique proposed by using an enhanced priority computation method and optimal patch selected for inpainting the target region with grey wolf optimization (GWO) [18] and modified artificial bee colony algorithm [19]. This method outperformed the existing techniques of video inpainting in terms of metrics PSNR, SSIM and edge similarity.

From all the researchers, one can understand that the proper alignment of source frames to the target frame plays a vital role in video inpainting to maintain spatiotemporal coherence between the frames of the reconstructed video sequence. In this paper, an advanced homography-based registration method to attain an efficient alignment of frames to the target frame is proposed from the inspiration of work [20]. This registration technique is utilized for our proposed video inpainting.

The traditional homography-based methods were implemented in two steps: feature point extraction and estimation of feature points matching between the images with homography estimation. The feature points are extracted using scale-invariant feature transform (SIFT) [21] and speeded-up robust feature (SURF) [22]. Then, global correspondence is established with feature matching and estimating the transformation between the frames using a robust estimation method, random sample consensus (RANSAC) [23]. The

incorrect feature point correspondences due to illumination changes and viewpoint differences are called outliers, which leads to inaccurate homography estimation results. The robust estimation methods were required to remove the outliers obtained due to the feature points matching. Proper selection of inliers in the feature matching can avoid outliers. The advanced homography transformation method improves the quality of alignment of frames.

In this paper, the image registration method is carried out by extracting the feature points using the HALF-SIFT [24] method instead of the SIFT. The localization error occurred in SIFT due to the feature detector. The HALF-SIFT method extracts more accurate feature points and rectifies the localized error. The localization error is assumed to arise either from pixel intensity noise or the feature extraction method itself, which has zero mean and non-uniform distribution. This type of error is called an anisotropic and non-uniformly distributed localization error [25–28]. It is also called a random localization error [20]. This random localization error does not satisfy the assumptions of the RANSAC estimate. To estimate this type of error, the covariance matrices are used in the proposed video inpainting method.

Next, the selection of inliers and estimation of the homography matrix is done using covariance weighted maximum likelihood sample consensus (CW MLESAC). Finally, the homography matrix is refined with the covariance weighted Levenberg–Marquardt algorithm (CW L-M). In this video inpainting method, the last step is the target region filling. The target region filling is completed by minimizing the robust energy cost function globally using the expansion move algorithm [29–31]. To observe the effectiveness of the proposed method, the experiments are conducted on the densely annotated video segmentation (DAVIS) dataset [32].

The paper is organized as follows: The proposed video inpainting approaches are described in Sect. 2. Section 3 consists of experimental results from the proposed method and a comparison of metrics with an existing method. The concluding remarks are presented in Sect. 4.

## 2 Proposed Video Inpainting Method

The proposed method is implemented in two main steps; the first step is image registration, and the second step is target region filling. The target frame with a hole is inpainted using aligning the neighboring frames to the target frame. The hole is filled with the diffusion of pixels from aligned frames to the target region. The image registration is achieved with an advanced homography-based image registration method and the hole in the target frame called the target region is filled by globally minimizing the robust energy function. The flow of the proposed method is shown in Fig. 1.

## 2.1 Advanced Homography-based Registration Method

In the proposed image registration method, the feature points are extracted using the HALF-SIFT [24] method instead of the SIFT method. The HALF-SIFT method extracts

$$\hat{v} = \underset{v}{\operatorname{arg\,min}} \left\{ \sum_{i,j,p=-1,0,1} \|X(m_c \pm i, n_c \pm j, (2^{1/3})^p \alpha) - X_v(m_c \pm i, n_c \pm j, (2^{1/3})^p \alpha)\|_2^2 \right\} \tag{1}$$

more accurate feature points and rectifies the localized error obtained due to the SIFT feature detector. Later estimation of localization error was obtained due to pixel intensity noise with the covariance matrix. The feature matching between the frames is obtained with minimum Euclidean distance as a parameter. Next, the selection of inliers and estimation of the homography matrix with covariance weighted maximum likelihood sample consensus (CW MLESAC) is done, and finally, the homography matrix is refined with the covariance weighted Levenberg–Marquardt algorithm (CW L-M).

### 2.1.1 Localized Feature Points Extraction Using HALF-SIFT

The localization error of feature points in the SIFT method is due to the use of the parabolic interpolation method to estimate the coordinates of feature points. The interpolation accuracy is improved by using the HALF-SIFT method, which produces accurate localized feature points.

In order to extract the accurate localized feature point, HALF-SIFT used regression analysis for minimizing the distance between the sampling points near the feature points in the difference between the Gaussians pyramid and response model function. The output obtained from the difference of the Gaussian (DoG) filter by applying the Gaussian function as input is called the response model function. This response model function is described by the parameter vector  $v$ . The parameter vector  $v = (m_0, n_0, a, b, c, r)$  determines

the response model of the difference of Gaussians filter. Here,  $M_0 = (m_0, n_0)$  is the accurate position of the feature point. This parameter vector ( $v$ ) is optimized with the Levenberg–Marquardt algorithm. The optimized objective function [24] is taken as,

Difference of Gaussians (DoG) is a feature enhancement algorithm that involves the subtraction of one Gaussian blurred version of an original image from another, less blurred version of the original. The result is a set of images of a variety of sizes, each being a "difference of Gaussians." It is called as difference of the Gaussian pyramid, represented as  $X(\cdot); (m_c, n_c)$ . is the extracted local maximum int with SIFT, and  $X_v(\cdot)$ . is the response model function described by the parameter vector  $v$  and  $\alpha$ ., where  $i$ .,  $j$ . and  $p$  are the amplitudes of scale in SIFT.

