A MULTISCALE METHOD FOR AUTOMATED INPAINTING

R.J.CANT, C.S.LANGENSIEPEN School of Computing and Mathematics, Nottingham Trent University

Abstract: We present a novel, simple and general method for image inpainting. Current methods may be crudely divided into those that aim to continue edges by various energy minimisation techniques, and those that perform texture synthesis from local information, but both have their weaknesses. Our method searches the image for areas of similarity and uses these to inpaint. By analysing the image at multiple resolution scales we can find similar features and textures from anywhere in the image at a reasonable speed. We present results using some challenging images where both features (edges) and textures from non-local information are used to achieve plausible restoration. Keywords: Inpainting, image restoration

1 INTRODUCTION

Image manipulation has a long and not very illustrious history. Stalin not only had people executed, but had their images removed from photographs as if they had never existed [1]. These days, there are more reasonable reasons why images may be edited before publication, from the 'improvement' of someone's appearance to the clarification of a scene for journalistic effect. Usually this involves a substantial amount of work by the user, to make the changes blend in with the remainder of the image. This paper proposes a method of inpainting an image after the removal of a feature, which requires the minimum of human intervention. Its aim is to be able to reconstruct pertinent features and textures to generate a plausible image without explicitly searching for artefacts such as edges.

2 RELATED WORK

There has been considerable work in the field of digital inpainting, approaching the problem from the directions of noise removal (due to compression and transmission), film and video artefact removal, and texture synthesis. [11] provides a comprehensive summary of work done in this area. Early work such as [5] concentrated on image reconstruction after effects caused by the scanning device characteristics. Work of particular relevance includes the texture synthesis method [15] based on Heeger and Bergen's [13] pyramid texture analysis. This produced convincing inpainting of large areas of small-scale texture e.g. grass, but no features or edges were included. [14] removed image noise while retaining line continuity, but at the cost of requiring manual choice of regions for spatial and frequency samples. De Bonet [8] used multiple resolution methods to reconstruct texture, showing that such a method could synthesise more difficult textures where overall feature directionality had to be maintained. More recently, [20] provided a multi-resolution method of impressive speed for infilling areas of texture. However, this method was purely on a single texture, and is not appropriate to more general inpainting, since image segmentation would also be needed. Methods which concentrated on edge, line, curvature continuity included work on the TV model [6] and the PDE approach of [3]. The latter produced impressive results, though the authors concede that, because it concentrates on achieving isophote continuity, it would have trouble with areas of texture. Their later work [2][4] includes what appears to be even more impressive inpainting of the standard Lena image. However, their method includes the use of field information directly from the undamaged image to direct the inpainting of the damaged area, so can really only be considered appropriate where there is such an image available e.g. film frames.

3 THE METHOD

Our method is based on work first done as an undergraduate student project [16] on black and white images. This method used pixels from other similar areas of the image to inpaint the area. If a region was to be inpainted, the process was started with the outermost pixels – those adjacent to parts of the image that were to be retained. The pattern of levels of pixels in the patch surrounding it was then compared with the remainder of the image to find the closest equivalent pattern match. 'Closest' was simply defined by the sum of the grey values for the patch. The pixel in the equivalent position within this patch was then used to replace the pixel to be reconstructed.

Although this method appeared quite successful, a few problems were evident. It was quite good at generating 'plausible' reconstructions of amorphous areas such as foliage, and of simple repeating structures such as identical windows in a distant building. However it could not cope with more distinct edges as in individual leaves, or repeating structures with ranges of scales within them. It was also extremely slow, and experiments had to be limited to small images. An apparently identical technique was later used by [9], who also commented on its slowness.

