An Artificially Evolved Vision System for Segmenting Skin Lesion Images

Mark E. Roberts and Ela Claridge*

School of Computer Science, University of Birmingham, B15 2TT, UK

Abstract. We present a novel technique where a medical image segmentation system is evolved using genetic programming. The evolved system was trained on just 8 images outlined by a clinical expert and generalised well, achieving high performance rates on over 90 unseen test images (average sensitivity 88% , average specificity 96%). This method learns by example and produces fully automatic algorithms needing no human interaction or parameter tuning, and although complex, runs in approximately 4 seconds.

1 Introduction

In many areas of medicine, images are used as a diagnostic aid, but images in themselves only partially contribute. Crucially, input comes from interpretation of the image by an expert using the power of the human visual system. This human system works in real time, does not need carefully tuned parameters, and perhaps most importantly, is able to learn by example to recognise general image features. These qualities provide the inspiration for this work. We present here a method which learns by example, and produces fully automatic, parameter-free algorithms to identify given features.

The diagnosis of malignant melanoma at the primary care level is difficult because at the early stages it may look similar to innocent pigmented lesions – "moles". Moderate diagnostic rates achieved by dermatologists [1] confirm this difficulty and for this reason there is a growing body of work on using image analysis methods to aid the diagnosis of melanoma [2]. The diagnosis first requires the segmentation of the lesion from the surrounding skin, which is a difficult task, mainly due to the variability in lesion appearance. Some lesions are well delineated and make good contrast with the skin, whereas others are indistinct, variegated and difficult to see by an untrained eye. The published methods use a variety of approaches, including threshold-based methods [3], colour clustering and distance functions in a colour space [3, 4], edge modeling [5, 6] and various combinations of these [7].

What these methods have in common is the fact that they all have been developed by image analysis experts, in most cases informed by clinical practitioners. This paper describes a very different approach, in which expertise in image analysis is not necessary for being able to create a well performing image processing system, in this case for lesion segmentation. This is achieved through the use of an evolutionary computation technique - genetic programming (GP) in which a lesion segmentation system is automatically evolved, purely on the basis of example segmentations provided by an expert clinician. The paper first outlines the concept of genetic programming. This is followed by the description of a study in lesion segmentation using GP, its results and discussion.

2 Materials and Methods

2.1 Pigmented Lesion Images

A set of 100 pigmented lesion images is used in this study. The images were acquired using a SIAscope [8], a device designed specifically for skin imaging, that takes a number of images of the same area of the skin at different wavelengths. In its normal mode of operation it uses an optical model of the skin to compute parametric maps showing the distribution and levels of individual histological components of the skin such as melanin, haemoglobin and collagen [8]. In this study images acquired in the blue band are used because of strong absorption by both melanin and blood makes the lesions stand out against the skin background. Image resolution is 40 microns per pixel and a circular area with radius of 280 pixels is used. A "ground truth" data set in the form of binary images, is created from outlines drawn by a clinical expert at Addenbrooke's Hospital, Cambridge, UK.

2.2 Genetic Programming

Genetic Programming (GP) is a powerful extension to the genetic algorithm (GA) paradigm which evolves populations of computer programs as opposed to the simplistic binary strings used in GAs. These programs are repre-

*Email: {M.E.Roberts,E.Claridge}@cs.bham.ac.uk

sented as tree structures and are initially created randomly from sets of functions and terminals. The programs are run on a problem, and a fitness value is assigned based on how well they perform. These fitness values are used to implement "survival of the fittest" procedures which select, and then adapt, the fitter individuals by means of mutations (random changes to a single individual) and crossover (creating new offspring influenced by two parents). With programs represented as trees, mutation replaces a randomly chosen sub-tree with a randomly generated new sub-tree. Crossover simply selects random points from two trees and swaps over the sub-trees beneath them to generate two new children. Over many generations, better and better solutions to the problem emerge. Effectively GP creates programs to solve a problem, without being told how to solve it, or knowing anything about its underlying nature. The programs produced are often quite novel, as the process is free from any human preconceptions about the problem or what constitutes a good solution.

However, the huge computational expense of running thousands of complex programs for many generations means that only recently have we reached the stage where imaging problems can be tackled by using image processing operations in the function set, and input images in the terminal set. Figure 1 shows some very simple examples of GP image trees, the output they produce, and how a crossover operation would create two random children.

