Efficient new-view synthesis using pairwise dictionary priors

O. J. Woodford I. D. Reid
Department of Engineering Science,
University of Oxford

A. W. Fitzgibbon
Microsoft Research,
Cambridge, U.K.

Abstract

New-view synthesis (NVS) using texture priors (as opposed to surface-smoothness priors) can yield high quality results, but the standard formulation is in terms of large-clique Markov Random Fields (MRFs). Only local optimization methods such as iterated conditional modes, which are prone to fall into local minima close to the initial estimate, are practical for solving these problems.

In this paper we replace the large-clique energies with pairwise potentials, by restricting the patch dictionary for each clique to image regions suitable for that clique. This enables for the first time the use of a global optimization method, such as tree-reweighted message passing, to solve the NVS problem with image-based priors.

We employ a robust, truncated quadratic kernel to reject outliers caused by occlusions, specularities and moving objects, within our global optimization. Because the MRF optimization is thus fast, computing the unary potentials becomes the new performance bottleneck. An additional contribution of this paper is a novel, fast method for enumerating color modes of the per-pixel unary potentials, despite the non-convex nature of our robust kernel. We compare the results of our technique with other rendering methods, and discuss the relative merits and flaws of regularizing color, and of local versus global dictionaries.

1. Introduction

The new-view synthesis (NVS) problem is this: given a set of images $\mathcal{I}_1$ to $\mathcal{I}_n$ of a 3D scene, compute the image $\mathcal{V}$ which would be obtained by placing the camera at a given new viewpoint that is not in the original set. It is a poorly constrained inverse problem and therefore effective priors are important if a good solution is to be obtained. The priors considered in this paper will all be defined in terms of energy functions applied to overlapping windows, or patches, in $\mathcal{V}$. The parameter which most strongly controls the tractability of these problems is the size of these patches (the *order* of the prior). Historically, there has been a tradeoff in the selection of priors: high order priors can model the complex structures of the natural world (hair, trees, etc.) but tend to result in very difficult inference problems [6, 14]; while tractable (low order) priors can model only simple scene classes such as piecewise smooth shapes [11, 10]. The contribution of this paper is to exploit the structure of the NVS problem to represent natural scenes using tractable priors.

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

Figure 1. **New view synthesis**. (a) A synthesized, maximum-likelihood view, showing a large region of error. (b) Best result with high-order cliques [6], which uses discriminative, 5×5 patches, but with obligatory local optimization. While the optimum solution to this problem may well fix the errors, the optimizer cannot reach it, given (a) as start position. (c) Using 2×1 cliques with a sequence-specific, global dictionary. While a strong optimum to this problem can be found, the regularizer is not discriminative enough to correct the error. (d) Using 2×1 cliques with a local dictionary. The regularizer provides powerful discrimination, while enabling the optimizer to find a strong minimum.

This is not to say that good results cannot currently be obtained for NVS, under appropriate conditions. In stereo reconstruction for example, the assumption of piecewise smooth scene geometry can be expressed in an energy minimization framework which encodes the constraint that the depths of neighboring pixels in the solution should be similar, so the prior is of order 2. Depending on the precise form of the prior, a global optimum (or strong local optimum) of the energy can be computed efficiently using powerful inference algorithms[1] such as tree-reweighted message passing [10] and graph cuts [1, 11]. When the piecewise smoothness assumption is approximately valid for the scene, these solutions yield high-quality synthesized views. However, as can be seen in figure 5, these methods produce unrealistic, blocky artifacts when applied to natural scenes.

[1] Throughout the paper we shall use the term "global optimizer" to refer to algorithms of this type, which, although not necessarily guaranteed to find global optima, find strong optima in practise

1-4244-1180-7/07/$25.00 ©2007 IEEE

It is possible to encode natural-world priors using patch dictionaries, which give excellent results for tasks such as constrained texture synthesis [12], inpainting [5] and new-view synthesis [6]. However, energy minimization under these priors does not allow the use of global optimizers, so techniques such as iterated conditional modes (ICM) or simulated annealing must be used, with the associated poor tolerance to local minima or high computational cost respectively. The computational burden can be reduced by improving patch lookup [13, 18], but this does not fix the convergence problem. Criminisi *et al*. [3] use much smaller, local patch dictionaries (by restricting patch search to be near epipolar lines), which allows real-time computation of the prior, but again rely on ICM to impose the prior. Roth and Black's "fields of experts" framework [14] replaces dictionary lookup with a continuous, filter-based prior, so that ICM may be replaced by gradient descent. In all cases, however, the dependence on local optimization remains, as does the concomitant requirement that the initial estimate of the solution be close to a good optimum. Woodford *et al*. [19] showed that simulated annealing does little to improve matters—if ICM converges to a poor solution, this typically means that large coherent search steps must be made to reach another optimum, and simulated annealing has a vanishingly small likelihood of making those steps.

