REGISTRATION AND OCCLUSION DETECTION IN MOTION BLUR

Abhijith Punnappurath [1] *A. N. Rajagopalan* [1] *Guna Seetharaman* [2]

[1] Department of Electrical Engineering, IIT Madras, Chennai, India
[2] Information Directorate, AFRL/RIEA, Rome NY, USA
jithuthatswho@gmail.com, raju@ee.iitm.ac.in, guna@ieee.org

ABSTRACT

We address the problem of automatically detecting occluded regions given a blurred/unblurred image pair of a scene taken from different viewpoints. The occlusion can be due to single or multiple objects. We present a unified framework for detecting occluder(s) that is reasonably robust to non-uniform motion blur as well as variations in camera pose (without the need for deblurring). We assume that the occluded pixels occupy only a relatively small area and that the camera motion trajectory is sparse in the camera motion space. We validate the performance of our algorithm with experiments on synthetic and real data.

Index Terms— Occlusion, non-uniform blur, registration, sparsity

1. INTRODUCTION

Detecting occluded regions in images is an extensively studied problem in image processing and computer vision due to its applicability to a vast range of areas such as tracking, surveillance, object recognition, inpainting [1], [2], [3], [4], [5] etc. The objective, in a typical setting, is to automatically detect occlusion(s) given a pair of images taken from different view points and at different times. The occlusions themselves may have been caused by the entry or disappearance of objects in the scene within the time-span of the two observations. A common approach is to first compensate for the variations in pose by registering the two images with respect to each other followed by differencing to reveal changes in the scene. For small occlusions, the images can be aligned even using standard registration techniques [6], [7] that do not account for occlusions. This is because the matching of unoccluded pixels can be expected to sufficiently outweigh any possible degradation arising from attempting to match occluded pixels with unoccluded pixels and vice versa. However, larger occlusions warrant methods that detect the occluded pixels and exclude them from the registration process [8]. This challenging problem of detecting occlusions becomes even more ill-posed if one of the images in the pair is blurred due to the presence of camera shake. This is often the case when a quick fly-through is attempted for the recoverage of a particular geographic area, for which detailed surveillance images (i.e., latent images) are already available. Moreover, if the revisit is made at a time when the luminance is weak [9], then the exposure time needs to be increased, thereby increasing the chances of motion blur. Detecting occlusions is important for revealing changes in infrastructure, deployment of military units, modification/introduction of equipment etc. As pointed out in [10], traditional registration methods such as direct and feature-based approaches cannot be used in such a case due to photometric inconsistencies introduced by the blur. The alignment approach presented in [10] is based on the convolution model and applies only to the restrictive uniform blur case. However, in the case of general camera motion, the blur incurred can be significantly non-uniform across the image, and a space-varying formulation becomes necessary to describe the blurring process [11], [12], [13], [14]. It is this compounded scenario that we address in this work as depicted in Fig. 1. Note that there can be more than one occluder.

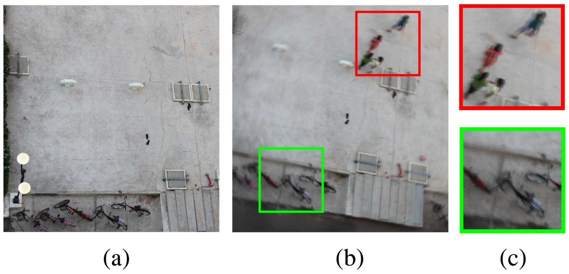


Fig. 1. An aerial view of a parking lot (best viewed electronically). (a) Latent (unblurred) image, (b) blurred and occluded observation taken from a different view point, and (c) zoomed-in regions from Fig. 1(b) showing the presence of non-uniform blur.

