Temporal Priors for Novel Video Synthesis

Ali Shahrokni, Oliver Woodford, and Ian Reid

Robotics Reseach Laboratory, University of Oxford, Oxford, UK
http://www.robots.ox.ac.uk/

Abstract. In this paper we propose a method to construct a virtual sequence for a camera moving through a static environment given an input sequence from a different camera trajectory. Existing image-based rendering techniques can generate photorealistic images given a set of input views, though the output images almost unavoidably contain small regions where the colour has been incorrectly chosen. In a single image these artifacts are often hard to spot, but become more obvious when viewing a real image with its virtual stereo pair, and even more so when when a sequence of novel views is generated, since the artifacts are rarely temporally consistent.
To address this problem of consistency, we propose a new spatio-temporal approach to novel video synthesis. The pixels in the output video sequence are modelled as nodes of a 3–D graph. We define an MRF on the graph which encodes photoconsistency of pixels as well as texture priors in both space and time. Unlike methods based on scene geometry which yield highly connected graphs, our approach results in a graph whose degree is independent of scene structure. The MRF energy is therefore tractable and we solve it for the whole sequence using a state-of-the-art message passing optimisation algorithm. We demonstrate the effectiveness of our approach in reducing temporal artifacts.

1 Introduction

This paper addresses the problem of reconstruction of a video sequence from an arbitrary sequence of viewpoints given an input video sequence. In particular, we focus on the reconstruction of a stereoscopic pair of a given input sequence captured by a moving camera through a static environment. This has application to the generation of 3-D content from commonly available monocular movies and videos for use with advanced 3-D displays.

Existing image-based rendering techniques can generate photorealistic images given a set of input views. Though the best results apparently have remarkable fidelity, closer inspection almost invariably reveals pixels or regions where incorrect colours have been rendered, as illustrated in Fig. 1. These are often, but not always, associated with occlusion boundaries, and while they are often hard to see in a single image, they become very obvious when a sequence of novel views is generated, since the artifacts are rarely spatio-temporally consistent. We propose to solve the problem via a Markov Random Field energy minimisation over

a *video sequence* with the aim of preserving spatio-temporal consistency and coherence throughout the rendered frames.

Two broad approaches to the novel-view synthesis problem are apparent in the literature: (i) multi-view scene reconstruction followed by rendering from the resulting geometric model, and (ii) image-based rendering techniques which seek simply to find the correct colour for a pixel. In both cases a data likelihood term $f(C, z)$ is defined over colour C and depth z which is designed to achieve a maximum at the correct depth and colour. In the multi-view stereo reconstruction problem the aim is generally to find the correct depth, and [1] was the first to suggest that this could be done elegantly for multiple input views by looking for the depth that maximises colour agreement between the input images.

Recent approaches such as [2, 3] involve quasi-geometric models for 3–D reconstruction where occlusion is modelled as an outlier process. Approximate inference techniques are then used to reconstruct the scene taking account of occlusion. Realistic generative models using quasi-geometric models are capable of rendering high quality images but lead to intractable minimisation problems [3].

More explicit reasoning about depth and occlusions is possible when an explicit volumetric model is reconstructed as in voxel carving such as [4, 5]. The direct application of voxel carving or stereo with occlusion methods [6–8] to our problem of novel *video* synthesis would, however, involve simultaneous optimisation of the MRF energy with respect to depth and colour in the *space-time* domain. The graph corresponding to the output video then becomes highly connected as shown in Fig. 2-a for a row of each frame. Unfortunately however, available optimisation techniques for highly connected graphs with non-submodular potentials are not guaranteed to reach a global solution [9].

In contrast, [10] marginalise the data likelihood over depth and thus have no explicit geometric reasoning about the depth of pixels. This and similar methods rely on photoconsistency regularised by photometric priors [10, 7] to generate photorealistic images. The priors are designed to favour output cliques which resemble samples in a texture library built from the set of input images.

