Super-resolution of a 3-dimensional Scene from Novel Viewpoints

Kyle Nelson, Asim Bhatti, Saeid Nahavandi
Centre for Intelligent Systems Research
Deakin University
Geelong, Victoria, Australia

Abstract—Super-resolution is a method of post-processing image enhancement that increases the spatial resolution of video or images. Existing super-resolution techniques apply only to images captured of a planar scene. This paper aims to extend super-resolution concepts from the 2D domain to the 3D domain, drawing on ideas from both super-resolution and multi-view geometry, two fields of research that until now have predominantly been studied in isolation. 2D super-resolution methods are not without their complexities and limitations. However, once multiple views of a scene are considered within a super-resolution framework, a new range of issues arise that must also be resolved. For example, when input images of a scene with variation in depth are considered, it is no longer clear how and where the images should be registered. This paper describes the use of sparse 3D reconstruction in order to ‘register’ the input images, which are then transferred to a novel image plane and combined to increase the perceived detail in the scene. Experimental results using real images captured from generally positioned input cameras are presented.

Keywords—super-resolution, 3-dimensional, multi-view, uncalibrated camera, novel view, sparse reconstruction

I. INTRODUCTION

Super-resolution is a method of post-processing image enhancement that increases the spatial resolution of video or images. Multi-frame super-resolution techniques improve image resolution by fusing information from multiple low resolution frames. Super-resolution algorithms require several slightly different perspectives of the same planar scene, a non-integer pixel shift between images, which are then aligned via an image registration process and intelligently combined to create a high-resolution image of the image plane. Multi-view geometry on the other hand uses multiple images of a 3-dimensional scene to extract the scene geometry. Given several low resolution images of a 3-dimensional scene, this paper investigates whether multi-view geometry and super-resolution techniques can be used to not only extract the geometry of a 3-dimensional scene, but to enhance an observer’s view of the scene at the same time.

To this point, there has been limited research in applying super-resolution ideas to 3-dimensional scenes. The algorithm described in this paper is designed to estimate a single super-resolved image of a 3-dimensional scene, which exploits the spatial diversity of multiple low resolution images captured from varying, unknown viewpoints. While the focus of this paper is to generate high resolution 3D visual information, the output of this research is not intended to be a 3D image. The desired output is a super-resolution image, constructed from a novel viewpoint, of a scene with depth. That is, given multiple low resolution images of a 3-dimensional scene, this algorithm creates an output image that not only exceeds the spatial resolution of the input frames, but views the scene from an angle different to the observed views.

Using super-resolution techniques to create a high resolution image in this work does not imply mosaicing images together to create a larger field of view; high resolution in this context implies greater spatial detail in the same field of view. Additionally, this research is not aimed at creating a high resolution 3D-mesh or any other type of surface model of a scene. While estimating the 3D structure of a scene from multiple low resolution views is a beneficial step in creating a high-resolution view of a 3D scene, there are many other areas of research that are more advanced in terms of 3D reconstruction.

In order to extend super-resolution techniques from 2 dimensions to 3 dimensions, this paper aims to remove some of the assumptions and restrictions imposed on the super-resolution problem in past research. Often in the field of super-resolution, researchers assume that camera parameters are known and in multi-view geometry research it is often assumed that camera locations, or the geometry relating cameras, is known. An aim of this research is to lift these assumptions. In addition, researchers sometimes vary the illumination of a scene, deliberately blur frames or use training images to simplify the super-resolution problem. Such techniques do not feature in this research.

This paper utilizes normal image frames from a standard digital camera. A key element in this research is the assumption of a completely uncalibrated camera. No prior knowledge of the camera parameters, its motion, optics or photometric characteristics is assumed. While the technique described in this paper applies to generally positioned input cameras, the viewpoint variation between frames is constrained by the use of 8DOF homographies to transfer input images to a novel image plane. If the difference between input viewpoints is large, the 8 parameter projective model is less applicable and this may affect the quality of the output image.
The presented research does not combine input from various measurement systems as others have done and no additional imaging or measurement hardware of any kind is required. We also assume that input images belong to a static 3-dimensional scene. The content of the scene, object locations and lighting were not altered during image capture. It is also assumed that, although images are captured from different viewpoints, images belong to the same static scene. No further assumptions about the scene are made, unlike previous research in which the depth map of a scene is assumed known.