We modified the technique by handling features in a way that seems to match human perception. If one reduces the resolution of an image by a factor of 2, 4 etc., one can still see the grossest features. One observes this effect when watching the 'pixellated' faces used to hide identities on screen – the facial shape, nose etc can still be seen if one 'squints' at it. By using a reduced resolution version of the image, the areas to be inpainted could be matched up with similar areas elsewhere on the image – where similarity is assessed at this reduced resolution. In other words, the method is finding features visible at this resolution which are relevant to the reconstructed area. Each resolution scale acts as a filter which selects out features at that scale for comparison. This gives a comparatively quick method of finding the best matches for the pixels to be reconstructed. Firstly, the picture resolution is reduced by a factor of 2**N. As in the original process, the whole image (at this resolution) is scanned to find the best matches for the patch surround the pixel to be inpainted. The set of best matches (currently defined purely as a fixed number of matches) is then used as the starting set (rather than the whole image) for a similar matching process at a higher resolution. This process will match on features at a smaller image scale. Regions around each of the best matches are then examined at the higher resolution to again find the best match with the original. This limits the search space, yet allows for some variation in best match from one scale to another. The process is repeated until the final target image resolution, whereupon the best match is used to generate the reconstructed pixel (Figure 1). At each resolution level, each pixel in the candidate patch is compared with its equivalent in the original patch (in RGB), and the differences summed to give the overall measure for that candidate patch. Each candidate patch is also reflected about x and y axes and rotated by a range of angles for comparison, since a feature may be rotated/reflected elsewhere in the image relative to the one being reconstructed.

This method has a number of advantages over the original method discussed and other work. Firstly: speed. The process of exhaustive comparison of each pixel in the image against the one to be filled (particularly as one has to look at a patch of pixels

Proceedings 17th European Simulation Multiconference
(c) SCS Europe BVBA, 2003 ISBN 3-936150-25-7

surrounding it) can be prohibitively slow. By performing the exhaustive search only at a much lower resolution, the overall process becomes much faster., as the search space used for the higher resolutions are then heavily culled. However, even with this improvement, images of 640*480 and 24 bit colour (which we used for convenience) could take from a few minutes to a couple of hours on a reasonable PC, depending on the size of area to be filled. The method of [3] is apparently faster, because apart from the initial smoothing, it examines pixels local to the inpainting area only but does not handle texture.

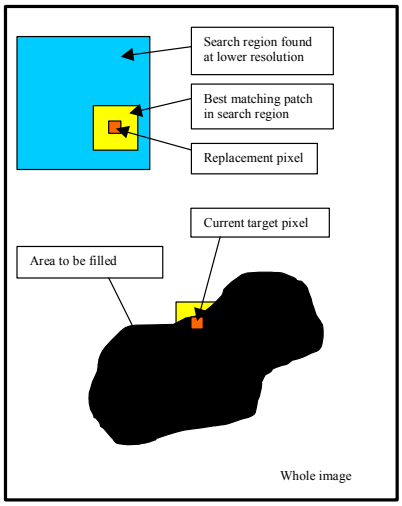

Figure 1 One stage in multiscale process

Secondly: feature identification. As discussed by DeBonet, human vision is very good at seeing features at a range of scales. That is why Gothic architecture is so appealing; it has repetition and variation at many different scales within a single building. By using a process that attempts to mimic the way we 'pan and zoom', the process has a better opportunity of finding the features that would affect the plausibility of the reconstruction. Thus this method, unlike [3] can also reconstruct texture, which is visible only at the higher resolutions.

Thirdly: generality. No explicit edges or lines are detected; the pixel colours in the patch are simply compared with those in the candidate patches. This means that one does not have to start looking for 'special cases' such as curved or straight edges. No explicit weighting is done to favour pixels closest to the one to be reconstructed, as otherwise some choice would have to be made as to how to weight. The effect of weighting is achieved by the use of lower resolutions, since each resolution scale essentially averages a different number of pixels.

