Using 3D information for classification of non-melanoma skin lesions

Steven McDonagh[1], Robert B. Fisher[1], Jonathan Rees[2] *

[1]School of Informatics, [2]Dermatology, University of Edinburgh

Abstract. New sensors allow simultaneous acquisition of 3D shape and colour data of skin at resolutions theoretically approaching cellular structures. We investigate whether the addition of 3D depth information increases classification rates relative to only using colour information for 5 non-melanoma skin lesions. The paper demonstrates that there is 6% increase in classification rates.

1 Introduction

There has been much image analysis research dedicated to the recognition of malignant melanoma from the mid-1980's (e.g. [1]) and reported classification sensitivities and specificities are now typically over 90% [2]. However, melanoma is not the only form of skin cancer, or indeed skin lesion, and there has been almost no image analysis on these other conditions. This paper presents the results of a classification study of 5 common classes of skin lesions:

AK - Actinic Keratosis	**BCC** - Basal Cell Carcinoma	**ML** - Melanocytic Nevus / Mole
SCC - Squamous Cell Carcinoma	**SK** - Seborrhoeic Keratosis	

We omit melanoma for two reasons: 1) other diagnostic methods, particularly the ELM-based methods briefly discussed below, are very successful already, and 2) melanoma is actually a rather rare, but quite dangerous, condition whereas other skin cancers are much more common. As we only had a few sample cases of melanoma, we chose to focus on other classes.

The classification study used two types of information: 1) RGB colour images from a digital SLR camera, and 2) registered 3D (x, y, z) data at each colour pixel. The addition of 3D data was motivated by the experience of clinical dermatologists, who touch lesions as part of their examinations. With the addition of depth data, skin shape properties can also be extracted that are potentially of benefit in the classification process. This leads to the key hypothesis of this paper: **A classification process that uses depth and colour image features has a higher classification rate than the same process using only colour features.** We will show that the classification rate increases from 77% to 83% and that the null hypothesis can be rejected with significance 0.3. More details can be found at [3].

2 Background

While there is much previous research on automated skin cancer diagnosis (e.g. [4]) that research is primarily concerned with melanoma, because of its danger, and because of the high likelihood of recovery given early detection. The three main techniques used for automated classification are based on: 1) the ABCD visual diagnostic system, 2) Epilumiunescent Microscopy (ELM) and 3) physics-based skin modelling. A good survey of classification techniques is given in [5], from which it can be seen that melanoma classification sensitivity and specificity rates above 90% are possible.

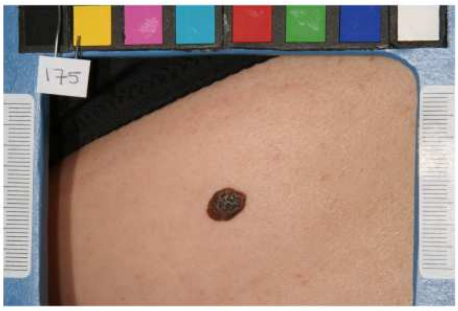
The ABCD system was originally developed for clinical use, standing for Asymmetry, Border structure, Colour variegation and Diameter of lesions. Several automated classification processes based on different implementations of the ABCD parameters have been reported (e.g. [4], with sensitivity 87% and specificity 92%). From about 1995, there has been research into automated classification systems based on ELM imagery, a noninvasive microscopic technique that uses oil on the skin to make the epidermis more transparent and thus allow observation of deeper skin structures. For example, [6] achieved sensitivity 93% and specificity 95%. Claridge et al [7] developed a physics-based model of tissue colouration linking spectral composition to skin structure and the physics of light interacting with skin. Using several (quantity unclear) measurements in the 400-1000 nm range, they reported 80% sensitivity and 82% specificity for melanoma.

