HOKKAIDO UNIVERSITY

Title	Video Inpainting Based on Multi-Layer Approach
Author(s)	Chang, I-Cheng; Hsu, Chia-We
Citation	Proceedings : APSIPA ASC 2009 : Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, 200-207
Issue Date	2009-10-04
Doc URL	http://hdl.handle.net/2115/39674
Type	proceedings
Note	APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference. 4-7 October 2009. Sapporo, Japan. Oral session: 3D Computer Graphics and Applications (5 October 2009).
File Information	MP-L1-4.pdf

Instructions for use

Video Inpainting Based on Multi-Layer Approach

I-Cheng Chang* and Chia-We Hsu
Department of Computer Science and Information Engineering
National Dong Hwa University, Hualien, Taiwan
Email: *icchang@mail.ndhu.edu.tw Tel: 886-3-8634022

Abstract—**The objective of video inpainting is to fill the removed region originally occupied by the selected object. Because the selected object may occlude or be occluded by other objects in the video, it is necessary to determine the sequence of these objects to fill the removed region. In the paper, we present a new framework for video inpainting which can repair the selected region overlapped by multiple objects. Experimental results demonstrate that we can arbitrarily remove any object in a video and repair the removed area correctly.**

I. INTRODUCTION

The topics of image inpainting and video inpainting attract researchers' attention since digital videos become more popular. The major difference between these two techniques is that a single frame lacks of the temporal information. If a video contains the moving regions to be filled, it is not possible to repair these regions by using image inpainting technique. In the previous works, most of video inpainting algorithms focus on repairing the removed region of a single moving object in a video. In this paper, we present a video inpainting approach to fill the target video with region occluded by multiple objects.

A. Related Work

Bertalmio [3] is the first pioneer to video inpainting. The author repaired a video using image inpainting frame by frame. Raphae1[4] used the proposed image inpainting algorithm to repair a damaged frame, and then exploited it to repair other damaged frames through frame alignment. Both of them used PDE to find the structure information. For repairing large miss area of a video, Wexler [5] defined an optimal function to search the best matching patch in a foreground and use the found patch to repair the moving foreground. This algorithm filled the video frame patch by patch. Jiaya [6] combined the inpainting algorithm ([1]) and novel sample to repair the background and foreground respectively. Snakjeev [7] adopted the image inpainting algorithm of [2] to repair a video. Because of the color of a video is easily interfered with the lighting, therefore, the author used Fast Fourier Transform (FFT) to reduce the noise of a video. In order to reduce the computation, Patwardhan [8]-[11] established the foreground mosaic for repairing the moving foreground. In addition, the works also adopted the algorithm [2] to repair the background. Different from above video inpainting system, the system [14] used the probability model to repair the moving object.

B. System Framework

In the paper, we propose a novel video inpainting system. Our system adopts the temporal information and uses database of moving objects to repair removed region. Especially, if there are multiple objects overlaid with the same region, we are able to determine the present order of these objects and rebuild the videos.

The framework of multi-layer video inpainting and composition system is shown in Fig.1. This system consists of two major parts: generation of depth map sequences and multi-layer video inpainting. The first part aims at the object segmentation and depth map computation. We use the enhanced snake model to find the boundary of ROI, and then the result of ROI is used to compute the depth map. Most of the video inpainting algorithms cannot repair the removed regions of multiple objects. Because it is difficult to differentiate which object should be filled firstly when several objects locate at the same region. Getting the depth information of the corresponding objects is one way to solve the problem. In the second part, we construct a database of moving object, and then combined the depth information to repair the removed region which is occluded by object. The depth information can be used in video inpainting.
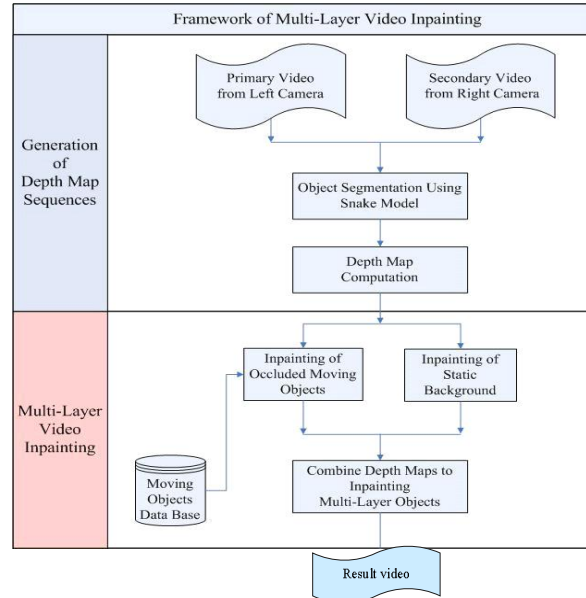

Fig. 1 System Framework of multi-layer video inpainting and composition.

II. GENERATION OF DEPTH MAP SEQUENCES

Depth information can be used to differentiate the occluded objects; however, it is a time-consuming task to get the depth information of the moving object for a video. Instead of the traditional depth computation, our system incorporates the object segmentation and the displacement vector field to get the video depth information.