The process does have some free parameters. As will be discussed later, choice of the size of patch to use and number of resolution scales can affect the quality of the reconstruction. Choice of region size mainly affects the process speed, since it is used to limit the search space. However a larger search space around the notional best match can help for local texture. A less important constraint, made solely for performance reasons, is that we limited the choice of rotation angles used for the comparison. Since the built landscape contains features that may vary in angle due to perspective, the matching process should rotate the comparison patches by small angles from the notional horizontal and vertical to find the best match. However, in the natural landscape, features may occur at any angle, so the rotation should be through a wider range of angles – we chose to use every 90 degrees (ie 8 rotations and tests for every patch including reflections). With more CPU speed, both sets of rotations could be tried for every image, and the comparison process itself would eliminate the worst matches.

4 RESULTS

4.1 The Best

In Figure 2 of a building in Prague, there are a challenging range of features of differing scales and textures. Figure 2 shows the original lamppost in the scene, Figure 3 shows the image forming the starting point for the reconstruction, after removal of the lamppost, and Figure 4 shows the resultant reconstruction.

Figure 2 Original Image

Note that the process has managed to reconstruct the vertical glazing bars automatically, as well as the triangular shapes between the windows and the smaller scale pavement texture. We believe this capability in feature and texture is unique to our multiresolution method. The vertical glazing bar in the bottom window would not have been inpainted by the edge continuity techniques, because it was completely removed by the mask. It would not have be found by the texture method [16] without massive enlargement of the patch size, which would have ensured the pavement texture would have degenerated into garbage as discussed in [9]. [20] found that there tended to be discontinuities across the inpaint boundary in their multiresolution method unless they modified the process. However, they were only using multiple scales to enlarge the local region. In contrast, we are using the lower resolutions to sample and include non-local regions of the image, and find that continuity across the boundary is reasonable at the higher resolutions. Ashikhmin[1] who modified the Wei-Levoy method to achieve some elegant synthesis of natural textures, commented that multiresolution did not appear to assist much in his work on single textures – we have found it essential for images with a range of features and textures.

For this image, the most plausible reconstruction was obtained using a patch size of 9 pixels, search region of 4 pixels around the 4 best candidate patches, and 2 lower resolution scales before the final processing. Patches were compared at their original angles and at slight deviations from the horizontal and vertical as mentioned earlier. The latter process seems to have resulted in the best matches for the glazing bars that are not quite vertical – further work needs to be done in exploring whether the best matching pixel should be modified if the angle is not as the original. In the lower window, the left hand horizontal glazing bar is not reconstructed, though the right hand one is. This is

plausible because one of the other windows (bottom left) also exhibits only one horizontal bar, since the other half window is open. The process has used this area in its reconstruction of the rightmost window.

Figure 3 Area masked for infill

A problem involving natural rather than built artefacts is this dolphin image. Figure 5 shows the original image, Figure 6 the masked out dolphin, and Figure 7 the best infilling attempt. For this image, the most plausible infilling occurred with a patch size of 3 pixels, a search region of 8 pixels, and 2 lower resolution scales. In this case, patches were compared at a wide range of rotation angles rather than just slight variations off vertical. The wave shape and local texture appear to be plausible.

Most of the images upon which we have attempted the technique have provided the most plausible reconstructions when using 2 lower resolution scales. However, this is partially related to the image size. Higher resolution versions of the same images need one or more additional low resolution stages to attain the equivalent level of image inpainting.

4.2 The Worst

Figure 8 shows the problems associated with trying to inpaint a relatively large area within an image. The process begins by trying to find the best matches to those pixels on the perimeter of the reconstruction area that have the most neighbours in the original image. It then fills those with the next highest count of 'real' neighbours until the whole perimeter has been covered, before moving inwards. This means that as one gets closer to the centre of a large area for reconstruction, more and more of the pixels used for comparison have actually been reconstructed themselves. As commented by Wei and Levoy [20], this spiralling inwards is necessary to avoid directional bias caused by the obvious artificiality of following scan lines. Furthermore, because of the freedom of matching, adjacent pixels in the area to be filled may not be sourced from contiguous parts of the original image. This tends to cause a 'smoothing' of texture in the centre of a large fill area, and this phenomenon can be seen where the Mountie has been replaced in Figure 10. Potentially, 'noisiness' could be added to the interior of the region, perhaps by the method of [13] or [15] but would need automated analysis to identify the degree to which the texture encloses the region for inpaint. A better solution might be to allow the user to indicate that noisiness was required.