It has recently been shown [11] that using small 2-pixel patch priors from a *local* texture library can be as effective as the larger patches used in [10]. [11] converts the problem of optimising over all possible colours, to a discrete labelling problem over modes of the photoconsistency function, referred to as colour modes, which can be enumerated *a priori*. Since the texture library comprises only pairs of pixels, the maximum clique size is two, and tree-reweighted message passing [12] can be used to solve for a strong minimum in spite of the non-submodular potentials introduced by enumerating the colour modes.

We closely follow this latter, image-based rendering approach, but extend it to sequences of images rather than single frames. We propose to define suitable potential functions between spatially and *temporally* adjacent pixels. This, and our demonstration of the subsequent benefits, form the main contribution of this paper. We define an MRF in space-time for the output video sequence, and optimise an energy function defined over the entire video sequence to obtain a solution for the output sequence which is a strong local minimum of the energy

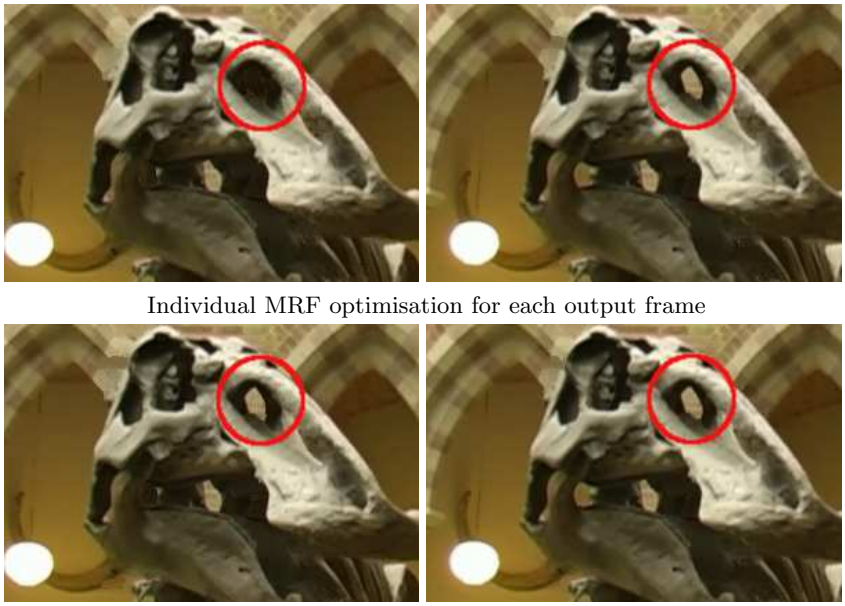


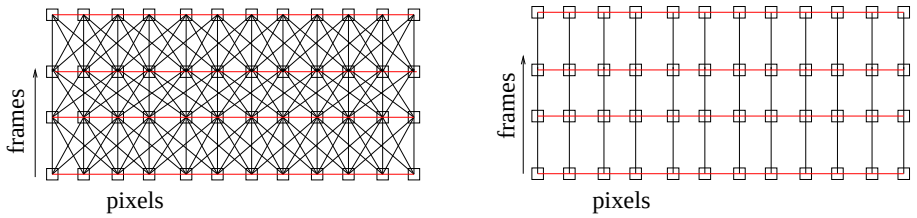
Fig. 1. A pair of consecutive frames from a synthesised sequence. Top row: individual MRF optimisation for each output frame fails to ensure temporal consistency yielding artifacts that are particularly evident when the sequence is viewed continuously. Bottom row: Using temporal priors, as proposed in this paper, to optimise an MRF energy over the entire video sequence reduces those effects. An example is circled.

function. Crucially, in contrast to methods based on depth information and 3-D occlusion, our proposed framework has a graph with a *depth-independent* vertex degree, as shown in Fig. 2-b. This results in a tractable optimisation over the MRF and hence we have an affordable model for the temporal flow of colours in the scene as the camera moves.