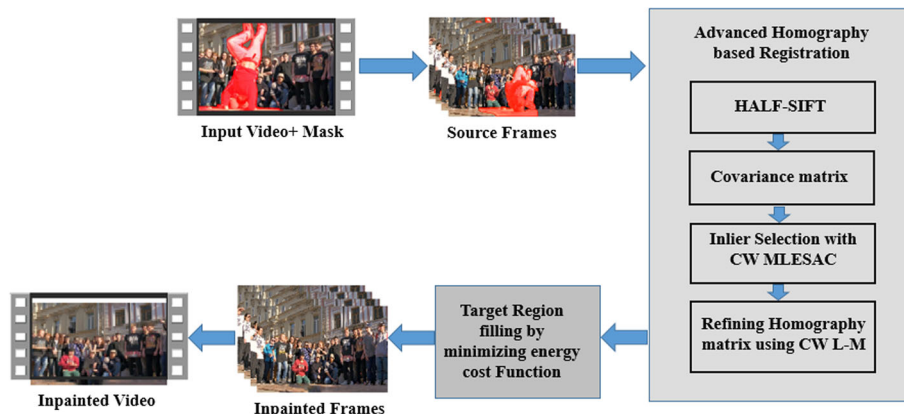
### 2.2 Estimation of Localization Error Using the Covariance Matrix

The localization error of the feature points depends on the distribution of the pixel intensity values near the feature points. Due to the inaccurate distribution of pixel intensity values, anisotropic and non-identical localization error occurs. The covariance matrix is used to indicate the anisotropic and non-identical error [25], which is represented as

$$\Lambda = \sigma \cdot G_r \cdot \begin{bmatrix} \mu & 0 \\ 0 & 1 - \mu \end{bmatrix} \cdot G_r^T \tag{2}$$

where  $G_r = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$  is the rotation matrix that represents the anti-clockwise rotation by an angle  $\theta$ .  $\sigma \in [0, \infty)$

Fig. 1 Flow of Proposed Video Inpainting Method



is the scale,  $\mu \in (0, 1)$  is the eccentricity, and  $\gamma \in [0, \pi)$  is the rotation angle of the covariance matrix  $\Lambda$ .

In this work, the random localization error of feature points is assumed as a bilateral Gaussian model and expressed in the covariance matrix, which is estimated as

$$\Lambda^{(\alpha_v)} = \left( \sum_{(k,l) \in N_v} u(k,l) \cdot \begin{bmatrix} X_{mm}(k,l,\alpha_v) & X_{mn}(k,l,\alpha_v) \\ X_{mn}(k,l,\alpha_v) & X_{nn}(k,l,\alpha_v) \end{bmatrix} \right)^{-1} \tag{3}$$

$$\Lambda = \Lambda^{(\alpha_v)} \cdot \left( \frac{\text{res}(\alpha_0)}{\text{res}(\alpha_v)} \right)^2 \tag{4}$$

where  $u(k, l)$  is the weighted coefficient of Gaussian distribution,  $X(\cdot, \alpha_v)$  is the particular layer of the scale-space pyramid,  $N_v$  is a small neighborhood near the feature point  $(k, l, \alpha_v)$  is the scale parameter of the present layer,  $\text{res}(\alpha_0)$  and  $\text{res}(\alpha_v)$  is the image resolution of the layer whose scale parameter is.

$\alpha_0$  and  $\alpha_v$  in a scale-space pyramid. The sub-indices  $m$  and  $n$  in Eq. (3) represent the local maxima point positions obtained from the SIFT method.

### 2.2.1 Selection of Inliers with CW MLESAC

The residuals are calculated in traditional feature point matching methods like LMedS and RANSAC using Eq. 5. For this calculation, the inliers correspondences vector is considered as  $C = \{(c_{1i}, c_{2i}) | i = 1, 2, 3, \dots, n\}$ , where  $c_{1i}, c_{2i}$  are pixel positions of feature points of two images.

$$r_i = \sqrt{\|e\|_2^2} = \sqrt{\|c_{2i} - \hat{c}_{2i}\|_2^2} \tag{5}$$

where  $e$  is the re-projected error vector and  $\hat{c}_{2i}$  is the re-projected pixel position of  $c_{2i}$  with homography transform  $H$ . The residuals calculated using the above formula do not include the anisotropic and non-identical properties of localization error, so this residual calculation leads to inaccurate homography estimation. This is rectified by selecting the inliers with the CW MLESAC method. This method is implemented by using normalized covariance weighted residuals (NCWR); this includes the properties of localization error of feature points.

The covariance matrix is decomposed as

$$\Lambda = U \cdot \begin{bmatrix} \varepsilon_1 & 0 \\ 0 & \varepsilon_2 \end{bmatrix} \cdot U^T \tag{6}$$

Here,  $\varepsilon_1, \varepsilon_2$  are eigenvalues of the covariance matrix, and  $U = [u_1, u_2]$  are eigenvectors.

In the proposed CW-MLESAC method the residual.

$r_i$  is replaced with the NCWR formula  $\bar{r}_i$  as,

$$\begin{aligned} \bar{r}_i &= \sqrt{(\bar{\varepsilon}_{\varepsilon_1})^2 + (\bar{\varepsilon}_{\varepsilon_2})^2} = \sqrt{\frac{2\varepsilon_1}{\varepsilon_1 + \varepsilon_2} (e_{u_1})^2 + \frac{2\varepsilon_2}{\varepsilon_1 + \varepsilon_2} (e_{u_2})^2} \\ &= \sqrt{\frac{2\varepsilon_1 (u_1^T \cdot (c_{2i} - \hat{c}_{2i}))^2}{\varepsilon_1 + \varepsilon_2} + \frac{2\varepsilon_2 (u_2^T \cdot (c_{2i} - \hat{c}_{2i}))^2}{\varepsilon_1 + \varepsilon_2}} \end{aligned} \tag{7}$$

Finally, the best inliers correspondences  $C_{inlier}$  and homography matrix  $H_{inlier}$  are selected using CW MLESAC.

### 2.2.2 Refining of Homography Matrix with CW L-M

Considering.

$C_{inlier} = \{(c_{1i}, c_{2i}) | i = 1, 2, 3, \dots, n\}$  as the inlier correspondences and pixel positions of images as.

$c_{1i}, c_{2i}$ . Then, the objective function ( $F$ ) taken in the existing L-M method is

$$F = \arg \min_H \sum_{k=1}^n \|e\|_2^2 = \arg \min_H \sum_{k=1}^n \|c_{2i} - H(c_{1i})\|_2^2 \tag{8}$$

here,  $e$  is the re-projected error vector,  $H(\cdot)$  is the homography transform.

This objective function did not include the anisotropic and non-identical properties of localization error hence the homography estimation was not accurate from this objective function. This problem was solved by using a new homography matrix refinement method called the CW L-M algorithm.