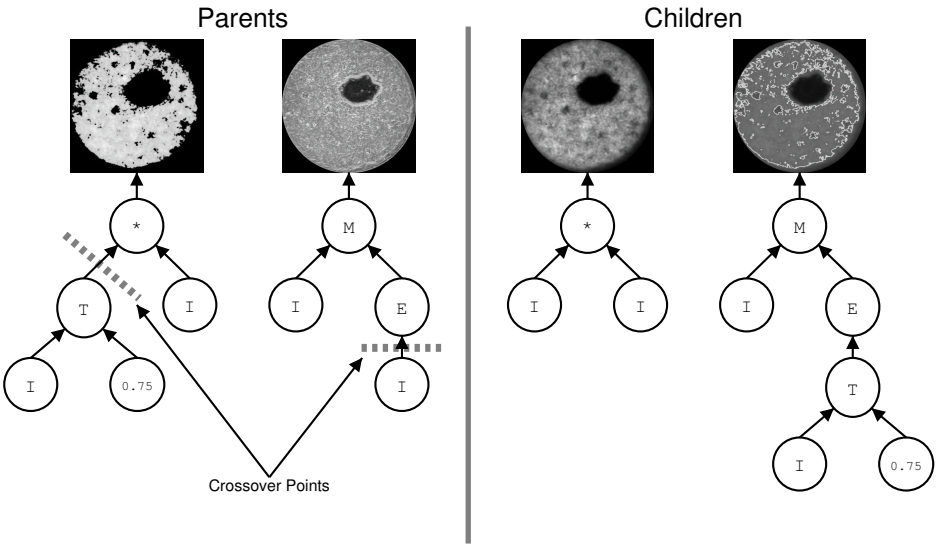

Figure 1. A demonstration of GP evaluation and crossover. The trees represent hierarchical programs. *I* represents the input image, *M* is a 50/50 image merge, * multiplies images, *T* thresholds the image to the given value and *E* performs an edge detection. Crossover can be seen to generate children with some of each parent's properties.

2.3 Experimental Setup

The data was divided into a training set containing 8 examples and a test set containing 92 examples. The 8 images were chosen as they represented most of the variation found in the dataset. A population of programs is then randomly created from the function and terminal sets. The function set contains imaging operations such as thresholds, quantisation, morphological operations, logical operations, region intensity functions (mean, min, max), edge filtering, merging etc. The terminal set consists of the input image, and numerical and coordinate values. More information about this type of system can be found in [9].

Every generation, each program in the population is run on each of the images in the training set. The fitness of each program is measured, and then used to influence the selection procedures to decide which ones are adapted and put into the next generation. A population of 5000 programs was used, and the system was run for 75 generations.

2.3.1 Fitness Function

The fitness function is key to the success of the evolution. It should provide proportional feedback to the GP system in a way that correctly captures what it means for a solution to be better or worse than another. The fitness function used in this case is a modification of a function proposed by Poli [10] for similar segmentation problems, summed over all of the N images in the training set. It is defined using measures of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) which are used in this fitness function. It is important to use these

measures instead of simple pixel difference counts for obvious reasons. The function used is:

$$f = \sum_{i=1}^{N} \frac{FP}{FP+TN} + \frac{FN}{TP+FN} \exp\left(10\left(\frac{FN}{TP+FN} - \alpha\right)\right) \quad (1)$$

where α is a parameter (in this case set at 0.4) allowing the relative importance of sensitivity and specificity to be varied. Also, as in Poli's work, a wrapper function is used which thresholds the image before the fitness calculation. So, the system is actually trying to find the solution that, when thresholded, best matches the target output.

3 Results

The binary image that the program outputs is compared to the correct outline and each pixel is classified as a TP, FP, TN, or FN. From these classifications, measures of sensitivity and specificity are produced. These are shown for each of the 92 test images the a 2-axis histogram in Figure 2. A four colour image produced showing these classifications as light grey, black, white, and dark grey respectively. Some examples are shown in Figure 3.

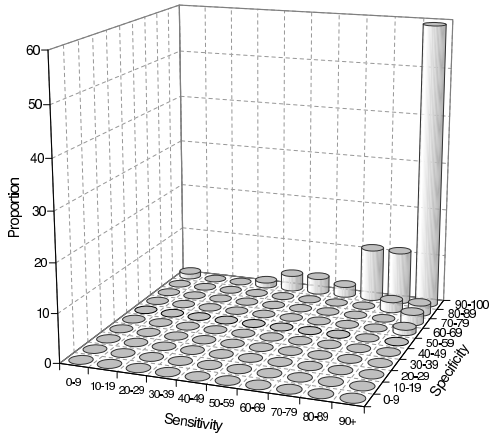

Figure 2. Sensitivity-specificity histogram showing the percentage of results in each performance category for the 92 unseen test images images

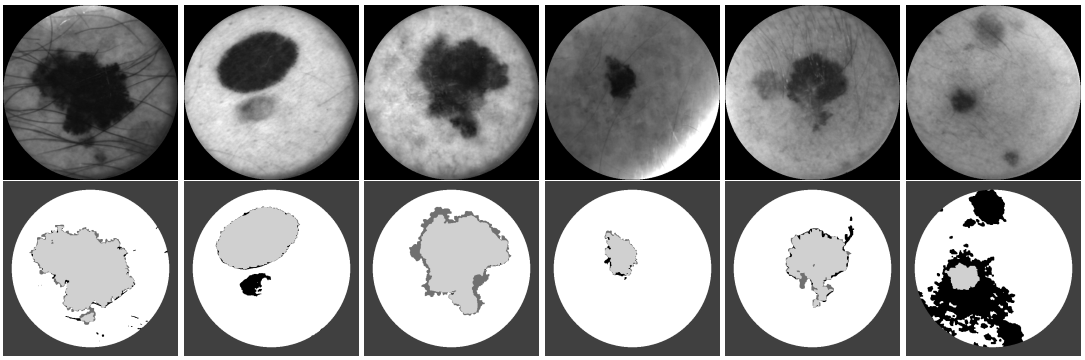

Figure 3. Examples of inputs and performance on images from the unseen test set. Light-grey=TP, Black=FP, Dark-grey=FN, White=TN. Shows examples of hair removal, detection of similar non-outlined lesions, good segmentations in spite of irregularity, and a complete failure.