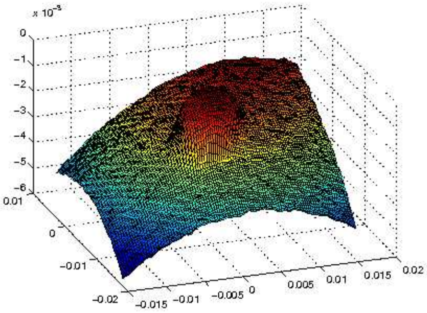
We have not identified any previous non-Edinburgh research using 3D depth data for skin cancer analysis. The DERMA [8] project used registered depth and colour data, where the depth data was obtained from a laser range sensor. The DERMA project measured the depth of wounds to investigate the time evolution of wounds. Ravindranath [9] made initial investigations into the simultaneous use of colour plus lesion height. While that study had only 84 samples, best

[*] rbf@inf.ed.ac.uk, jonathan.rees@ed.ac.uk. We thank the Foundation for Skin Research for their support.

results were achieved (84.6% accuracy) on classifying malignant *versus* non-malignant. No evidence for a benefit from the single depth feature was found. Round et al [10] used the intensity patterns observed from shape texture to seperate melanoma from melanocytic naevi (moles).

Range and colour data are acquired using a Dimensional Imaging dense stereo image capture system built around a pair of Canon EOS 350D SLR cameras. Each camera acquires a 3456x2304 image. Given camera placement, lenses and patient placement, each pixel corresponds to about 0.03 mm skin sample separation. Measurements have determined an RMS depth error also of about 0.03 mm. An example of the colour (left) and 3D shape (right) capture is:



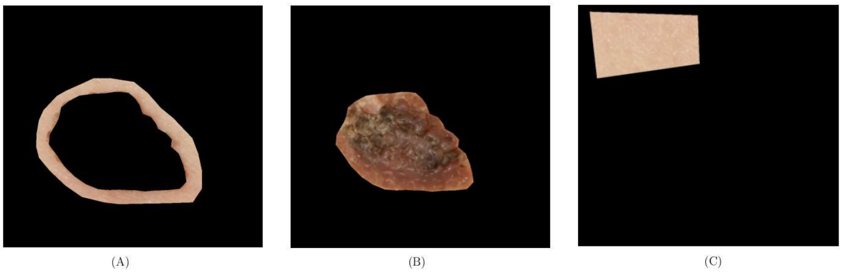


3 Method

A Bayes Classifier with a unimodal multidimensional Gaussian class model was used. Thirty different colour or depth based features were computed from the image data. A greedy feature selection method was used to identify the best $N = 10$ features for the colour and range+colour classifiers.

Preprocessing consisted of 2 stages:

1. **Segmentation of lesion and normal regions**: Most of the features require some relative comparison between properties of the normal skin and the lesion. There has been much research into automatic segmentation of melanoma lesions [4], but we chose to not reimplement this to simplify the research. Three regions were marked by hand using an interactive image markup tool: 1) lesion, 2) "uncertain" boundary region around the lesion, 3) normal skin. An example of the "uncertain" (left), skin lesion (middle) and "normal skin" (right) masked areas is here:



(A) (B) (C)

2. **Rotation to fronto-parallel**: As the patient images could be captured with some slant relative to the sensor, the 3D data was rotated to have the normal skin surface parallel to the image plane. This was done by fitting a plane to the 3D data of the 'uncertain' (left patch in the figure above) skin patch and then rotating that plane until its surface normal was facing the viewer.

Six families of features were extracted from the colour (22 features) and depth (8 features) image data, for a total of 30 features. Only the features ultimately used are listed here.

Relative Colour Brightness Features (9): Features 2-10 are the ratios of colour intensities within the lesion relative to the normal skin. This family of features was implemented in an attempt to automatically compute quantities which emulate the "colour" aspect of the ABCD clinical diagnosis rule criteria discussed previously. The ratio of mean colours is used to normalise for global illuminant intensity. The features are $\frac{\mu_{a,S}}{\mu_{b,T}}$, where μ is the mean value of channel $a, b \in \{R, G, B\}$ and S=lesion and T=Normal skin patches. The features are numbered (ab): 2:RR, 3:RG, 4:RB, 5:GR, 6:GG, 7:GB, 8:BR, 9:BG, 10:BB.

