TOWARDS STEREO FROM SCALE TREES

K L Moravec, J Ruiz Hidalgo, R Harvey and J A Bangham

University of East Anglia, UK.

The image trees described in this paper hierarchically organize image segments according to scale, with the coarsest scale, the scale of the image itself, as the root of the tree and the finest scales as the leaves. The segmentation algorithm used to form the nodes of the tree is the *sieve*, a nonlinear morphological scale-space operator.

We use a conventional matching criterion, the sums of squares of differences, SSD, which is statistically meaningful when ergodicity of its matching regions is assumed. Here, simple statistical similarity measures are applied to the scale tree to simplify it into a set of relatively homogeneous regions. This simplified scale tree is then used to generate matching regions which frequently satisfy the ergodicity condition. This approach reduces the errors in the resulting disparity map particularly within sharp-edged regions with low texture – where conventional methods fail.

INTRODUCTION

Tree data structures are widely used in computer science, and they facilitate common operations such as searching and ordering of data. They have also been applied to computer vision as a way to order extracted features from an image, as in Kliot and Rivlin (1) and Fu and King (2) . Information about enclosure, scale and intensity can be encoded in the structure of the tree. Furthermore, image trees form a key part of the proposed MPEG-4 standard in which images are composed of audio-visual objects.

The scale tree is an image tree, organized hierarchically by scale, which uses a non-linear morphological operator, the sieve, to generate the nodes of the tree. The sieve operates by recursively removing local maxima and minima of a certain scale in an image. The many aspects of the sieve have been described in detail in other papers, including a 1-dimensional implementation, Bangham et al (3) , Bangham et al (4) , an n-dimensional implementation (here we use the area-sieve, which is a 2-dimensional implementation) Bangham et al (5) , and descriptions of its robustness, Harvey et al (6) , Harvey et al (7) .

The maxima and minima removed by the sieve form connected level-sets called *granules*, which then become nodes in a tree. As scale increases, the size of the granules increase, and the granules of smaller scale enclosed by the granule of large scale become the children of the larger granule. The sieve is a good choice for tree segmentation because it does not introduce artifacts into the image, the original image can be recovered by adding up all the nodes of the tree, and the resulting tree is relatively invariant to viewpoint changes. The scale tree bears a close relationship to the objects in an image, and has been used to filter, segment and detect motion in an image.

In many cases real objects are represented by extremal regions so nodes represent image objects but sometimes this is not the case so in this paper we augment the scale-tree by explicitly representing the regions implicitly associated with non-leaf nodes. The new segments are the *non-extremal* level-sets and are illustrated in Figure 1 . On the left is an image and its associated scale tree. On the right is the augmented tree with two additional nodes, F and G, that represent the image background (a rectangle with a hole in the centre) and the face region not covered by the eyes and mouth.

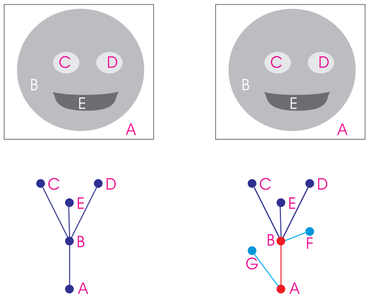

Figure 1: Lower left panel shows a simple scale tree with $A \subset B \subset \{C, D, E\}$. On the right, the level-set tree with additional nodes $G = A \bigcap \bar{B}, F = B \bigcap \overline{E \bigcup C \bigcup D}$.

SIMPLIFYING THE TREE

Because the sieve decomposes images by connected grey levels, flat zones within the image, it is very good at find-

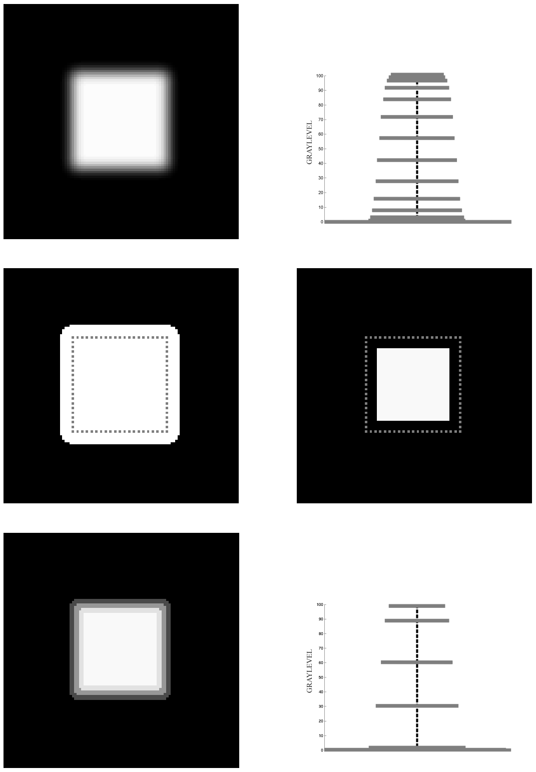

Figure 2: A simple blurred square and its resulting scale tree (top left and right). Granules collapsed too low, and granules collapsed too high (middle left and right), where the original border of the square is shown as a dotted line. The square after collapsing similar nodes and its resulting scale tree (bottom left and right)