4 Discussion and Conclusions

4.1 Analysis of Performance

On the unseen test data the program performs very well on the majority of the examples, as can be seen from the histogram in Figure 2. Most of the examples are clustered in the very high accuracy regions of the histogram. The examples on which the program performed badly are generally those which are highly irregular which were not fully represented in the training set. Future training should use more of these irregular images.

Although the segmentations produced are not perfect, the algorithm would be a good first step to a system such as [5], which can perform a more detailed analysis of the lesion borders but needs first to locate the centre of the lesion. This system could quite easily produce this sort of input.

4.1.1 Programs produced

The best program produced after 75 generations contained 330 nodes and executes in about 4 seconds. Pruning and optimisation of the program tree could easily reduce this down to real time. The run that produced the program took approximately 24 hours to complete, running on a cluster of computers of varying specifications. This amount of time may at first make this approach seem prohibitive, but this is a one-off expense in search of just one single, and actually quite simple program which takes only 4 seconds to run. The obvious question arises from this work; what does the resultant program do that makes it so good? This is a very difficult question to answer because of the complexity of the programs and the often unconventional steps they use, and this analysis is an ongoing task.

4.2 Summary and Future Work

We have presented preliminary results of a system which uses GP to evolve a program to segment pigmented skin lesions. The method presented has several important benefits over more traditional segmentation methods. The most important is that it can be used by non-experts. All the system needs is input images and target outputs. Secondly, the fact that the system learns by example makes it applicable to many more problems than model based approaches which can be too specific. Generalisation is a key feature of this method and the fact that the system performs well when trained on only 8 images demonstrates this. Also, the programs produced are free from human preconceptions about the problem, and pick up on aspects of the problem that humans may miss.

There is enormous scope for future work in this area. The method could be applied to almost any binary segmentation problem and a few modifications to the paradigm could make the system applicable to non-binary problems. All that is required is for an expert to provide hand segmented examples for training. Specific future work includes using outlines drawn by multiple experts in order to reduce the intra and inter-expert ambiguities which confuse the learning process.

Acknowledgments

The authors would like to thank Mr. Jonathan Powell, Addenbrooke's Hospital, for supplying the outlined dataset.

References

1. C. Morton & R. MacKie. "Clinical accuracy of the diagnosis of cutaneous malignant melanoma." *British Journal of Dermatology* **138**, pp. 283–287, 1998.
2. G. Day & R. Barbour. "Automated melanoma diagnosis: where are we at?" *Skin Research and Technology* **6**, pp. 1–5, 2000.
3. L. Xu, M. Jackowski & et. al. "Segmentation of skin cancer images." *Image and Vision Computing* **17(1)**, pp. 65–74, January 1999.
4. P. Schmid. "Segmentation of digitized dermatoscopic images by two-dimensional color clustering." *IEEE Transactions on Medical Imaging* **18(2)**, pp. 164–171, February 1999.
5. E. Claridge & A. Orun. "Modelling of edge profiles in pigmented skin lesions." In A. Houston & R. Zwiggelaar (editors), *Proceedings of Medical Image Understanding and Analysis 2002*, pp. 53–56. 2002.
6. J. Gao, J. Zhang, M. Fleming et al. "Segmentation of dermatoscopic images by stabilised inverse diffusion equations." In *Proceedings of the International Conference on Image Processing*, pp. 823–827. Chicago, October 1998.
7. H. Ganster, A. Pinz, R. Rohrer et al. "Automated melanoma recognition." *IEEE Transactions on Medical Imaging* **20(3)**, pp. 233–239, March 2001.
8. E. Claridge, S. Cotton, P. Hall et al. "From colour to tissue histology: Physics based interpretation of images of pigmented skin lesions." In T. Dohi & R. Kikinis (editors), *Proceedings of MICCAI'2002*, pp. 730–638. Springer, 2002.
9. M. Roberts. "The effectiveness of cost based subtree caching mechanisms in typed genetic programming for image segmentation." In S. e. a. Cagnoni (editor), *Applications of Evolutionary Computation, Proceedings of EvoIASP 2003*, volume 2611 of *LNCS*. Springer-Verlag, Essex, UK, April 2003.
10. R. Poli. "Genetic programming for feature detection and image segmentation." In T. Fogarty (editor), *Proceedings of the AISB'96 Workshop on Evolutionary Computation*, volume 1143 of *Lecture Notes in Computer Science*, pp. 110–125. Springer, April 1996.