II. BACKGROUND

In the literature to this point the fields of super-resolution and multi-view geometry have been predominantly studied in isolation. While several researchers identify and acknowledge the close relationship between the fields, few have addressed the issues concerned in the context of super-resolution [1-3].

The research presented here is motivated by the fact that the application of super-resolution techniques to a 3D environment is, to this point, essentially untested. Several authors acknowledge that existing super-resolution research has been confined to the 2D domain and that very little research has been conducted regarding a 3D scene where the input cameras are generally positioned [2,4-6]. Fanaswala [2] consistently draws attention to the fact that super-resolution of a 3D scene has not received adequate attention, thus this paper directly examines super-resolution using multi-view images and studies the inherent difficulties of this specific scenario.

While the problem of super-resolved 3D scenes is far from solved there are a limited number of research publications that are relevant to this issue. However, in existing research applying super-resolution to a 3D scenario, often difficulties are bypassed by employing a range of hardware solutions. In [7] a compound imaging system is used, a microlet array in [8], while Nava and Luke [9,10] employ a plenoptic camera. Kim and Zhu [11] combine the use of standard cameras and two time-of-flight depth cameras and in [12] a pair of low resolution high-speed cameras and a single high resolution low-speed camera are used. Finally, Caner, Tekalp and Heinzelman [13] mount multiple cameras together and in [14] input is drawn from both low resolution and high resolution cameras. As outlined previously, we use only images from a standard digital camera, no additional imaging or measurement hardware of any kind is required.

In addition to introducing complex hardware, researches also apply significant restrictions in existing work such that only a very specific subset of the 3D super-resolution problem is addressed. Joshi and Chaudhuri [15] vary the illumination of a scene, Berthod et al. [16] alternately assume either the high resolution image or depth map of a scene are known, while in [16-19] the input images are restricted in such a way that super-resolution is applied to an essentially 2D plane. Others have focused their work on specific domains such as aerial and satellite imaging [16-19], 3D television [14,20,21] and computer graphics [22,23], imposing restrictions such that the research has limited applicability in other areas. Another subset of related publications employ super-resolution techniques to solve a problem in the 3D domain, but the focus is not the use of super-resolution [3,15-17,24-26]. Rajagopalan and Chaudhuri [24] limit their super-resolution algorithm to scenes with only a small depth of field, in general there is a paucity of research involving scenes with significant depth variation and varying perspectives of a 3D scene.

The earliest work related to the application of super-resolution techniques in a 3D setting is that of Berthod et al. [16,17]. The authors propose an iterative method to recover the albedo as well as the depth of a scene. In order to recover a high resolution image, Berthod et al. assume that the depth map for the scene is known. To recover the high resolution depth map they then assume the high resolution image is known. It is also assumed that the focal point is a significant distance from the scene, limiting the applicability of the algorithm to essentially 2D scenes.

In [18] Shekarforoush et al. are the first to refer explicitly to a ‘3D super-resolution algorithm’. Their paper discusses the use of Papoulis’ generalized sampling theorem in a Bayesian framework using MRF, solving depth and albedo simultaneously. The authors acknowledge that accurate image registration is required and that camera parameters are assumed known.

Rajan and Chaudhuri utilize the Depth From Defocus (DFD) technique in [25]. Their algorithm uses a single camera to capture two images of a scene with different but known camera parameters; the depth in the scene is then calculated based on the image focus. The authors assume there is no relative motion between the scene under observation and the camera, hence the input images are essentially taken from the same viewpoint. This greatly simplifies the problem as an accurate image registration procedure is no longer required, however this also implies that super-resolution could be achieved using an existing 2D technique.

