Proceedings of the International Conference on Electrical and Computer Systems
Ottawa, Ontario, Canada, 22-24 August 2012
Paper No. 191

Object Removal in Multi-view Images using Disparity Layering and Piecewise Homography

Tan H. Ho, Richard M. Dansereau
Carleton University, Dept. of Systems & Computer Engineering
1125 Colonel By Drive Ottawa, Ontario K1S 5B6
tanhuyho@yahoo.com; rdanse@sce.carleton.ca

Eric Dubois
University of Ottawa, School of Electrical Engineering and Computer Science
800 King Edward Avenue, Ottawa, Ontario, Canada K1N 6N5
edubois@uottawa.ca

Abstract - In this paper, object removal in multi-view images is considered where occluded image information is synthesized from other views through disparity estimation to build layered depth maps and then a piecewise homography is used to transform occluded information from other views. Experimental results show improved performance over single homography transforms in synthesizing the occluded image information after the object is removed.

Keywords**:** Object Removal, Piecewise Homography, Disparity Layering, Multi-View

1. Introduction

The topic of view synthesis is concerned with constructing a new view based on a number of images taken from different viewpoints. In this application, it is possible that some of these images contain unwanted objects which should be removed in the constructed new view. To achieve this, information from those images which either do not contain the unwanted objects or are only partially occluded can be used to remove the unwanted objects in the new view. Another application is virtual navigation through a scene where unwanted occluding objects can be removed. This paper focuses specifically on removing an object from the target image, replacing it with information from a set of reference images, taken from different viewpoints, which contain the occluded image information.

Through our survey of similar works, Bhavsar and Rajagopalan (2010) propose a method for inpainting both image and depth of a scene using multiple stereo images. The method exploits the fact that the information missing in some images may be present in other images due to the motion cue. Criminisi et al. (2004) propose an algorithm for removing large objects from images. The method combines advantages of both texture synthesis algorithms for generating large image regions from sample texture, and inpainting techniques for filling in small image gaps to perform exemplar-based image inpainting. Lee et al. (2008) propose an inpainting algorithm for multi-view video sequences. The method assumes a number of frames taken simultaneously by spatially separated cameras and that the multiple cameras are arranged in parallel. By using disparity estimation and block matching, the information divided into grids from the reference frame is used to paste into the desired frame to remove the object. Patwardhan et al. (2005) propose an inpainting technique for removing an object and filling in the background information of a video sequence. This method does not make use of multiple views. Zitnick et al. (2004) describe a system for high-quality view interpolation between relatively sparse camera viewpoints.

To the best of our knowledge, our proposed method for object removal is novel in that it uses multi-view images and makes use of disparity layering and piecewise homography to fill in for the occluded region. The techniques proposed by Criminisi et al. (2004) and Patwardhan et al. (2005) are concerned

with filling in the missing region(s) using information from neighbouring pixels not occluded by the unwanted objects. In both of these methods, no other images from different viewpoints are used. The techniques proposed by Bhavsar and Rajagopalan (2010) and Lee et al. (2008) make use of information from other images from different viewpoints to fill in the missing region(s). Bhavsar and Rajagopalan (2010) constructs a depth map of the image and considering the depth map information, extracts pixel information from the reference image to be used for inpainting and filling in the missing region(s) in the desired image. Lee et al. (2008) divides the image into grids and by using block matching between the images, the missing information grids from one image are cut-and-pasted into the desired image to remove the unwanted object. The proposed approach for object removal described in this paper is similar to Bhavsar and Rajagopalan (2010), and Lee et al. (2008) in that it uses multiple images from different viewpoints, and disparity information. Unlike Bhavsar and Rajagopalan (2010) which is an inpainting technique and fills in the missing regions at the pixel level, and unlike Lee et al. (2008) which uses block matching of grid regions for cut-and-paste into the desired image, the proposed method removes large object from an image using piecewise homography with the disparity layering to fill in multiple segments of the occluded region.