As noted above, priors such as piecewise smoothness in depth can be defined as the sum of energies defined on 2-pixel patches, which enables very efficient inference. Priors over intensity rather than depth images, however, do not admit such a compact definition. From studies of the statistics of natural scenes [9, 15] the distribution of 2-pixel intensity patches is known to be well modeled by a t-distribution. When converted to a prior, this ultimately means that the most correct possible second order prior for natural intensity images is simply piecewise constant color—a poor regularizer for textured natural scenes. This means that to model general natural scenes we must go to larger patch sizes and hence to intractable inference.

In new-view synthesis, however, the prior need not model all natural scenes; rather it should bias the output view to look like the input sequence. Thus one might try to learn a second order prior just over the input images. Figure 2 shows that this restriction is still equivalent to imposing piecewise smoothness, and we shall see that it fails to give a sufficiently powerful prior. However, narrowing the training samples further, to small regions of the input sequence, does usefully regularize the problem. We compare the new local prior to previous approaches on some image sequences containing complex geometry, and show that it achieves better solutions in considerably less time than previous methods.

The main contribution of this paper then is to show how patch priors on NVS can be reduced to pairwise priors, which allows for global inference. We argue that this is the first such reduction: although patch-based methods have previously been expressed using pairwise energies [4, 7], this is only for special problems in which the patches overlap only in pairs. This is because the number of unary terms is smaller than the number of output pixels: in super-resolution [7], the number of unary terms is equal to the number of low-resolution input pixels; in texture quilting [4] the unary terms exist only at the boundaries of the region to be painted. In the NVS problem, there must be one patch per output pixel, with dense overlap. Reducing the overlap would require using a larger patch size when computing photoconsistency, which would further require the assumption of piecewise smooth scene geometry.

In overview, our algorithm has two main steps: 1. the continuous problem of determining color at every output pixel is converted to a discrete problem by computing a small number of modes of the photoconsistency likelihood at every pixel. 2. One of these modes is selected at every pixel in order to maximize a combination of photoconsistency with the prior term, which prefers that patches of the output image look like patches from the input sequence. These steps are discussed in §3 and §4, after which §5 describes experimental comparisons of the new method with the state of the art.

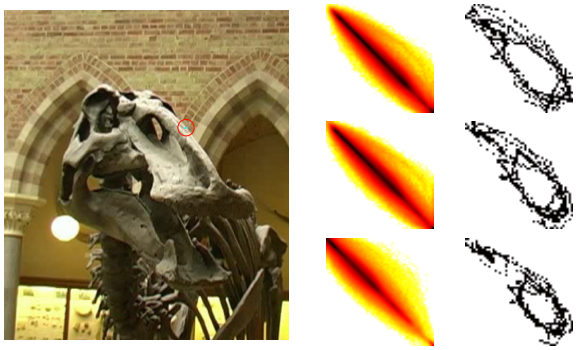
2. Notation

The new view $\mathcal{V}$ is a set of pixel colors $\{V(i)\}_{i=1}^{M}$, defined in some appropriate color space, say $\mathbb{R}^3$. Pixels are indexed by integers, in raster-scan order. A *neighborhood*, or *clique*, is a set of indices. For example, for a $W \times H$ image, the 4-connected neighbors of a non-boundary pixel i are the set $\mathtt{N} = \{i - W, i - 1, i + 1, i + W\}$. The set of neighboring colors may then be written $V(\mathtt{N})$. A *neighborhood system* is a set of neighborhoods, $\{\mathtt{N}_j\}_{j=1}^{N}$ where N is the total number of neighborhoods in the image. We shall use a variety of neighborhood systems:

- The *patch neighborhood* is denoted $\mathtt{P}_j$, which for concreteness we shall say is a set containing the indices of pixels in the 5×5 window centered at pixel j. Again $V(\mathtt{P}_j)$ is the patch viewed as a 5×5 image. Ignoring boundary effects, $N = M$, i.e. the number of cliques is the number of pixels.

- The 4-connected neighbor system is the set $\{\mathtt{C4}_j\}_{j=1}^{2M}$, with two cliques per pixel, again ignoring boundary bookkeeping, which might comprise the "north" cliques $\{i, i - W\}$ and the "east" cliques $\{i, i + 1\}$.

- The 8-connected neighbor system is the set $\{\mathtt{C8}_j\}_{j=1}^{4M}$, with four cliques per pixel which add to those of the 4-connected system the "north-east" clique $\{i, i-W+1\}$ and the "south-east" clique $\{i, i + W + 1\}$.



(a) (b) (c)

Figure 2. **Pairwise color priors**. (a) Part of an image from the "Edmontosaurus" sequence. (b) Pairwise, negative log histograms of all 2x1 patches $[I(i), I(i+1)]$ in the input sequence, where I is (top to bottom respectively) the red, green and blue channel of the input patches. The dominant diagonals show that the global prior derived from the image sequence is still non-specific, effectively imposing the piecewise smoothness constraint that $I(i) \approx I(i+1)$. (c) Binary histograms for the red, green and blue channel of horizontal 2x1 patches in the local dictionary generated for the pixel highlighted in (a). We show that local pairwise priors are more specific, producing better results.

