Video Completion for Indoor Scenes

Vardhman Jain and P.J. Narayanan

Center for Visual Information Technology
International Institute of Information Technology
Hyderabad, India
{vardhman@research., pjn@}iiit.ac.in

Abstract. In this paper, we present a new approach for object removal and video completion of indoor scenes. In indoor images, the frames are not affine related. The region near the object to be removed can have multiple planes with sharply different motions. Dense motion estimation may fail for such scenes due to missing pixels. We use feature tracking to find dominant motion between two frames. The geometry of the motion of multiple planes is used to segment the motion layers into component planes. The homography corresponding to each hole pixel is used to warp a frame in the future or past for filling it. We show the application of our technique on some typical indoor videos.

1 Introduction

Segmenting and removing objects from images or videos is of much current interest. Object removal leaves the image or video with unknown information where the object was earlier placed. Missing information recovery in images is called *inpainting*. This is accomplished by inferring or guessing the missing information from the surrounding regions. For videos, the process is termed as *completion*. Video completion uses the information from the past and the future frames to fill the pixels in the missing region. When no information is available for some pixels, inpainting algorithms are used to fill them. Video completion has many applications. Post-production editing of professional videos in creative ways is possible with effective video completion techniques. Video completion is perhaps most effective with home videos. Video can be cleaned up by removing unnecessary parts of the scene. Inpainting and video completion is often interactive and involve the users as the objective is to provide desirable and appealing output.

Image inpainting inevitably requires approximation as there is no way of obtaining the missing information. For videos, the missing information in the current frame may be available from nearby frames. Significant work has been done on inpainting and professional image manipulation applications and tools exist to accomplish the task to various degrees. The solution to the problem of object removal in video depends also on the scene complexity. Most video completion work has focused on scenes in which a single background motion is present such as an outdoor scene. In scenes with multiple large motion, motion layer segmentation methods are used to obtain different motions layers. A particular layer

P. Kalra and S. Peleg (Eds.): ICVGIP 2006, LNCS 4338, pp. 409–420, 2006.

can be removed by filling the information with the background layers. Scenes with multiple motion, such as indoor scenes, are challenging to these algorithms. For scenes with many planes, motion model fitting may not be suitable as the boundaries between the layers are not exact. This is especially problematic for video completion as the region being filled could straddle these boundaries. Periodicity of motion is also often used by techniques which fill the holes by patching from some other part of the video.

In this paper, we present a technique for video completion for indoor scenes. We concentrate on scenes where the background motion consists of two or three planes in the neighborhood of the object to be removed. The main contribution of this paper is the use of the geometry of intersecting planes in multiple views for motion segmentation, without applying a dense motion segmentation in the image. We also show that segmentation of only the nearby background around the missing region is sufficient for the task of video completion. Full-frame motion segmentation can thus be avoided. The geometric nature of the method ensures accurate and unique background assignment to the pixels in the unknown region, which to the best of our knowledge is not possible with other video completion methods. We particularly concentrate on scenes where the neighborhood around the object to be removed is planar in nature.

The rest of the paper is organized as follows. In Section 2, we describe relevant previous work. Section 3 discusses various stages of our algorithm in detail. Results are shown in Section 4. Conclusions and ideas for future work follow in Section 5.

2 Previous Work

The work presented here is closely related to a few well studied problems. *Image inpainting* fills-in the unknown regions (or holes) in an image based on the surrounding pixels. Structure propagation and texture synthesis are the two basic approaches for image inpainting. Structure propagation methods propagate the structure around the unknown region progressively to inside it. Bertalmio *et al* [1] proposed a method for filling-in of the image holes by automatic propagation of the isophotes (lines of similar intensity) in the image. Texture synthesis [2,3] methods assume the existence of a pattern in the image and fill the pixels in the missing region by finding a patch matching the neighboring texture in the whole image. Texture synthesis has been done at pixel level [2] as well as block level [3,4]. Structure propagation methods work well only on small holes, whereas texture synthesis methods require texture in the image. Methods combining both structure propagation and texture synthesis have been proposed in recent years and show impressive results [5,6]. These image inpainting methods calculate the values of unknown regions. These can only be an approximation of original data, however.