While conventional approaches to detecting occluder(s) would require one to follow the deblur-register-difference pipeline, we present a unified framework which directly solves for the occluder(s) by accounting for the non-uniform blur and the changes in camera pose given a blurred/unblurred image pair. We show that direct registration of the pair is possible without the need for deblurring. Registration, in this context, tantamounts to estimating the set of warps which when applied on the focused image aligns it with the blurred image in the region of overlap. The elegance of our method lies in the fact that registration and occlusion detection turn out to be a natural fallout of our blur estimation process. We assume that the occluded pixels occupy only a relatively small portion of the image and that the camera motion trajectory is sparse in the camera motion space. We also assume the scene to be sufficiently far away so that depth variations can be ignored. We use a multiscale approach in which the image resolution is varied from coarse-to-fine, thus rendering the algorithm efficient both in terms of computational time and memory requirements.

Summary of contributions:

1. We propose a method for registering a blurred/unblurred image pair in the presence of non-uniform blur.
2. The approach is direct in the sense that it can detect occlusions without deblurring. It is robust to motion blur and variations in camera pose.
3. We derive an objective function with sparsity constraints that when minimized yields the location and the intensity of the occluded pixels as well as an estimate of the camera motion.

2. PROJECTIVE MOTION BLUR MODEL

In this section, we briefly review the non-uniform blur model for a far-away planar scene. The projective model discussed in [11], [12], [13], [14] assumes that the blurred image is the weighted average of warped instances of the latent image. In the discrete domain, this can be represented as

$$b\,(i,j) = \sum_{k\in\mathbf{T}} \omega\,(k)\, l\,(\mathcal{H}_k\,(i,j)) \qquad (1)$$

Here $l(i,j)$ denotes the latent image of the scene, $b(i,j)$ is the blurred observation, and $\mathcal{H}_k(i,j)$ denotes the image coordinates when a homography $\mathcal{H}_k$ is applied to the point (i,j). The parameter $\boldsymbol{\omega}$, also called the *transformation spread function* (TSF), depicts the camera motion, where $\omega(k)$ denotes the fraction of the total exposure duration for which the camera stayed in the position that caused the transformation $\mathcal{H}_k$. The TSF $\boldsymbol{\omega}$ is defined on the discrete transformation space $\mathbf{T}$ which is the set of sampled camera poses. The transformation space is discretized in such a way that the difference in the displacements of a point light source due to two different transformations from the discrete set $\mathbf{T}$ is at least one pixel. Akin to a PSF, $\sum_{k\in\mathbf{T}} \omega\,(k) = 1$.

The homography $\mathcal{H}_k$ in equation (1) in terms of the camera parameters is given by

$$\mathcal{H}_k = K_v \left(R_k + \frac{1}{d_0} T_k [0 \;\; 0 \;\; 1] \right) K_v^{-1}$$

where T_k=$[T_{x_k} \;\; T_{y_k} \;\; T_{z_k}]^T$ is the translation vector, and d_0 is the scene depth. The rotation matrix R_k is parameterized [11] in terms of θ_X, θ_Y and θ_Z, which are the angles of rotation about the three axes. The camera intrisic matrix K_v is assumed to be of the form $K_v = \text{diag}(v, v, 1)$, where v is the focal length. Six degrees of freedom arise from T_k and R_k (three each). In this discussion, we assume that v is either known or can be extracted from the camera's EXIF tags.

3. SPARSITY, REGISTRATION AND OCCLUSION HANDLING

If $\mathbf{l}$, $\mathbf{b}$ represent the latent image and the blurred image, respectively, lexicographically ordered as vectors, then, in matrix-vector notation, equation (1) can be expressed as

$$\mathbf{b} = A\boldsymbol{\omega} \qquad (2)$$

where A is the matrix whose columns contain projectively transformed copies of $\mathbf{l}$, and $\boldsymbol{\omega}$ denotes the vector of weights $\omega(k)$. Note that $\boldsymbol{\omega}$ is a sparse vector since the blur is typically due to incidental camera shake and only a small fraction of the poses in $\mathbf{T}$ will have non-zero weights in $\boldsymbol{\omega}$.