The remainder of this paper is organised as follows. In Section 2, we introduce the graph and its corresponding energy function that we wish to minimise, in particular the different potential terms. Section 3 gives implementation details, experimental results and a comparison of our method with (i) per-frame optimisation, and (ii) a naïve, constant-colour prior.

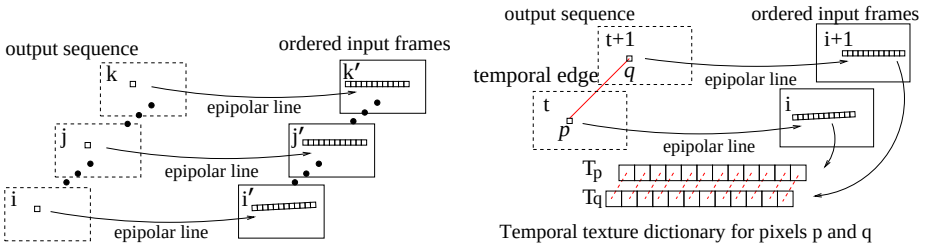
2 Novel Video Synthesis Framework

We formulate the MRF energy using binary cliques with pairwise, texture-based priors for temporal and spatial edges in the output video graph. Spatial edges in the graph exist between 8-connected neighbourhood of pixels in each frame. Temporal edges link pixels with the same coordinates in adjacent frames as shown in Fig. 2-b. Therefore, the energy of the MRF can be expressed in terms of the unary and binary potential functions for the set of labels (colours) $\mathcal{F}$ as



(a) (b)

Fig. 2. Temporal edges in an MRF graph for video sequence synthesis. a) Using a 3–D occlusion model all pixels on epipolar lines of pixels in adjacent frames must be connected by temporal edges (here only four temporal edges per pixel are shown to avoid clutter). b) Using our proposed temporal texture-based priors we can reduce the degree of the graph to a constant.



(a) (b)

Fig. 3. a) Local texture library is built using epipolar lines in sorted input views $\mathcal{I}$ for each pixel in the output video sequence. b) Local pairwise temporal texture dictionary for two output pixels p and q connected by a temporal graph edge.

follows.

$$E(\mathcal{F}) = \sum_p \phi_p(f_p) + \lambda_1 \sum_p \sum_{q \in \mathcal{N}_s(p)} \psi_{pq}(f_p, f_q) + \lambda_2 \sum_p \sum_{q \in \mathcal{N}_t(p)} \psi_{pq}(f_p, f_q) \quad (1)$$

where f_p and f_q are labels in the label set $\mathcal{F}$, ϕ is the unary potential measuring the photoconsistency and ψ encodes the pairwise priors for spatial and temporal neighbours of pixel p denoted by $\mathcal{N}_s(p)$ and $\mathcal{N}_t(p)$ respectively. λ_1 and λ_2 are weight coefficients for different priors. The output sequence is then given by the optimal labelling $\mathcal{F}^*$ through minimisation of E:

$$\mathcal{F}^* = \underset{\mathcal{F}}{\operatorname{argmin}}\{E(\mathcal{F})\} \quad (2)$$

Next, we first discuss the texture library for spatial and temporal terms and introduce some notations and then define the unary and binary potentials.

2.1 Texture Library and Notations

To calculate the local texture library, we first find and sort subsets of the input frames with respect to their distance to the output frames. We denote these subsets by $\mathcal{I}$. The input frame in $\mathcal{I}$, which is closest to the output frame containing