The paper is structured as follows. Section 2 describes the proposed novel approach for object removal with piecewise homography and disparity layering. The simulation results and analysis are presented in Sec. 3. Finally, conclusions from the current work are presented in Sec. 4.

2. Proposed Approach for Object Removal

Our proposed approach for object removal makes uses of multi-view images using piecewise homography and disparity layering. In this paper, the target image is defined as the image with the unwanted object to be removed. The desired region is defined as the image region in the target image which is being occluded by the unwanted object. The reference images, taken from different camera positions, are images that do not have the unwanted object occluding the desired region and from which image information is to be extracted and used to fill in the desired region occluded by the unwanted object in the target image. The ground truth target image without the unwanted object is also available in our testing phase for the purpose of computing the structural similarity (SSIM) index for the resulting image for performance evaluation. Our approach assumes that there are at least two reference images available, the scene is static, the images can be colour or greyscale, lighting condition between the target image and the reference images are relatively similar, and the baseline between the images is moderate to ensure relatively dense corresponding point matching.

The fundamental concept involves three main processing steps: (*i*) finding corresponding points, (*ii*) establishing disparity layering, and (*iii*) inserting occluded image information with piecewise homography transformations. The flow diagram of the overall approach is shown in Fig. 1. Fig. 1 shows the simplistic case which has a target image and only two reference images, where reference image #1 is used to extract image information to fill in for the occluded desired region in the target image, and reference image #2 is used to provide the required information to allow the segmentation of the desired region into sub-regions for piecewise homography as explained next. First, for the reference#1-reference#2 image pair, corresponding points (CPs) are determined and the CPs are clustered into different disparity layers, and similarly the same is done between the target-reference#1 image pair. With the CPs from the target-reference#1 image pair, an approximate homography matrix is computed. Using this approximate homography matrix and the disparity layering result from the reference#1-reference#2 image pair, the desired region is segmented into sub-regions. For each sub-region, the optimal CP cluster from the target-reference#1 image pair is selected. Then for each sub-region, the chosen CP cluster is used to compute the homography matrix to perform piecewise homography to transform the sub-regions from reference image #1 to the target image. The transformed sub-regions are then pasted into the target image to produce the resulting image with the unwanted object removed. The SSIM index is then computed between the ground truth image and the resulting image to assess the quality of the final image. The following sub-sections describe the key steps in more detail.