We shall also consider the depth map $\mathcal{Z}$, where $z(i)$ is the depth of the scene at pixel i. We cast the problem in the framework of energy minimization, so the goal is to find the $\mathcal{X}$ (either $\mathcal{V}$ or $\mathcal{Z}$ or both) which minimizes an energy function

$$E(\mathcal{X}|\phi_{1..M}, \psi_{1..N}) = \underbrace{\sum_i \phi_i[X(i)]}_{\text{unary energy}} + \underbrace{\sum_j \psi_j[X(\mathtt{N}_j)]}_{\text{prior (or clique) energy}} \tag{1}$$

where the functions ϕ_i and ψ_j can be computed as a function of the input data. In [6] for example, $\phi_i : \mathbb{R}^3 \to \mathbb{R}^+$ is a measure of photoconsistency, and $\psi_j : \mathbb{R}^{5\times5\times3} \to \mathbb{R}^+$ gives the squared distance of the patch $\mathbf{X} = V(\mathtt{P}_j)$ to the closest patch in a dictionary $\mathbb{T}$ of exemplar patches, $\psi(\mathbf{X}) = \min_{\mathbf{T}\in\mathbb{T}} \|\mathbf{T} - \mathbf{X}\|^2$. Note that in this case ϕ varies with i, while ψ is independent of j. As another example, piecewise smooth regularization of depth has a similar unary term, but the neighborhood system is 4-connected ($\mathtt{N}_j = \mathtt{C4}_j$), and the prior term takes the form $\psi(\{z, z'\}) = \varrho(|z - z'|)$ where $\varrho(\cdot)$ is a robust kernel, for example the truncated quadratic $\varrho(t) = \min(t^2, 1)$.

As discussed above, clique size is the parameter which has most effect on tractability of the minimization, but a second important factor is the discretization of V. Although the above energies are written in terms of continuous variables, V and z, efficient optimization under arbitrary priors is possible only for discrete variables. To directly discretize V space—about 10^7 values for 8-bit RGB images—would be impractical. Our approach is to maintain a small set of potential colors at each pixel (see §3.1). This set is computed offline, so the minimization, rather than being over $\mathcal{V}$, is over a label image $\mathcal{L}$, with the energies ϕ and ψ being redefined with appropriate bookkeeping.

3. Unary energy: photoconsistency

We begin by defining the unary energy ϕ which measures photoconsistency at every pixel. We use the following photoconsistency term, from [6]:

$$E_{\text{photo}}(V(i), z(i)) = \sum_{k=1}^{n} \rho(\|\mathbf{C}_i(k, z(i)) - V(i)\|) \tag{2}$$

where $\mathbf{C}_i(k, z)$ is the color (bilinearly interpolated) of the pixel in image $\mathcal{I}_k$ corresponding to pixel i in $\mathcal{V}$ at depth z, where depth is measured in the coordinate system of image $\mathcal{V}$. We assume that all camera projection matrices are known and that the reader is familiar with the projection of points between views [8]. The robust kernel $\rho(\cdot)$ will generally be the truncated quadratic model

$$\rho(x) = \min(x^2, \tau^2) \tag{3}$$

where τ is a tuning parameter of the algorithm. This model assumes that pixels are generated either using an inlier process, whereby the input image samples are normally distributed, noisy measurements of some true color, or an outlier process, which models all other samples. It is similar to the generative model based approach of Strecha *et al*. [16].

For pure new-view synthesis, we are interested only in the color at each pixel, in which case E_{photo} is a function only of $V(i)$. Conversely, when doing multi-view reconstruction of depth, it is a function only of z. Therefore we define two "overloads" of E_{photo} for these cases:

$$E_{\text{photo}}(V(i)) = \min_{z_{\min}<z<z_{\max}} E_{\text{photo}}(V(i), z) \tag{4}$$

$$E_{\text{photo}}(z(i)) = \min_{V\in\mathbb{R}^3} E_{\text{photo}}(V, z(i)) \tag{5}$$

where $z_{\min}$ and $z_{\max}$ represents a bounding volume in the scene, determined during the camera calibration stage.

3.1. Discretization

As noted above, both color V and depth z are naturally continuous variables, which must be discretized for efficient optimization. We discretize z into a range of 50 to 100 steps between $z_{\min}$ and $z_{\max}$, spaced so that steps of one quantum in z correspond to steps of at most one pixel in the input images. Thus, z in the following may be considered an integer index.

Following [6] we enumerate the local minima of $E_{\text{photo}}(V)$ at each pixel, typically around five to twenty colors. We refer to these minima as *modes*, i.e. maxima of the pseudo-likelihood $p(V) = \exp(-E_{\text{photo}}(V))$.