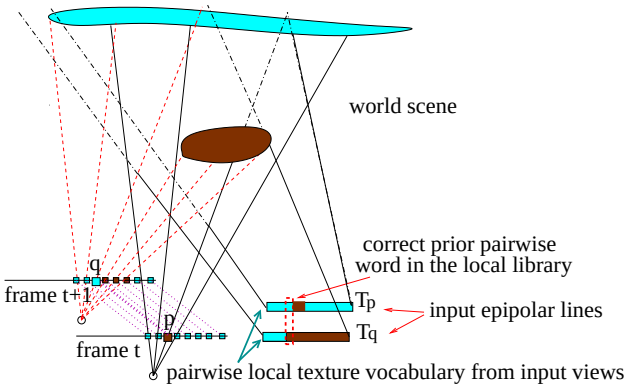


Fig. 4. The temporal transition of colours between pixels in two output frames. A constant colour model between temporally adjacent output pixels p and q is clearly invalid because of motion parallax. On the other hand, there is a good chance that the local texture vocabulary comprising colour pairs obtained from the epipolar lines T_p and T_q (respectively the epipolar lines in the corresponding input view of the stereo pair) captures the correct colour combination, as shown in this case.

pixel p is denoted by $\mathcal{I}(p)$. Then for each pairwise clique of pixels p and q, the local texture library is generated by bilinear interpolation of pixels on the clique epipolar lines in $\mathcal{I}$ as illustrated in Fig. 3. For a pixel p the colour in input frame k corresponding to the depth disparity z is denoted by $C_k(z, p)$. The vocabulary of the library is composed of the colour of the pixels corresponding to the same depth on each epipolar line and is defined below.

$$\mathcal{T} = \{(C_i(z,p), C_j(z,q)\ ) \mid z = z_{min}, \ldots, z_{max}\ ,\ i = \mathcal{I}(p), j = \mathcal{I}(q)\} \tag{3}$$

we also define T_p as the epipolar line of pixel p in $\mathcal{I}(p)$,

$$T_p = \{C_k(z,p) \mid z = z_{min}, \ldots, z_{max}\ ,\ k = \mathcal{I}(p)\}. \tag{4}$$

2.2 Unary Potentials

Unary potential terms express the measure of agreement in the input views for a hypothesised pixel colour. Since optimisation over the full colour space can only be effectively achieved via slow, non-deterministic algorithms, we use instead a technique proposed in [11] that finds a set of photoconsistent colour modes. The optimisation is then over the choice of *which mode*, i.e. a discrete labelling problem. These colour modes are denoted by f_p for pixel p and using their estimated depth z the unary potential is given by the photoconsistency of f_p in a set of close input views V:

$$\phi_p(f_p) = \sum_{i \in V} \rho(||f_p - C_i(z,p)||) \tag{5}$$

where $\rho(.)$ is a truncated quadratic robust kernel.

2.3 Binary Potentials

Binary (pair-wise) potentials in graph-based formulation of computer vision problems often use the Potts model (piece-wise constant) to enforce smoothness of the output (e.g. colour in segmentation algorithms, or depth in stereo reconstruction). While the Potts model is useful as a regularisation term, its application to temporal cliques is strictly incorrect. This is due to the relative motion parallax between the frames as illustrated in Fig. 4. In general, the temporal links marked by dotted lines between two pixels p and q for example do not correspond to the same 3–D point and therefore colour coherency assumption using the Potts model is invalid.

Instead, we propose to use texture-based priors to define pairwise potentials in temporal edges. As shown in Fig. 4, a local texture library given by Eq. 3 for the clique of pixels p and q is generated using epipolar lines T_p and T_q defined in Eq. 4 in two successive input frames close to the output frames containing p and q. This library contains the correct colour combination for the clique containing p and q corresponding to two distinct 3-D points (marked by the dotted rectangle in Fig. 4. This idea is valid for all temporal cliques in general scenes provided that there exists a pair of successive input frames throughout the whole sequence which can see the correct 3–D points for p and q.