Relative Variability Features (4): Features 11-14 assess the variability of colours/shape inside the lesion relative to the normal skin patch. The relative ratio normalises for scale and human variations. The features are $\frac{\sigma_{c,S}}{\sigma_{c,T}}$ where σ is the standard deviation of channel $c \in \{R, G, B, Z\}$ and S=lesion and T=Normal skin patches. The features are numbered (c): 11: R, 12:G, 13:B, 14:Z.

Peak and Pit Density Features (12): Image data is convolved with a Gaussian filter with 3 different scales to remove small features. A peak/pit is defined as a pixel whose value was larger/smaller than the 8 nearest neighbours. The ratio $v_{c,s} = \frac{\#peaks_c + \#pits_c}{lesionarea}$ is computed where $c \in \{R, G, B, Z\}$ and $s \in \{0.5, 1.0, 2.0\}$ is the Gaussian filter standard deviation. The features are numbered (c,s): 15:R,0.5, 16:G,0.5, 17:B,0.5, 18:Z,0.5, 19:R,1.0, 20:G,1.0, 21:B,1.0, 22:Z,1.0, 23:R,2.0, 24:G,2.0, 25:B,2.0, 26:Z,2.0.

3D Shape Moment Invariant Features (3)

Features 28:J_1, 29:J_2, 30:J_3, characterise the mass distribution of the lesion volume, treating the 3D lesion as the volume that lies away from the ground plane [11].

Figure 1 left shows the distribution of feature values for the 5 classes for feature 5. There is clearly some difference in the distributions, but it is also clear that there is considerable overlap in the ranges of feature values. Hence, a statistical classifier was used.

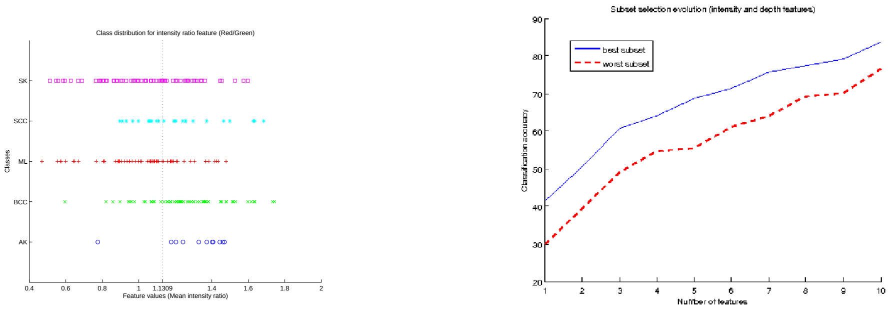


Figure 1. Left: Pre-normalisation distribution of feature values for the 5 classes for feature 5. Right: top/blue: Increase in classification rate as new features are added. bottom/red: rate achieved if worst feature is added.

A standard Bayesian classifier with unimodal multivariate Gaussian observation model was used. Visual examination of the feature value distributions suggests that the Gaussian model is reasonable. All features were normalised so that their feature values were zero mean, unit standard deviation. Distribution means and covariances were estimated from the training data subsets. *A priori* class probabilities were estimated using the incidence rates in the training data.

The only non-standard aspect to the classification was the calculation of the average classification rate. Dividing the total correct classifications by total classifications biases the results towards the classification rates of the larger classes. Here, some classes (AK,SCC) did not have many samples. So, instead, the classification rate is computed for each of the five classes and the average classification rate is the average of these 5 rates.

Although 30 features were implemented, only 10 of these were used because the AK class had only 11 examples. Trying all feature combinations ($choose(30, 10) \doteq 3 \times 10^7$) gives optimal feature selection, but at 15 seconds per combination, this is computationally infeasible. The efficient but suboptimal *Sequential forward selection* (SFS) [12] algorithm was used. Starting with an empty set of features, SFS iteratively adds the single feature that maximises classifier performance when using that feature plus all those previously selected, continuing until the required dimensionality is achieved. This algorithm considers only 245 combinations. Figure 1 (right) shows the increase in classifier performance as each new feature is added (top blue curve) and what it would have been if the worst feature was added to the previous set of best features (bottom red curve).