Kang *et al* [7] proposed a technique for inpainting or region filling using multiple views of a scene. Their technique is based on finding the appropriate region

in the second view and then mapping the pixels back to the first view using the affine projection calculated using the correspondence in the two views. Similar methods are used in video completion as discussed below.

Object removal in videos has received attention in recent years. Two types of techniques have been proposed. The first type finds out the missing data by searching for a patch matching the neighborhood of the hole in the video. The match is defined in terms of spatial and temporal feature similarity. Periodicity in motion is a common assumption for these techniques. Space time video completion [8] uses a five dimensional sum of squared differences to find the appropriate patch for filling the holes where the matrices include the three color values and velocity along x and y direction. *Video Repairing* proposed by Jia *et al* [9] recovers the missing part of foreground objects by *movel* sampling and alignment using tensor voting to generate loops of motion by connecting the last frame to the first frame. Motion field interpolation based methods have also been developed recently. Kokaram *et al* [10] perform object removal by using the motion information to reconstruct the missing data by recursively propagating data from the surrounding regions. Matsushita *et al* [11] proposed *motion inpainting* where the inference of the unknown pixels information is based on the optical flow vectors which are in turn interpolated based on the flow of the surrounding pixels.

In the second scenario, explicit use of the geometry of multiple views is made to infer the information missing in the current frame from the nearby frames. This is directly related to the problem of disocclusion in computer vision. The fact that two views of a plane are related by a perspective transformation defined using a Homography matrix, forms the basis of most such approaches. Jia *et al* [9] proposed the repairing of the static background by the use of planar layered mosaics. The layers are assumed to be available from initial manual segmentation followed by tracking using the mean shift algorithm. Similar approach has been demonstrated by Zhang *et al* [12]. They use an automatic layer extraction approach followed by layered mosaicing. If some holes still remain an image inpainting approach is used in frame-wise manner based on a graph cuts formulation.

When the camera is far from the background, the nearby frames of the background can be approximated to be related by an affine or projective transformation. This approximation is used by some methods [9]. Such methods will fail for indoor scenes where multiple background motion exists. In general, it would be impossible to identify every single plane in the scene and apply layer mosaicing on each of them individually, automatically and accurately.

Structure from motion problems employ some techniques that are relevant to this problem. Vincent and Laganire [13] discuss the problem of dividing the image into planes. They start with a set of point correspondence and apply the RANSAC algorithm with an optimal selection of the four initial points to maximize the chance that the points are on same plane. All the other points in the image are declared to belong to the plane whose homography gives least reprojection error. Fraundorfer *et al* [14] find the interest regions in the two views

on which affine region matching is performed. The affine matching is helpful in removing the non-planar regions from considerations. On the matched region the homography is determined and a region growing is performed around the region to include regions which match the homography well. During the region growing step the homography is updated to include the new interest points inside the region for the estimation. At the termination of the region growing, the scene is segmented into a set of planar regions. Wills *et al* [15] proposed a graph cuts formulation for motion segmentation. First a set of dominant motions in the two views is obtained. The energy terms in the graph are based on the re-projection error due to each motion model and the smoothness term is defined based on color similarity between the pixels.

The work presented in this paper combines many of these ideas to perform video completion indoor scenes with multiple background motions.

3 Video Completion for Indoor Scenes

In this paper, we address the problem of object removal and video completion for indoor scenes where the transformation of the background is non trivial and variable. An overview of the process is shown in Figure 1. We track the foreground (the object to be removed) interactively using our earlier work [16] to