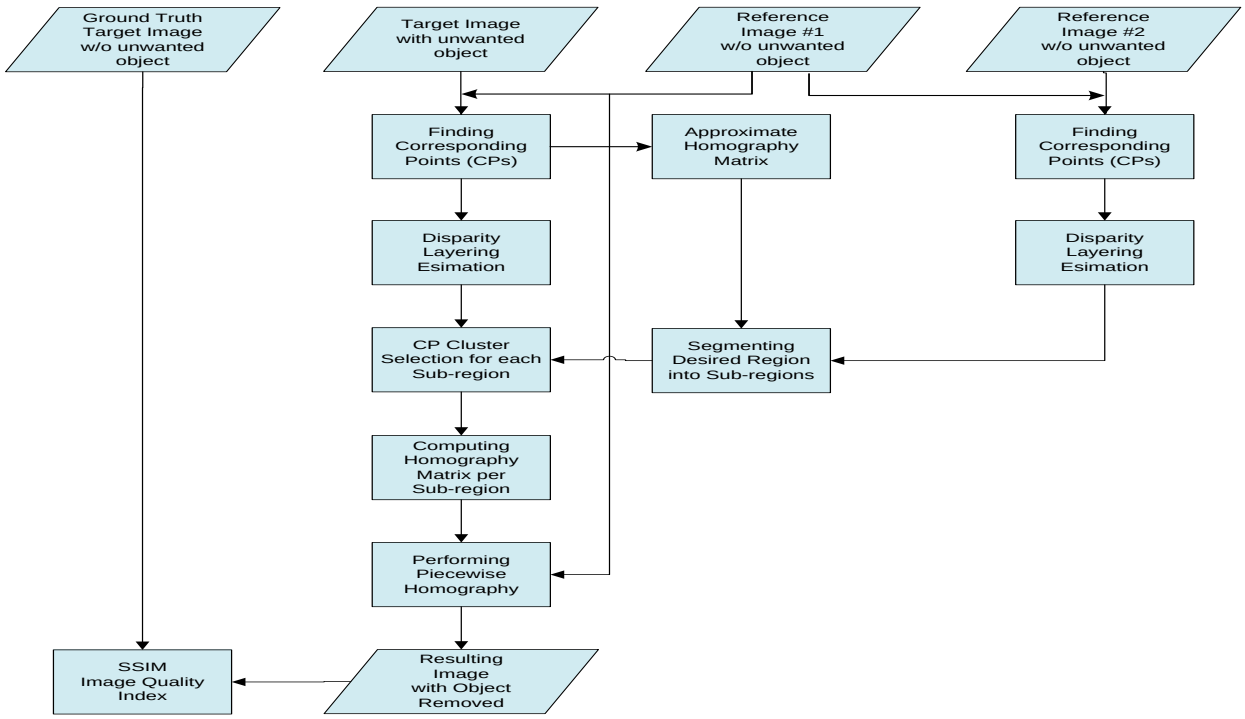


Fig. 1 Information flow diagram of the proposed approach

2.1. Finding Corresponding Points

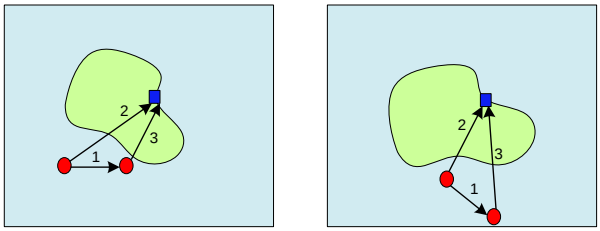
In this study, affine-SIFT (ASIFT), as proposed in Morel and Yu (2009), is used for finding CPs. ASIFT has been demonstrated to produce dense corresponding points between stereo images with wide baseline. The CPs between the images in our research are obtained using the demo software provided at Web-1.

2.2. Disparity Layer Estimation

Since objects within the image can have different depths from the camera, building up a map of depth layers is needed to allow the separation of CPs into clusters. These depth layers are determined based on disparity differences between image pairs. This is essential because a homography transform using CPs with different disparity will result in significant image distortion, since a homography matrix can deal with only affine transforms and not perspective transforms. In addition, the CP clusters provide information on how the desired region should be optimally segmented for piecewise homography. This concept is explained in the next section. Disparity layering is required to be done for the target-reference#1 image pair, and for the reference#1-reference#2 image pair.

The disparity layering of CPs between two images is done by first computing the vector in pixel-coordinate between each point to every other point within the first image. Each point with one other point forms a pair; therefore, for N points in the image, there are $(N-1)$ x $(N-2)$ pairs. Next, the same is done for all pairs of points in the second image. Fig. 2(a) and (b) show the two images and are used to illustrate the concept of disparity layering and CP clustering. For each pair of points in the first image there is a corresponding pair of points in the second image (i.e., due to the fact that the points in the first image are corresponding points to those in the second image). In Fig. 2(a), the first image, there are three pairs of points with the three vectors shown as 1, 2, and 3. These corresponding pairs of points and vectors are shown in Fig. 2(b) which is the second image. Then, for each corresponding vectors from the two images the vector magnitude difference is computed. For a particular pair, the greater the vector magnitude difference, the greater the disparity difference is between these two points. If the vector magnitude difference is less than a predefined threshold, then these two points are determined to be in the same CP cluster or disparity layer. As shown in Fig. 2, the vector magnitude difference between vector

#1 in the first image and the corresponding vector #1 in the second image is small (i.e., less than the predefined threshold), then these two points indicated as red circles are determined to be in the same disparity layer. On the other hand, the vector magnitude difference for vector #2 or vector #3 shown is more significant, and thus, the blue-square point is determined to not belong to the same disparity layer as the red-circle points.