ing and preserving sharp edges over many scales. This makes it complementary to linear decompositions such as Gaussian and wavelet pyramids, where large scale objects have blurred edges. The sieve is less well matched to blurred images, as Figure 2 shows. A blurred object has a tree of many nodes, each having a single parent and child, and often differing from its immediate relatives by only a few pixels. For easy manipulation, these nodes may be collapsed into one node, but this must be done carefully. In Figure 2 , blurring has converted a simple two level image into the extended tree shown on the top right. Below are masks formed from the smallest node in the tree (right), and the largest non-root node (left). Neither are entirely accurate representations of the original, unblurred square (outlined with the dotted line) but both are much simpler than the original. What is needed is a tunable method to selectively merge similar parent/child nodes in a principled way.

For each node in the tree, we compare the homogeneity of the statistics of the node under consideration with those of its children. Specifically it is assumed here that either all regions are drawn from the same unimodal Gaussian distribution or they are are drawn from separate distributions.

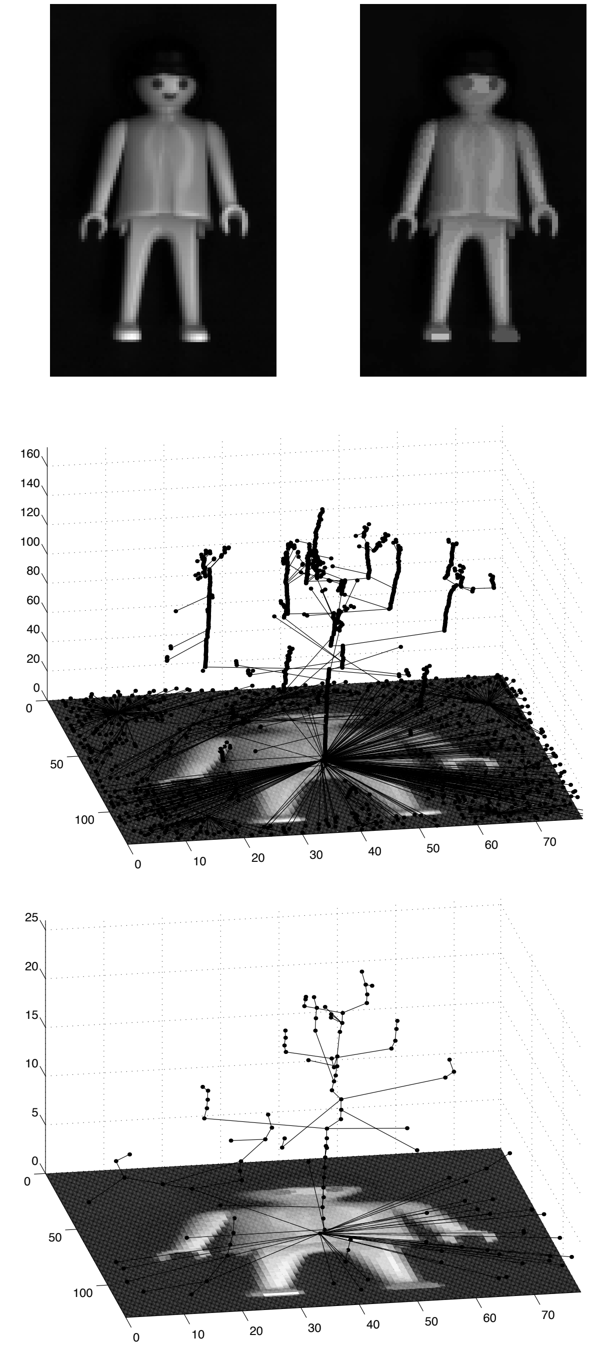

Figure 3: A test picture of a doll, before and after tree simplification (top). The original tree, containing 1510 nodes (middle) and the simplified tree, containing 121 nodes (bottom).