This uses the covariance weighted objective function ( $\bar{F}$ ), which is taken as

$$\bar{F} = \arg \min_H \sum_{k=1}^n \left[ \frac{(u_{1i}^T \cdot (c_{2i} - H(c_{1i})))^2}{\varepsilon_{1i}} + \frac{(u_{2i}^T \cdot (c_{2i} - H(c_{1i})))^2}{\varepsilon_{2i}} \right] \tag{9}$$

here,  $\varepsilon_{1i}, \varepsilon_{2i}$  are eigenvalues of the covariance matrix of the  $i_{th}$  feature point correspondence in inliers with  $C_{inlier}$ .  $U = [u_{1i} u_{2i}]$ , eigenvectors concerning eigenvalues.

From Eq. (9), the new objective function includes the anisotropic properties of the localization error by re-projected error vector axes are rotated toward the covariance matrix–vector direction. Also divided with the eigenvalues of the covariance matrix, it indicates different contributions of the different feature points. From all these, the localization error of the feature points becomes isotropic and identically distributed. Then, L-M method produces the optimal solution.

### 2.3 Target Region Filling

The hole ( $\Phi_t$ ) in the target frame ( $I_t$ ) is called the target region is filled with similar pixels from the group of

frames aligned to the target frame by maintaining spatiotemporal coherence. The missing regions in the target frame are inpainted by minimizing an energy function globally using the expansion move algorithm [29–31]. If we consider the  $M$  number of past and  $M$  number of future neighbor frames, then every pixel in the hole is inpainted with best-suited pixels from  $2M$  (past  $M$  + future  $M$ ) number of registered frames  $\hat{I}_i, i = 1 \dots 2M$ . Let  $S_v$  and  $S_w$  denote the category of matching pixel values near pixel  $v$  and  $w$ , respectively.  $S^*$  value is calculated by minimizing the energy cost function for all the pixel values of the hole in the target frame.

$$S^* = \arg \min_S \xi(S) \tag{10}$$

with

$$\xi(S) = \sum_{v \in \Phi_t} E_d(S(v)) + \gamma \sum_{(v,w) \in N(\Phi_t)} E_{sm}(S(v), S(w)) \tag{11}$$

Where,  $E_d(S(v))$  is called the data term, which indicates the stationary background surrounding the pixel  $v$ . The data term is divided into three terms

$$E_d(S(v)) = E_0(\hat{I}_{S(v)}) + E_1(\hat{I}_{S(v)}) + E_2(\hat{I}_{S(v)}) \tag{12}$$

where first-term  $E_0(\hat{I}_{S(v)})$  is the sum of the squared difference between the current target frame ( $I_t$ ) and the frames registered to the target frame ( $\hat{I}_{S(v)}$ ) taken as,  $E_0(\hat{I}_{S(v)}) = \log(1 + SSD(I_t, \hat{I}_{S(v)}))$ . This term is used to find the pixels in the best-aligned frames, which gives less alignment error. The logarithm is used to limit the dynamic range of this term.

Second term  $E_1(\hat{I}_{S(v)})$  is a constant term calculated as

$$E_1(\hat{I}_{S(v)}) = \frac{1}{2M|\Omega_{S(v)}|} \sum_{k=-M}^M \Omega_{S(v)} - \Omega_k^2, \text{ where } \Omega_{S(v)}, \text{ the patch is centered with the pixel } v \text{ in the registered frame } \hat{I}_{S(v)} \text{ and } |\Omega_{S(v)}| \text{ is the number of pixels in the patch and.}$$

$\Omega_k$  represents the patch in the registered frames of  $\hat{I}_{S^*(v)}$ . This term is responsible to maintain temporal consistency in the inpainting results.

In the third term,  $E_2(\hat{I}_{S(v)})$  represents the similarity of the pixels in the target frame to the pixels in the source frame centered at  $v$ , which is calculated with  $E_2(\hat{I}_{S(v)}) = I_s(v) - \hat{I}_{S(v)}(v)^2$ , the value of  $I_s(v)$  is calculated using spatial inpainting. This is the term responsible for spatiotemporal coherence in the video inpainting results.

The second term ( $E_{sm}(S(v), S(w))$ ) in the Eq. (11) is called the smoothness term calculated between each pair of

adjacent pixels in the hole of the target frame, which is taken as

$$E_{sm}(S(v), S(w)) = \hat{I}_{S(v)}(v) - \hat{I}_{S(w)}(v)^2 + \hat{I}_{S(v)}(w) - \hat{I}_{S(w)}(w)^2 \tag{13}$$

The smoothness term is used to maintain the spatial consistency by inpainting the pixels in the target region with similar pixel values in the neighboring registered frames.  $N(\Phi_t)$ , is the 4-neighbors of pixel  $v$ . The value of  $\gamma$  in Eq. (11) is chosen as 10 to maintain the balance between the data term and the smoothness term. Equation (10) is minimized by using expansion-move algorithm [23–31].