A. Object Segmentation based on Enhanced of Snake Model

Snake model, also known as active contour model, is one of the techniques which are commonly used to find the boundary of region of interesting (ROI). Here, we improve the snake model to perform the object segmentation. Snake model needs the user give initial curve $v(s)$ of the ROI initially, and find the contour by the minimizing energy function,

$$E^*_{snake} = \int_0^1 E_{snake}(v(s))ds$$
$$= \int_0^1 (E_{int}(v(s)) + E_{ext}(I(x,y))ds \tag{1}$$

The Snake model consists of two major components: internal force E_{int} and external force E_{ext}. The internal force is defined in Eq. (2).

$$E_{int} = (\alpha |v'(s)|^2 + \beta |v''(s)|^2) \tag{2}$$

where $v'(s)$ and $v''(s)$ are the slope and curvature of the contour, and α and β are the weighted parameters. The weighting parameters regulate the tension and rigidity of the slope and curvature, respectively. And the external force is defined in Eq. (3).

$$E_{ext} = -|G_\sigma \times \nabla I(x,y)|^2 \tag{3}$$

where the gradient operator ∇ is used to calculate the edge information of an image, and G_σ is a two dimensional Gaussian function with standard deviation σ. If the value of σ is increased, the boundary of the image will be more blurry. The equation is represented as follows.

$$E_{snake} = \int_0^1 \frac{1}{2}(\alpha |v'(s)|^2 + \beta |v''(s)|^2) + E_{ext}(I(x,y))ds \tag{4}$$

To minimize the energy function in Snake model, E_{snake} should satisfy the Euler equation in Eq.(5).

$$\alpha v''(s) - \beta v''''(s) - \nabla E_{ext} = 0 \tag{5}$$

In Eq. (5) , it depicts that the internal force and the external are balancing. As the external potential force E_{ext} pushing the active contour toward the desired boundary of the image, the internal force E_{int} will control the smoothness of the active contour, and prevent the active contour from being too crooked. By dynamic programming, the iterative formula (6) deforms the active contour, and finally finds the boundary of ROI.

$$v_t(s,t) = \alpha v''(s,t) - \beta v''''(s,t) - \nabla E_{ext} \tag{6}$$

The Snake model finds the boundary of ROI by using intensity of gradient. If the difference of gradient between the ROI and background is too small, the found boundary will be too crooked. Increasing the parameter α of internal force can improve the condition, but it may induce a restriction on the external force and make the active contour be not close to the concave parts of ROI. It is not easy to adjust the proper values of α and β.

In order to overcome the disadvantage of traditional Snake model, we modify the definition of the external force. Firstly, Canny edge detector is used to derive the edge map from the image to make up a deficiency of gradient, and then add this edge map to external force. The modified external force is written as Eq. (7)

$$E_{ext} = -G_\sigma \times (|\nabla I(x,y)|^2 + EM_I(x,y)) \tag{7}$$

where EM_I is the edge map of image I and computed by Canny edge detector.

The proposed system modifies Eq. (7) by adding the foreground mask. This system uses the background subtraction to get the foreground mask. In generally, the method of background subtraction uses the frames to subtract the background by RGB color space, and the foreground will be show up. However, the intensity will be influenced by the light and the result in RGB color space will cause the fragmental foreground. Therefore, this system uses YCbCr color space to procede the background subtraction. The result is shown in Fig. 2(e) that foreground derived from proposed background subtraction. The foreground of Fig. 2 (e) is more complete than Fig. 2 (c), as YCbCr color space to our method, especially when the color of foreground is similar to the color of background. After the background subtraction, we use the median filter to reduce the noises (Fig. 2 (d) and (f)).
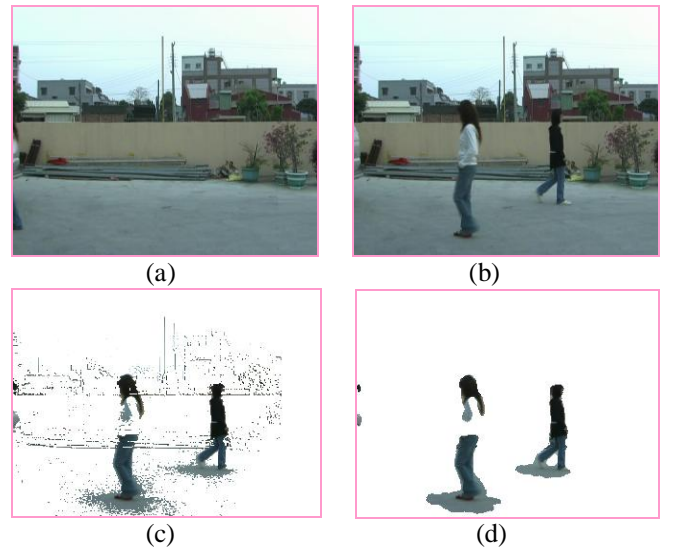

(a)   (b)

(c)   (d)

(e)   (f)

Fig. 2 Results of foreground mask. (a) Background image. (b)The frame contained foreground. (c)The result of background subtraction is derived from using RGB color model. (d) The result of noise removal (c). (e) The extracted foreground by proposed method in YCbCr color model. (f) The result of noise removal of (e).