Fitzgibbon *et al*. [6] propose a gradient descent method for finding the modes. However, this method is not only slow (they quote a time of 0.1 seconds per pixel) but we have found it to fail occasionally in finding all modes. Instead, we propose a simple, deterministic method for finding color modes over all depths, given color modes at a particular depth (addressed in §3.2). Let us, writing V for $V(i)$, define V' as a mode of $E_{\text{photo}}(V)$. The requirement of our modes is

$$E_{\text{photo}}(V') < E_{\text{photo}}(V' + \delta V) \tag{6}$$

for all sufficiently small $\delta V \in \mathbb{R}^3$. From equation 2 it is possible to define the depth at which the mode V' can be found thus

$$z' = \underset{z_{\min}<z<z_{\max}}{\operatorname{argmin}} \; E_{\text{photo}}(V', z) \tag{7}$$

Consider that if

$$E_{\text{photo}}(V', z') \geq E_{\text{photo}}(V' + \delta V, z'), \tag{8}$$

then equation 6 cannot be true. As a result, it can be seen that a color mode over all depths has the following properties: a) it must necessarily be a color mode of the depth z', and b) the depth z' must also be the depth at which $E_{\text{photo}}(V', z)$ is lowest. A deterministic method of finding color modes is therefore:

1. For each $z(i)$, enumerate all minima of $E_{\text{photo}}(V, z(i))$.
2. For each such minimum, denoted $(V_{\min}, z(i))$, reject it as a mode if $E_{\text{photo}}(V_{\min}, z(i)) > \min_z E_{\text{photo}}(V_{\min}, z)$.

3.2. Enumerating color modes

The problem of enumerating all minima of $E_{\text{photo}}(V, z)$ at a given pixel i and depth z depends on the form of the robust kernel ρ. For example, a quadratic kernel is trivially shown to have a single mode: the mean. This closed form computation means that modes over color for this kernel can be computed quickly and reliably using the algorithm of the previous section. Indeed, the reliability of our method in finding modes over color depends only on the ability to find them at each depth. It is therefore generally optimal for all convex kernels (*e.g*. quadratic, absolute, Huber), which produce a single mode at each depth, as those modes can always be computed using a standard non-linear optimizer.

Our robust kernel is a truncated quadratic (which is not convex), a mode of which can be shown to be the mean of inlier samples. We therefore use a mean shift algorithm [2] to find these modes. Starting with an initial estimate V_0, we iterate the following update function:

$$V_{t+1} = \frac{\sum_{k=1}^{n} \mathbf{C}_i(k,z) g(\|\mathbf{C}_i(k,z) - V_t\|^2)}{\sum_{k=1}^{n} g(\|\mathbf{C}_i(k,z) - V_t\|^2)} \tag{9}$$

$$g(x) = \begin{cases} 1 & \text{if } x < \tau^2, \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

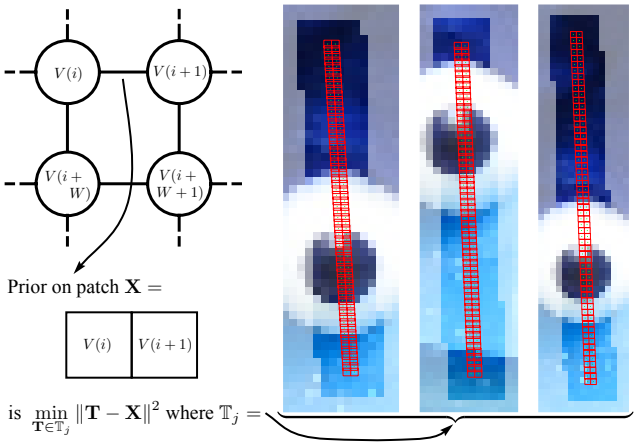


Figure 3. **Local patch dictionary**. The prior on the 2-pixel clique $\mathtt{N} = \{i, i+1\}$ is defined using the local patch dictionary $\mathbb{T}_{\mathtt{N}}$. The set of patches making up the dictionary described in §4.1, for $D = 3$, is the set of all 1×2 patches in the unshaded regions of the images. The set of patches in the dictionary described in §4.2 is given by the reprojected patches, represented by the red outlines and sampled at each of the green pairs of points, which lie on the epipolar lines corresponding to pixels i and $i+1$.

The algorithm, which stops when $V_{t+1} = V_t$, is guaranteed to converge on a mode. As our aim is to reduce the number of labels we must choose from at each pixel, we reject any color modes which have only one inlier (i.e. $\sum_{k=1}^{n} g(\|\mathbf{C}_i(k,z) - V\|^2) = 1$), as these may be many, and exhibit no consistency between input images. While this may result in the rejection of the correct color, this only occurs when it is visible in only one view, which is rarely the case.

Given that we are trying to find modes for which $\sum_{k=1}^{n} g(\|\mathbf{C}_i(k,z) - V\|^2) \geq 2$, we initialize our mean shift algorithm at the following points:

$$V_0 = \frac{\mathbf{C}_i(k,z) + \mathbf{C}_i(j,z)}{2} \tag{11}$$

for all $k, j \in \{1, .., n\}$ for which $k \neq j$ and

$$\|\mathbf{C}_i(k,z) - \mathbf{C}_i(j,z)\|^2 < (2\tau)^2 \tag{12}$$

Using these starting positions does not guarantee that all minima will be located but in practice performs well and is efficient to compute.

After the above procedure, we have a list of modes at each pixel, denoted

$$\text{Modes}(i) = \{V(i,m)\}_{m=1}^{\text{Nmodes}_i} \tag{13}$$

so that the optimization problem is to choose a label $l(i)$ at every pixel to minimize equation 1. The unary energy then takes the form

$$\phi_i[\text{label}] = E_{\text{photo}}(V(i, \text{label})). \tag{14}$$

We now proceed to define the clique energy which imposes the prior.