Under these assumptions it is fairly easy to derive a restricted likelihood test (as in Basman et al (8) and Silvey (9)) in which one hypothesis, homogeneity, is a special case of the other. The log of the likelihood, λ of regions 1 and 2, is well known to be:

$$\log \lambda^2 = N_{12} \log \sigma_{12}^2 - N_1 \log \sigma_1^2 - N_2 \log \sigma_2^2 \qquad (1)$$

where (N_1, σ_1^2), (N_2, σ_2^2) and (N_{12}, σ_{12}^2) are the areas and variances of region 1, region 2 and the combined regions respectively.

Of course for a grey level segmentation, it is incorrect to model pixels from level-sets as Gaussian variants– the very fact that they are level sets implies a variance of zero or, more realistically, $q^2/12$, where q is the quantization step. However, for larger scales where the regions may contain many children, we find the Gaussian approximation satisfactory. Furthermore, if the regions are to be merged using other features such as color, the Gaussian approximation is more justified. The merge parameter is not the likelihood but the confidence of the likelihood which may be computed as:

$$c = 1 - \lambda \tag{2}$$

where c is in $(0, 1)$.

Figure 3 shows an example image and its associated tree before and after merging all zones that have a confidence $c < 1 - N \log(\sigma_N)$ where σ_N is the standard deviation of all N pixels.

USING THE TREE FOR STEREO MATCHING

In stereo vision, the features in two or more images are matched. By computing the distortion of a feature between two images, if certain camera parameters are known, the scene geometry can be recovered. Dense matching techniques attempt to do this for every visible point in the scene, but suffer from errors due to occlusion, false matches, noise, low texture and repeating texture. Others have shown that by varying the shape of matching regions, the matching is improved, Moravec et al (10) , Harvey et al (11) , Kanade and Otukumi (12) and Fusiello et al (13) . The technique reported here is to use the scale tree to generate the regions and hence use the tree structure to assist in matching.

Similarity measures such as SSD (Sum of Squared Differences), SAD (Sum of Absolute Differences), MAD (Mean of Absolute Differences), Barnard and Fishler (14) , cross correlation and min correlation, Maragos (15) are well known ways to densely match signals and images. In the case of stereo images the conventional technique is as follows:

1. Two images of a scene are obtained and calibrated such that the epipolar lines are known.

2. Regions in the first image are matched with a number of candidate regions lying along the epipolar line in the second image. The regions to be matched (which must have the same area and contour) may be denoted as $f_1(v)$ and $f_2(v)$ where v is a particular pixel Here $f_1(v)$ is the intensity of the vth pixel in the first image and $f_2(v)$ the intensity of the vth pixel in the right-hand image. The similarity of two pixels may be measured by, Haralick and Shapiro (16) :

$$e(p, q) = \frac{\mathrm{var}\left[X_p - X_q\right]}{\sqrt{\mathrm{var}\left[X_p\right]\mathrm{var}\left[X_q\right]}} \tag{3}$$

   where $X_{p,q}$ are random variables sampled from $f_1(p)$ and $f_2(q)$ where p and q are pixel labels. In practice we usually have only one sample of $f_1(p)$ and $f_2(q)$ so and sample means and variances are computed over windows, W_1 and W_2 which are fixed regions centred around p and q. Further, if the position vector of each pixel is $\boldsymbol{x}(p), p \in V$ then, provided W_1 and W_2 have identical shape, it is possible to have a set of p and q such that

$$\boldsymbol{x}(p) + \boldsymbol{d} = \boldsymbol{x}(q), p \in W_1, q \in W_2 \tag{4}$$

   In which case the variance, (3), may be computed as

$$e(\boldsymbol{d}) = \frac{\sum_{i \in W_1}\left(\tilde{f}_1(i) - \tilde{f}_2(j)\right)^2\Big|_{\boldsymbol{x}(i)+\boldsymbol{d}=\boldsymbol{x}(j)}}{\left[\left(\sum_{i \in W_1}\tilde{f}_1^2(i)\right)\left(\sum_{i \in W_2}\tilde{f}_2^2(i)\right)\right]^{1/2}} \tag{5}$$

   where $\tilde{f}_{1,2}(i)$ are the intensities in regions 1 and 2 after the sample mean intensity computed in that region has been removed and N is the number of pixels in W_1 and W_2. Of course a substantial limitation of this technique is that W_1 and W_2 are fixed and usually do not correspond to statistically homogeneous regions in either image.

3. The offset $\boldsymbol{d}_{\mathrm{min}} = min(e(\boldsymbol{d}))$ is the best match for that region and is called the disparity. The disparity is assigned to all or part of the region in the corresponding disparity image.