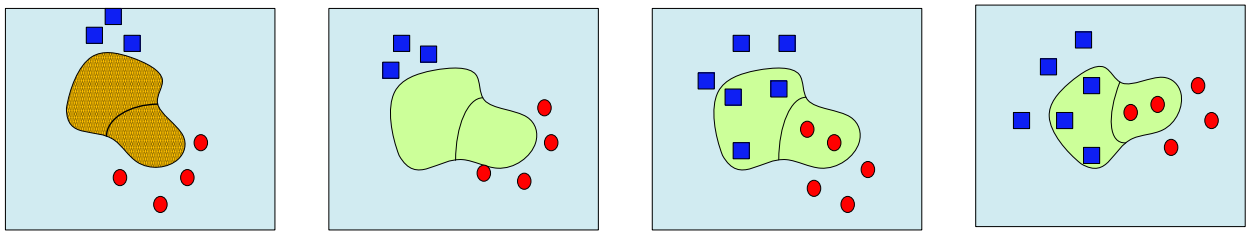


(a) (b)

Fig. 2 Disparity layering and CP clustering

2.3. Desired Region Segmentation

The desired region in the target image which is occluded by the unwanted object could consist of image sub-regions with different disparity. In this case, it is essential that the desired region is segmented into sub-regions according to the disparity layers so that an optimal homography matrix can be used for each sub-region. The process of segmenting the desired region into sub-regions is described next.



(a) Target Image (b) Reference Image #1 (c) Reference Image #1 (d) Reference Image #2

Fig. 3 (a) and (b) are CP clusters in Target-Reference#1 image pair, and (c) and (d) are CP clusters in Reference#1-Reference#2 image pair

As an example, Fig. 3(a) and (b) show the two CP clusters in the target-reference#1 image pair; (c) and (d) show the two CP clusters in the reference#1-reference#2 image pair. In order to identify the desired region in reference image #1, an approximate homography matrix relating the target image to the reference image #1 is computed from only CPs located around the desired region in the target image. (It is considered an approximate homography matrix because the CPs used in the computation could have different disparity; however, by using CPs located around the desired region only, the probability is higher that these CPs are similar in disparity as compared to points within the desired region.)

Fig. 3(a), the target image, shows the unwanted object as the hashed orange region. Fig. 3(b), (c), and (d), which are the reference images, show in green the desired region which is occluded in the target image. Since the unwanted object is not present in the reference image #1, there are no CPs detected in the desired region for the target-reference#1 image pair. Therefore, the CP clusters in the reference#1-reference#2 image pair are used to provide information about the disparity layers within the desired region. As shown in Fig. 3(c) and (d), the two CP clusters in the desired region indicate that the desired region should be segmented into two sub-regions as shown. The next step is to perform piecewise homography which is described in the next section.

2.4. Piecewise Homography

Piecewise homography is needed in the case that the desired region consists of multiple sub-regions with different disparity. In this case, if a single homography matrix is used to transform the entire desired region from the reference image to the target image, the resulting transformed desired region will be significantly distorted. The idea of piecewise homography is to compute a unique homography matrix for each sub-region. The optimal CP cluster to be used in the computation of the homography matrix for a particular sub-region contains those points which are in the same disparity layer as that of the sub-region. The method for selecting the optimal cluster of CPs is described next.

As discussed in the previous section, Fig. 3(c) and (d) show the segmentation of the desired region into two sub-regions. In Fig. 3(c), the first sub-region on the left is in the same disparity layer as the blue-square CP cluster according to the location of these CPs. When comparing Fig. 3(b) and (c), which are the same reference image, it can be seen that the blue-square CPs in Fig. 3(c) are in the same disparity layer as the blue-square CPs in Fig. 3(b) according to the location of these CPs. Therefore, the blue-square CP cluster is chosen to be used to compute the homography matrix for this first sub-region. A similar method is applied for the second sub-region on the right and the red-circle CP cluster is chosen. Once the CP cluster is chosen for a particular sub-region, the process of performing the piecewise homography is described next.