We incorporate the foreground mask into the external force, and the external force is defined as:.

$$E_{ext} = -G_\sigma \times (|\nabla I(x,y)|^2 + EM_I(x,y) + EM_f(x,y)) \quad (8)$$

where the $EM_f(x,y)$ is the edge map of foreground mask. The boundary found by enhanced Snake Model is shown in Fig. 3(d). The foreground mask of the external force can effectively reduce the influence of background intensity on active contour.

(a)   (b)

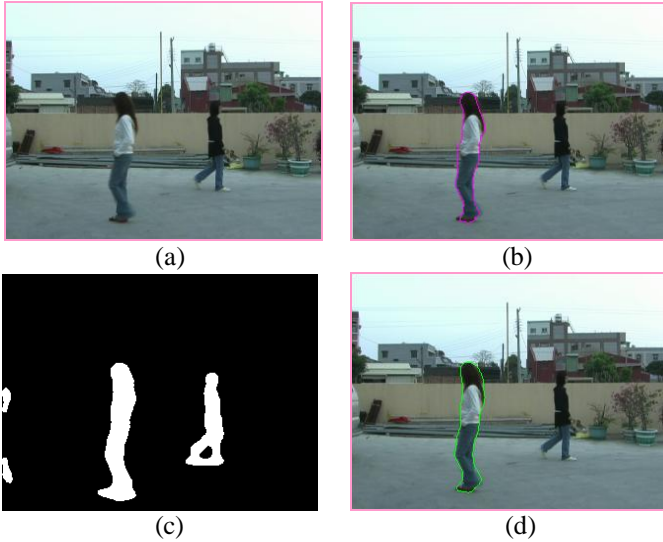(c)   (d)

Fig. 3 Find the boundary of ROI by using enhancement of Snake Model. (a) Original image. (b) The boundary of ROI is found by only adding the Canny edge to external force. (c) The foreground mask. (d) The boundary is found by using enhanced Snake model.

B.   Depth Map Computation

Two methods are commonly used to get depth information. One approach is to use a hardware system to find the distance between the camera and the objects by calculating the time of reflected when the light is projected onto objects. The other method used the stereo system to capture the images from the target different views, and then to find the displacement of the object. But the displacement would not be so accurate, especially if the distribution of the vectors of the color is smooth. Anantrasirichai [12] established a system with three cameras to find the relation among objects. Lawrence Zitnick [13] used dynamic block to increase the accuracy. The block size is depended on the result of color segmentation during the processing.

In order to effectively get the depth information, we apply the displacement vector field with the result of object segmentation. Two cameras, primary and secondary cameras, form a stereo system. The disparity between stereo images is computed by using block matching. The image captured by the primary camera is divided into $N \times N$ blocks, and then we search the correspondence block in the image captured by secondary camera. And, the sum absolute difference (SAD) is used to measure the similarity of these blocks. The disparity $Dis(x,y)$ of a block $P(x,y)$ is defined as follows:

$$Disp(x,y) = |P(x,y) - \hat{P}(x+i,y+j)|$$
$$\hat{P}(x+i,y+j) = \arg\min_{i\in h, j\in v} (P(x,y), \hat{S}(x+i,y+j))$$
$$\hat{S}(x+i,y+j) = \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} |P(x+k,y+l) - S(x+k+i,y+l+j)| \quad (9)$$

where $Disp(x,y)$ is the disparity of a block and (x,y) is the position of the block center. $\hat{S}(x+i,y+j)$ is the difference from block $P(x,y)$ and block $S(x,y)$ in the corresponding frame. The coordinate of most similar block is record in $\hat{P}(x,y)$. The result of block matching is shown in Fig. 4(d). If the color of a pixel is closer to red and the value of disparity of a pixel is bigger, then the pixel is closer to camera. Fig. 4 (d) presents the rough depth information of Fig. **4**(a). We use the result of object segmentation to find the ROI displacement, and count the displacement of the ROI to find the maximum value as the depth of ROI. Fig. 4 (e) shows the final result. The method can get the depth map sequence of a video and also reduce the noise influence on the disparity maps.