In the scenario that we consider, one of the images in the pair is not only blurred because of camera jitter but can also contains occluder(s). To deal with this situation, we modify the linear model of (2) as

$$\mathbf{b_{occ}} = \mathbf{b} + \mathbf{o} \qquad (3)$$

where $\mathbf{b_{occ}}$ is the blurred and occluded observation. In the image formation model, the occlusion happens first followed by blurring, i.e., $\mathbf{b_{occ}}$ is the weighted average of warped instances of an unknown focused image containing occlusions. The non-zero entries of $\mathbf{o}$, therefore, model the blurred occluder(s) in $\mathbf{b_{occ}}$. Since the occluder(s) can have arbitrary intensities, techniques designed for small noise cannot be used here. The locations of occlusion differ for different input images and are not known a priori to the algorithm. But we assume that the occluded pixels occupy only a relatively small portion of the image. Therefore, the occlusion vector $\mathbf{o}$, in the same vein as the vector $\boldsymbol{\omega}$, has sparse non-zero entries [15]. Since $\mathbf{b} = A\boldsymbol{\omega}$, we rewrite equation (3) as

$$\mathbf{b_{occ}} = \begin{bmatrix} A & I \end{bmatrix} \begin{bmatrix} \boldsymbol{\omega} \\ \mathbf{o} \end{bmatrix} = B\mathbf{x} \qquad (4)$$

Here $B = [A\ I] \in \mathbb{R}^{N\times(N_\mathbf{T}+N)}$, where N is the total number of pixels in the image, $N_\mathbf{T}$ is the total number of transformations in $\mathbf{T}$, and I is an $N \times N$ identity matrix. Hence, the system $\mathbf{b_{occ}} = B\mathbf{x}$ is always underdetermined and does not have a unique solution for $\mathbf{x}$. We, therefore, attempt to recover $\mathbf{x}$ as the sparsest solution to the system $\mathbf{b_{occ}} = B\mathbf{x}$. Note that in the absence of occlusion, $\mathbf{x} = \boldsymbol{\omega}$ and our problem reduces to the special case of estimating a sparse TSF.

With the occlusion model incorporated, the energy function to be minimized takes the form

$$E(\mathbf{x}) = ||\mathbf{b_{occ}} - B\mathbf{x}||_2^2 + \beta||\mathbf{x}||_1 \qquad (5)$$
$$\text{s.t} \quad \forall k \in \mathbf{T}, \omega(k) \geq 0 \quad \text{and} \quad \textstyle\sum_{k\in\mathbf{T}} \omega\,(k) = 1.$$

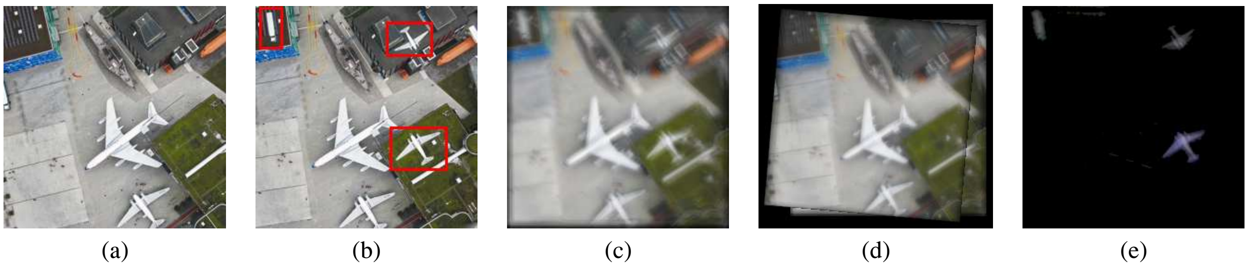


Fig. 2. (a) Latent image, (b) latent image from a different camera pose and with synthetically added occluders, (c) blurred and occluded observation, (d) latent image reblurred using the estimated camera motion and overlaid on the blurred and occluded observation, and (e) residual image.

where $\mathbf{x} = [\boldsymbol{\omega}\ \mathbf{o}]^T$, and $\boldsymbol{\omega}$ is non-negative. In the presense of an occluder, $\mathbf{o}$ can be positive or negative depending on whether the occluder causes the intensity at that pixel to increase (bright occluder) or decrease (dark occluder), respectively. Hence, non-negativity cannot be imposed on the entire vector $\mathbf{x}$ because then the model will not be able to handle dark occluder(s). Instead, in our implementation, we impose non-negativity and sparsity on $\mathbf{x}$, but change the form of the identity matrix I as discussed below.