Note, however, that the process has construct a plausible beach texture, and continued the edges at the lakeside and distant shoreline successfully. The 'Connectivity Principle' presented by [7] and discussed by [11] with regard to the Mumford-Shah model's failings is preserved with this technique, as well as the restoration of texture provided the area is not too large

Figure 13 shows the problems associated with the reconstruction of areas in distinct predictable lines. The human eye/brain combination is very good at pattern finding, and so the narrow sections where the wires have been removed can still be seen in the middle left of the image. As commented by Ashikhmin, the human eye will filter out discontinuities in an image with high frequency content – this is the manner in which the simple version of chaos mosaic texture synthesis [12] achieves plausibility. The bear image is not 'busy' and so artefacts would be comparatively visible. There are real vertical features in the right hand area of water, so to an observer who had not seen the original, the image would still be plausible. Interestingly, the horizontal 'replaced wires' are far harder to see, despite both inpainting areas being of similar dimensions. However the bear fur texture, body edge, stones and water are reasonably well reconstructed.

Figure 4 Lamppost filled in

Figure 5 Original Image

Figure 6 Dolphin masked out

Figure 7 Image after processing

.

Figure 8 Original image of Mountie

Figure 9 Masked image

Figure 10 Best inpainting

Figure 11 Original image of bear

Figure 12 Masked image

4.3 Analysis of results

Our experiments with a range of images showed that there were noticeable differences in the visual quality of reconstruction depending on the values of patch and region size chosen for a particular image and on the number of resolution scales used for the earlier stages. A small patch size means that the filled pixel correctly reflects small-scale variation i.e. texture, but may not achieve the correct overall shade. A large patch size may provide a better match to the local shade, but loses the opportunity to show texture, and 'smoothes' the image. As mentioned earlier, the optimum number of resolution scales tends to relate to the original image size, but some images did need more or fewer scales.

Figure 13 Best inpainting

These points can be illustrated by the example shown in Figure 14. This picture of lilies has its most plausible reconstruction following the removal of one stem (as shown in Figure 15) by using only 1 lower resolution scale (i.e. a factor of 2 in x and y) – see Figure 16. When using 3 lower resolution scales (Figure 17) the program failed to continue the horizontal wire correctly across the infill area. In both cases the patch and region sizes were the same.

It became apparent during the course of experiments that the experimenter could make a reasonable guess at the best parameters to use for the reconstruction from simply viewing the image. Further investigations are thus required to understand how this is occurring. For this inpainting method to work, parts of the area being reconstructed must be replicated (approximately) elsewhere in the image. The number of resolution scales used acts as a crude filter into frequency bands, whereas the patch size itself relates directly to the scale of the feature. The region size used for the local best match relates to the 'wave packet' scale, the distance within which the highest frequencies are part of some common feature. Initial Fourier analyses of our test images have not yet led to any clear relationships between the image scales and the parameters.

5 FUTURE WORK AND CONCLUSIONS

In order to derive the relationship mentioned above, further experiments need to be conducted. Because at present plausibility is so difficult to define mathematically, we will have to use human testing to achieve at least a statistical measure. For example, although [19] calculate an error measure for their magnification process, they found images with similar error values could have markedly different visual appeal. In a similar vein [7] found that some algorithms generated solutions which caused lines to be terminated, while human perception much prefers continuity. As a result of these difficulties, much of the work discussed earlier presents the images to the reader without evaluation or objective comparison (as do we). As the processes for inpainting improve, we have to start measuring the plausibility of the resultant images, in order to assess whether any new technique is a genuine improvement.