(a)   (b)

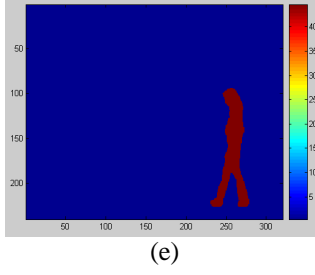(c)   (d)

(e)
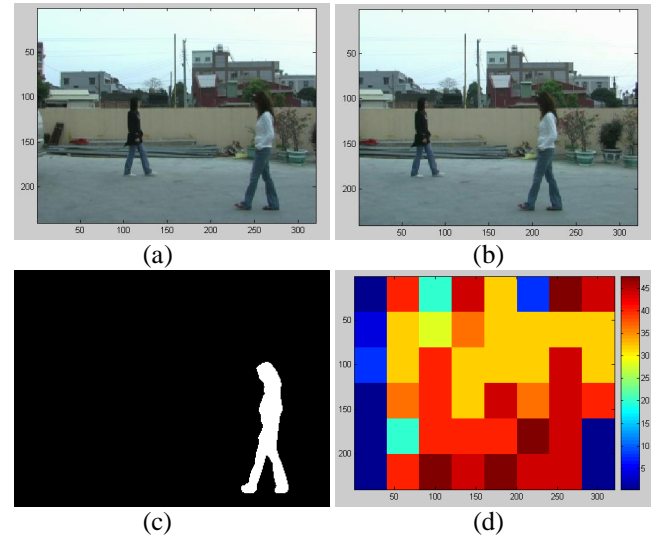
Fig. 4 Generation the depth map. (a) This frame is captured from primary camera.(b) This frame is captured from secondary camera.(c)The ROI mask.(d)The result of block matching and the block size is 40×40 .(e)Recalculating the depth information of ROI.

If there are multiple objects in a video, the system is able to combine the ROI tracked by Snake model together with the result of block matching to get the depth information of all the moving objects in the video (Fig. 5).
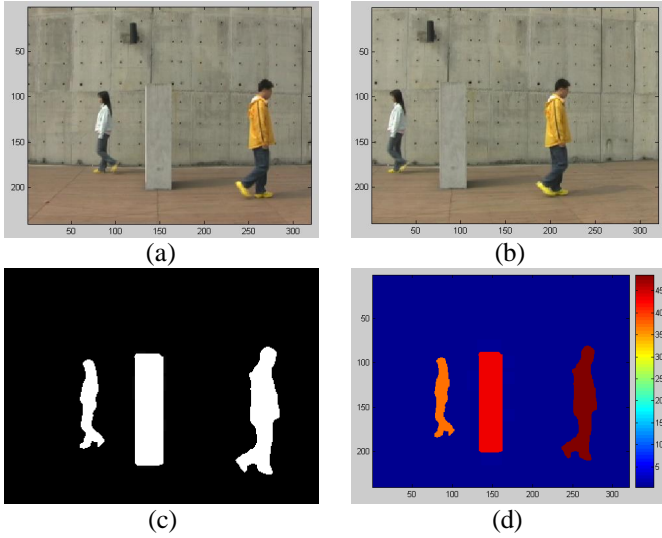

Fig. 5 The depth information of multiple objects (a) This frame is captured by primary camera.(b)This frame is captured by secondary camera. (c)The mask of ROI. (d) Depth map.

III. MULTI-LAYER VIDEO INPAINTING

The video inpainting of our system consists of two parts: inpainting of the stationary background and repair of the region occluded by objects. In previous work, most of video inpainting algorithms were used to repair small damaged static area. Wexler [5] defined an optimal function to repair occluded moving objects, which is a patch-based algorithm. In recently year, some researchers ([8] and [11]) repaired the occluded object by establishing foreground mosaic and used foreground mosaic to repair the damaged object. The methods of [5], [8] and [11] were designed to repair a single moving object. Assume that the damaged area is marked with white as Fig. 6(b). For object A of Fig. 6 (a), the object A would be correct repaired. However, for object C, since there are two moving object needs to be repaired, it is needed to know whether object A or object C should be filled firstly. For repairing multiple objects, a solution [10] used the ground line of object to determine filling order of these objects. But, it is difficult to find the accurate ground line of object, because the shadow of object is difficult to separate from object. Instead of the ground line approach, we apply the depth map of a video to repair multi-layer occluded objects.
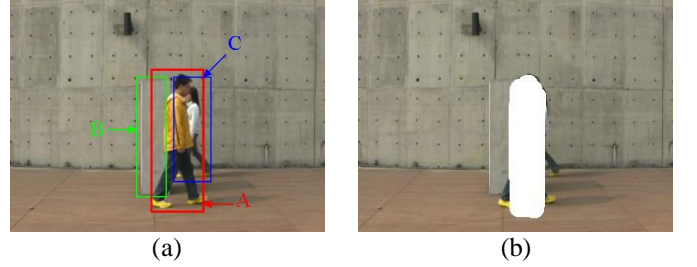

Fig. 6 An example of multi objects of a video. (a) Original frame. (b) The miss region is marked with white.

A. Depth Map Computation

According to the types of static background, we perform the region inpainting using two different approaches.