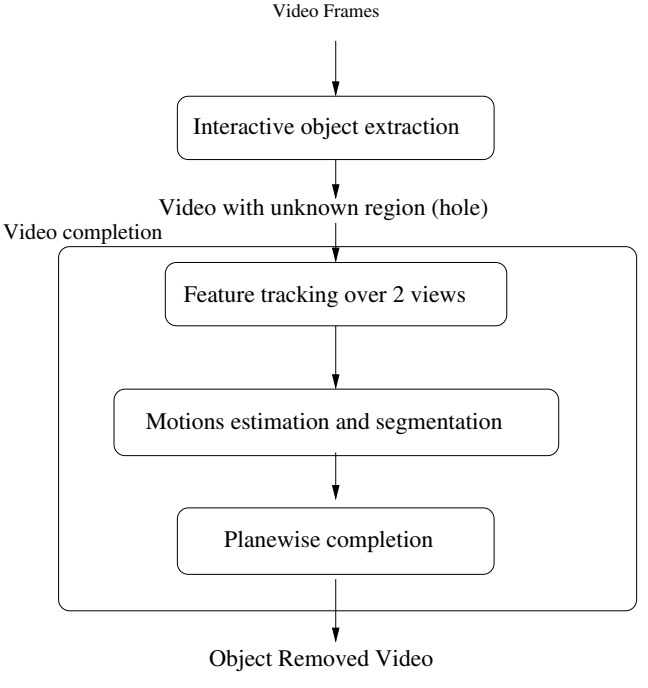
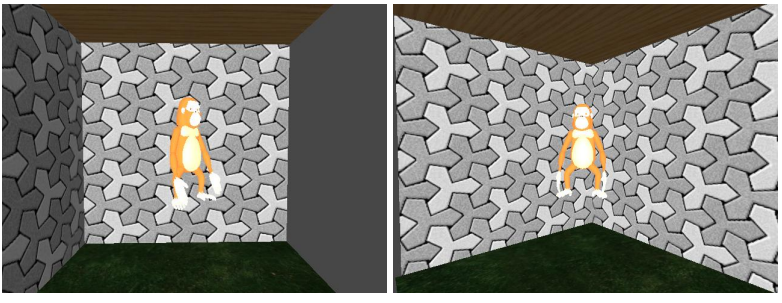


Fig. 1. The overview of the various steps of our system

track the objects across the video. For this paper, we assume that the background has a maximum of 2 planes around the object to be removed in two adjacent views. The region around the object is segmented into one or two planes, using dominant motion model estimation followed by an optimal boundary detection algorithm. We then apply the respective homography to recover the unknown pixels from the neighboring frames. These steps are explained below.



(a) (b)

Fig. 2. Two different cases of object removal (a) The local background around the object is a single plane (b) The local background around the object is spread over more than 1 plane. Due to the local nature of the plane segmentation technique the first case (a) doesn't need any motion segmentation. Motion segmentation in the second case (b) is also local in nature and even though there are more planes in the image only the two planes which constitute the object's background would be segmented.

3.1 Object Segmentation

The segmentation step provides the masks of the object to be removed across the video frames. Unlike image inpainting techniques, getting this mask from the user in each frame is not feasible. We use an interactive method of object extraction using graph cuts and feature tracking to generate the mask across the video sequence.

The user gives a binary segmentation of the first frame, marking the foreground and the background. We track features points in the segmented frame to the current frame (unsegmented) and set them as seed points in the 3D graph constructed with the two frames. A graph cuts optimization on the graph gives the segmentation for the current frame. The user can mark extra stroke and run the iterative graph cut to improve the segmentation before proceeding to next frame. Our method has the advantage of being fast and interactively driven. This allows us to have complex object or object with complex motion segmented across the video. This method is similar to Video object cut and paste [17].

After running through the frames of video, we get the object mask in each frame. This mask defines the region to be filled in using the video completion algorithm.

3.2 Video Completion

Our algorithm's basic assumption is the existence of a piecewise planar background in local neighborhood of the object to be removed. Our video completion algorithm can be divided into following major sub-steps.

Feature tracking in two views: The first step is finding the corresponding feature points in the two frames of the video. We use the KLT tracking for tracking point features across the frames. The method involves finding trackable features in the first image, which are then matched in the second image. We find the features selectively in only local neighborhood of the hole, this is to ensure that we only consider useful correspondences for our motion estimation and completion steps. We call the region around the hole where we do the selective matching as the Region of Interest (ROI). Figure 3 (b) shows the optical flow vectors calculated in the ROI. The ROI can be obtained by dilating the object mask with an appropriate thickness.