The work by Mudunagudi et al. [6] presents experimental results using images from slightly different viewpoints of a 3D scene. Their research focuses on the generation of novel super-resolved views of a 3D scene with no restrictions on the scene geometry or camera positions, however the camera viewpoints are assumed known in this case. The authors note that due to the depth variation in a 3D scene and the arbitrary placement of input cameras, the registration problem is far more complex than the 2D case. They assert that in order to accurately register input images, dense depth estimation at each pixel is required. However, no explicit or accurate depth recovery is used in this paper; registration is accomplished using the automatic tracking and calibration software Boujou.
The master’s thesis by Fanaswala [2] is primarily focused on super-resolution using multi-view images, captured by the ProFUSION 25 compound camera array. Fanaswala uses a 6 parameter affine model to relate the different views from the compound camera array, which is only applied to small regions of the input images, later combined using a smoothness constraint across the entire image. This registration algorithm uses an iterative coarse-to-fine method that takes 90 minutes to register two 480×480 images.

In [21] and [20] Knorr, Kunter and Sikora present a technique that uses super-resolution in the creation of 3D video from 2D video, for use with stereoscopic displays. Like Mudenagudi et al. [6], the authors do not use dense depth maps, instead employing photo-consistency measures to avoid computationally expensive depth estimation.

Goldluecke and Cremers apply super-resolution to image-based texture reconstruction in [22] and [23]. In order to render an object from varying virtual views or under different lighting conditions high resolution texture maps are required. Hence, the authors propose the use of super-resolution to re-create surface textures as accurately as possible. Goldluecke and Cremers assume that the surface geometry of their target object is known and assume that in their calibrated multi-view camera arrangement most of the surface of an observed object has been captured; they do not consider noise in the input images. Also, the authors do not perform any explicit image registration; hence the success of their technique is dependent on the assumed surface geometry and camera calibration. Goldluecke and Cremers’ algorithm takes around 3 hours to converge to a result.

### III. 3D SUPER-RESOLUTION FRAMEWORK

#### A. Detection and Matching Key Image Features

As with any super-resolution algorithm, our framework begins by first detecting and matching key image features. Currently input frames are presented as a temporal sequence, however ideally the algorithm would accept input frames of a 3D scene in any order. Key features are detected in the first 2 temporal frames using FAST high speed corner detection [27,28]. FAST corner detection was selected for use in this research due to its ability to locate a high number of interest features. This becomes especially important during reconstruction and in the latter stages of the algorithm. Harris corner detection [29] or a Harris-Laplace detector [30] allowing for scale changes may also produce adequate results. Putative matches are established based on maximal correlation between windows surrounding the interest points.

#### B. Sparse 3D Reconstruction

Historically the second stage in traditional super-resolution algorithms involves registering input frames to a common image plane, as without accurate registration super-resolution algorithms are redundant. However, when applying super-resolution concepts to scenes with depth, the space in which to register images is no longer clear. Consider the situation in which multiple cameras are positioned around a scene with unknown position and orientation. The images captured no longer necessarily belong to the same planar surface as is the case in traditional super-resolution approaches. Due to the variation in depth between the input images, the registration problem becomes much more complicated, as observed by Mudenagudi et al. [6].

While consideration was given to various manifolds on which input images could be registered, sparse 3D reconstruction was selected as the standout ‘registration’ method. The creation of a high resolution mesh or any other surface model of a scene is by no means the focus of this algorithm. A sparse 3D point cloud in this case acts only as an intermediate device facilitating accurate ‘registration’ of camera position and orientation.

The fundamental matrix between the first two input views in the image sequence, $F_{12}$, is first found using RANSAC [31]. Putative matches are also updated at this time, only inlying matches from the RANSAC stage are accepted. Singular Value Decomposition is then used to estimate the epipole, $e_{12}$, and the projection matrices for each frame are assigned according to

$$P_1 = [I_{3\times3}]e_1,$$  
$$P_2 = ([e_{12}]\times F_{12}[e_{12}]).$$

Given the two projection matrices, $P_1$ and $P_2$, and the set of image point pairs, the corresponding scene points are triangulated using Direct Linear Transformation (DLT) [32]. The world frame is aligned with $P_1$. The cheirality of the triangulated points is then calculated to determine their direction relative to the cameras $P_1$ and $P_2$. The cheirality of a point $X = (X, Y, Z, T)^T$, with respect to a camera $P$ is given by