Each pairwise potential term measures how consistent the pair of labels for pixels p and q is with the (spatio-temporal) texture library. The potential is taken to be the minimum over all pairs in the library, viz:

$$\psi_{pq}(f_p, f_q) = \min_z \left\{ \rho(||f_p - T_p(z)||) + \rho(||f_q - T_q(z)||) \right\}. \tag{6}$$

Note that the use of a robust kernel $\rho(.)$ ensures that cases where a valid colour combination does not exist are not overly penalised; rather, if a good match cannot be found a constant penalty is used.

As explained above, exploiting texture-based priors enables us to establish a valid model for temporal edges in the graph which is independent of the depth and therefore avoid highly connected temporal nodes. This is an important feature of our approach which implies that the degree of the graph is independent of the 3–D structure of the scene.

3 Implementation and Results

We verified the effectiveness of temporal priors for consistent novel video synthesis in several experiments. We compare the generated views with and without temporal priors. In all cases, the spatial terms for all 8-connected neighbours in each frame in the MRF energy were similarly computed from texture-based priors. Therefore the focus of our experiments is on the texture-based temporal priors. We also show results from using a the simpler constant-colour prior (the Potts model).

The energy function of Eq. 1 is minimised using a recently introduced enhanced version of tree-reweighted max-product message passing algorithm known

(a) frame #2 (b) frame #5 (c) magnified part of #2 (d) magnified part of #5

Fig. 5. Top row, results obtained using our proposed texture-based temporal priors. Middle row, using the Potts temporal priors. Bottom row, individual rendering of frames. Columns (c) and (d) show the details of rendering. It can be noted that the Potts model and individual optimisation fails on the sharp edges of the leaves.

as TRW-S algorithm [12] which can handle non submodular graph edge costs and has guaranteed convergence properties. For an output video sequence with n frame of size $W \times H$, the spatio-temporal graph would have $n \times W \times H$ vertices and $(n-1) \times W \times H$ temporal edges in the case of using texture-based temporal priors or the Potts model. This is the minimum number of temporal edge for a spatio-temporal MRF and any other prior based on depth with number of disparities z would require at least $z \times (n-1) \times W \times H$ temporal edges, where z is of the order of 10 to 100. Typical run time to process a space-time volume of $15 \times 100 \times 100$ pixels is 600 seconds on a P4 D 3.00GHz machine. The same volume when treated as individual frames takes $30 \times 15 = 450$ seconds to process.

The input video sequence is first calibrated using commercial camera tracking software[1]. The stereoscopic output virtual camera projection matrices are then generated from input camera matrices by adding a horizontal offset to the input camera centres. The colour modes as well as unary photoconsistency terms given by Eq. 5 for each pixel in the output video are calculated using 8 closest views in the input sequence. We also compute 8 subsets $\mathcal{I}$'s for texture library computation as explained in Section 2.3 with the lowest distance to the ensemble

[1] Boujou, 2d3 Ltd.

of the n output camera positions. Finally in Eq. 1 we set λ_1 to 1 and λ_2 to 10 in our experiments.

Fig. 5 shows two synthesised frames of a video sequence of a tree and the details of rendering around the leaves for different methods. Here, in the case of temporal priors (textured-based and Potts) 5 frames of 300×300 pixels are rendered by a single energy optimisation. In the detailed view, it can be noted that the quality of the generated views using texture-based temporal priors has improved especially around the edges of the leaves.

As another example, Fig. 6 shows some frames of the novel video synthesis on the Edmontosaurus sequence using different techniques. Here the temporal priors are used to render 11 frames of 200×200 pixels by a single energy optimisation. The first row shows the results obtained using our proposed texture-based temporal prior MRF. Using the Potts model for temporal edges generates more artifacts as shown in the second row in Fig. 6. Finally the third rows show the results obtained without any temporal priors and by individual optimisation of each frame. It can be noted that the background is consistently seen through the holes in the skull, while flickering artifacts occur in the case of the Potts prior and individual optimisation. Here the output camera matrices are generated by interpolation between the first and the last input camera positions. Finally Fig. 7 show the entire stereoscopic frames constructed using temporal priors over 15 frames.

