# From Visual Query to Visual Portrayal

Ali Shahrokni† Christopher Mei† Philip H.S. Torr‡ Ian D. Reid†

† Robotics Research Group
University of Oxford

‡ Department of Computing
Oxford Brookes University

**Abstract**

In this paper we show how online images can be automatically exploited for scene visualization and reconstruction starting from a mere visual query provided by the user. A visual query is used to retrieve images of a landmark place using a visual search engine. These images are used to reconstruct robust 3–D features and camera poses in projective space. Novel views are then rendered corresponding to a virtual camera flying smoothly through the projective space by triangulation of the projected points in the output view. We introduce a method to fuse the rendered novel views from all input images at each virtual view point by computing their intrinsic image and illuminations. This approach allows us to remove the occlusions and maintain consistent and controlled illumination throughout the rendered sequence. We demonstrate the performance of our prototype system on two landmark structures.

## 1 Introduction

In this paper we show how a minimal amount of pictorial information about a landmark can be enriched using online information to yield an explorable 3–D representation of the landmark. The abundance of online information has dramatically changed the way modern technologies work. Recent advances in visual query-based search engines [2] enable us to retrieve images from online resources based on their resemblance to the user provided visual query. Likewise, advanced computer vision systems have emerged that are capable of exploiting the vast source of online information. One such system was introduced by Hays and Efros [5] which uses online images to interactively remove undesirable regions and fill them seamlessly with data from ranked similar images. Another example of such applications is scene reconstruction from online photos of landmarks proposed by Snavely *et al.* [9]. This system introduces a new fashion for navigation through a large set of images of a site manually collected from the Internet in which the photos are arranged and accessed based on their computed view points and added annotations.

We aim at taking these ideas further and investigate how, starting from a small visual query associated to a place, the ensemble of online images of that site can be exploited to visualize the queried scene. This idea is illustrated in Fig. 1. A visual query is defined in the form of a distinctive image region selected from a picture of the landmark as shown in the top left image in Fig. 1. The query is then passed to a visual search engine to obtain a set of images of the site.

The queried landmark is chosen to be a place of significant popularity among visitors and photographers. Therefore, we expect to have a large number of photos of that place available online. The retrieved images are then used to reconstruct a set of 3–D scene points and measure the camera parameters for each input image. This reconstruction in turn enables navigation through the scene and synthesis of novel views of the scene with controlled illumination as shown at the bottom row of Fig. 1. This task is challenging because we assume no constraints on the input views. The retrieved images are taken using different cameras from scattered view points under different lighting conditions and often suffer from presence of occlusions (e.g. cheerful tourists). Hence, robust and innovative techniques are required to obtain a 3–D representation of the scene as well as synthesized novel views which are temporally coherent and are free of occlusions and varying illumination.

State-of-the-art multi-view reconstruction algorithms assume as input a set of calibrated images usually taken by the same camera. Different approaches based on Bayesian inference [4], expectation maximization [10] and graph cuts and message passing energy minimization [11, 14] have been used to minimize a global photo-consistency cost function with priors imposed to regularize the results. The priors are typically in the form of visual hull or surface smoothness constraints. An alternative strategy was proposed by Furukawa and Ponce [3] to gradually enforce local photometric as well as global visibility constraints. They showed how a sparse set of matched points can be used as seed to match, expand and filter a dense cloud of points using the measured input views. In the absence of calibration data and other constraints used in the above techniques, we proceed by reconstructing robust 3–D features and cameras in projective space, and therefore avoid precise estimation of the internal parameters of the cameras which requires additional resources such as picture metadata and further priors [9].

Novel views are then rendered corresponding to a virtual camera flying smoothly through the projective space by triangulation of the projected points in the output view. Lhuillier and Quan [6] proposed joint view triangulation method to interpolate views using a quasi-dense set of point matches. They introduce a greedy algorithm to match constrained triangulations in two views based on connected component boundaries. This method works well for two-view interpolation between narrow-baseline images. However it is not straightforward to generalize it to multi-view joint triangulation. Furthermore, the planar patch matching used in conjunction with joint view triangulation interpolations is prone to error due to occluding objects in the scene as well as wide-baseline images. Instead of optimizing a single snapshot for each virtual view given all the input data, we adopt a simple triangulation-based method to create a novel view based on each measured input view at each virtual camera location.