(a) Temporal approach

Some frames of a video (Fig. 7(b)) are damaged due to noise or other factors. Our system uses temporal information to repair damaged frame. We search the position of missing area of the image and copy the corresponding area from nearest undamaged fames (ex: Fig. 7 (a)) to repair the damage one.
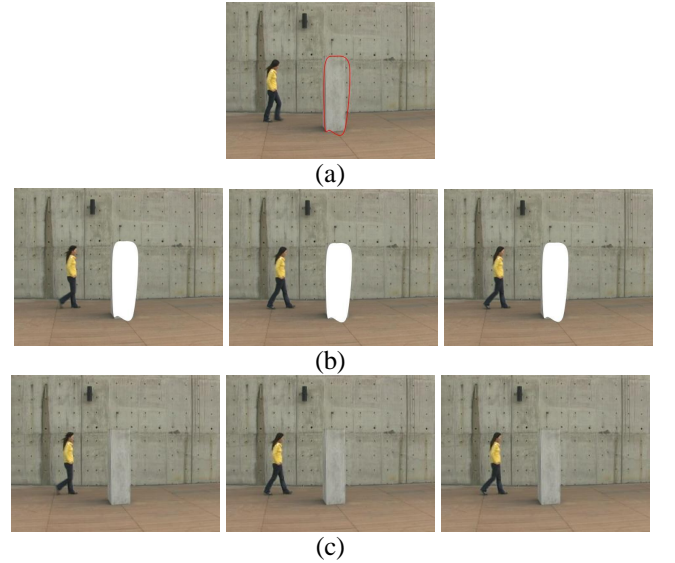

Fig. 7 The result of static background inpainting using temporal information to inpainting static background. (a)The red area is the removed area. (b) The removed area of image is marked whit white. (c) The result of inpainting background.

If the moving foreground is the selected area to be removed as Fig. 8(b), we can use object segmentation to remove unwanted moving object. And the removed region can be filled by temporal information. The results of filling dynamic removed area in damaged frames are showed in Fig. 8 (c).
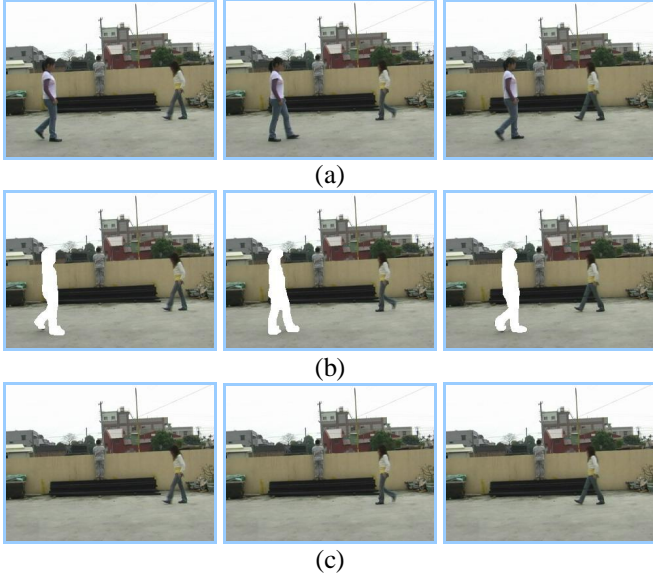
Fig. 8. Remove the unwanted moving object and filling the dynamic miss area of damaged frame. (a)The original sequences frame # 10, #20 and #30. (b) The object is removed and masked with white. (c)The images are shown the results of the inpainting dynamic miss area of damaged frames.

(b) Spatial approach

If the static object or damaged area occurs at the same position during the whole video, there is no temporal information for reference. For instance, all frames have a pillar in the same location in a video (Fig. 9), where the background information is unavailable. In the situation, the temporal information is not enough to repair the damaged frame. Therefore, the image inpainting algorithm is used to reconstruct the missing area. And other damaged frames are filled by copied the corresponding area from repaired background, so as to maintain consistent background throughout whole video.
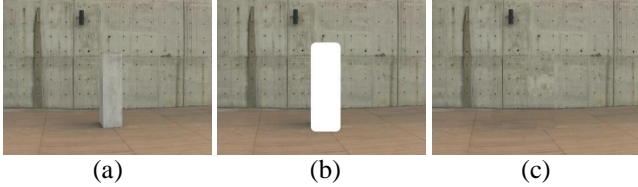

Fig. 9 The result of using inpainting algorithm to reconstruct stationary background. (a)Original image. (b)The unwanted area is masked with white. (c)The result of reconstructed background.

B. Appropriate Template Selection from Moving Object Database

The background information is generally not appropriate to repair the region occupied by moving object. We use undamaged frames to establish the moving object database, where the damaged moving object can be repaired rapidly and accurately. Fig. 10(a) shows the trajectory of a running kid and we remove the pillar from background as Fig. 10 (b). Because the kid is occluded by the pillar in some frames, we search the kids in the other frames and use these images to repair the lost area. In order to build up the database of all

moving object in the video, we record the size of minimum boundary rectangle (MBR) of all candidates, and then use their MBR to save all candidates as Fig. 10 (c).
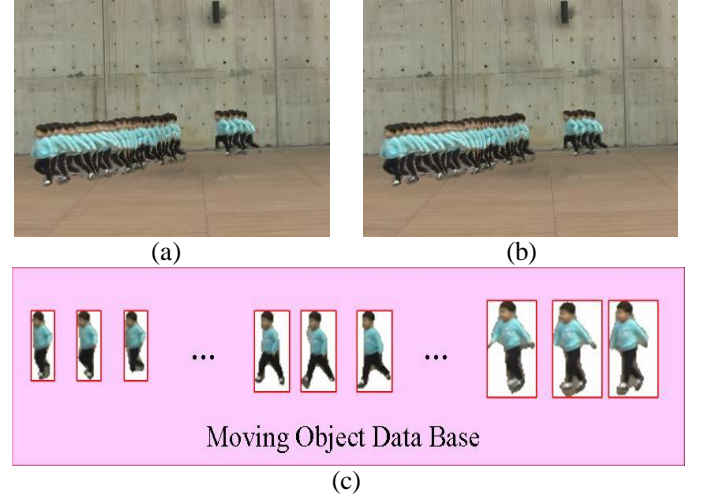