In the absence of an occluder, the convex combination of the elements of a particular row, say i, of A produces the intensity of the blurred pixel at the i^{th} location in the image. If the observed intenstiy (in $\mathbf{b_{occ}}$) at the i^{th} pixel is greater than the maximum intensity of the elements of the i^{th} row, then, by convexity, we can deduce that it is the presence of a bright occluder that causes the intensity at that pixel to increase. A positive value in $\mathbf{o}$ will then explain the observed intensity at that pixel. On the other hand, if the observed intensity at the i^{th} pixel is less than the minimum intensity of the elements of the i^{th} row, we conclude that the occluder is dark. In this case, we replace the '1' at the corresponding location in I with a '-1'. This change in sign permits us to impose non-negativity on $\mathbf{x}$ because the residual can now take both positive and negative values. Thus B now becomes $[A\ I_{mod}]$ where I_{mod} is a diagonal matrix (with +1 and -1 along the diagonal) obtained after verifying the above condition. It is to be noted that the blurred pixel at the i^{th} location receives contributions only from the neighbourhood of the i^{th} location in the latent image, where the neigbhourhood size varies with the spatial location and the warps applied (since the blur is non-uniform). Therefore, the minimum and maximum values of the rows of A are not global constants. If a dark occluder pixel occurs with an intensity greater than the minimum intensity of the elements of the corresponding row in A, then it would render the above model erroneous. However, this is an unlikely event since we expect the occluder pixel's intensity to be significantly different from that of the background. The optimization problem in equation (5) can be solved using the nnLeastR function of the Lasso algorithm [16] which considers the additional $l1$-norm constraint.

3.1. Multiscale Implementation

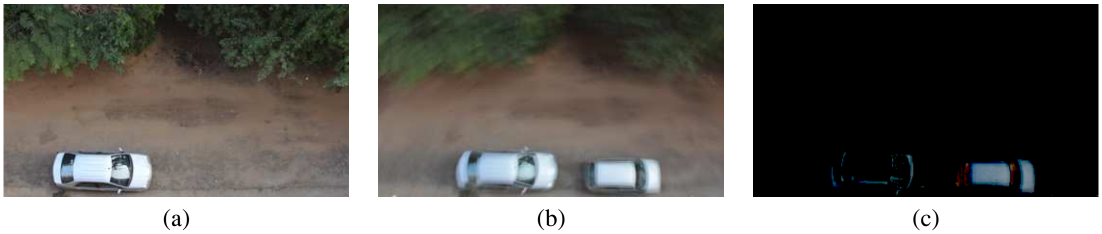
Solving equation (5) directly at full resolution would require storing all the transformed copies of $\mathbf{l}$ simultaneously in A and allowing for occlusion at each and every pixel in the image. For general 6D camera motion, with thousands of poses in the TSF space, this might mean that the computer's memory capacity will be exceeded even for moderately sized images. We, therefore, use a multiscale approach similar to [11]. We start with a coarse representation of the image, the TSF and the occlusion, and repeatedly refine the estimated TSF and occlusion at higher resolutions.

We first build Gaussian pyramids for both $\mathbf{l}$ and $\mathbf{b_{occ}}$. At the coarsest scale, the matrix A is built for the whole transformation space $\mathbf{T}$ and we allow for occlusion at every pixel since the location of the occluder(s) is unknown. It must be mentioned that downsampling the images has the effect of reducing the blur, thereby decreasing the space of allowed transformations in $\mathbf{T}$. At each scale, we find the optimal TSF $\boldsymbol{\omega}$ and the occlusion $\mathbf{o}$ by minimizing equation (5). We then upsample $\boldsymbol{\omega}$ and $\mathbf{o}$ to the next scale using bilinear interpolation and find the non-zero entries in both. For $\boldsymbol{\omega}$, this process gives us several 6D non-zero regions inside the transformation space. When finding the optimal TSF at the next stage, we only search for valid homographies which lie within these non-zero regions. Likewise, at the next scale, we look for occluded pixels only at those locations in $\mathbf{o}$ which have non-zero entries. This corresponds to discarding many columns of B, reducing both the computation and memory demands of the algorithm. We repeat this process at each scale until the optimal TSF and the occluder(s) have been estimated at the highest resolution.