Furthermore, we fuse the rendered novel views from all input views at each virtual view point by computing their so-called *intrinsic image* [12] and illuminations. This approach allows us to remove the occlusions and insert consistent and controlled illumination throughout the rendered sequence. Illumination analysis and manipulation using image of the same scene under variable illumination has been used in the context of foreground layer recovery, removing shadows and reflections [1], as well as estimating intrinsic images [1, 8, 12]. To the best of our knowledge this is the first time that image light decomposition has been applied to novel views based on images from different points of view to compute illumination and remove occlusions.

Our contribution is therefore to show how a mere visual query provided by the user

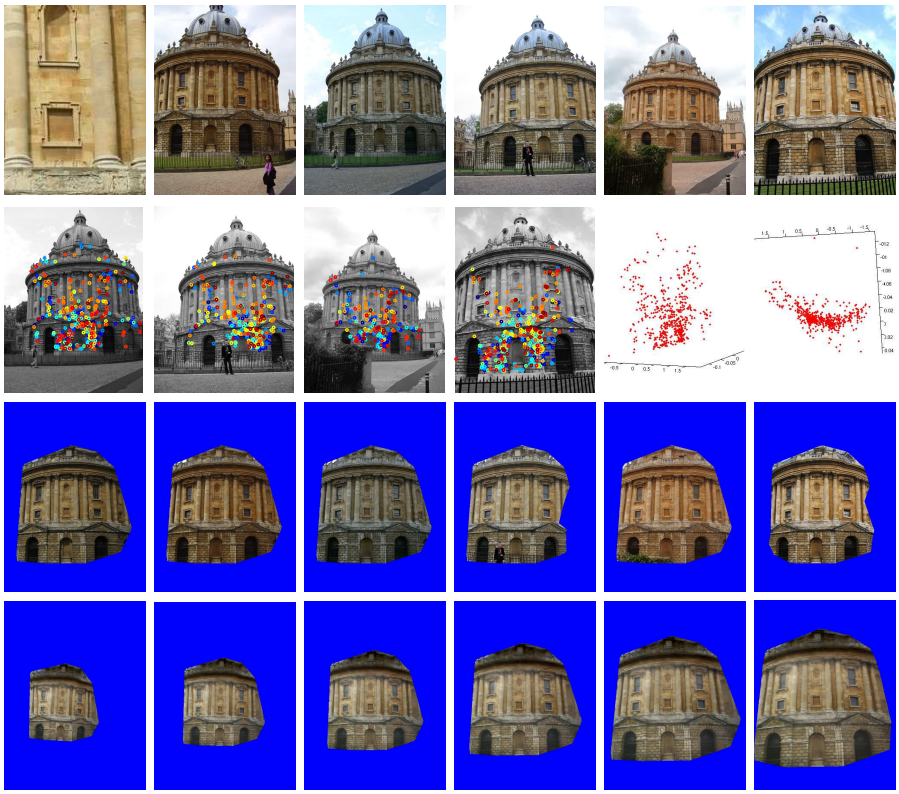


Figure 1: What information can be obtained from a visual query. Top row: The user provided query is shown in the top left image and is used to automatically retrieve images of the relevant building using a visual search engine [2]. Second row: robust 3–D feature and camera reconstruction is done in projective space. Here the projected 3–D features in some input views and affine reconstruction of the 3–D points are shown. Third row: Input views are used to render multiple novel views for each point of view in the projective space. These novel views have occluding objects and varying illuminations. Bottom row: Novel views based on different input views are fused together to remove occlusion and artifacts and render the scene with controlled illumination.

can be used to generate a visual representation of the scene corresponding to the query. This space can be explored through virtual views which are rendered without occluding elements present in the retrieved views. Furthermore, we maintain consistent and controlled illumination across the entire virtual views by computing the intrinsic image and light in the rendered views.

The remainder of this paper is organized as follows. Section 2 discusses the details of different steps of our prototype system. Results and analysis are presented in section 3. Finally we summarize the merits and limitations or our system in section 4.