4. Clique energy: texture prior

Our proposed algorithm uses the same non-parametric, nearest-neighbor patch-lookup regularizer as [6], but with a local patch dictionary for each clique, as introduced in [3], and with a change of clique size, from 5×5 to 2×1. For clique j containing the two pixels, s and t, our clique potential can therefore be expressed as:

$$E_{\text{texture}}(V(s), V(t)) = \min_{\mathbf{T} \in \mathbb{T}_j} \|\mathbf{T} - [V(s), V(t)]\|^2 \quad (15)$$

The patch dictionary for this specific clique, $\mathbb{T}_j$, is obtained in two ways, which we call "local patch dictionary" and "local projected patch dictionary", obtained as follows.

4.1. Local patch dictionary

The formation of the local patch dictionary is illustrated in figure 3. With every pixel in the output view is associated an epipolar line segment in each input view. If clique j contains pixels s and t, then we create the dictionary from all 2-pixel patches in the input sequence which are within a threshold distance D of any of the epipolar lines associated with pixels s and t. Typical settings for D range from zero—meaning that only patches which intersect an epipolar line are included—to 3 pixels.

4.2. Local projected patch dictionary

The above strategy implicitly assumes that the new view and the input images are of the same scale and orientation. Another way to generate the patch prior is to use corresponding samples on the epipolar lines—the patch $\{s, t\}$ is assumed to be fronto-parallel, and projected into all the input images at all depths. The input images are sampled at each of these sets of points (in fact, the samples are the same as those in $\mathbf{C}_s$ and $\mathbf{C}_t$), to make the patch dictionary, thus

$$\mathbb{T}_j = \{[\mathbf{C}_s(k, z)\ \mathbf{C}_t(k, z)]\ \forall k, z\} \quad (16)$$

At first sight, this texture dictionary might simply appear to encode the prior that 2-pixel cliques are fronto-parallel in the new view. However, at occlusion boundaries, the single-pixel modes (the minima of the unary energy) are not corrupted by sampling from either side of the boundary, and the dictionary will include M examples of the transition across that boundary, permitting the correct reconstruction. In textured areas, the dictionary includes several samples from the texture at a variety of offsets, so that the dictionary performs just as hoped, encouraging the reconstructed view to have the same texture as the input. In textureless areas the dictionary will encourage piecewise smoothness, which is consistent in the absence of any image information to the contrary.

Note also that projection of the patch $[\mathbf{C}_s(k, z)\ \mathbf{C}_t(k, z)]$ into input view k is an explicit imposition of the assumption that the patch is fronto-parallel in $\mathcal{V}$. Traditionally, stereo methods have simply taken a window around the reprojection of the patch's center pixel. When the images in question are rectified, this amounts to the same thing; however, when they are not, the full reprojection is required to enforce the assumption that the patch is fronto-parallel.

4.3. Summary

In terms of the label map, where conversion from label to color is given by equation 13, the prior term in the MRF is then as follows: for clique j, with the neighborhood $\mathtt{N}_j = \{s, t\}$, we have

$$\psi_j[L(\mathtt{N}_j)] = \psi_j[\{l_s, l_t\}] = E_{\text{texture}}(V(s, l_s), V(t, l_t)). \quad (17)$$

Combining the two terms, we have the following energy which must be minimized over the unknown labels $l(i)$:

$$E(\mathcal{L}) = \sum_i \phi_i(l(i)) + \lambda \sum_j \psi_j[L(\mathtt{N}_j)] \quad (18)$$

where λ is a tuning parameter of the algorithm which controls the influence of the prior on the final solution. We reiterate that j indexes cliques, and that E_{texture} for clique j uses a local patch dictionary $\mathbb{T}_j$.

5. Experiments

We tested the new algorithm on three freely available test sequences: "Monkey", "Plant & Toy" and "Edmontosaurus". We do a leave-one-out test, where one view is selected to be synthesized using the 8 input images whose camera centers are closest to that of the novel view. By comparing the rendered view to the original we can obtain ground-truth comparisons.

Color modes are precomputed, requiring about 7 minutes for a 640×480 output image, and a further 7 minutes to compute all possible cliques. In all our experiments we set the threshold τ on the robust kernel ρ to 50, and the prior influence parameter λ to 1, unless otherwise stated. The energy is minimized using a publicly available implementation[2] of the TRW algorithm [10]. Note that the pairwise energies ψ_j are different for every clique meaning that memory requirements are relatively high (approximately 800 bytes per pixel), but fit within 1GB of main memory. Though TRW does not guarantee to find the global minimum, it provides a lower bound on the energy, allowing

[2] TRW-S: http://www.adastral.ucl.ac.uk/∼vladkolm/papers/TRW-S.html

us to estimate how close our solution is to the global minimum. In our experiments we found our solutions ranged from being within 0.01% to 2% of the lower bounds.
In our experiments we compared several algorithms:

"**ML**", *No prior*: equivalent to setting $\lambda = 0$.