Assuming that human tests can give us a relationship for our free parameters to the image feature scales, the process of digital inpainting can then be made more automated. Given an image with a masked region, the program could use an initial analysis to determine plausible values for the 3 parameters, perform the inpainting using these and some nearby values, and thus provide a small set of reconstructed images from which the user could make the final selection

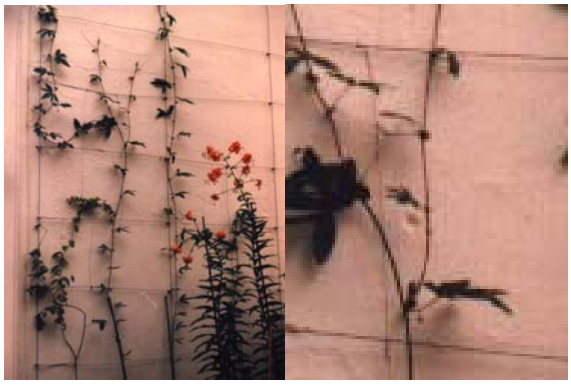

Figure 14 Original image and enlarged region

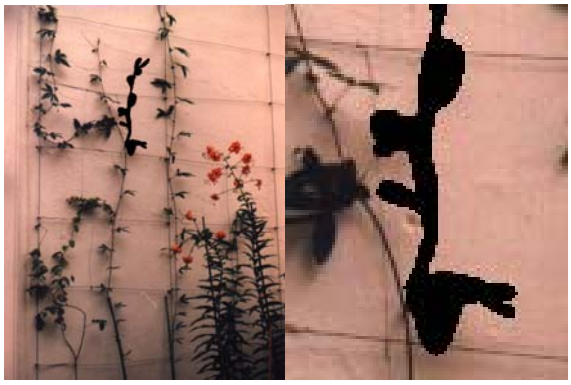

Figure 15 Masked image and enlarged region

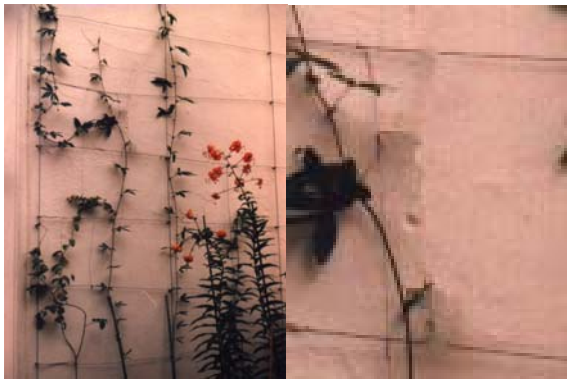

Figure 16 Best infill, and enlarged region

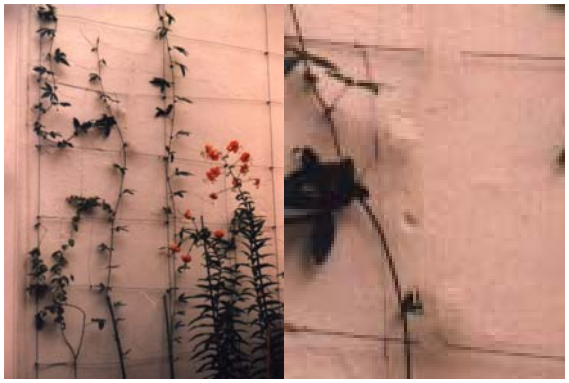

Figure 17 Infill using more lower resolution stages

If human tests could further assist in identifying which technique was most suitable for a given class of image, one could extend the process. Apart from simple Fourier or wavelet analysis, some segmentation could be carried out eg as in [18] to categorise the areas of texture to be inpainted. The most appropriate techniques could then be selected and used. Such a tool could free the user from the drudgery of manual editing, while still giving them the freedom to decide on the best image to use.

6 ACKNOWLEDGEMENTS

Thanks to .P. Bown, M.R.Cant for use of some images.

References

[1] M. Ashikhmin. Synthesizing Natural Textures. Proceedings of 2001 ACM Symposium on Interactive 3D Graphics, Research Triangle Park, North Carolina pp. 217-226, March 2001.