Motion Segmentation: Given the point correspondences in the two images, our aim is to find the planar segmentation of the ROIs. Figure 2 shows the two possible scenarios. In Figure 2(a) the ROI around the object is a single plane, while in Figure 2 (b) the ROI includes two different planes. We use a combination of two approaches to robustly estimate the segmentation of the points inside the ROI into multiple planes. The algorithm proceeds by finding the dominant motions in the ROI using the set of correspondences. We use the RANSAC [18] algorithm to determine the dominant motion. RANSAC algorithm has the advantage of being robust to outliers, which are indeed present in our correspondence pairs due to the existence of multiple planes.

To begin with, we use all the correspondence pairs to determine the dominant motion. The features which are inliers for the current dominant motion are then removed from the set and the step is repeated to find the next dominant motion. To avoid RANSAC algorithm from choosing wrong set of initial four points, we modify the selection phase to accept the set of points only if they are within a set threshold distance. The points which are declared inliers to the RANSAC algorithm are then used for a least square error fitting estimate of the homography using the normalized DLT algorithm [19]. This fitting gives us the final homography for the set of points. Figure 3 (c,d) shows the automatically determined first and second dominant motions as cluster of optical flow vectors which are their inliers.

Optimal boundary estimation: Optimal boundary estimation is needed to actually separate the ROI into two different planes. This information is later used during the filling-in process. Note that unlike other methods [13,14] we cannot depend on the region growing method to give us the boundaries of the planes because we can not estimate these boundaries in the unknown region. We assume the intersection of the two planar regions to be a line. Let H_1 and

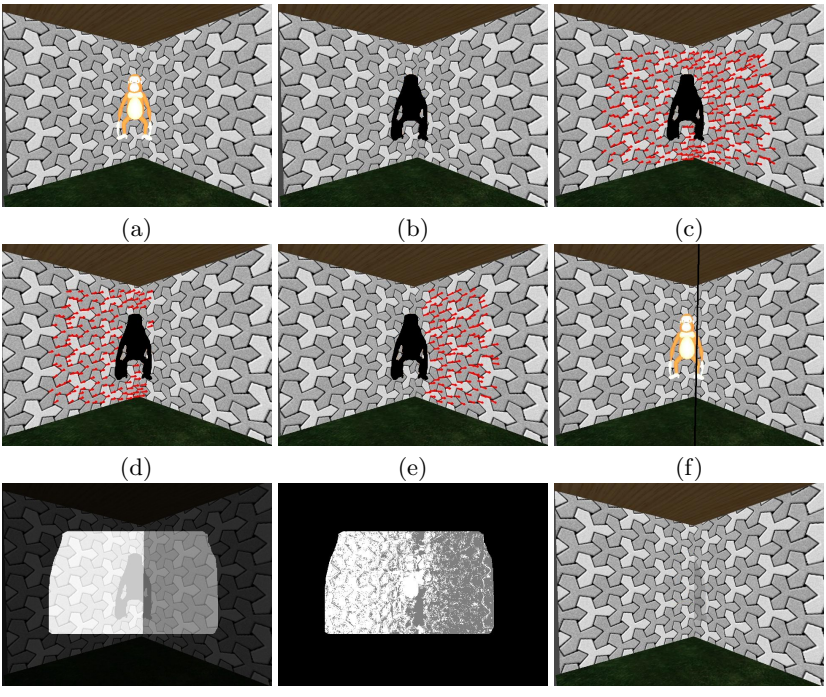


Fig. 3. Intermediate outputs at the various stages of the algorithm (a) Input image (second frame is not shown) (b) The object to be removed is masked out and region is shown in black (c) Sparse optical flow vectors on the image (shown in red, in twice the original size to make them visible) (d,e) First and Second dominant motion vectors clustered respectively (f) Line of intersection of the two planes calculated as detailed in Section 3.2. (g) The surrounding background of the region is segmented into two planes (h) Output of graph cuts based binary partitioning of the segments, shown for comparison (i) The results of the completion on this frame.