## 2 Visual Scene Representation

Once a set of input images is retrieved by the visual query search engine, the highest ranking ones are selected and the resulting set is denoted by $\mathbf{I}$. We refer to the $i^{th}$ image in $\mathbf{I}$ by $I_i$. The first task is then to measure the cameras used to capture those images and as well as a set of robust 3–D features in the scene. In this section we explain how wide-base line feature matching can be used to reconstruct the features and cameras. This reconstruction is then used to render virtual views based on input views which are then fused to obtain a coherent and occlusion free fly-trough sequence in the projective space. These steps are explained in detail below.

### 2.1 Projective Reconstruction

Since we make no assumptions about the internal parameters of the images in $\mathbf{I}$, we reconstruct the scene in projective space. Each camera $P_i$ corresponding to image $I_i$ is therefore represented by a $3 \times 4$ matrix of rank 3 with 11 degrees of freedom and the task is to accurately estimate $P_i$ for all $I_i \in I$ as well a set of triangulated features.

We use SIFT features [7] due to their robustness to scale and view point variance. For all image pairs $I_i$ and $I_j$ in $\mathbf{I}$, SIFT features, $f_i \in I_i$ are matched against SIFT features $f_j \in I_j$ to yield an initial set of matches $m_{ij}^0$. These matched features contain outliers and must be further processed before proceeding with the scene reconstruction. We robustly compute a fundamental matrix for each image pair using the initial set of feature matches. The estimated fundamental matrix is then used to filter out the outliers in the set $m_{ij}^0$ and obtain a refined set of matches denoted by $m_{ij}$.

Given the set of refined robust matches in all image pairs, we build feature tracks across all images in $\mathbf{I}$. Consistency is enforced in the set of tracks by eliminating tracks that contain conflicting matches across different images. These tracks are then used to perform robust bundle adjustment to yield a set of accurate 3–D features $X$ and camera matrices $P_i$ for all images in $\mathbf{I}$. The second row of Fig. 1 shows the projection of 3–D features $X$ using the estimated camera matrices $P_i$ in input views and their 3–D affine visualization.

### 2.2 Novel View Synthesis

Virtual cameras can be defined in a number of ways in the projective space. In order to compute a virtual camera sequence between all input views we directly interpolate between projection matrices $P_i$ to avoid explicit decomposition of the matrices into rotation and translation in projective space. The projectively interpolated virtual camera positions are shown in Fig. 2 in affine space for better viewing. It can be seen that the projective sampling corresponds to an intuitive camera interpolation in the affine space and that all the virtual cameras point at the scene. The sequence rendered from these cameras gives a smooth transition from one input frame to the next as can be seen in section 3.

Given the virtual camera matrices $Q$ and the measured 3–D points $X$ and input views $\mathbf{I}$ we can render virtual snapshots $\mathbf{I}'_q$. While multiple depth maps and global optimization techniques can be used to compute each novel view as a global minimum of a high-dimensional energy function, these methods are computationally expensive and require specialized optimization algorithms to handle the involved MRF energy function [13].

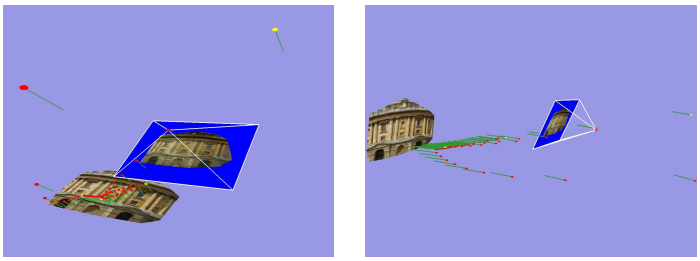
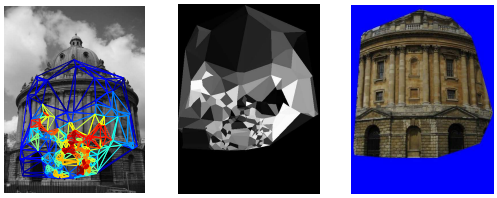


Figure 2: Affine 3–D visualization of the computed projective scene. Input cameras are marked by the yellow circles. The red circles show virtual cameras and their line of sight. The camera matrices are interpolated in projective space.