[2] C.Ballester, M.Bertalmío, V.Caselles, G.Sapiro and J.Verdera. Filling-In by Joint Interpolation of Vector Fields and Gray Levels. IEEE Trans. Image Processing, 10(8),pp1200-1211,August 2001.

[3] M. Bertalmío, G. Sapiro, V. Caselles and C. Ballester. Image Inpainting. Proceedings of SIGGRAPH 2000, pp 417-424, New Orleans, USA, July 2000.

[4] M. Bertalmío, A. Bertozzi, G. Sapiro. Navier-Stokes, Fluid-Dynamics and Image and Video Inpainting. IEEE CVPR 2001, Hawaii, USA, December 2001.

[5] T.E.Boult and G.Wolberg. Local Image Reconstruction and Sub-pixel Restoration Algorithms. CVGIP, Graphical Models and Image Processing, 55(1), pp63-77, 1993.

[6] T.F.Chan and J.Shen. Mathematical Models for local non-texture inpainting. SIAM, J.Appl.Math,63(2) pp1019-1043,2001.

[7] T.F.Chan and J.Shen. Non-texture inpainting by curvature driven diffusion (CDD). J.Visual Comm. Image Rep, 12(4), pp436-449, 2001.

[8] J.S.DeBonet. Multiresolution sampling procedure for analysis and synthesis of texture images. SIGGRAPH 97, pp361-368, 1997..

[9] A.Efros, T.Leung. Texture synthesis by non-parametric sampling. Proc. IEEE International Conference on Computer Vision, pp1033-1038, Corfu, Greece, September 1999.

[10] A. Efros W.T. Freeman. Image Quilting for Texture Synthesis and Transfer, Proceedings of SIGGRAPH '01,pp 341-346 Los Angeles, California, August, 2001.

[11] S.Esedoglu, J.Shen. Digital inpainting based on the Mumford-Shah-Euler model. European J.Appl.Math, 13, pp353-370, 2002.

[12] B. Guo, H. Shum, Y.-Q. Xu. Chaos Mosaic: Fast and Memory Efficient Texture Synthesis. Microsoft research paper MSR-TR-2000-32.

[13] D.Heeger, J.Bergen. Pyramid based texture analysis/synthesis. SIGGRAPH 95 pp229-238, 1995.

[14] A.N.Hirani, T.Totsuka. Combining frequency and spatial information for fast interactive image noise removal. SIGGRAPH '96, pp269-276, New Orleans, LA, 1996.

[15] H.Igehy and L.Pereira. Image replacement through texture synthesis. Proceedings of 1997 IEEE International Conference on Image Processing, pp 186-189, Santa Barbara, Oct 1997.

[16] J.Keen. Image reconstruction after object removal. BSc thesis, Nottingham Trent University 1997.

[17] D. King. The Commissar Vanishes, Holt,Henry & Co. ISBN: 0805052941.

[18] J. Malik, S.Belongie, T.Leung, J.Shi. Contour and Texture analysis for image segmentation. International Journal of Computer Vision, 43(1), pp7-27, 2001.

[19] B.S.Morse, D.Schwartzwald. Image Magnification and Level-Set Reconstruction. Computer Vision and Pattern Recognition 2001 (CVPR'01).

[20] L.-Y.Wei and M. Levoy. Fast texture synthesis using tree-structured vector quantization. Proceedings of SIGGRAPH 2000, pp479-488.

Richard Cant started life as a theoretical physicist, then moved into industry, where he spent 9 years as a system designer working on computer generated imaging for military training systems. His current research interests continue this theme and include use of neural nets for evaluation of image algorithms, including their validity for training.

Caroline Langensiepen started as a theoretical physicist, moved into industry as a nuclear physicist, then spent 8 years on system /software design of high reliability comms. systems & training simulators. She followed this with a period as an independent consultant for very large mission-critical systems. Her current interests include image processing and real-time sensor information analysis.