Using the selected CP cluster, the homography matrix relating the reference image #1 to the target image is computed for each sub-region. The transformed sub-regions are then pasted into the target image to remove the unwanted object.

3. Simulation Results and Analysis

This section presents the preliminary results of our proposed object removal technique on a target image with a vehicle that we wish to remove from the scene, and a second scene with some people to be removed. In order to provide a quantitative assessment of the performance of the object removal technique, the structural similarity (SSIM) index is used. The target image without the unwanted object is available to provide the ground truth for the computation of the SSIM index. The SSIM index is computed only for the desired region and not on the entire image in order to avoid bias in the case that the desired region is much smaller than the entire image.

In terms of the computing environment used, the ASIFT CPs are first obtained by uploading images to Web-1 and using the provided demo software to provide the CPs – the results are available typically within 10-15 seconds. The CPs are processed by our Matlab algorithm running on a modest computer (HP Elitebook, Duo CPU at 2.53GHz with 4GB SRAM). The typical execution time of the Matlab algorithm with approximately 1000 CPs and three piecewise homography operations take approximately 10-15 seconds. The processing time increases to about 25 seconds for 3000 CPs and three piecewise homography operations. Each piecewise homography operation requires about one third of this total 25 second time. The processing time increases approximately linearly with the number of piecewise homography operation.

Fig. 4 and Fig. 5 show the results for the scene with the vehicle to be removed. Fig. 4 shows the CP clusters after the disparity layering step for the target-reference#1 image pair, and for the reference#1-reference#2 image pair. Each CP cluster is shown using different colours and symbols. According to our proposed approach, from Fig. 4(c) it can be seen that the desired region consists of two CP clusters, green CP cluster and red CP cluster, which subdivide the desired region into two sub-regions (i.e., top and bottom). Therefore, the desired region is segmented accordingly to two sub-regions. From Fig. 4(b) and (c), the top sub-region is determined to be in the same disparity layer as the green CP cluster. For the bottom sub-region, it is unfortunately not obvious from this result which CP cluster to use since the red CP cluster in Fig. 4(c) does not coincide any of the CP clusters in Fig. 4(b). However, from visual observation, it is decided that the red CP cluster (white CP cluster can also be used) in Fig. 4(b) will be used since it is observed to be approximately in the same disparity layer as the bottom sub-region. It is expected that if this data set contains more distinct features then the ASIFT algorithm would be able to find a denser set of corresponding points to allow the proposed method of selecting optimal CP cluster for

the bottom sub-region to work. The disparity layering process eliminates erroneous corresponding points which significantly helps to reduce image distortion since these erroneous CPs are not used in the computation of the homography matrix.