"**Depth**", *Piecewise smooth depth prior, 4-connected*: Patch neighborhood system $\mathtt{C4}_j$. Optimization over discrete depths z with unary cost (5), and prior $\psi_j[\{z_s, z_t\}] = \max(\alpha|1/z_s - 1/z_t|, 1)$, where α is scene dependent, $\lambda = 700$. This is optimized using alpha-expansion graph cut [1].

"**5x5**", *Large-clique prior, global dictionary* [6]: Patch neighborhood system $\mathtt{P}_j$. Texture dictionary $\mathbb{T}_j$ equal to the global dictionary $\mathbb{T}$, i.e. the set of all 5×5 patches in the input sequence. This is optimized using an approximation of ICM.

"**5x5Local**", *Large-clique prior, local dictionary* [3]: as 5x5, but the dictionary is obtained as described in §4.1.

"**2Global**", *Small-clique prior, global dictionary, 4-connected*: Patch neighborhood system $\mathtt{C4}_j$. Texture dictionary $\mathbb{T}_j$ equal to a global dictionary $\mathbb{T}$, which in this case comprises all 2-pixel patches in the input sequence. This is optimized using TRW.

"**2GMM**", *Small-clique prior, global GMM, 4-connected*: Patch neighborhood system $\mathtt{C4}_j$. Texture energy is computed using a Gaussian mixture model (GMM) with full covariance matrices in 6D trained, using freely available software[3], on all 2-pixel patches in the input sequence. The number of Gaussians used was typically 8–10. This is optimized using TRW with $\lambda = 10$ and, having learnt the GMM, is orders of magnitude faster than the global dictionary version above.

"**2Local4**", *Small-clique prior, local dictionary, 4-connected*: Patch neighborhood system $\mathtt{C4}_j$. Texture dictionary $\mathbb{T}_j$ restricted to epipolar lines as defined above (§4.2). This is optimized using TRW.

"**2Local8**", *Small-clique prior, local dictionary, 8-connected*: as 2Local4, but with neighborhood system $\mathtt{C8}_j$.

For the small cliques there are two patch orientations in $\mathtt{C4}_j$ and four in $\mathtt{C8}_j$. Separate libraries are constructed for each patch orientation and used as appropriate.

The two success metrics in which we are interested are speed and accuracy. Figure 4 provides a summary of the speed results on the "Edmontosaurus" and "Plant & Toy" sequences. The fastest techniques are "ML" and "Depth", followed by "2GMM" and "2Local8", which are about five times slower. An order of magnitude slower again is "5x5", with the global patch dictionary, "2Global", a further two orders of magnitude slower.

More important, however, is the quality of the results, also summarized in figure 4. Three evaluation metrics are useful: RMS pixel difference from ground truth, number of pixels with gross errors, and a visual assessment. The quantitative measures tend to favor the "Depth", "2Global", "2Local4" and "2Local8" techniques, though percentage pixel errors alone ranks "2Local8" less highly, due to the larger area of blue feather replaced with background texture. A qualitative inspection of the images reveals that the "2Global" method fails to recover from the gross region error in the "Edmontosaurus" new view, leaving only "Depth" as a suitable competitor to the "2Local*" methods.

Figure 5 demonstrates the relative benefits of "2Local8" over "Depth". The surface-smoothness regularizer does not perform well around fine features such as fur, or at depth discontinuities where the background is faintly textured, while the texture regularizer behaves correctly in these circumstances.

6. Discussion

We have investigated the use of small cliques in an image-based prior for the new view synthesis problem. Finding an effective small-clique formulation is of interest, because it allows the use of global optimizers. The advantages are twofold. First, there is a speed advantage to solving this class of problem—once all possible clique potentials have been precomputed, the optimizer can find a strong minimum relatively quickly. More importantly, however, global optimizers are able to correct large regions of error in a maximum-likelihood image.

We have shown that regularization is improved by restricting the training data for the prior to local regions within the sequence. This improves the results and confers a further, considerable speed advantage. Bringing image-based priors for the first time into competition with pairwise depth priors, it is instructive to compare the two on this problem. We show that the piecewise-smooth-world assumption inherent in pairwise depth priors generates artifacts when presented with complex natural structures. Because the depth priors must be image-independent (there's no way to learn them from the input sequence), they must of necessity be generic smoothing priors, even if trained on databases of natural scenes, just as generic image priors must. Therefore the second advantage of image-based priors is reinforced: they can be conditioned on the input data, so that regularization is tuned to the task at hand.

It would be interesting to see if depth priors can be made conditional on the sequence characteristics (rather than, say, conditional on the image gradient), but this is beyond the scope of this work.