4 Experiments and Evaluation

There were 234 samples over the 5 classes, distributed as seen in Table 1. Classes AK and SCC are underrepresented in the experiments, but were limited by patient availability. These samples were selected from a pool of patients at the Dermatology Clinic at the Edinburgh Royal Infirmary. Ground-truth classification was made by one of the authors (Rees), a consultant dermatologist, based on clinical observation and, in some instances, histopathology. Samples were excluded if: diagnosis was ambiguous, depth recovery failed or colour image was unsatisfactory.

Because of the small pool of cases, the cases were not separated into independent training and test sets. Instead, leave-one-out k-fold cross validation was used (i.e. each classifier is trained on all of the available skin lesions *apart from the one that is to be classified*). The system was trained 234 times on all data except for one sample and a prediction is made for that sample. This affords us the maximum mileage possible from the available data in terms of model training.

As there were only 11 samples in class AK, the number of features used in the classification was limited to 10. Features were selected as described above, selecting the feature at each stage that produced the lowest classification error using leave-one-out cross validation. The features that were selected follow. Depth based features are in bold.

Feature set	Feature pool	Feature subset
I	Colour only	{10,9,8,3,5,25,4,16,13,7}
II	Colour and depth	{10,9,8,3,**22**,**30**,21,25,15,**26**}

With these features, the final leave-one-out cross validation experiment was performed. Tables 1 and 2 show the classification results for the colour and depth+colour classifiers respectively. Adding the depth features clearly improves the combined (using the combination method described above) classification accuracy from 77.3% to 83.7%. The statistical significance of this difference is discussed below.

The class specific results in Tables 1 and 2, show that the two classifiers performed similarly with the classes AK, BCC, ML and SK with both feature sets recognising all AK samples correctly and achieving similar accuracies for the other three classes. Both classifiers had some trouble recognising the dangerous SCC class, displaying their lowest individual class accuracies. The classifier using depth+colour features is able to outperform the colour features by a considerable 32% recognition rate. Typical SCC lesions were noted to have a "crater like" appearance with raised surroundings and a central depression. The three features utilising depth as a modality in feature set (II) may be at least partially helpful in extracting these characteristics. The most severe problems in these classifications are the cases in which malignant lesions (BCC,SCC) have been categorised as benign lesions (ML,SK) such as the 8 BCC and 5 SCC samples in Table 1. In a similar fashion, Table 2 exhibits 4 BCC and 2 SCC samples categorised as benign lesions. Thus, using the depth+colour classifier has also reduced the potential cost of dangerous misclassifications.

		Diagnosis						
		AK	BCC	ML	SCC	SK	Number	Rate
True	AK	11	0	0	0	0	11	100%
	BCC	0	55	3	2	5	65	84.6%
	ML	0	7	49	0	5	61	80.3%
	SCC	0	9	2	11	3	25	44%
	SK	0	7	8	1	56	72	77.7%
Overall accuracy							234	77.3%

Table 1. Confusion matrix for feature set (I)

		Diagnosis						
		AK	BCC	ML	SCC	SK	Number	Rate
True	AK	11	0	0	0	0	11	100%
	BCC	0	57	1	4	3	65	87.6%
	ML	0	4	48	2	7	61	78.6%
	SCC	0	4	1	19	1	25	76%
	SK	0	4	8	5	55	72	76.3%
Overall accuracy							234	83.7%

Table 2. Confusion matrix for feature set (II)

McNemar's test [13] was used to test whether the difference in results is statistically significant. McNemar's test is a relatively simple measure that can be applied to dual system classification experiments. Let **s1** be the classifier based on Colour features only and **s2** be the classifier based on Depth and Colour features. Let n_{10} be the number of examples misclassified by **s2** but not by **s1** and n_{01} be the number of examples misclassified by **s1** but not by **s2**. The **null hypothesis** is: **the two systems have the same error rate**. McNemar's test is based on a χ^2 test and essentially computes a goodness of fit that compares the distribution of counts expected under the null hypothesis to the observed counts. This statistic:

$$\phi = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

is distributed (approximately) as χ^2 with 1 degree of freedom. Here, $n_{10} = 27$ and $n_{01} = 19$, so $\phi \doteq 1.065$, which means that the null hypothesis can be rejected at the 0.3 confidence level. Thus, there is evidence that a classification system using a combination of colour and depth based image features may be more successful than a system that uses colour based features alone. A larger test set would give more power to the conclusion.