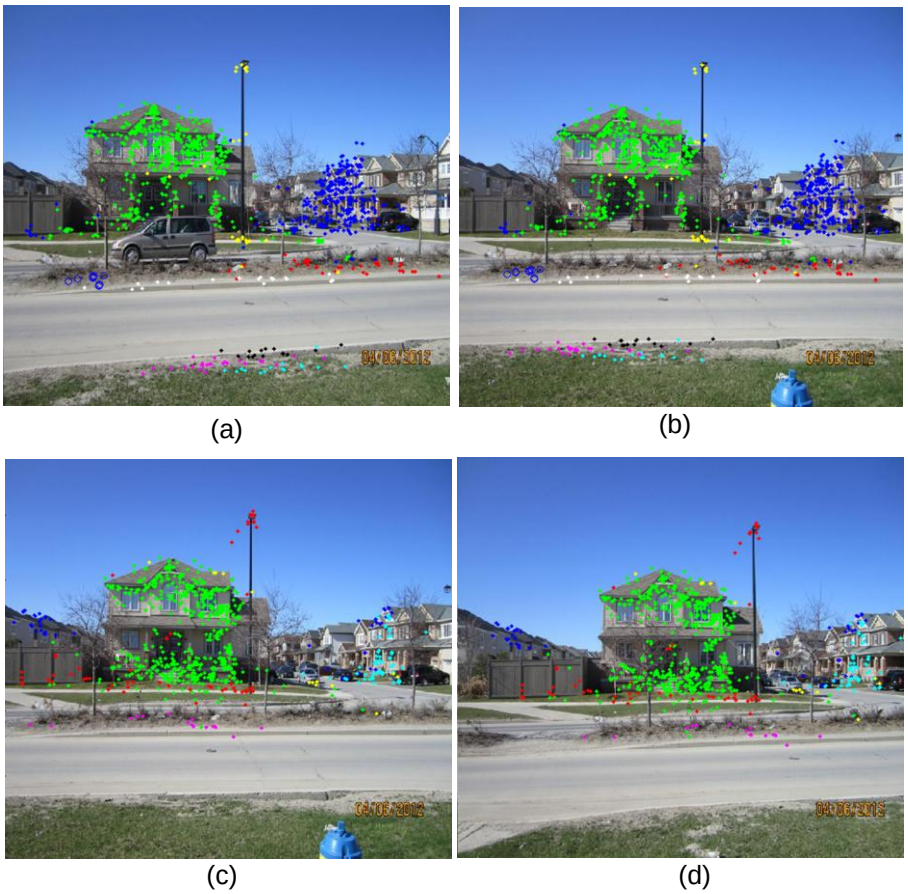
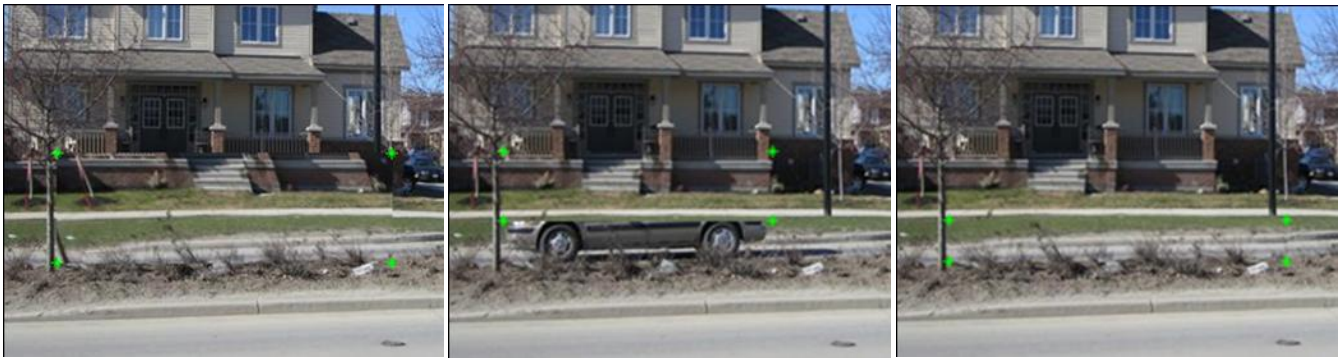


Fig. 4 (a), (b) are CP clusters in Target-Ref#1 image pair, and (c), (d) are CP clusters in Ref#1-Ref#2 image pair

Fig. 5(a) shows the final resulting image using the single homography matrix. For visual purposes, the desired region is marked with asterisks. The SSIM index is 0.85 – an image with index of less than 0.90, typically, has very visible artefacts. From visual inspection, it can be seen that the transformed desired region in the target image is significantly distorted. Fig. 5(b) and (c) show the resulting image using the piecewise homography matrix with (b) being the image after the first sub-region is removed and (c) being the final image with both sub-regions removed. The SSIM index is 0.97 – an image with index of greater than 0.95, typically, is very good from visual inspection with non-obvious artefacts. Therefore, the index of 0.97 is considered significantly better than the case using the single homography matrix. Note that the shadow cast by the vehicle as seen at the bottom and rear of the vehicle is also removed.