$$\text{Cheirality}(X) = wT\det(M),$$

where $M$ is the left hand $3\times3$ block of $P$, $x$ is the projection of point $X$ by camera $P$ and $PX = wx$. If required, the orientation of the cameras $P_1$ and $P_2$ are then corrected to ensure that all triangulated points lie in front of both cameras, thus creating a feasible scene reconstruction.

Once an initial point cloud has been established, all remaining images in the sequence must be added to the reconstruction in order to ‘register’ their position. FAST corner detection is again used to identify interest features in the next temporally adjacent view and feature matches between neighboring views are established by maximal correlation. Regardless of the feature matching method applied, occasionally a situation arises whereby multiple points in one image may be mapped to the same location in
another image. Consider the situation where two interest points in image 1, \( x_{11} \) and \( x_{12} \), are both matched to the same point in image 2, \( x_{21} \). When a third image is included in the reconstruction, point \( x_{31} \) is matched to point \( x_{21} \). In this situation it is unclear if the point in the third image, \( x_{31} \), relates to \( x_{11} \) or \( x_{12} \) in the first image leading to problems with regards to triangulation and reconstruction. Therefore, such ambiguous points are eliminated from the set of interest features to ensure a unique solution.

Once putative correspondences between an additional frame and the preceding image have been established, the subset of interest features common to all three views is identified, along with the corresponding scene points that were previously triangulated. Given the set of image points and scene points common to three adjacent frames, the projection matrix for the third frame can be found. This overdetermined problem is solved using Direct Linear Transformation (DLT). The projection matrix for each additional image in the input sequence is found in the same manner. The chirality of scene points with respect to each camera added to the reconstruction is found, enabling camera orientation to be corrected if required. With the addition of each new camera the reconstructed point cloud and camera properties are refined using bundle adjustment [33,34].

As the projection matrix for each image is estimated and refined using bundle adjustment, feature points from that image are added to the sparse scene reconstruction. Given a newly estimated projection matrix, \( P_i \), the projection matrix for the previous frame, \( P_{i-1} \), and the set of putative correspondences between the two views, the corresponding scene points can be triangulated. Those points that are unique to this image pair are identified and added to the scene reconstruction point cloud.

### C. Novel View Calculation

The output of the 3D super-resolution framework is not intended to be a 3D image, but rather a super-resolved image of a scene with depth. Therefore, if a user is to be presented with a single high resolution view of a 3D scene, an appropriate virtual viewpoint must first be determined. The ideal novel perspective would be the virtual view encompassing the greatest scene detail. Currently the novel view is selected by calculating the mean of input camera positions. The projection matrix for the novel camera is derived from the previously estimated input views. An initial sparse view from the novel perspective is obtained by projecting scene points from the metric reconstruction using the calculated novel projection matrix.

The principal axis vector for a camera, \( P \), is given by

\[
v = \det(M)m^3,
\]

where \( m^3 \) is the third row of \( M \), the left hand 3×3 block of \( P \). The principal axes of all input views are calculated and the angle between each and the principal axis of the novel view is found in order to identify the spatially neighbouring views to the novel camera. The closest \( \lfloor \frac{n}{2} \rfloor \) views, where \( n \) is the number of input frames, are then used in the next stage of the algorithm to obtain an estimate of the novel view.

### D. Input View Transfer

In order to transform input frames to the perspective of the novel view, putative correspondences must first be established between the sparse representation of the novel view and the immediately neighboring views. Due to the sparse nature of the current novel view estimate, feature matching algorithms based on descriptors or windows surrounding interest features are not successful. Hence, interest features are matched to the sparse novel view using non-rigid point cloud registration algorithms [36,37].
With the establishment of interest point matches, the transform relating each neighboring view to the novel view can be found. Each homography is first estimated through a quasi-linear method then optimized by minimizing the re-projection error. The closest $$\frac{n}{3}$$ input views are then transferred to the novel image plane through application of the relevant homography. To ensure absolute alignment, the transformed images are registered to sub-pixel accuracy using the iterative method of Keren, Peleg and Brada [38]. A temporal average of the transformed frames is found by taking the median of the registered images, resulting in a complete estimate of the novel view, thus replacing the earlier sparse representation.