### 3 Experimental Results and Discussion

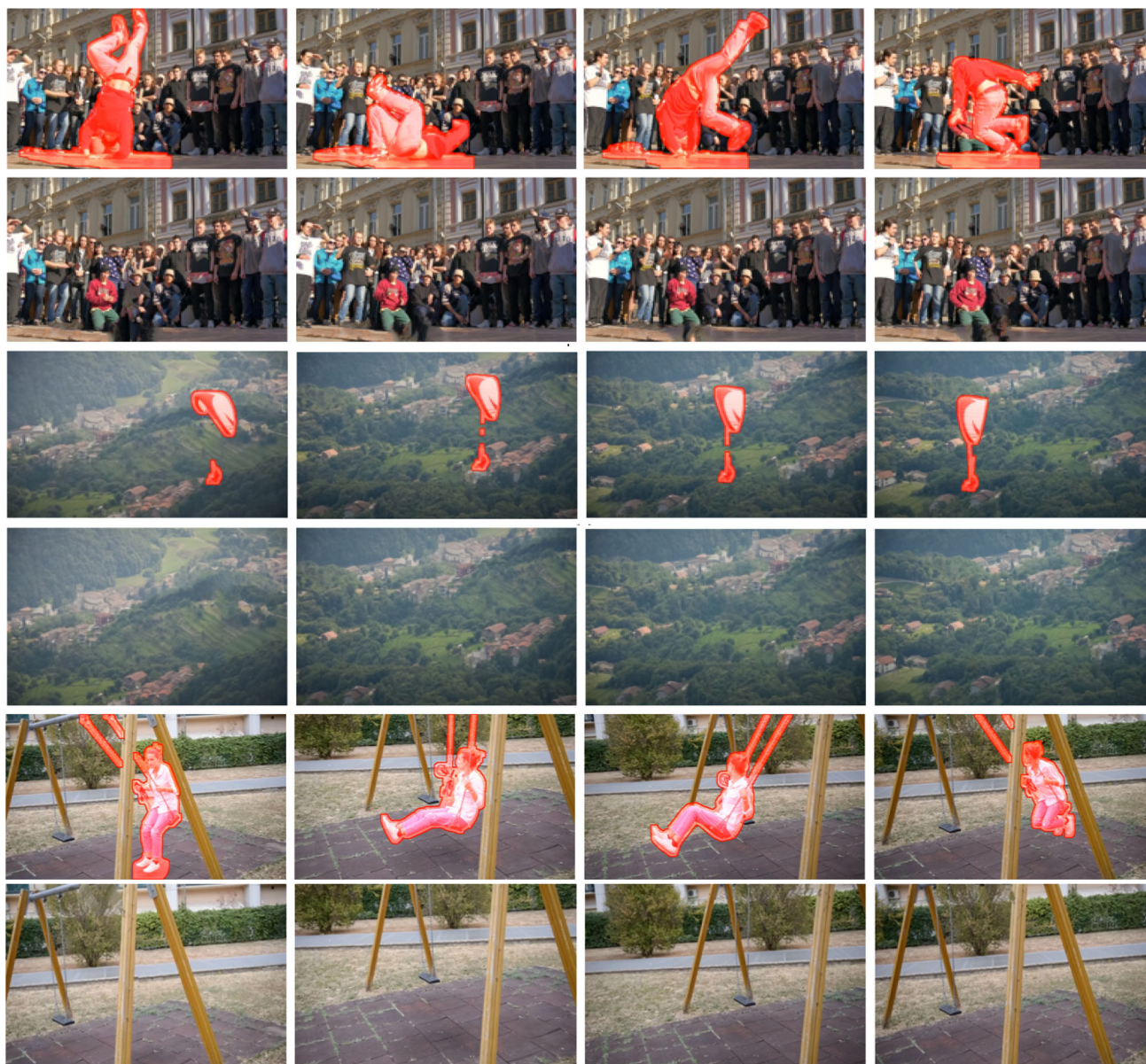
The proposed video inpainting algorithm is implemented in MATLAB. The process of inpainting for a video with around 100 frames takes at most 3 h with an Intel Core i3 processor of 8 GB RAM. The proposed algorithm tested with multiple complex video sets includes motion blur, camera movement, dynamic background and complex hole shapes. The results are compared with the existing video completion methods. The results obtained from the proposed method are visually feasible in comparison with existing methods of video inpainting.

In the proposed advanced homography-based video inpainting method, the experiments are carried out for the quantitative analysis in terms of PSNR and SSIM.

The PSNR and SSIM values are determined as follows:

- A separate data set of composite videos is created for 11 videos. These 11 videos are treated as ground truth videos.
- The composite video is created by adding one mask to the ground truth video, the mask is taken from the DAVIS data set.
- This created composite video is treated as an input video for inpainting.
- The proposed inpainting method is applied to the composite video to remove the added mask object.
- The resulted video is compared with the corresponding ground truth video to calculate the PSNR and SSIM values.
- Similarly, 11 composite videos inpainting is done and metrics are determined.
- In the same way, the existing video inpainting techniques are applied to composite videos to determine the PSNR and SSIM values.

The video inpainting results of a few video sequences are shown in Fig. 2; odd rows represent the mask corresponding to an object to be removed in the video frames and even rows



**Fig. 2** Object Removal; Odd row: Input Video frames (Videos 1, 2, 3), Even rows: Frames using the proposed inpainted method

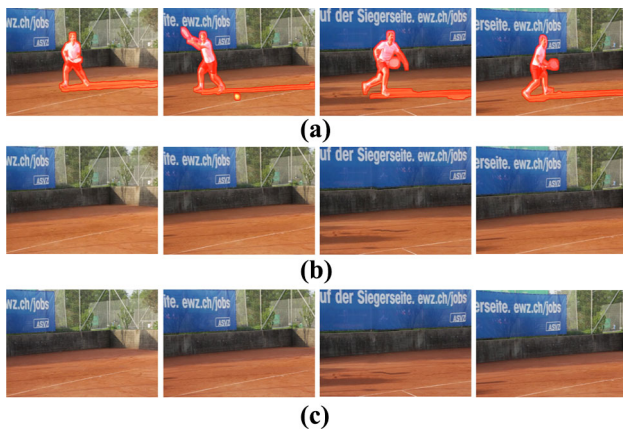
represent the resultant frames after removing the object using the proposed method.

### 3.1 Comparison with Existing Methods

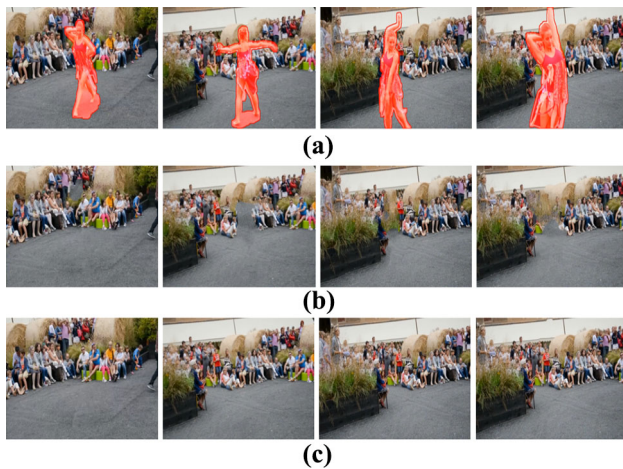
The video inpainting approach by Granados et al. [10] is quite similar to the proposed method in this paper. The homography-based image registration is used in [10] to align the input video frames to the target frame. The missing pixels in the video frames are filled with the information taken from registered frames. The cost function minimization is used to find the best pixel values to fill in the target region.

This entire process takes more time even for low-resolution videos and long-duration sequences.