Most melanoma classification results are reported in terms of specificity and sensitivity, both of which are now better than 90% typically. However, these criteria are only appropriate for two class decision algorithms (e.g. melanoma *vs* non-melanoma). Here, the specificity (0.95 average) and sensitivity (0.84 average) of the range+colour classifier is computed for each class individually, treating all other class samples as true negatives.

Diagnosis	AK	BCC	ML	SCC	SK	Average
Specificity	1.00	0.93	0.95	0.95	0.94	0.95
Sensitivity	1.00	0.88	0.79	0.78	0.76	0.84

5 Discussion

The key point to make is that the addition of depth features almost certainly gives improved classification results (83.7%) compared to simply colour features (77.3%). We were limited to the use of 10 features by having only 11 samples of the AK class. Figure 1 is still showing a pronounced increase in classification rates at 10 features, suggesting that increased performance could be achieved by using additional features. Principal Component Analysis might be usable to reduce the dimensionality of the feature space without significant loss of discrimination ability. The features that were used were somewhat generic and improved discrimination between commonly confused classes (e.g. ML & SK) could be possible with additional features designed specifically for separating these two classes. Another family of features that were not explored were those using shape and colour simultaneously, which might draw out any correlations between the properties. Given the clinical use of shape and texture in the classification of lesions, the results from our preliminary study give weight to the hypothesis that depth measurements can also be useful in automated skin lesion classification systems. A second important point is that this seems to be the first study to consider a wider class of lesion than simply malignant melanoma *versus* moles. Hence, it has a broader impact. While performance is not yet near the 90% threshold of melanoma research, this paper highlights one source of information that might help achieve that level.

References

1. A. Dhawan. “An expert system for the early detection of melanoma using knowledge-based image analysis.” *Anal Quant Cytol Histol* **10**, pp. 405–416, 1988.
2. D. Gutkowicz-Krusin, M. Elbaum, P. Szwaykowski et al. “Can early malignant melanoma be differentiated from atypical melanocytic nevus by in vivo techniques?” *Skin Res Technol* pp. 3:15–22, 1997.
3. S. McDonagh. “Skin Cancer Surface Based Classification.” *Undergraduate Thesis, School of Informatics, University of Edinburgh* 2008.
4. H. Ganster, A. Pinz, R. Rohrer et al. “Automated melanoma recognition.” *IEEE Transactions on Medical Imaging* **20**, pp. 234–239, 2001.
5. G. R. Day & R. H. Barbour. “Automated melanoma diagnosis: where are we at?” *Skin Research and Technology* **6**, pp. 1–5, 2000.
6. S. Seidenari, G. Pellacani & P. Pepe. “Digital videomicroscopy improves diagnostic accuracy for melanoma.” *J Am Acad Dermatol* **39**, pp. 175–1181, 1998.
7. E. Claridge, S. Cotton, P. Hall et al. “From colour to tissue histology: Physics based interpretation of images of pigmented skin lesions.” *Medical Image Analysis* **7(4)**, pp. 489–502, 2003.
8. M. Callieri, P. Cignoni, P. Pingi et al. “Derma: Monitoring the evolution of skin lesions with a 3D system.” In *VMV*, pp. 167–174. 2003.
9. H. Ravindranath. “Skin Spot Classification using 3D data.” *MSc Thesis, School of Informatics, University of Edinburgh* 2006.
10. A. Round, A. Duller & P. Fish. “Lesion classification using skin patterning.” *Skin Research and Technology* **6(4)**, pp. 183–192, 2000.
11. A. Sadjadi & E. Hall. “Three-dimensional moment invariants.” *IEEE Trans. on Pattern Analysis and Machine Intelligence* **2(2)**, pp. 127–163, March 1980.
12. P. Devijver & J. Kittler. *Pattern Recognition: A Statistical Approach.* Prentice-Hall, 1982.
13. T. Dietterich. “Approximate statistical test for comparing supervised classification learning algorithms.” *Neural Computation* **10(7)**, pp. 1895–1923, 1998.