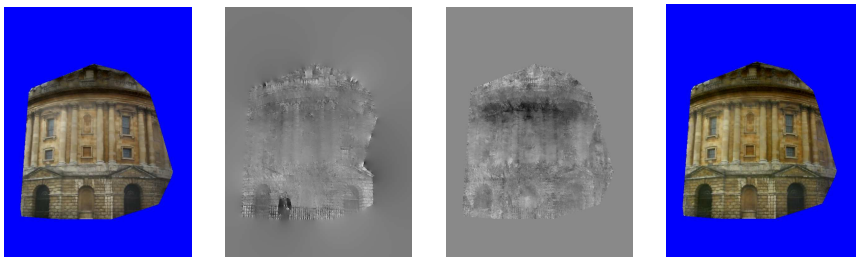
(a) (b) (c)

Figure 3: Novel view synthesis is done by warping the corresponding triangles in input view $I_i$, shown in (a), to the triangulated output view, (b). The rendered result, $I_i^q$ is shown in (c).

Instead we adopt a simple triangulation-based method for rendering novel views which, combined with our fusion technique, can generate desirable intermediate novel views. Instead of optimizing a single snapshot for each virtual view given all the input data, we create a novel view based on each measured input view $I_i$ at each virtual camera location to obtain a set of views $\mathbf{I}'_q = \{I_i^q = F_i^q(I_i) | I_i \in I\}$. The function $F_i^q$ transforms input view $I_i$ to the virtual view $I_i^q$. This function is defined through Delaunay triangulation of the projection of reconstructed 3–D points in the virtual camera view. The texture map for each triangle $\Delta_q^m$ in the virtual view is obtained by warping the corresponding triangle $\Delta_i^m$ in input view $I_i$ such that $\Delta_q^m = W_m(\Delta_i^m)$ as illustrated in Fig. 3. The resulting set $\mathbf{I}'_q$ consists of the views rendered from the same virtual point of view but from each input view independently. Therefore, the occlusions and the particular illumination of each input view is directly transformed into the corresponding virtual view $I_i^q$. Moreover, artifacts due to triangulation of non visible 3–D points in $I_i$ are also introduced in $I_i^q$. Another step is therefore necessary to deal with these problems.

## 2.3 Novel View Fusion

In this section we show how the rendered views $\mathbf{I}'_q$ can be fused to reduce the artifacts and occlusions and maintain consistent illumination. $\mathbf{I}'_q$ contains a set of $N$ (number of input views) intensity images which we refer to as $I(x,y,t)$, $t = 1,\ldots,N$, rendered from



(a) (b) (c) (d)

Figure 4: Input views are used to obtain a set of rendered novel views $\mathbf{I}'_q$ from the same virtual camera $q$. These novel views are decomposed into a reflectance image (a) and a set of illumination images $l(x,y,t)$, one of which is shown in (b). Note that the standing person is removed from the reflectance image and but appears in the illumination image. (c) Median filtering of the illumination images removes the artifacts. The filtered light can then be added to the reflectance image to re-light the resulting novel view (d).

the same point of view under varying light conditions. This set can be decomposed into a single reflectance image $R(x,y)$ and the corresponding illumination images $L(x,y,t)$. The logarithm of these images, denoted by lower case letters, are related by:

$$i(x,y,t) = r(x,y) + l(x,y,t) \quad t = 1,\ldots,N. \tag{1}$$

Weiss [12] showed that the maximum likelihood (ML) estimate of the gradient of the reflectance image is given by the median of the gradient input images $i(x,y,t)$. Agrawal *et al.* [1] further extended this approach to non Lambertian scenes by first estimating the reflectance image $r$ as proposed by Weiss and then removing scene texture edges from illumination images efficiently as follows. For each image $i$, a cross projection tensor $D^r$ is estimated using $r$ and $i$. $D^r$ is used to transform the gradient field $\nabla i$. This transformation removes all edges in $i$ which are present in $r$. The illumination images $l(x,y,t)$ are then computed by 2–D integration of the modified gradient field $\nabla l(x,y,t) = D^r \nabla i(x,y,t)$ for all $t$. We use this approach to compute at each novel view point a reflectance image and a set of illuminations images corresponding to the input views rendered from the virtual camera view through our triangulation method.