In this work, different advanced and robust homography estimation as compared to [10] is proposed to align the input video frames to the target frame. Further, the inpainting quality is improved by introducing data term in the energy function to fill the target region. The author has not published the code to compare with his work. It is difficult to reproduce this work without making any errors. Hence, this work is not compared with the proposed work in terms of performance metrics. The patch-based video inpainting algorithm [13] used spatiotemporal sampling to fill the target region. This work [13] is compared with the proposed work. The



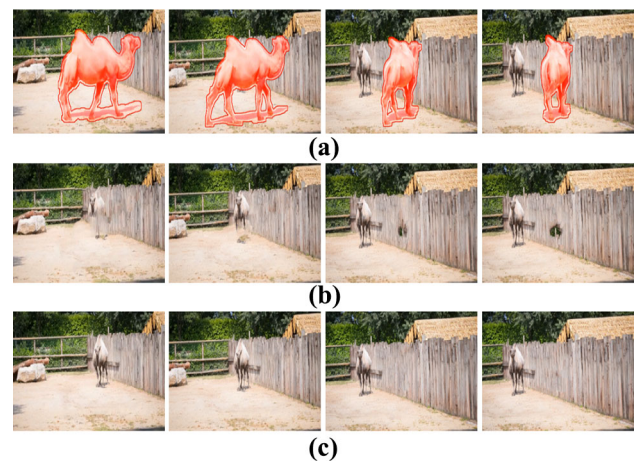
**Fig. 3** Comparison of Results with [13] **a** Input Video Frames with an object to be removed (Video 4) **b** Results from [13] **c** Results from the proposed method



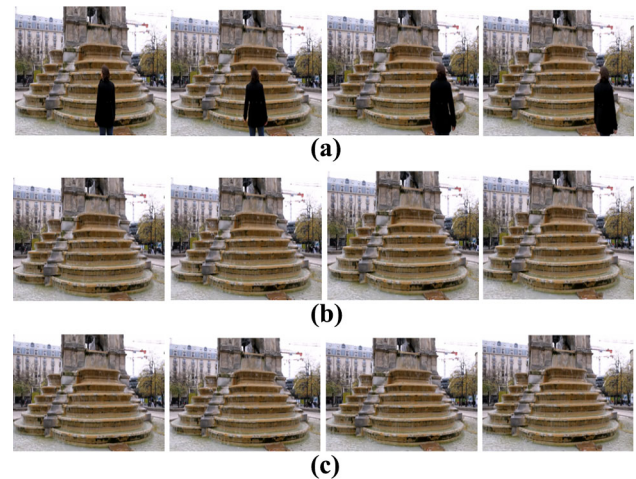
**Fig. 4** Comparison of results with [13]: **a** Input Video Frames with an object to be removed (Video 5); **b** Results from [13]; **c** Results from the proposed method

frames of the video sequence with an object to be removed and the inpainted result are shown in Figs. 3, 4 and 5. From these results, we can identify that some artifacts occurred due to the incorrect flow of pixels from the known region to the hole in the target frame and incorrect alignment of frames. The proposed method achieves more reliable results compared to [13] due to the proper alignment of source frames to the target frame.

Compared with the method [16], it is computationally complex due to the forward and backward flow fields in the image sequence. As mentioned by the author in their paper, this method produces some noticeable artifacts in the videos contain a dynamic camera, foreground and background. By comparing, our method reconstructed the frames without any artifacts in all above cases due to the inclusion of high accurate feature points calculation to match the pixels in the



**Fig. 5** Comparison of results with [13]; **a** Input Video Frames with an object to be removed (Video 6); **b** Results from [13]; **c** Results from the proposed method

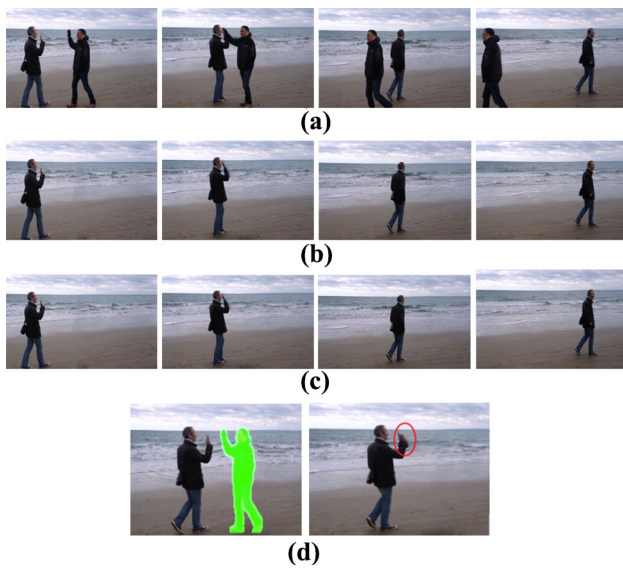


**Fig. 6** Comparison of Results with [16]; **a** Input Video Frames with an object to be removed (Video 7); **b** Results from [16]; **c** Results from the proposed method

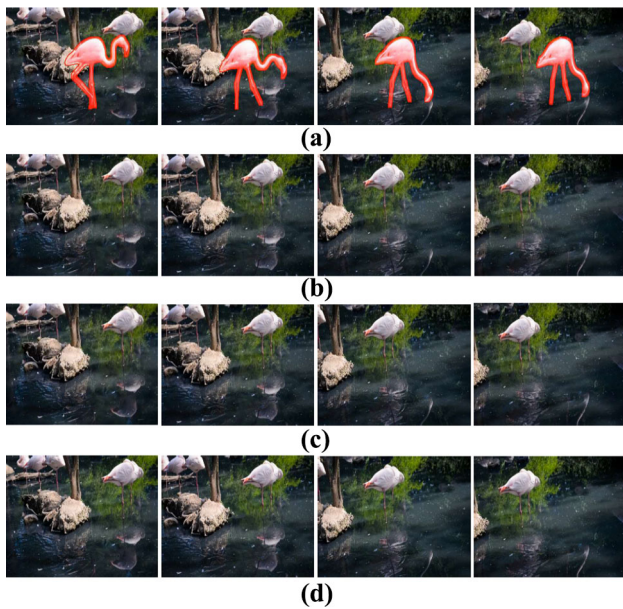
frames. The comparison of inpainted results of the proposed method is shown in Figs. 6 and 7.

The results obtained from the proposed method are compared with [13] and [16] as shown in Figs. 8, 9 and 10.