Fig. 10. Use the max size of MBR to save all candidates in database. (a)The trajectory of moving object. (b) The pillar is removed and we need to repair the moving object in this area.(c) All candidates are saved with their MBR.

In Fig. 10, it is obvious that the MBR size of ROI is different. Because the trajectory of person is not parallel with image plane, all candidates are normalized by using Eq.(10) before finding the best match template.

$$L\_MBR'_j = L\_MBR_j \times \gamma_x \ , \ W\_MBR'_j = W\_MBR_j \times \gamma_y$$

$$\gamma_x = \frac{L\_MBR_j}{L\_MBR_i} \ , \ \gamma_y = \frac{W\_MBR_j}{W\_MBR_i}$$

$$(10)$$

where, the γ_x and γ_y are the scale operators for the length and width of image. L_MBR_i and L_MBR_j are the length of MBR related to object i and object j, respectively. W_MBR_i and W_MBR_j are the width of MBR of object i and object j. The sizes of all candidates were multiplied by γ_x and γ_y. Additionally, in order to calculate the objects shift information, we also record the coordinates of all candidates in the original video.

Because human walking is periodic motion, the position at one point in some walking cycle can be found in another cycle walking cycle. However the motion in the two half periods will not be the same since the movement of left and right legs is slightly different. For example, Fig. 11(a) and (b) show the human walking for different leading legs, where the body center in the MBR is moving forth and back. If the quality q is defined as the distance between the body center and the left boundary of MBR, the variation of q during a walking period can be represented as Fig. 11(c).
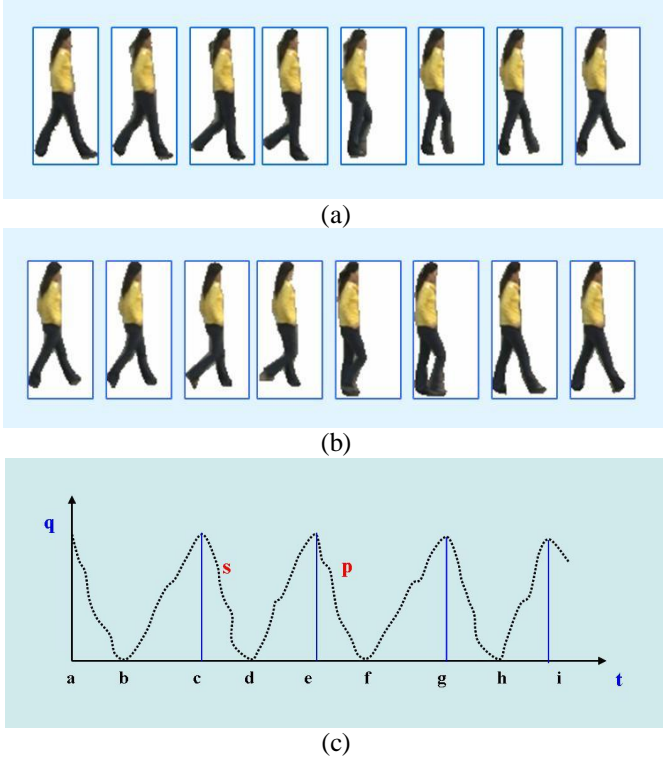
(a)

(b)

(c)

Fig. 11 Human walking. (a)Leading led is right. (b) Leading leg is left.(c)The variation of position of body center.

After building up the database of moving object and selecting the target template, our system find the best matching templates by using the measure function described as follows.

$$B(O_j) = \arg\max_{j \in t} d(M_i, M_j)$$
$$d(M_i, M_j) = \sum_{w=1}^{3} \sum_{x \in m, y \in n}^{m,n} \mathrm{XOR}(M_{i+w}(x,y), M_{j+w}(x,y)) \quad (11)$$

This measure function is used to ensure the consistency of object motion from frame to frame. We select continuous three templates to be as the target template (slide window =3) from a video before the occluded happen. If the object is occluded at time t, the target templates are denoted as O_{t-1}, O_{t-2} and O_{t-3} respectively. M_{i+w} and M_{j+w} are the corresponding binary masks of target templates and candidate templates, where m and n is the size of the template rectangle.