4 Conclusion

We have introduced a new method for *novel video rendering* with optimisation in space-time domain. We define a Markov Random Field energy minimisation for rendering a video sequence which preserves temporal consistency and coherence throughout the rendered frames. Our method uses a finite set of colours for each pixel with their associated likelihood cost to find a global minimum energy solution which satisfies prior temporal consistency constraints in the output sequence.

In contrast to methods based on depth information and 3–D occlusion we exploit texture-based priors on pairwise cliques to establish a valid model for temporal edges in the graph. This approach is independent of the depth and therefore results in a graph whose degree is independent of scene structure. As a result and as supported by our experiments, our approach provides a method to reduce temporal artifacts in novel video synthesis without resorting to approximate generative models and inference techniques to handle multiple depth maps. Moreover, our algorithm can be extended to larger clique texture-based priors while keeping the degree of the graph independent of the depth of the scene. This requires sophisticated optimisation techniques which can handle larger cliques such as [13, 14] and will be investigated in our future work. Quantitative analysis of the algorithm using synthetic/real stereo sequences is also envisaged to further study the efficiency of temporal priors for video synthesis.

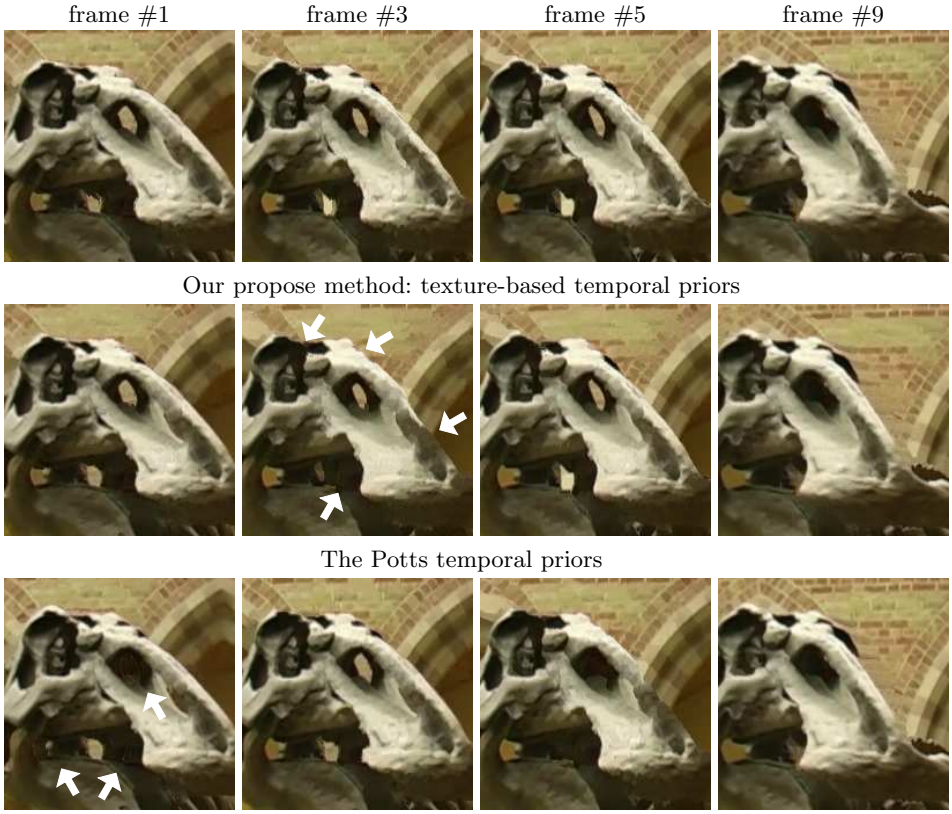


Fig. 6. Synthesised Edmontosaurus sequence. First row, results obtained using our proposed texture-based temporal priors. Second row, using the Potts temporal priors creates some artifacts (frame #3). Third row, individual rendering of frames introduces artifacts in the holes (the nose and the jaw). Also note that the quality of frame #5 has greatly improved thanks to the texture-based temporal priors.