H_2 be the homography due to π_1 and π_2 between the two views. We find the *generalized eigenvectors* of the pair (H_1, H_2) by solving the equation,

$$H_1 v = \lambda H_2 v.$$

The eigenvectors obtained have the property that two of them are the projections of two points on the line of intersection of the two planes π_1, π_2 on to the image plane I_1 and third one is the epipole in the image I_1. The two eigenvectors corresponding to the points on the plane can be identified due to the equality of their corresponding eigenvalues. The reader is referred to Johansson [20] for a proof of this fact.

Using the homogeneous coordinates of the two points on the image plane, we can obtain the exact line of intersection in the image. In fact we need this line only over the ROI. Thus, we have the planar layers for the ROI. We warp

these layers in the neighboring frame to the frame to be fill-in the unknown region. The correspondence between layers obtained in two views is established by measuring the percentage of the tracked points that are part of the layer in previous frame. In the ongoing discussion we use the word *label* of a pixel to refer to the layer assigned it. Figure 3. (f) shows a line obtained by this method, (g) shows the plane segmentation in the ROI which is defined by the line.

The correctness of the line determined using the method needs to be ensured as small errors in homography calculation can lead to high errors in line determination. In fact the homography pair may have complex generalized eigenvalues and eigenvectors and may not yield a valid pair of points to obtain the line. We validate the correctness of the boundary line by ensuring that it partitions the correspondence pairs into different clusters depending on the homography to which they belong. In case the line is not determinable or validation fails we obtain the line from a neighboring frame where it was detected and verified by applying the underlying homography.

It should be noted that the methods which give good results for dense motion segmentation from multiple views are not suitable for segmentation of the frames with the missing region. Graph cuts based motion segmentation techniques [15,21] determine the dominant motion models in the scene and assign each pixel to one of the motion model based on an optimal graph cuts segmentation. The unknown pixel can never be accurately assigned to any particular label in these approaches due to lack of both color and motion information, which are used for determining the weights in the graph. We show the result of applying binary graph cuts partitioning in Figure 3(f), to illustrate this fact. We only apply a binary labeling in the graph, the white region shows points supporting first dominant motion and gray region shows points supporting second dominant region. Grey region of the image was not considered for the segmentation stage. Similarly methods like [14,13] which assign the pixels to the motion model or planes based on re-projection error measure can not assign the unknown pixels to any particular layer accurately.

3.3 Layer-Wise Video Completion

The line dividing the two planes gives a single confident label to each pixel in the ROI. Once the label is determined we can fill the hole by warping the nearby frames according to the homography related to the label. We build the mosaic of each plane using the neighboring frames. The missing pixels are assigned the color from the mosaic of the plane correspondence to their label. This method is in principle similar to the layered mosaic approaches [9,12]. The difference is that we have exact knowledge of which plane an unknown pixel belongs to and use only that corresponding plane (layer). The blending of homographies of multiple layers is not needed. As in case of layered mosaic approaches the intensity mismatch might occur due to combination of various frames, simple blending methods could be applied to circumvent the error due to this.

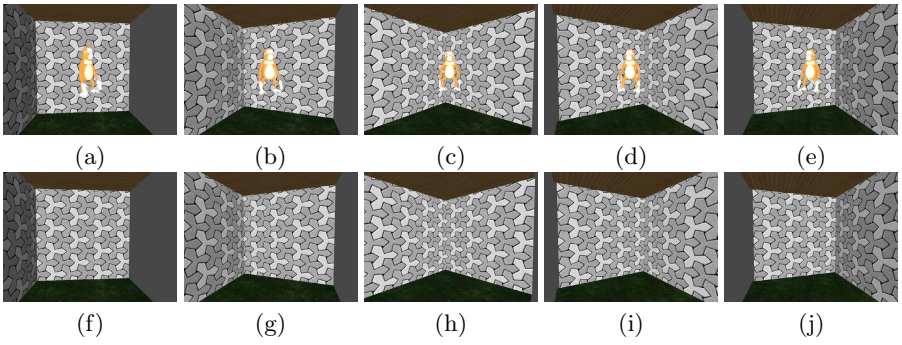


Fig. 4. The process applied on a synthetic sequence. (a-d) show the five frames of the sequence. (e-h) show the frames after completion. The monkey is removed from the original video. (a,e) have only one background plane, while in (b,c,d) two planes are present in the background.