A direct benefit of light decomposition approach applied to novel view synthesis is that the resulting reflectance novel view is robust to rendering artifacts and occlusion as shown in Fig 4-a. This is thanks to the median filtering of the gradient images. However, the occluding objects in the input views and the missing areas due to triangulation will still be present in some of the illumination images, as shown in Fig. 4-b, but they can be effectively removed by filtering the illumination images. A simple median filter of the illumination images is sufficient to create a photo-realistic light image that can be added to the reflectance image to re-light the novel view as shown in Fig. 4-d. The computed illumination images can also be used to interpolate lights to create artificial shadows and illuminations in the process of rendering a novel view sequence without explicitly simulating a 3–D light source and computing all surface normals. This is illustrated in the next section.

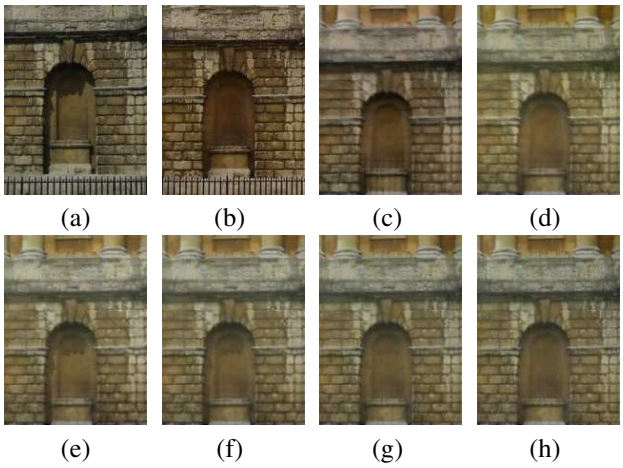



Figure 5: Illumination in novel views based on input views. (a) and (b) are the original input view segments. (c) Median novel view. Note the ghost effects of the bars as well as the imprinted shadow in the archway. (d) Fused novel view using our method with the shadows and the ghost effects removed. The second row shows controlled illumination, see text for details. Note the gradual dissolve of the added shadow in the archway.

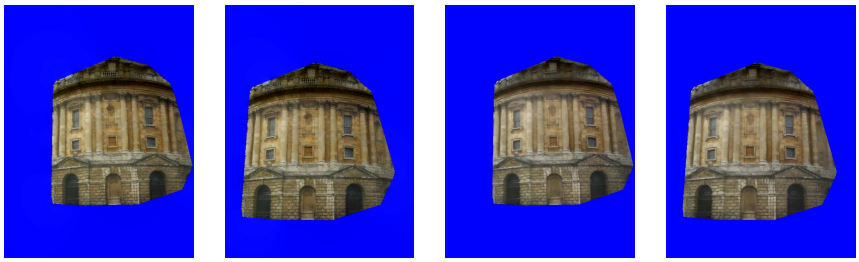
# 3 Experimental Results

In this section we demonstrate the performance of our prototype system on reconstruction and visualization of a landmark structure. The query shown in the top left image of Fig. 1 is passed to a functional visual search engine [2] that operates on online photo sharing websites such as Flickr. We keep the 7 top ranking retrieved images. The detected SIFT matches between all pairs of images are used to robustly estimate the fundamental matrices using RANSAC. The consistent tracks of all inlier matches across the input views are then used for bundle adjustment. We use a publicly available projective bundle adjustment code[1] to measure the projective structure and cameras.

Virtual camera trajectory is created by interpolation of 20 camera matrices between consecutive input cameras. The projections of the reconstructed 3–D points in each novel view are triangulated and used to re-render the input images. This yields a set of novel views, $\mathbf{I}'_q$, for each virtual camera as shown in the third row of Fig. 1. OpenGL implementation of the rendering algorithm is fast and the rendering time is in the order of a fraction of a second. The set of rendered views $\mathbf{I}'_q$ corresponding to virtual camera $Q$ is then decomposed into light and reflectance images. Matlab implementation of the cross projection tensor transformation and the 2–D integration of the gradient field from [1] was used to compute re-lit images. These operations take 4 minutes per frame on a 3.0GHz Pentium 4.