4. EXPERIMENTS

This section consists of two parts. We first evaluate the performance of our algorithm on synthetic data. Following this, we demonstrate the applicability of the proposed method on real images.

We begin with a synthetic example. A latent image of size 240 × 240 pixels of an airport bay is shown in Fig. 2(a). The

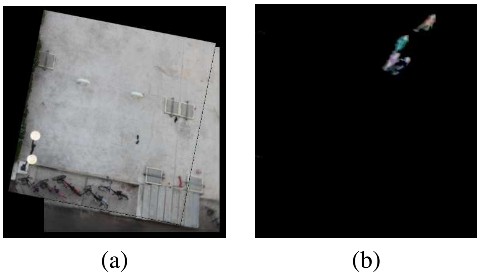


(a) (b) (c)

Fig. 3. (a) Real latent image, (b) blurred and occluded observation, and (c) residual image.

same scene from a different camera pose and with synthetically added occluders (enclosed in red boxes) is shown in Fig. 2(b). The TSF space is chosen as follows- in-plane translations: $T_x,\ T_y = [-7 : 1 : 7]$, in-plane rotation: $R_z = [-3^\circ : 1^\circ : 3^\circ]$, out-of-plane translation: $T_z = [0.95 : 0.05 : 1.05]$ and out-of-plane rotations: $R_x,\ R_y = [\frac{-4}{3}^\circ : \frac{1}{3}^\circ : \frac{4}{3}^\circ]$. To simulate the motion of the camera, we manually generate 6D camera motion with a connected path in the motion space and initialize the weights. The synthezied camera motion (TSF model) is applied on Fig. 2(b) to produce the blurred and occluded image (Fig. 2(c)). To evaluate the proposed method, we set the number of scales in the multiscale implementation to 3 and first coarsely align the latent image and the blurred and occluded image at the lowest resolution without accounting for occlusion. In this step, the transformation intervals are expanded to $T_x,\ T_y = [-40 : 1 : 40]$ and $R_z = [-8^\circ : 1^\circ : 8^\circ]$ to accommodate for the large change in pose between the two images. Note that this increase in the transformation intervals is not very demanding because we work at the lowest resolution of the image and the TSF in the multiscale algorithm. The 'dominant pose', i.e., the pose with the highest weight from the estimated vector $\boldsymbol{\omega}$ is used to align the latent image with the blurred image. The TSF is now built around this dominant pose and we minimize equation (5) using the multiscale approach but now by also taking occlusions into consideration. Fig. 2(d) shows the latent image reblurred using the estimated $\boldsymbol{\omega}$ and overlaid on the blurred and occluded observation. It is to be noted that the TSF model implicitly accounts for the change in pose between the two images. The residual image shown in Fig. 2(e) is the absolute difference between the blurred and occluded observation (Fig. 2(c)), and the reblurred latent image. Note that the occluders are correctly detected.

Next, we discuss results obtained using real data. We begin with an example in which the view point change is only due to in-plane translation of the camera. The roof-top images in Figs. 3(a) and 3(b), which represent the latent image and the blurred and occluded image, respectively, were captured using a Canon 60D DSLR camera with the same lens setting. The focal length was retrieved from the camera's EXIF data. Since an interval of time had elapsed before the second image was captured, the images were pre-processed using the approach in [17] to account for small global changes in illumination. The residual image, shown in Fig. 3(c), demonstrates our algorithm's ability to detect the occlusion even in the presence of significant amounts of blur. Yet another real example but with an appreciable change in view point was shown in Fig. 1. A zoomed-in view of the blurred occluders (in this case people) is shown in the red box in Fig. 1(c). The latent image reblurred using the estimated $\boldsymbol{\omega}$ and registered with the blurred and occluded observation is shown in Fig. 4(a). The residual image (Fig. 4(b)) reveals that the dark occluders have been accurately detected by the proposed method.