3.4 Inpainting

Some pixels may remain unknown after the layer-wise video completion due to absence of the information in the video. Pixels which are always covered by the object to removed belong to this set. As in case of image inpainting techniques we can only approximate the values of these pixels based on the surrounding information. The extra information however is the knowledge of which plane the pixel belongs to. We can restrict the filling algorithm to use values only from the corresponding plane.

4 Results

We demonstrate the application of our approach on two sequences. Figure 4 shows the results of our algorithm on a synthetic sequence. The sequence is set in a room with two wall, a roof and a ceiling i.e. four planes. Our approach removes the monkey as shown in the figure. Due to intensity difference on the wall during the motion the mosaicing of the wall over the views generate some intensity seams. Simple blending applied during the mosaic construction gives much better results. No application of inpainting was needed in this sequence.

Figure 5 demonstrate the result of the technique applied to a real sequence. Some black holes are present in the output due to unavailability of data. Inpainting is not being applied on the sequence as it is neither structure rich nor texture rich. Seams which are visible in the results can be removed by applying some blending approach.

The algorithm takes around 2 seconds per frame for the motion segmentation and plane matching step the completion step is dependent on number of neighboring frames used for creating the mosaic and takes around 1-2 seconds when 12 (6 forward and 6 backward) frames are used.

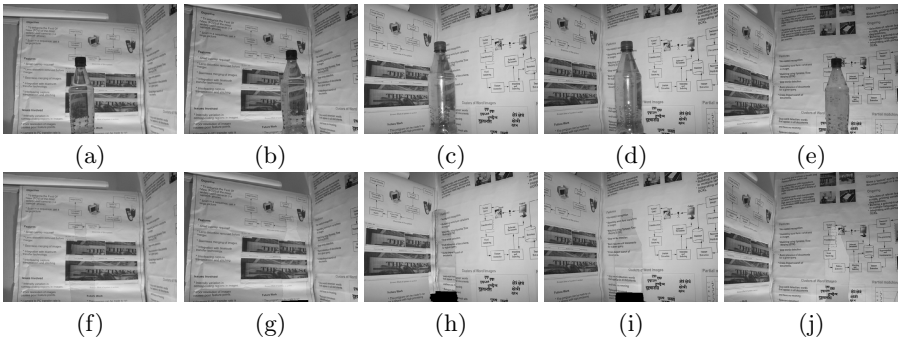


Fig. 5. The process applied on a real sequence, we remove the bottle from the video (a-e) shows five frames of the sequence. (f-j) shows the results of video completion algorithm on each input frame. Initial and final frames have only one background while frames in the middle have two background planes. The output has visible seams at the junction of the removed object due to very high intensity change in the scene.

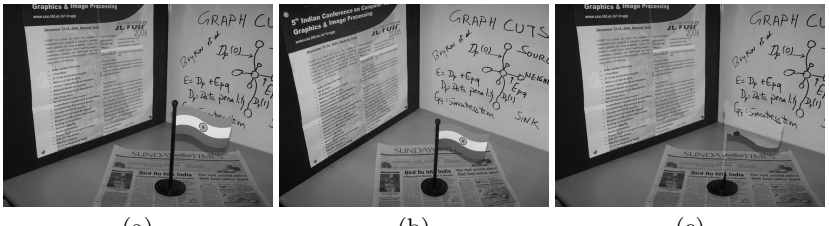


Fig. 6. Application of our approach to images. (a,b) two views of the scene containing 3 different background planes. (c) Image (a) is filled-in using information from image (b) to remove the hole created due to the removed flag. Note that the shadow of the flag is present in the completed image as shadow region was not selected for removal.

Our method can also be used for object removal in pairs of images. We demonstrate a simple example of this in Figure 6. The background of the flag object has three planes. Motion estimation gives us three different motion models. The intersection line is obtained for each pair of planes and used in same way as described as for videos for layer-wise completion of the unknown region. We used an affine region matching to determine the point correspondences as the inter-frame motion was large in this case. There is also significant change in illumination between the views, which is apparent after the flag is removed and the image is completed. Both images didn't see table in the region near the flag and in the region containing the flag's shadow. Thus, that information could not be filled in.