C. Interpolation of Missing Trajectory

After obtaining the best matching templates, we find the miss trajectory of moving object. Yuping [9] described the trajectory of moving object as a curve. The author adopted the spatial-temporal coherence error function to repair miss trajectory. This error function samples the points of propagated path, and then fills the appropriate foreground mosaic in the sampled points. However, the method of [9] for video inpainting is time-consuming. In order to efficiently repair the occluded region and get smoother miss trajectory, our system refers to the horizontal and vertical shift

information of found candidates to repair the miss parallel trajectory and unparallel trajectory, respectively. We introduce this method under the following two situations:

Situation 1 is the missing trajectory of object is parallel to image plane (Fig.12). We need to find the horizontal shift of walking human. Firstly, we assume the time of finding best matching template is time q and the object is occluded at time t. In our system, the value of subtracted the position of left boundary of object in the frame q from the position of left boundary of object in the frame $q+1$ is regarded as horizontal shift information. Then, the matched templates are shifted according to this horizontal shift information.

Fig. 12. Original Frame #55 ,#58 ,#61 ,and # 64 in video.

Situation 2 is the miss trajectory of object is not parallel to image plane (Fig.13). When filling the template in damage frame, this system considers not only the horizontal shift information of walking people but also the vertical shift information. The method of getting the vertical shift information is similar to the computation of the horizontal information. The vertical displacement is equal to the result of subtracting the top position of boundary of object in frame q from the top position of boundary of object in frame $q+1$ .When we filling the template in the damaged frame, the template is shifted according to horizontal and vertical displacement information.
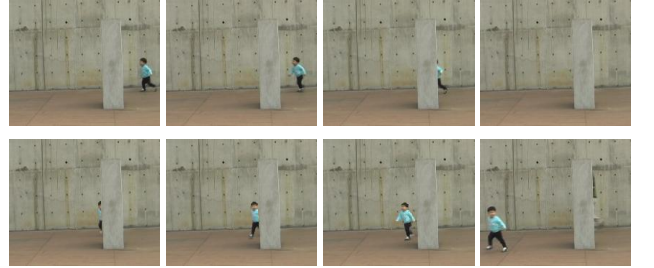

Fig. 13 An example of trajectory is not parallel to image plane.

IV.   EXPERIMENTAL RESULTS

The example videos used in the section are captured by two stereo cameras and the image resolution is 320*240 pixels.

Experiment 1

In the experiment, we remove the pillar of background and repair the occluded moving object. Fig.14 shows the result of repairing the occluded objects, where its trajectory is parallel to image plane. The missing area of background is filled well, and the moving object has been repaired successfully.
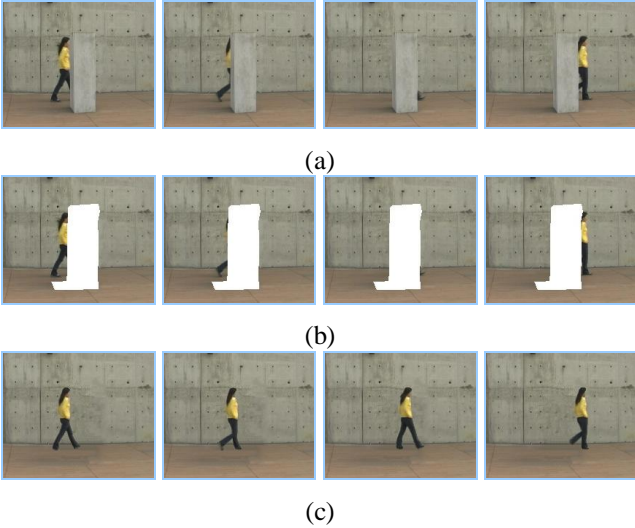
(a)

(b)

(c)

Fig. 14 Repair the moving foreground and static background. (a)The first row is original frames of a video.(b) The removed pillar is marked with white in middle row.(c) The result frames.

Experiment 2

We also remove the pillar of the background and repair the occluded kid. But, the trajectory is not parallel to image plane. Because the region of the object and the shift information changes as kid more and more closes to camera. The candidates need normalization before the procedures of finding matching point. And the displacement also needs adjustment in filing the miss trajectory too. In Fig. 15(b), the kid is running rapidly and it will increase the difficult of object segmentation. But the object is still segmented well by enhancement Snake model. And the repaired moving object is consistent in the sequences in Fig. 15 (c).


(a)
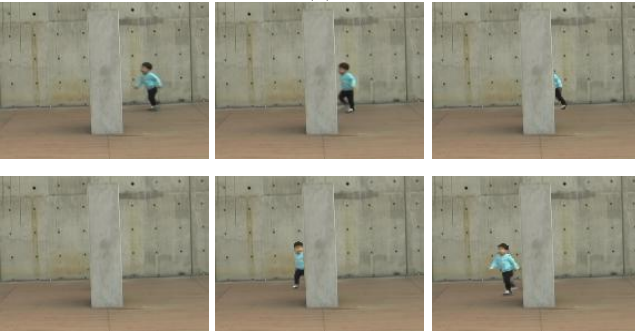
(b)

(c)

Fig. 15 Remove the static object and repaired occlusion. In addition, the trajectory of moving is not parallel with the image. (a) The removed object is mark with white .The first row is shown original frames. (b) The first row is shown original frames. (c)And the repaired frames are shown in last.