The PSNR and SSIM values are computed for the existing methods in the literature by implementing the code taken from the author's page. The proposed method is applied to the videos which are used in the existing methods. The experimental results such as PSNR and SSIM for different videos are given in Table 1. In this table, the state-of-art methods of inpainting proposed by Newton [13] and Huang [16] are compared with the proposed novel inpainting method with respect to PSNR and SSIM values for 11 input videos. The experiment is carried out on 11 standard input videos from the DAVIS dataset and the average of PSNR and SSIM is eval-

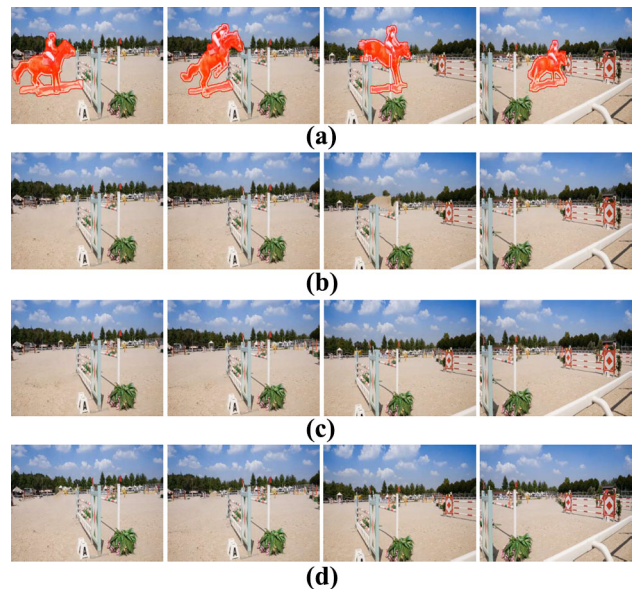


**Fig. 7** Comparison of results with [16]; **a** Input Video Frames with an object to be removed (Video 8); **b** Results from [16]; **c** Results from the proposed method; **d** Mask of the object to be removed for complex frame and artifact in the inpainted frame

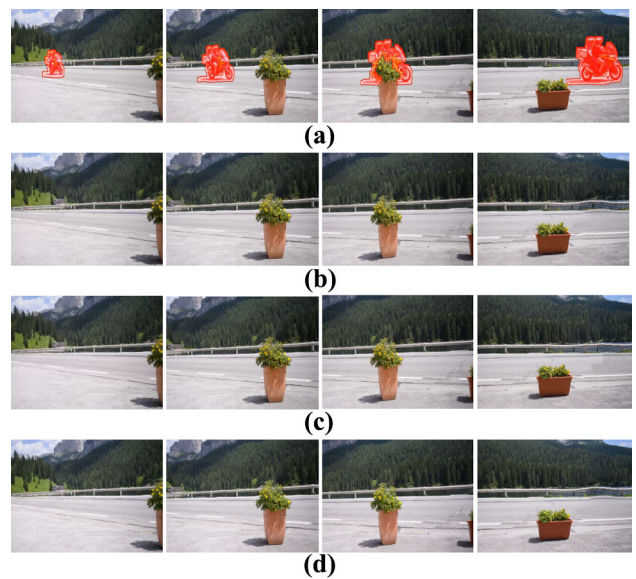


**Fig. 8** Comparison of results with [13] and [16]; **a** Input video frames with object to be removed (video 9); **b** Results from [13]; **c** Results from [16]; **d** Results from proposed method

uated to compare the performance of the proposed method with existing methods. The average PSNR value for the proposed method is 21.075, whereas for Newton [13] is 19.921 and for Huang [16] is 20.556. From these results, we can say that the proposed algorithm has enhanced PSNR by 6% and 3% when compared with [13] and [16] works, respectively. The average SSIM of the proposed method for 11 videos



**Fig. 9** Comparison of results with [13] and [16]; **a** Input video frames with object to be removed (video 10); **b** Results from [13]; **c** Results from [16]; **d** Results from proposed method



**Fig. 10** Comparison of results with [13] and [16]; **a** Input video frames with object to be removed (video 11); **b** Results from [13]; **c** Results from [16]; **d** Results from proposed method

is 0.932, which is an improved value when comparing the existing method values of 0.9 [13] and 0.907 [16].

Figure 11 represents the variation of PSNR and SSIM values for different videos from the DAVIS dataset. Video 2 has the highest PSNR value among all the videos, whereas video 8 has the lowest value. Hence the average PSNR is evaluated and compared with the state of artworks. The PSNR values are distinguished between the proposed method and state-of-art methods shown in Fig. 12. It is observed that

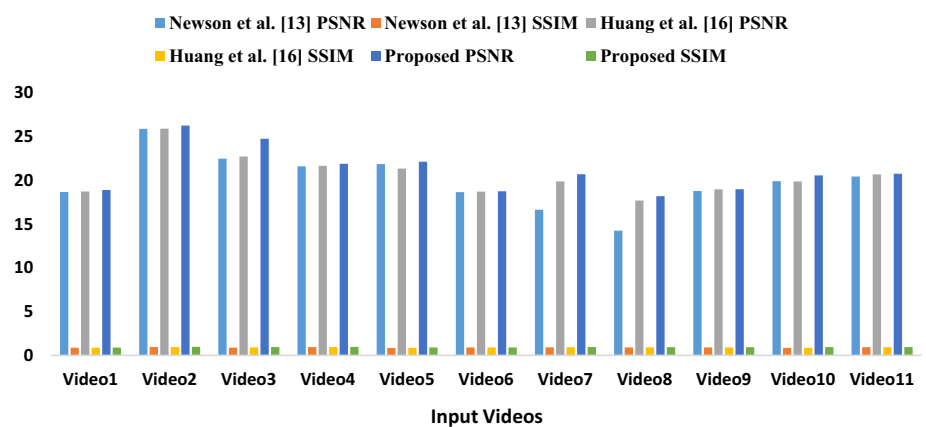


**Table 1** Comparison of PSNR and SSIM values of various videos with available methods in the literature [13, 16] and the proposed method