5 Conclusions and Future Work

In this paper, we address the problem of video object removal and completion for indoor scenes. Our method involves user interaction only for object selection and performs the rest of the operations without any user interaction. Ours is an attempt to use multiview information for scene inference and video completion. We showed results on scenes with piecewise planar background near the object to be removed. The technique can be easily extended to more planes as long as the dominant motion segmentation can be achieved.

The geometric information we used give better segmentation of multiple motions. The motions are segmented at the pixel level without region growing or interpolation, unlike the motion segmentation performed in the image space. Motion inpainting methods can work well for scenes with a multiple planes or non-textured surfaces. Combining the geometric information with motion inpainting will be the most promising one for scenes with multiple planes. We propose to investigate the problem further in that direction.

References

1. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques, New York, NY, USA, ACM Press/Addison-Wesley Publishing Co. (2000) 417–424
2. Efros, A., Leung, T.: Texture synthesis by non-parametric sampling. In: IEEE International Conference on Computer Vision, Corfu, Greece (1999) 1033–1038
3. Efros, A., Freeman, W.: Image quilting for texture synthesis and transfer. Proceedings of SIGGRAPH 2001 (2001) 341–346
4. Rother, C., Kolmogorov, V., Blake, A.: "grabcut": Interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. **23** (2004) 309–314
5. Bertalmio, M., Vese, L., Sapiro, G., Osher, S.: Simultaneous structure and texture image inpainting. In: CVPR03. (2003) II: 707–712
6. Criminisi, A., Perez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE transactions on Image Processing **13** (2004) 1200–1212
7. Kang, S., Chan, T., Soatto, S.: Landmark based inpainting from multiple views. Technical report, UCLA Math CAM (2002)
8. Wexler, Y., Shechtman, E., Irani, M.: Space-time video completion. In: CVPR04. (2004) I: 120–127
9. Jia, J., Wu, T., Tai, Y., Tang, C.: Video repairing: Inference of foreground and background under severe occlusion. In: CVPR04. Volume 1. (2004)
10. Kokaram, A., Collis, B., Robinson, S.: A bayesian framework for recursive object removal in movie post-production. In: ICIP03. (2003) I: 937–940
11. Matsushita, Y., Ofek, E., Tang.X., Shum, H.: Full frame video stabilization. In: CVPR05. (2005)
12. Zhang, Y., Xiao, J., Shah, M.: Motion layer based object removal in videos. WACV/Motion **01** (2005) 516–521
13. Vincent, E., Laganiere, R.: Detecting planar homographies in an image pair. In: Symposium on Image and Signal Processing and Analysis (ISPA01). (2001)

14. Fraundorfer, F., Schindler, K., Bischof, H.: Piecewise planar scene reconstruction from sparse correspondences. Image and Vision Computing (2006)
15. Wills, J., Agarwal, S., Belongie, S.: What went where. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. (2003) 37–44
16. Jain, V., Narayanan, P.: Layer extraction using graph cuts and feature tracking. In: In Proceedings of the third, International Conference on Visual Information Engineering. (2006) 292–297
17. Li, Y., Sun, J., Shum, H.: Video object cut and paste. ACM Trans. Graph. **24** (2005) 595–600
18. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communication of the ACM **24** (1981) 381–395
19. Hartley, R.: In defence of the 8-point algorithm. In: ICCV '95: Proceedings of the Fifth International Conference on Computer Vision, Washington, DC, USA, IEEE Computer Society (1995) 1064
20. Johansson, B.: View synthesis and 3d reconstruction of piecewise planar scenes using intersection lines between the planes. IEEE International Conference on Computer Vision **1** (1999) 54–59
21. Bhat, P., Zheng, K., Snavely, N., Agarwala, A., Agrawala, M., Cohen, M., Curless, B.: Piecewise image registration in the presence of multiple large motions. In: CVPR06. (2006) II: 2491–2497