Acknowledgements This work was supported by EPSRC grant EP/C007220/1 and by a CASE studentship sponsored by Sharp Laboratories Europe. The authors also wish to thank Andrew W. Fitzgibbon for his valuable input.

References

1. Okutomi, M., Kanade, T.: A multiple-baseline stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence **15**(4) (1993) 353–363
2. Strecha, C., Fransens, R., Gool, L.V.: Combined depth and outlier estimation in multi-view stereo. In: Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, IEEE Computer Society (2006) 2394–2401

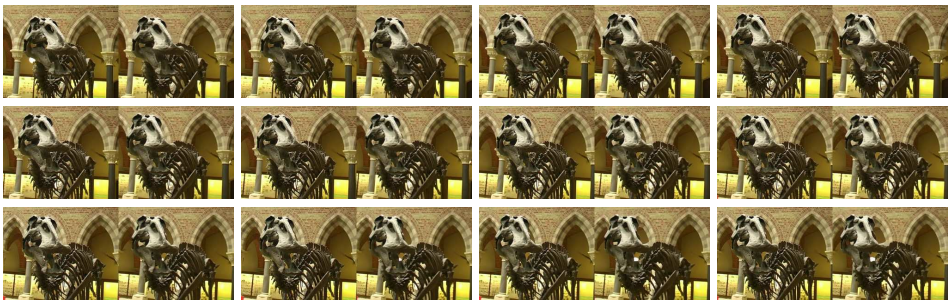


Fig. 7. Stereoscopic frames generated using texture-based temporal priors over 15 frames. In each frame, the left image is the input view (corresponding to the left eye) and the right image is the reconstructed right eye view.

3. Gargallo, P., Sturm, P.: 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. (2005) 885–891
4. Goesele, M., Seitz, S.M., Curless, B.: Multi-View Stereo Revisited. In: Conference on Computer Vision and Pattern Recognition, New York, USA (2006)
5. Kutulakos, K., Seitz, S.: A Theory of Shape by Space Carving. International Journal of Computer Vision **38**(3) (2000) 197–216
6. Kolmogorov, V., Zabih, R.: Multi-Camera Scene Reconstruction via Graph Cuts. In: European Conference on Computer Vision, Copenhagen, Denmark (2002)
7. Sun, J., Zheng, N., Shum, H.: Stereo matching using belief propagation. IEEE Transactions on Pattern Analysis **25** (2003) 1–14
8. Tappen, M., Freeman, W.: Comparison of graph cuts with belief propagation for stereo,using identical MRF parameters. In: International Conference on Computer Vision. (2003)
9. Kolmogorov, V., Rother, C.: Comparison of energy minimization algorithms for highly connected graphs. In: European Conference on Computer Vision, Graz, Austria (2006)
10. Fitzgibbon, A., Wexler, Y., Zisserman, A.: Image-based rendering using image-based priors. In: Proceedings of the International Conference on Computer Vision. Volume 2. (2003) 1176–1183
11. Woodford, O.J., Reid, I.D., Fitzgibbon, A.W.: Efficient new view synthesis using pairwise dictionary priors. In: Conference on Computer Vision and Pattern Recognition. (2007)
12. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. IEEE Trans. Pattern Anal. Mach. Intell. **28**(10) (2006) 1568–1583
13. Kohli, P., Kumar, M.P., Torr, P.H.: P3 & Beyond: Solving Energies with Higher Order Cliques. In: Conference on Computer Vision and Pattern Recognition. (2007)
14. Potetz, B.: Efficient Belief Propagation for Vision Using Linear Constraint Nodes. In: Conference on Computer Vision and Pattern Recognition. (2007)