For (5) to have a valid statistical interpretation the assumption of ergodicity must hold: the windows must not span image regions drawn from different distributions. In practice this will not hold and at the boundaries of regions there is a mixing of distribtions which manifests itself as a disparity image with ill-defined edges. This is minimised by using small windows but there is a cost: small windows do not allow much averaging. The new algorithm performs a dense segmentation of the image so that windows are as large as possible but do not span statistically inhomogeneous regions.

The scale tree disparity estimation algorithm examines nodes in *pre-order*, starting with the root node. For each node, the disparity estimate is computed by translating the region represented by that node along the epipolar line, calculating the position and error of the best match. This disparity is then assigned to the node. If the error of this node is lower than that of its parent then the disparity of this node is accepted in the support region for this node.

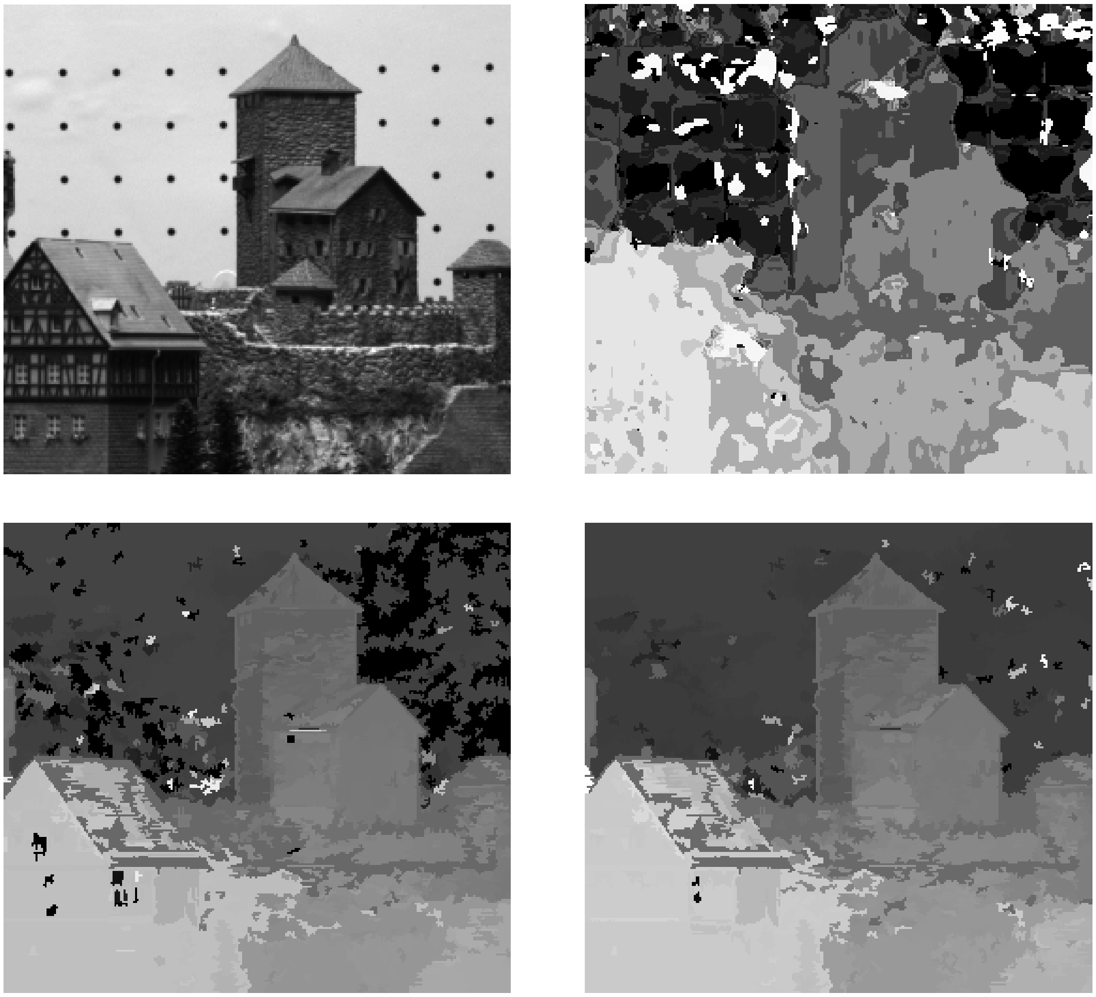

Figure 4: Model castle stereo picture from CMU test set [17] (top left). Conventional SSD disparity estimate (top right), tree-based estimate (bottom left), and simplified tree-based estimate (bottom right)

If the scale tree used is pruned by the likelihood test, the ergodicity assumption has already been tested for these nodes, their parents and children. The pruned scale tree should then have fewer errors than either a fixed window method an unsimplified tree method. The pruned scale tree also has the advantage of faster computation, as the nodes are, at least for the images so far tested, significantly less than the original scale tree.