(a) (b) (c)

Fig. 5 (a) Final image with object removed using single homography matrix, (b) image with first sub-region removed, and (c) image with both sub-regions removed

Fig. 6 and Fig. 7 illustrate the results for the Alysa-Daniel scene. Fig. 6(a) and (b) show the CP clusters in the target-reference image pair, and Fig. 6(c) shows the final image with the objects removed using a single homography matrix. The obvious artefacts in the resulting image are indicated with red arrows. Fig. 7 shows the final image where the objects are removed using our proposed approach. The artefact in Fig.7(c) indicated with the red arrow is present due to the fact that there are no corresponding points detected which have the same disparity as this part of the image.

(a) (b) (c)

Fig. 6 CP clusters in target-reference image pair for Alysa-Daniel scene (a) target image, (b) reference image, and (c) final image with objects removed using single homography matrix

(a) (b) (c)

Fig. 7 (a) Image with first part removed, (b) image with second part removed, and (c) image with third part removed

From this work, we observe that the denser the number of corresponding points the better our proposed method works overall since more CPs allow more granularity in disparity layering, more accurate segmentation of the desired region, more accurate selection of optimal CP cluster for each sub-region, and more accurate piecewise homography matrix. However, the wider the baseline between the target image and the reference image the more difficult it is for the image registration algorithm (i.e. ASIFT in our case) and thus the fewer the number of corresponding points.

From the data set presented in this paper, it can be observed that the desired region (i.e. the region occluded by the van) is a complex background with many features which are not possible to be derived from conventional inpainting and texture synthesis approaches.

5. Conclusion

The work presented in this paper has shown promising results in using multi-view images with disparity layering and piecewise homography for object removal. This approach allows the removal of small or large objects which occlude complex background consisting of sub-regions with disparity difference. The conventional inpainting and texture synthesis approaches would not be able to deal with this type of scenario.

In this paper we use the simple rectangular area for selecting the desired region, and the sub-regions. A more sophisticated contour area will be implemented in the future work allowing more sophisticated segmentation of the desired region. In addition, we present the case in which only two reference images are used and that these reference images do not contain the unwanted object. In future work, we will consider the case where more reference images are used and that these reference images may contain the unwanted object, but that the unwanted object in different reference image could partially occlude different parts of the desired region either due to the motion of the unwanted object or due to the viewpoint difference. Therefore, by using multiple reference images all parts of the desired region can be extracted and used to removed the unwanted object in the target image. Also, instead of relying on a cut-and-paste process for the entire sub-region, in future work we will also consider processing at a pixel level where necessary.

References

Bhavsar A. V., Rajagopalan A. N. (2010). Inpainting in Multi-image Video, "Proceedings of the 32nd DAGM Conference on Pattern Recognition."

Criminisi A., Perez P., Toyama K. (2004). Region Filling and Object Removal by Exemplar-Based Image Inpainting, "IEEE Transactions on Image Processing," Vol. 13, Issue 9.

Morel J.M., Yu G. (2009). ASIFT: An Algorithm for Fully Affine Invariant Comparison, "SIAM Journal on Imaging Sciences," 2(2): 438-469.

Lee S.Y, Heu J.H., Kim C.S., Lee S.U.. An Object Removal with Multi-view Sequence Inpainting Technique (2008). "International Conference on Intelligent Information Hiding and Multimedia Signal Processing."

Patwardhan K. A.., Sapiro G., Bertalmio M. (2005). Video Inpainting of Occluding and Occluded Objects, "IEEE International Conference on Image Processing," Vol. 2.

Zitnick C. L., Kang S. B., Uyttendaele M., Winder S., Szeliski R. (2004). High-Quality Video View Interpolation Using a Layered Representation, "ACM Transactions on Graphics (TOG) -- Proceedings of ACM SIGGRAPH," Vol. 23, Issue 3.

Web sites:

Web-1: http://www.ipol.im/pub/algo/my_affine_sift/, consulted 1 Apr. 2012.