Fig. 5 illustrates the effects of input illumination in novel views. Fig. 5-a and b are the original input view segments, $I_1$ and $I_2$. Fig. 5-c shows a novel view rendered by taking the median of intensity of all novel views, $\mathbf{I}'_q$, at the virtual camera location. Undesirable

[1] http://cmp.felk.cvut.cz/∼svoboda/SelfCal/



(a) (b) (c) (d)

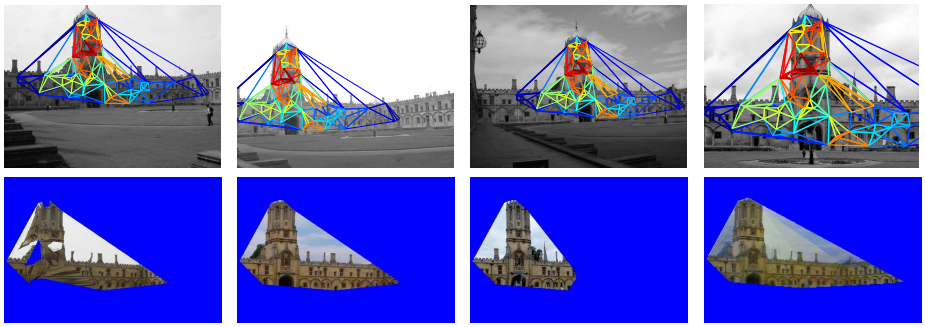
Figure 6: Novel views from various virtual camera positions. (a) and (b) are generated by adding the illumination computed from an input view to the reflectance image, this illumination contains shadows. (c) and (d) are the same generated novel views using the median of the input view illuminations to remove shadows.

artifacts such as ghost effects due to the bars visible in one input view as well as the imprinted shadow in the archway can be observed. Fig. 5-d shows the fused novel view using the median of the illumination images, $l_q(x,y,t)$ as proposed in this paper. Shadows and ghost effects are removed using this technique. The second row of Fig. 5 shows how illumination images corresponding to re-rendered input views $I_1^q$ and $I_2^q$ can be used to interpolate and simulate new illuminations. The generated illumination images are then added to the reflectance image $r_q(x,y)$ to obtain the final novel view.

Finally Fig. 6 shows some virtual snapshots at various locations along the virtual camera trajectory. Two types of illuminations have been used to re-light the computed reflectance image for each camera. The top row shows the results when using the illumination image corresponding to the re-rendered input image $I_1^q$. The second row shows the re-lit images using the median of the illumination images of the set $\mathbf{I}^q$.

In addition to novel view generation, the projective space can be upgraded to affine or metric by some user interaction to accommodate 3–D exploration of the scene. This procedure involves locating images of the plane at infinity (for affine reconstruction) or the absolute conic (for metric reconstruction) in a pair of images and reconstructing the vanishing points. The reconstructed affine 3–D features are shown in the second row of Fig. 1 and the textured map scene is shown in Fig. 2.

As a second reconstruction example, we used 20 images of Christ Church College in Oxford by selecting the top 20 results of the visual query search engine [2]. The projective structure and cameras were computed using robust bundle adjustment on SIFT features. We kept 9 camera matrices for which the reprojection error of the 3–D features were accurate. The top row of Fig. 7 shows the triangulated reprojection of the 3–D points in 4 of the 9 input images. Examples of the rendered novel views are shown in Fig. 7-(a-c). The set of 9 rendered novel views (from the same point of view) are then fused to obtain Fig. 7-d. Note that the rendered novel view in Fig. 7-a suffers from gross errors due to triangulation of erroneous 3–D points which project incorrectly into the corresponding input view. The triangulation error due to those points is negligible in other input views and therefore the final fused view (Fig. 7-d) does not suffer from those artifacts thanks to the decomposition and filtering scheme involved in the fusion algorithm.



(a) (b) (c) (d)

Figure 7: Reconstruction of the Christ Church college in Oxford from retrieved online photos. First row: some of the measure views and projection of measured 3–D points. Second row: (a)-(c) three rendered views from input views. view (a) contains artifacts due to projection of some errorneous 3–D points. (d) shows the fused view using 9 input images, note that the error in the rendered view (a) is corrected in the fused view.

