

Fuzzy Description of Skin Lesions

Nikolaos Laskaris^a, Lucia Ballerini^a, Robert B. Fisher^a, Ben Aldridge^b, Jonathan Rees^b

^aSchool of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh, UK;

^bDepartment of Dermatology, University of Edinburgh, Lauriston Place, Edinburgh, UK

ABSTRACT

We propose a system for describing skin lesions images based on a human perception model. Pigmented skin lesions including melanoma and other types of skin cancer as well as non-malignant lesions are used. Works on classification of skin lesions already exist but they mainly concentrate on melanoma. The novelty of our work is that our system gives to skin lesion images a semantic label in a manner similar to humans. This work consists of two parts: first we capture the way users perceive each lesion, second we train a machine learning system that simulates how people describe images. For the first part, we choose 5 attributes: colour (light to dark), colour uniformity (uniform to non-uniform), symmetry (symmetric to non-symmetric), border (regular to irregular), texture (smooth to rough). Using a web based form we asked people to pick a value of each attribute for each lesion. In the second part, we extract 93 features from each lesion and we trained a machine learning algorithm using such features as input and the values of the human attributes as output. Results are quite promising, especially for the colour related attributes, where our system classifies over 80% of the lesions into the same semantic classes as humans.

Keywords: skin lesion images, semantic labels, human perception, feature extraction, supervised learning

1. INTRODUCTION

Dermatologists usually perform the diagnosis of skin lesion based on their personal experience. The criteria to reach a diagnosis are often derived from qualitative properties they observe in the images, but they are not able to describe the lesions in a consistent way using quantitative measurements.

The goal of this work is to develop a system that can describe skin lesions in a manner similar to dermatologists. This is an attempt to create a model to simulate human perception, using image analysis and machine learning techniques. The system takes as input features extracted from skin lesion images and produces as output a description of some shape, colour and texture properties (i.e. regular, irregular, pale, dark, uniform, rough, etc.). In such a way, the presented system assigns descriptive labels to skin lesions in a similar manner as the visual perception of dermatologists.

Methods to interpret visual information carried in a digital image, describing it in ways that humans do, by “extracting” high-level semantics such as “people playing football” or “tiger hunting its prey” have been of great interest during the previous years.

Humans tend to use high-level features (concepts), such as keywords, text descriptors, to interpret images and measure their similarity. On the other hand, the features automatically extracted using computer vision techniques are mostly low-level features (colour, texture, shape, spatial layout, etc.). In general, there is no direct link between the high-level concepts and the low-level features.¹ More specifically, the discrepancy between the limited descriptive power of low-level image features and the richness of user semantics, is referred to as the “semantic gap”.²

A recent survey¹ shows that the state-of-the-art techniques in reducing the “semantic gap” include mainly five categories: (1) using object ontology to define high-level concepts; (2) using machine learning methods to associate low-level features with query concepts; (3) using relevance feedback to learn the user’s intention; (4) generating semantic templates to support high-level image retrieval; (5) fusing the evidence from HTML text and

Further author information: (Send correspondence to Lucia Ballerini)
Lucia Ballerini: E-mail: lucia.ballerini@ed.ac.uk, Telephone: +44 (0)131 651 5664

the visual content of images for WWW image retrieval. Major recent publications are included in that survey covering different aspects of the research in this area, including low-level image feature extraction, similarity measurement, and deriving high-level semantic features.

Some authors^{3,4} report that in recent years, content-based image retrieval research (CBIR) has shifted its focus towards bridging the gap between low-level features and high-level semantics.

This work can be seen as attempt to reduce the “semantic gap” that uses supervised learning methods to associate low-level features with high-level semantic concepts.

Most of the work in dermatology has focused on skin cancer detection. Different techniques for segmentation, feature extraction and classification have been reported by several authors. Concerning segmentation, Celebi et al.⁵ presented a systematic overview of recent border detection methods: clustering followed by active contours are the most popular. Numerous features have been extracted from skin images, including shape, colour, texture and border properties.^{6,7,8} Classification methods range from discriminant analysis to neural networks and support vector machines.^{9,10,11}

These methods are mainly developed for images acquired by epiluminescence microscopy (ELM or dermoscopy) and they focus on melanoma, which is actually a rather rare, but quite dangerous, condition whereas other skin cancers are much more common.

However, no work has been done previously on automatically annotating skin lesion images with a symbolic label and thus the novelty of this work consists in both inspecting the way people understand skin lesion images, and attempting to model human perception.