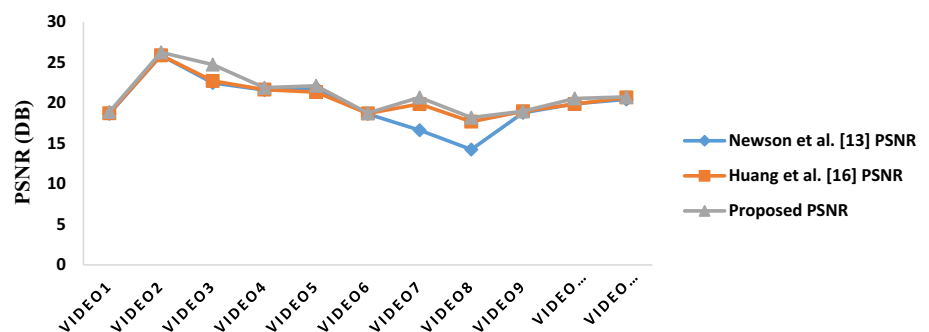
Input Video	Newson et al. [13]		Huang et al. [16]		Proposed	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Video 1	18.6653	0.87765	18.7247	0.88108	<b>18.8954</b>	<b>0.89182</b>
Video 2	25.8765	0.94843	25.8927	0.95107	<b>26.2525</b>	<b>0.97482</b>
Video 3	22.4714	0.89634	22.7196	0.90575	<b>24.7426</b>	<b>0.94165</b>
Video 4	21.5946	0.94385	21.652	0.94571	<b>21.8914</b>	<b>0.95698</b>
Video 5	21.8604	0.8475	21.3439	0.8590	<b>22.1308</b>	<b>0.89887</b>
Video 6	18.6504	0.89866	18.7116	0.9032	<b>18.7449</b>	<b>0.90746</b>
Video 7	17.7907	0.91267	21.6401	0.92634	<b>21.6871</b>	<b>0.94161</b>
Video 8	15.9602	0.89856	16.1205	0.91852	<b>16.9672</b>	<b>0.93097</b>
Video 9	18.784	0.89944	18.9701	0.90175	<b>18.9807</b>	<b>0.93236</b>
Video 10	19.9035	0.8539	19.8662	0.8562	<b>20.5579</b>	<b>0.94174</b>
Video 11	20.4396	0.93261	20.6817	0.93737	<b>20.7526</b>	<b>0.94357</b>

Bold indicates the proposed method values

**Fig. 11** Comparison graph of PSNR and SSIM values



**Fig. 12** Comparison of PSNR values for standard input videos

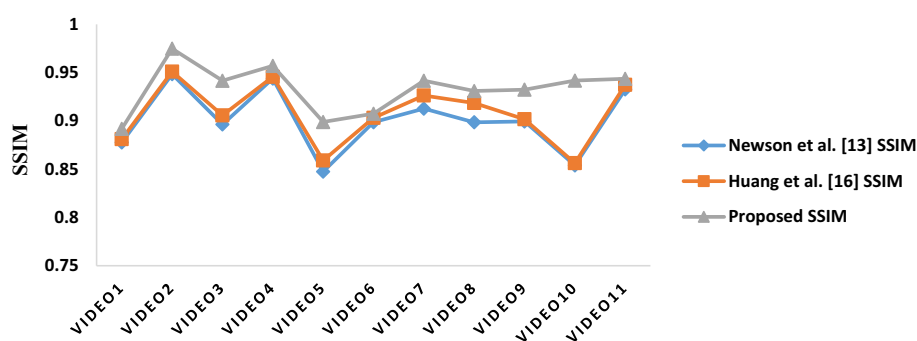


the PSNR of the proposed method is improved for all the standard input videos when compared with Newton [13] as well as Huang [16]. The variations of SSIM of the proposed method are clearly described in Fig. 13. It is noted that there is a significant improvement is achieved by the proposed method.

### 4 Conclusions

A novel video inpainting method using advanced homography-based image registration is proposed. In this homography-image registration, HALF-SIFT is used for proper feature point extraction, and a covariance matrix

**Fig. 13** Comparison of SSIM values



is used to estimate and remove the localization error. The proper selection of inliers to remove outliers is achieved using CW MLESAC. In order to get further refining the homography matrix, CW L-M was used. This entire process of image registration strengthens the spatiotemporal coherence in the video inpainting. Next, the inpainting of the hole in the target frame is accomplished by globally minimizing the energy cost function. The experimental video results are compared with two video inpainting methods available in the literature in the form of images. The comparison of inpainted videos or images shows that the proposed method produces high-quality results compared to the existing methods. The performance metrics, such as PSNR and SSIM values, are determined for the proposed method and compared with the results of existing methods. The average PSNR and SSIM values of the proposed methods for 11 videos are evaluated and compared with existing state-of-art inpainting methods. The improvement of 6% and 3% in average PSNR and SSIM is achieved by the proposed inpainting method when compared with existing inpainting methods.