A summary of the algorithm is as follows:

1. Decompose the image into a scale tree using the complement tree representation as illustrated in Figure 1.

2. Traverse the tree is postorder applying the confidence measure to each graph edege connecting a node and its parent. We test the image region supported by the node and the image region supported by the node's parent. If the confidence measure falls below some threshold the edge is removed by merging node and parent.

3. Progress preorder through the tree and for each node:

    (a) Generate a window from that node.

    (b) Find the best disparity and variance for that node using SSD.

    (c) If the variance of that node is less than that of its parent, reassign the disparity of that region in the disparity map to the new disparity.

A real calibrated image, Maimone and Schafer (17) , and its resulting disparity maps shown in Figure 4 . The map resulting from using the simplified scale tree (bottom right) has fewer errors, particularly in the background, where the repeating texture of the dots tends to confuse SSD algorithms. There is only sparse ground truth disparity for these images but at these points the new method's error is less than the ground truth error. The new method produces sharp–edged disparity regions and works well in regions of low texture.

DISCUSSION

A feature of this method is that although the method for producing possible windows is novel, the matching is conventional. Furthermore in our implementation the tree is used only in the first image and the matching is performed from image 1 to 2. Reversing the match by computing a tree from image 2 and matching from image 2 to 1 is a well known and obvious extension. A more interesting refinement would be to account for the projective effects between the images. The tree may be well suited to cases where projective distortions are significant. By extracting the projectively invariant features of tree nodes it should be possible to compute the three-dimensional structure of images by matching together the trees generated by such images. We are currently implementing such a method.

1 References

1. M. Kliot and E. Rivlin, 1998, “Invariant-based Shape Retrieval in Pictorial Databases”, Proc. European Conf. Computer Vision, 491–507.

2. K.S. Fu and Sun King, 1974, “Syntactic methods in pattern recognition”, Academic Press.

3. J.A. Bangham, R. Harvey and P.D. Ling, 1996, ”Morphological scale-space preserving transforms in many dimensions”, J. Electronic Imaging, 5(3),283–299

4. J.A. Bangham, P.D. Ling and R. Harvey, 1996, ”Nonlinear scale-space causality preserving filters”, IEEE Trans. Patt. Anal. Mach. Intell.,18,520–528

5. J.A. Bangham, R.W. Harvey, P.D. Ling and R.V. Aldridge, 1996, “Nonlinear scale-space in many dimensions”, European Conference on Computer Vision, 1, Kluwer Academic Press, 186–192

6. R.W. Harvey, A. Bosson and J.A. Bangham, 1997, “The robustness of some scale-spaces”, British Machine Vision Conference, 11 – 20,

7. R. Harvey, A. Bosson and J.A. Bangham, 1996, “A comparison of linear and nonlinear scale-space filters in noise”, Signal Processing VIII: Theories and Applications, III, 1777–1780

8. A. Basman, J. Lasenby and R. Cipolla, 1997, “The creep and merge segmentation system”, Technical Report No. CUEDDIF–INFENG/TR295, Cambridge University

9. S.D.Silvey, 1975, “Statistical Inference”, Chapman and Hall.

10. K.Moravec and R.Harvey and J.A.Bangham, 1998, “Connected-set filters to improve and obtain stereo disparity maps”, British Machine Vision Conference, 822–831.

11. R. Harvey, K. Moravec and J.A. Bangham, 1998, “Stereo vision via connected-set operators”, Proc. European Signal Processing Conference, 2, 613–616.

12. T. Kanade and M. Okutomi, 1994, “A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment” , IEEE Trans Patt. Anal. Mach. Intell., 1994, 16, 9, 920-932.

13. A.Fusiello and V.Roberto and E.Trucco, 1997, “Efficient stereo with multiple windowing, Computer Vision and Pattern Recognition, 858–863.

14. S. T. Barnard and M. A. Fishler,1982, ”Computational Stereo”, Computing Surveys, 14 (4), 553-572.

15. P. Maragos, “Pattern Spectrum and Multi scale Shape Representation”,1989, IEEE Trans Patt. Anal. Mach. Intell., 11 701–716.

16. Haralick, R.M and Shapiro, L.G., “Computer and Robot Vision”, 1992, Addison-Wesley.

17. M.Maimone and S.Shafer, “The CMU Calibrated Imaging Stereo Datasets”, http://www.cs.cmu.edu/People/cil/cil.html