The local dictionary may also be seen as locally learning

[3] Netlab: http://www.ncrg.aston.ac.uk/netlab/

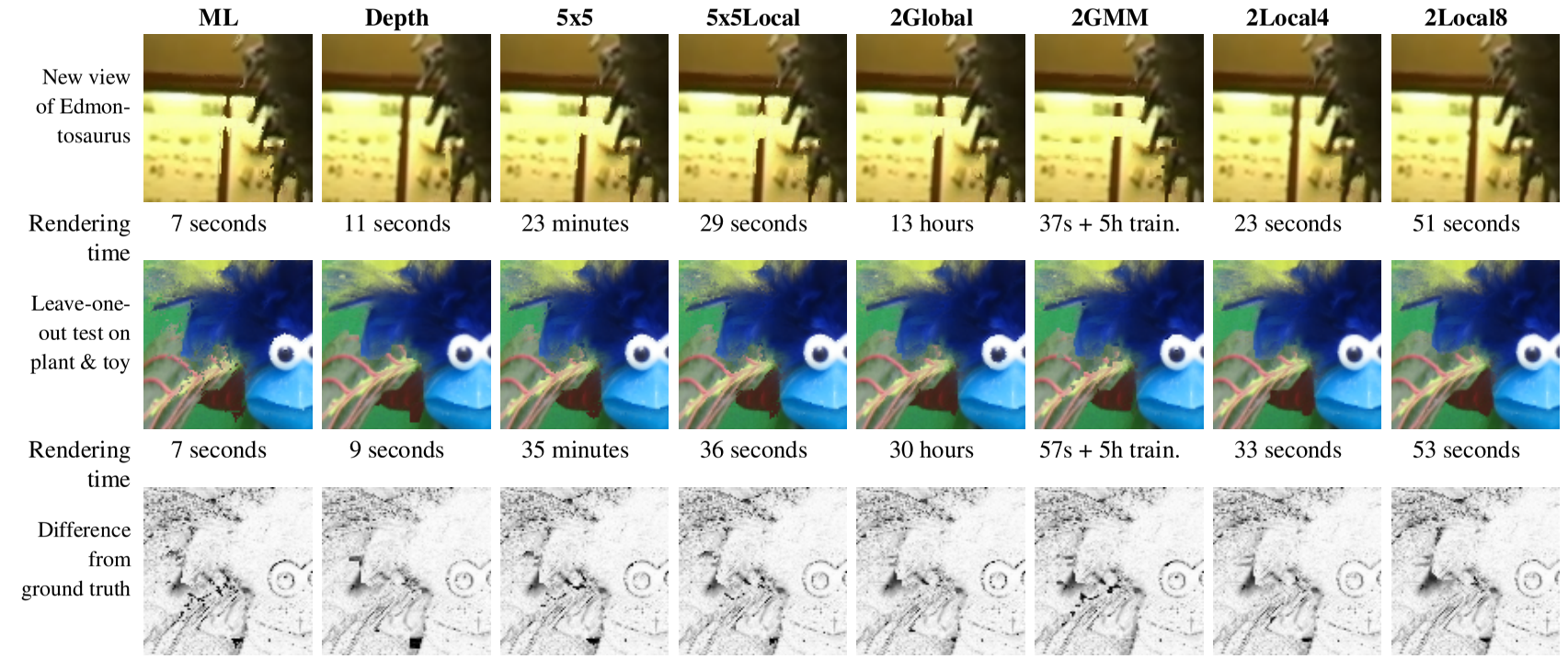


Figure 4. **Comparison of regularizers.** A table of images, rendering times and error metrics for zooms of a new view for the "Edmontosaurus" sequence, and a leave-one-out test on the "Plant & Toy" sequence, complete with difference from ground truth. The images were rendered using each of the methods described in §5. Pixel err. is the percentage of pixels which are grossly different from the ground truth color, G, determined by $\|V - G\|^2 > 1000$.

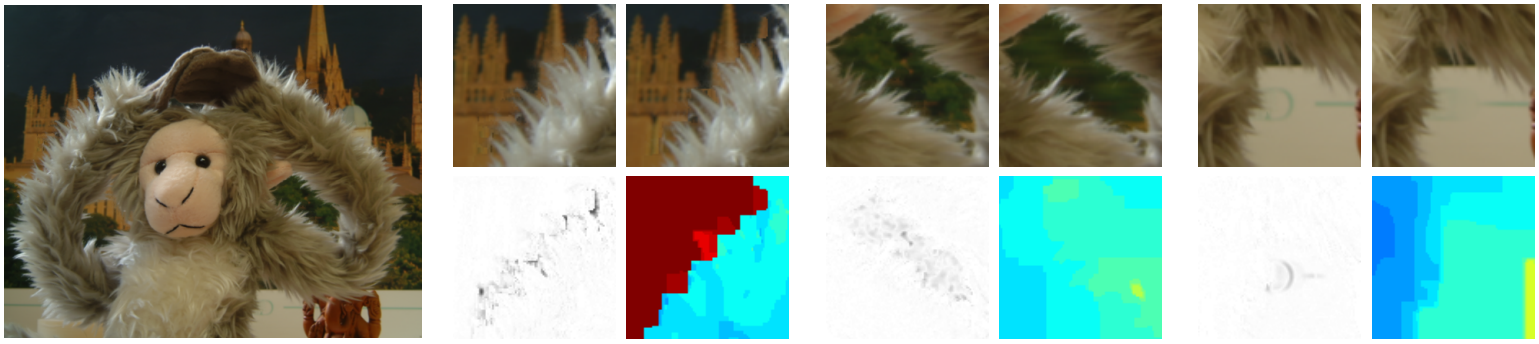


Figure 5. **Regularization of color vs. depth.** (a) A new view for the "Monkey" sequence, rendered using "2Local8". (b), (c) & (d) Each 2 by 2 grid of images shows, clockwise from top right, a zoom from (a), the same zoom from the equivalent image rendered using "Depth", the disparity image output by "Depth" for that zoom, and finally, the difference between the two rendered zooms. (b) highlights the box-like artifacts of regularization over depth around fine, foreground features. (c) and (d) show the smoothing over of holes and discontinuities in depth respectively generated by regularizing over depth.