(a) (b)

Fig. 4. Output results for the real example in Fig. 1. (a) Reblurred latent image registered with the blurred and occluded observation, and (b) residual image.

5. CONCLUSIONS

In this paper, we proposed a unified framework for registration and detection of occlusions from a blurred/unblurred image pair. The proposed method exploits the sparsity of the TSF and the occlusions, and is robust to non-uniform motion blur and variations in camera pose. Synthetic as well as real results were given for the purpose of validation. In the future, we plan to extend the scope of this work to include local changes in illumination and to handle occlusions with intensities close to that of the background.

6. ACKNOWLEDGEMENTS

A part of this work was supported by a grant from the Asian Office of Aerospace Research and Development, AOARD/AFOSR. The support is gratefully acknowledged. The results and interpretations presented in this paper are that of the authors, and do not necessarily reflect the views or priorities of the sponsor, or the US Air Force Research Laboratory.

7. REFERENCES

[1] R.S. Feris, B. Siddiquie, J. Petterson, Yun Zhai, A. Datta, L.M. Brown, and S. Pankanti, “Large-scale vehicle detection, indexing, and search in urban surveillance videos,” *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 28 –42, Feb. 2012.

[2] V. Ablavsky and S. Sclaroff, “Layered graphical models for tracking partially occluded objects,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 9, pp. 1758 –1775, sept. 2011.

[3] A. Criminisi, P. Perez, and K. Toyama, “Region filling and object removal by exemplar-based image inpainting,” *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1200–1212, Sept. 2004.

[4] K. Shafique and M. Shah, “A non-iterative greedy algorithm for multi-frame point correspondence,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 51–65, 2003.

[5] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the International Conference on Computer Vision-Volume 2*, 1999.

[6] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, 1981, pp. 674–679.

[7] M. Brown and D. G. Lowe, “Automatic panoramic image stitching using invariant features,” *International Journal of Computer Vision*, vol. 74, no. 1, pp. 59–73, 2007.

[8] M. McGuire and H. S. Stone, “Techniques for multiresolution image registration in the presence of occlusions.,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 3, pp. 1476–1479, 2000.

[9] L. Lelégard, E. Delaygue, M. Brédif, and B. Vallet, “Detecting and correcting motion blur from images shot with channel-dependent exposure time,” in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume I-3*, 2012, pp. 341–346.

[10] L. Yuan, J. Sun, L. Quan, and H.-Y Shum, “Blurred/non-blurred image alignment using sparseness prior,” in *Proceedings of International Conference on Computer Vision*, 2007, pp. 1–8.

[11] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce, “Non-uniform deblurring for shaken images,” *International Journal of Computer Vision*, vol. 98, no. 2, pp. 168–186, 2012.

[12] A. Gupta, N. Joshi, C. L. Zitnick, M. Cohen, and B. Curless, “Single image deblurring using motion density functions,” in *Proceedings of European Conference on Computer Vision*, 2010.

[13] C. Paramanand and A. N. Rajagopalan, “Inferring image transformation and structure from motion-blurred images.,” in *Proceedings of British Machine Vision Conference*, 2010, pp. 1–12.

[14] Z. Hu and M.-H Yang, “Fast non-uniform deblurring using constrained camera pose subspace,” in *Proceedings of British Machine Vision Conference*, 2012.

[15] J. Wright, A. Y. Yang, A. Ganesh, S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[16] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning with Efficient Projections*, Arizona State University, 2009.

[17] J. Yin and J. R. Cooperstock, “Color correction methods with applications to digital projection environments,” *Journal of WSCG*, vol. 12, pp. 1–3, 2004.