Experiment 3

Our system can not only remove the static object in a video, but also remove the moving objects of multiple layers. The moving object on the first layer is removed and the occluded object is repaired. Moreover, we use the object segmentation to find the ground line of this object and remove its shadow. The results of repaired frames are show in Fig.16.
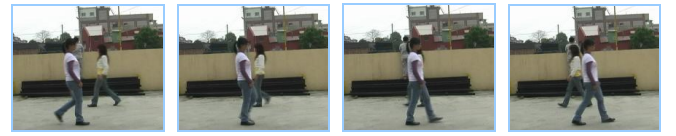
(a)

(b)

(c)

Fig. 16 Remove the first layer moving and repair the occluded moving object.(a) The original frames .(b) The removed object is depicted as white area. (c) The result repaired frames.

V. CONCLUSIONS

In the paper, we present a new system framework for video inpainting. Different from the previous works, our system can repair the damaged video with more than two moving objects. Instead of repairing the object patch by patch, we reconstruct the occluded foreground using the object database. The approach can complete a damaged video effectively. In experimental results, we demonstrate that the objects can be arbitrary removed in a video and the system can repair the removed area well.

In the system, the method of finding besting point only depends on the shape of the candidates. Although we have constraints on selection of the target template; however, if the result of the object segmentation is not accuracy enough, it will increase the error rate of finding best matching point. In the future, we want to add other features in finding the leading leg of walking people to reduce the error. Furthermore, we

will extend the application of video inpainting to the videos with camera motion.

REFERENCES

[1] J. Jia and C.-K. Tang, "Image Repairing: Robust Image Synpaper by Adaptive ND Tensor Voting." , *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* ,Volume 1, Page(s):I-643 - I-650, June 2003.

[2] A. Criminisi, P. P´erez, and K. Toyama, "Region Filling and Object Removal by Exemplar-Based Image Inpainting" , *IEEE Transactions on Image Processing* ,Volume 13, Issue 9, Page(s):1200 - 1212 ,Sept. 2004.

[3] M.Bertalmio, A.L.Bertozzi, and G..Sapiro, "Navier-Stokes, Fluid Dynamics, and Image and Video Inpainting." IEEE Proceedings of Computer Society Conference on *Computer Vision and Pattern Recognition*, Volume 1, Page(s): I-355 - I-362 ,Dec 2001.

[4] Raphaël Bornard, Emmanuelle Lecan, Louis Laborelli, and Jean-Hugues Chenot, "Missing Date Correction in Still Images and Image Sequences." *Proceedings of the tenth ACM international conference on Multimedia*, Dec. 2002.

[5] Y. Wexler, E. Shechtman, and M. Irani, "Space-Time Video Completion." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* ,Volume 1, 27 June-2 July 2004.

[6] Jiaya Jia, Tai-Pang Wu, Yu-Wing Tai, and Chi-Keung Tang, "Video Repairing: Inference of Foreground and Background under Severe Occlusion." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* ,Volume 1, Page(s):I-364 - I-371 ,July 2004.

[7] Sanjeev Kumar, Mainak Biswas, Serge J. Belongie, and Truong Q. Nguyen, "Spatio-Temporal Texture Synpaper and Image Inpainting For Video applications." *ICIP IEEE International Conference on Image Processing*, Volume 2, 11-14, Sept. 2005.

[8] K.A. Patwardhan, G. Sapiro, M. Bertalmio, "Video Inpainting of Occluding and Occluded Objects." *IEEE International Conference on Image Processing,* Volume 2, Sept. 2005.

[9] Yuping Shen, Fei Lu, Xiaochun Cao, Hassan Foroosh, "Video Completion for Perspective Camera under Constrained Motion.", *IEEE International Conference on Pattern Recognition,* Volume 3, Page(s):63 – 66, 2006.

[10] Sen-Ching S. Cheung, Jian Zhao, M.V. Venkatesh, "Efficient Object-Based Video Inpainting", *IEEE International Conference on Image Processing,* Page(s):705 – 708, Oct. 2006

[11] K.A. Patwardhan, G. Sapiro, and M. Bertalmio, "Video Inpainting under Constrained Camera Motion." *IEEE Transactions on Image Processing ,* Volume 16, Issue 2, Feb. 2007.

[12] N. Anantrasirichai, C.N. Canagarajah, D.W. Redmill, and D.R. Bull, "Dynamic Programming for Multi-View Disparity Depth Estimation." *ICASSP Proceedings. IEEE International Conference on Acoustics, Speech and Signal Processing ,* Volume 2, 2006.

[13] C. Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski, "High-Quality Video View Interpolation Using A Layer Representation. " *ACM transactions on Graphics,*Volume 23, Issue 3 , August 2004.

[14] V. Cheung, B.J. Frey, N. Jojic, "Video Epitomes." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,*Volume 1, Page(s):42 - 49 , June 2005.