MRF clique potentials. To do so at every pixel using conventional methods would not be computationally feasible, nor would there be enough training data. Effectively we replace this learning with a nearest-neighbor lookup, and it appears that this implicitly avoids the over-counting that pairwise learning normally entails [17].

Of course, it would be useful to incorporate both depth and texture priors, but a difficulty with our current approach is memory use. We must have a separate $L_s \times L_t$ edge cost matrix for each clique $\{s, t\}$, where L_i is the number of labels at node i, but in order for these matrices to fit in memory each edge can have of the order of only 100 label combinations. This provides the potential for a few color modes at each pixel, but not enough for modes spaced densely over depth (required for regularization of depth). This means, due to the computational expense required by a global optimizer, we can only regularize over either depth or color, but currently not both.

It is important to recognize that restricting the possible colors in this way can lead to the correct color for some pixels being omitted, through either the color being visible in only one input view, the color not being a mode, or the method of §3.2 failing to find the mode. However, these situations are generally rare. A further shortcoming of our approach is that it can lead to temporal flickering in synthesized video sequences, as pixels can switch between similarly plausible, but very different, color modes between frames with impunity. This suggests the need for a prior on the temporal, as well as spatial, texture domain.

6.1. Conclusion

In this work we have demonstrated that powerful inference techniques are necessary to fix large areas of error in the maximum-likelihood solution to the NVS problem. The fact that these techniques can, in practise, only be applied to 2-pixel patches poses a problem, as such small cliques tend to lack the discriminative power required for regularizing texture. We have shown how to construct a clique-specific patch dictionary that can be used to overcome this problem, providing cutting-edge image quality at a practical rendering speed. In addition, we have introduced a fast algorithm for enumerating color modes of E_{photo}, even using the non-convex, truncated quadratic kernel.

Acknowledgements We thank Carsten Rother, Victor Lempitsky and Ali Shahrokni for their generous help. Research supported by EPSRC and Sharp.

References

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE PAMI*, 23(11):1222–1239, 2001.

[2] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE PAMI*, 24(5):603–619, May 2002.

[3] A. Criminisi and A. Blake. The SPS algorithm: patching figural continuity and transparency by split-patch search. In *Proc. CVPR*, volume 1, pages 342–349, Jun 2004.

[4] A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In *Proc. ACM SIGGRAPH*, pages 341–346, Aug 2001.

[5] A. Efros and T. Leung. Texture synthesis by non-parametric sampling. In *Proc. ICCV*, pages 1039–1046, Sep 1999.

[6] A. Fitzgibbon, Y. Wexler, and A. Zisserman. Image-based rendering using image-based priors. In *Proc. ICCV*, volume 2, pages 1176–1183, Oct 2003.

[7] W. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. In *IJCV*, Jul 2000.

[8] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[9] J. Huang and D. Mumford. Statistics of natural images and models. In *Proc. CVPR*, pages 1541–1547, 1999.

[10] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE PAMI*, 28(10):1568–1583, Oct 2006.

[11] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE PAMI*, 26(2):147–159, 2004.

[12] V. Kwatra, I. Essan, A. Bobick, and N. Kwatra. Texture optimization for example-based synthesis. In *Proc. ACM SIGGRAPH*, volume 24, pages 795–802, Jul 2005.

[13] L. Liang, C. Liu, Y.-Q. Xu, B. Guo, and H.-Y. Shum. Real-time texture synthesis by patch-based sampling. *ACM Trans. Graph.*, 20(3):127–150, 2001.

[14] S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *Proc. CVPR*, pages 860–867, 2005.

[15] A. Srivastava, A. Lee, E. Simoncelli, and S. Zhu. On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision*, 18(1):17–33, Jan 2003.

[16] C. Strecha, R. L. Fransens, and L. Van Gool. Combined depth and outlier estimation in multi-view stereo. In *Proc. CVPR*, pages 2394–2401, 2006.

[17] C. Sutton and A. McCallum. Piecewise training of undirected models. In *Uncertainty in Artificial Intelligence*, page 568ff, 2005.

[18] O. Woodford and A. W. Fitzgibbon. Fast image-based rendering using hierarchical image-based priors. In *Proc. BMVC.*, volume 1, pages 260–269, 2005.

[19] O. Woodford, I. D. Reid, P. H. S. Torr, and A. W. Fitzgibbon. Fields of experts for image-based rendering. In *Proc. BMVC.*, volume 3, pages 1109–1108, 2006.