# 4 Conclusion

We have proposed a novel methodology to use various computer vision techniques in an innovative way to build a prototype system that is capable of 3–D visualization of a scene based on online images starting from a small visual query associated to that site. Our approach can easily be integrated into advanced 3–D visualization technologies such as Microsoft Live Labs' Photosynth [9], where they currently use an offline image dataset of a landmark as well as estimations of the internal camera parameters for the reconstruction. Moreover, in their work the transition between images in 3–D space is based on simple view morphing.

In our approach, input images are automatically retrieved using visual search engines. In the absence of calibration data and other priors about the retrieved images, we proceed by reconstructing robust 3–D features and cameras in projective space using bundle adjustment. The scene can then be explored through virtual views generated in the projective space. Furthermore, the projective space can be upgraded by some user interaction to accommodate 3–D visualization of the scene in affine or metric space. The structure and camera matrices can be further exploited to obtain a quasi-dense cloud of points in the scene. A dense set of 3–D points improves the novel view synthesis results and can be considered as future improvements to the system.

Optimizing a single snapshot for each virtual view given all the input data is not feasible due the scattered data, occlusions and varying illumination. Instead, we have proposed a simple triangulation-based method to create novel views corresponding to virtual camera parameters based on the measured input images. We fuse the re-rendered views at each virtual camera by computing their intrinsic image and the corresponding illumination images. This technique effectively removes the occluding elements present in the retrieved views. Our fusion technique elegantly decouples the problem of occlusion handling from rendering, furthermore, it accommodates consistent and controlled illumination across the sequence of virtual views. The fusion step also increases robustness to outliers in the

input views retrieved by the visual search engine. This will be further investigated in our future work.

**Acknowledgements** We gratefully acknowledge the support of the EPSRC through grants EP/C007220/1, EP/D0370771/1 and GR/T24685.

# References

[1] Amit K. Agrawal, Ramesh Raskar, and Rama Chellappa. Edge suppression by gradient field transformation using cross-projection tensors. In *Conference on Computer Vision and Pattern Recognition*, pages 2301–2308, 2006.

[2] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proceedings of the 11th International Conference on Computer Vision, Rio de Janeiro, Brazil*, 2007.

[3] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multi-view stereopsis. In *Conference on Computer Vision and Pattern Recognition*, 2007.

[4] Pau Gargallo and Peter Sturm. Bayesian 3d modeling from images using multiple depth maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, California*, volume 2, pages 885–891, jun 2005.

[5] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (SIGGRAPH)*, 26(3), 2007.

[6] M. Lhuillier and L. Quan. Image interpolation by joint view triangulation. In *Conference on Computer Vision and Pattern Recognition*, pages 139–145, 1999.

[7] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 20(2):91–110, 2004.

[8] Y. Matsushita, K. Nishino, K. Ikeuchi, and M. Sakauchi. Illumination normalization with time-dependent intrinsic images for video surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1336–1347, 2004.

[9] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3d. In *ACM Transactions on Graphics (SIGGRAPH)*, pages 835–846, New York, NY, USA, 2006. ACM Press.

[10] Christoph Strecha, Rik Fransens, and Luc Van Gool. Combined depth and outlier estimation in multi-view stereo. In *Conference on Computer Vision and Pattern Recognition*, pages 2394–2401, Washington, DC, USA, 2006. IEEE Computer Society.

[11] G. Vogiatzis, P. H. S. Torr, and R. Cippola. Multi-view stereo via volumetric graph-cuts. In *Conference on Computer Vision and Pattern Recognition*, pages 391–399, 2005.

[12] Y. Weiss. Deriving intrinsic images from image sequences. In *International Conference on Computer Vision*, 2001.

[13] O. J. Woodford, I. D. Reid, and A. W. Fitzgibbon. Efficient new view synthesis using pairwise dictionary priors. In *Conference on Computer Vision and Pattern Recognition*, 2007.

[14] O. J. Woodford, I. D. Reid, P. H. S. Torr, and A. W. Fitzgibbon. On new view synthesis using multiview stereo. In *British Machine Vision Conference*, 2007.