FAST corner detection is used once again to locate interest features in the complete estimate of the novel view. Also, as the novel view is no longer represented by a sparse planar point cloud, features located in the novel view can be matched to all other input views using maximal correlation techniques. As was done for the immediately neighboring views, the homographies relating each input view to the novel image plane can then be calculated based on putative matches. All input frames are then transferred to the novel perspective of the scene. Finally, all transformed images are aligned to sub-pixel accuracy in preparation for super-resolution.

E. Super-Resolution

At this point the framework enabling super-resolution of a 3-dimensional scene is essentially complete. Input images capturing different perspectives of a scene with depth have been transformed and aligned to an initial estimate of a novel view; the plane in which super-resolution is to be applied. Potentially any existing super-resolution algorithm designed to operate on images of a planar scene could now be applied. In the experiments described below, an adaptation of the Bayesian super-resolution algorithm 3.1 in [39] was used, which employs the following texture-preserving non-sparse simultaneous autoregressive (SAR) prior:

$$p(x|\alpha) \propto \alpha^{LN/2}\exp\left\{-\frac{\alpha}{2}\|Cx\|^2\right\}.$$  (5)

where $$x$$ is the unknown high-resolution image, $$L$$ is the resolution enhancement factor, $$N$$ the number of pixels in the observed low-resolution images, $$\alpha$$ a model parameter and $$C$$ is the Laplacian operator.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section we present some results from our 3D super-resolution algorithm. All experiments use datasets of real images captured using a handheld digital camera. The first experiment demonstrates our method's ability to generate a super-resolved novel view using the ‘Games’ dataset. Fig. 2 shows 4 of the 10 input frames; the relative motion between images is most obvious when inspecting the rear game in the centre of the image. A top view of the sparse reconstruction used to calculate the previously unknown camera locations and orientations can be seen in Fig. 1. Fig. 3 shows a super-resolved image with a magnification of ×2, which provides a novel perspective of the scene.
In order to compare a generated super-resolved view with ground truth data a second experiment is presented using the ‘Textbooks’ dataset. All 10 low-resolution 320×240 pixel frames are initially provided to the algorithm as input. Once the sparse reconstruction stage is completed and all camera locations and orientations have been determined, one of the 10 frames is selected as the target view and all information about this frame is disregarded apart from its projection matrix. The super-resolution framework then uses the remaining 9 input frames to recreate a super-resolved estimate of the missing target view.

Fig. 4 shows four low-resolution input frames from the ‘Textbooks’ dataset. Camera motion between images can be observed by examining the top right of the frames or the occlusion in the bottom right of the central black book. Fig. 5(a) shows the ground truth target view upscaled using bilinear interpolation, while Fig. 5(b) shows the super-resolved estimate generated using only the remaining 9 frames in the sequence. The RMS error, PSNR and SSIM between the interpolated ground truth frame and the super-resolved estimate are 19.3193, 22.4110 dB and 0.7130 respectively. Fig. 5(c) and (d) show close up comparisons of 90×90 pixel regions from the interpolated ground truth and super-resolved images.
V. Conclusion

In this paper we have presented an algorithm capable of generating super-resolved images of 3-dimensional scenes, from existing or novel viewpoints. We propose creating a sparse 3D reconstruction in order to "register" input images, which are transferred to a novel image plane in order to perform super-resolution. This versatile framework allows any existing 2D super-resolution method to generate spatially enhanced views of a scene with depth variation. We present experimental results using real images captured from generally positioned input cameras. Initial work to facilitate greater variation between input viewpoints has shown promising results when using the trifocal tensor rather than a homography for image transfer.

REFERENCES