## References

- Sridevi, G., Kumar, S.S.: Image inpainting and enhancement using fractional order variational model. *Defence Sci. J.* **67**(3), 308–315 (2017). <https://doi.org/10.14429/dsj.67.10665>
- Sridevi, G., Srinivas Kumar, S.: Image inpainting based on fractional-order nonlinear diffusion for image reconstruction. *Circuits, Syst. Signal Process.* (2019). <https://doi.org/10.1007/s00034-019-01029-w>
- Janardhana Rao, B., Chakrapani, Y., Srinivas Kumar, S.: Image inpainting method with improved patch priority and patch selection. *IETE J. Educ.* **59**(1), 26–34 (2018). <https://doi.org/10.1080/09747338.2018.1474808>
- Criminisi, A., Perez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process.* **13**(9), 1200–1212 (2004). <https://doi.org/10.1109/TIP.2004.833105>
- Lee, J., Lee, D.K., Park, R.H.: Robust exemplar-based inpainting algorithm using region segmentation. *IEEE Trans. Consum. Electron.* (2012). <https://doi.org/10.1109/TCE.2012.6227460>
- Matsushita, Y., Ofek, E., Ge, W., Tang, X., Shum, H.-Y.: Full-frame video stabilization with motion inpainting. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(7), 1150–1163 (2006). <https://doi.org/10.1109/TPAMI.2006.141>
- Patwardhan, K.A., Sapiro, G., Bertalmio, M.: Video inpainting under constrained camera motion. *IEEE Trans. Image Process.* **16**(2), 545–553 (2007). <https://doi.org/10.1109/TIP.2006.888343>
- Shih, T.K., Tang, N.C., Hwang, J.-N.: Exemplar-based video inpainting without ghost shadow artifacts by maintaining temporal continuity. *IEEE Trans. Circuits Syst. Video Technol.* **19**(3), 347–360 (2009). <https://doi.org/10.1109/TCSVT.2009.2013519>
- Shih, T.K., Tan, N.C., Tsai, J.C., and Zhong, H.-Y.: “Video falsifying by motion interpolation and inpainting.” In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, (Jun. 2008), pp. 1–8. DOI: <https://doi.org/10.1109/CVPR.2008.4587701>
- M. Granados, J. Tompkin, K. I. Kim, J. Kautz, and C. Theobalt, “Background inpainting for videos with dynamic objects and a free moving camera.” In: *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 682–695. [https://doi.org/10.1007/978-3-642-33718-5\\_49](https://doi.org/10.1007/978-3-642-33718-5_49)
- Whyte, O., Sivic, J., and Zisserman, A.: “Get out of my picture! Internet based inpainting.” In: *Proc. Brit. Mach. Vis. Conf.*, (2009), pp. 1–11. doi:<https://doi.org/10.5244/C.23.116>
- Chen, X., Shen, Y., and Yang, Y. H.: “Background estimation using graph cuts and inpainting.” In: *Proc. Graph. Inter.*, (2010), pp. 97–103.
- Newson, A., Almansa, A., Fradet, M., Gousseau, Y., Pérez, P.: Video inpainting of complex scenes. *SIAM J. Imag. Sci.* **7**(4), 1993–2019 (2014). <https://doi.org/10.1137/140954933>
- Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patch Match: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* **28**(3), 24:1–24:11 (2009). <https://doi.org/10.1145/1576246.1531330>
- Ebdelli, M., Le Meur, O., Guillemot, C.: Video inpainting with short term windows: application to object removal and error concealment. *IEEE Trans. Image Processing* **24**(10), 3034–3047 (2015). <https://doi.org/10.1109/TIP.2015.2437193>
- Huang, J.B., Kang, S.B., Ahuja, N., Kopf, J.: Temporally coherent completion of dynamic video. *ACM Trans. Graph.* (2016). <https://doi.org/10.1145/2980179.2982398>
- Janardhana Rao, B., Chakrapani, Y., Srinivas Kumar, S.: Hybridized cuckoo search with multi-verse optimization-based patch matching and deep learning concept for enhancing video inpainting. *Comput. J.* (2021). <https://doi.org/10.1093/comjnl/bxab067>
- Janardhana Rao, B., Chakrapani, Y., Srinivas Kumar, S.: An enhanced video inpainting technique with grey wolf optimization for object removal application. *J. Mobile Multimedia* **18**(3), 561–582 (2022). <https://doi.org/10.13052/jmm1550-4646.1835>
- Janardhana Rao, B., Chakrapani, Y., Srinivas Kumar, S.: MABC-EPP: Video in-painting technique with enhanced priority function and optimal patch search algorithm. *Concurr. Computat. Pract. Exper.* (2022). <https://doi.org/10.1002/cpe.6840>

20. Zhao, C., Zhao, H.: Accurate and robust feature-based homography estimation using HALF-SIFT and feature localization error weighting. *J. Vis. Commun. Image Represent.* **40**, 288–299 (2016). <https://doi.org/10.1016/j.jvcir.2016.07.002>
21. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004). <https://doi.org/10.1023/B:VISI.0000029664.99615.94>
22. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Surf: speeded up robust features. *Comput. Vis. Image Understand. (CVIU)* **110**(3), 346–359 (2008). <https://doi.org/10.1016/j.cviu.2007.09.014>
23. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **24**(6), 381–395 (1981). <https://doi.org/10.1016/B978-0-08-051581-6.50070-2>
24. Kai Cordes, Oliver Müller, Bodo Rosenhahn, Jörn Ostermann: “HALF-SIFT: high accurate localized features for SIFT”, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Miami, U.S.A., (2009), pp. 31–38. DOI: <https://doi.org/10.1109/CVPRW.2009.5204283>
25. Brooks, M.J., Chojnacki, W., Gawley, D.: “What value covariance information in estimating vision parameters?”. In: IEEE International Conference on Computer Vision, Vancouver, Canada, (2001), pp 302–308. DOI: <https://doi.org/10.1109/ICCV.2001.937533>
26. Steele, R.M., Christopher, J.: Feature uncertainty arising from covariant image noise. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, (2005), pp. 1063–1070.
27. Abdel-Hakim, A.E., Farag, A.A.: A novel stability quantification of detected interest points in scale-space. In: International Conference on Pattern Recognition, Tampa, FL (2008), pp. 124–127
28. Zeisl, B., Georgel, P.F., Schweiger, F.: Estimation of location uncertainty for scale invariant feature points. In: Proceedings of the British Machine Vision Conference, London, United Kingdom (2009)
29. Kolmogorov, V., Zabini, R.: What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(2), 147–159 (2004). <https://doi.org/10.1109/TPAMI.2004.1262177>
30. Boykov, Y., Veksler, O., Zabini, R.: Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(11), 1222–1239 (2001). <https://doi.org/10.1109/34.969114>
31. Boykov, Y., Kolmogorov, V.: An experimental comparison of mincut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(9), 1124–1137 (2004). <https://doi.org/10.1109/TPAMI.2004.60>
32. Perazzi, Federico, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung: “A benchmark dataset and evaluation methodology for video object segmentation.” In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 724–732. (2016). DOI: <https://doi.org/10.1109/CVPR.2016.85>



**B. Janardhana Rao** received his B.Tech from Nagarjuna University and M.Tech from Nagarjuna University. Presently, he is pursuing his Ph.D at JNT University, Kakinada, and currently working as an Associate Professor in the Department of ECE, CVR College of Engineering, Hyderabad, Telangana. His areas of interest include image inpainting, video inpainting, image restoration and enhancement.



video signal processing.

**Y. Chakrapani** received his B.Tech in ECE from JNT University Anantapur. He completed his Master Degree from NIT Warangal. He completed his Ph.D in Image Processing from JNTU Anantapur. He has 30 years of teaching experience. He worked as a professor and HOD of ECE in G. Pulla Reddy Engineering College, Kurmool; presently, he is working as a professor of ECE in ACE Engineering College, Hyderabad. His research interest includes image processing and



digital image processing, computer vision and the application of artificial neural networks and fuzzy logic to engineering problems.

**S. Srinivas Kumar** is working as a Professor in ECE Department, UCEK, JNTUK, Kakinada, Andhra Pradesh, India. He received his M.Tech from JNTU, Hyderabad, India. He received his Ph.D from E&ECE Department, IIT, Kharagpur. He has 35 years of experience in teaching and research. He has published more than 125 research papers in National and International Journals, and also in proceedings of reputed conferences. His research interests are digital

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

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.