To our knowledge, there is only a proposal for a CBIR system of skin lesion images that incorporates human perception. The system proposed by Celebi et al.¹² incorporates human perception to guide the search for an optimum similarity function. They designed an experiment to measure the perceived similarity of each image with every other other in the database. However, they focus only on shape similarity.

The novelty of our work is that our system gives to skin lesion images a semantic label in a manner similar to humans. Our system classifies 31 lesions into the same colour description as the most common selected label with 19% error.

The structure of the paper is as follows. Section 2 describes the two parts of our work, i.e. the experiment to gather user description of lesions, and the machine learning system implementation. Results are presented in Section 3, together with their evaluation. Conclusion and possible future direction follows.

2. METHOD

This work consists of two parts: the first step is to capture how users perceive each image, the second step is to train a machine learning system that simulates how people describe images.

Images used in this work are acquired using a Canon EOS 350D SRL camera, having a resolution of about 0.03 mm. A subset of 31 pigmented lesions, having different characteristics, were chosen by dermatologists for this part of the study. Some images are shown in Figure 1.

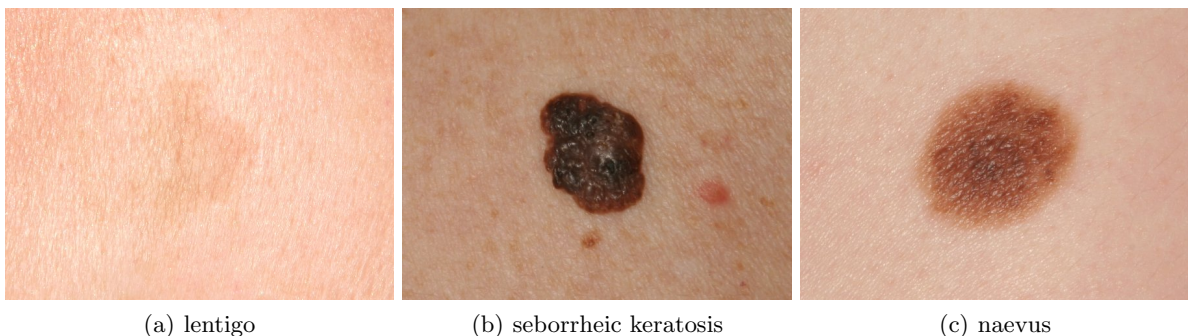


Figure 1. Examples of skin lesion images used in this work

2.1 Gathering User Perception

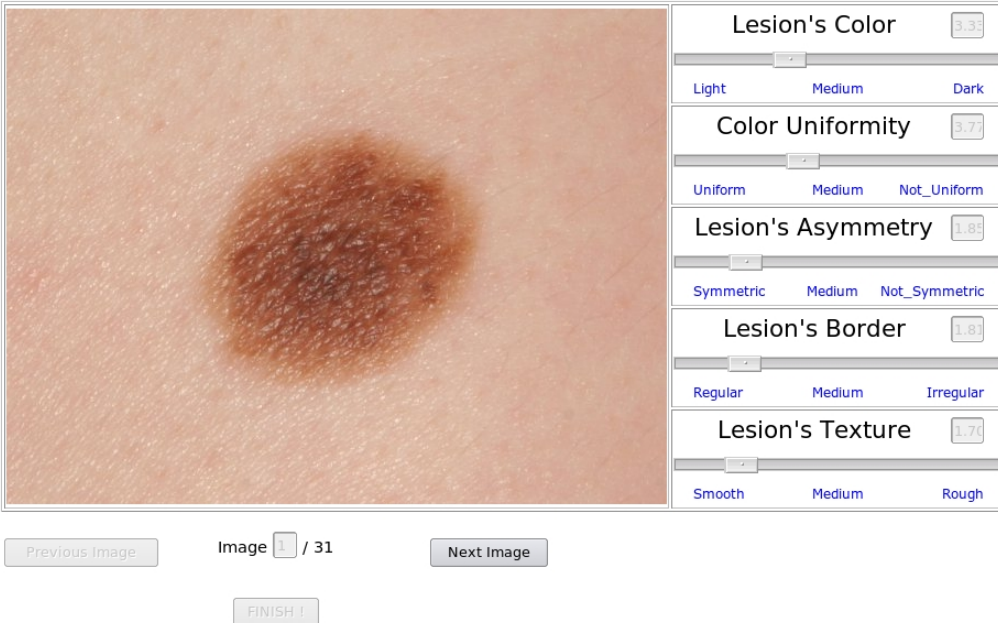
A pilot study showed that people tend to describe the lesion with concepts they are familiar with, e.g. “it looks like a pizza”. Moreover, there was huge variability in the description of the same image using natural language.

For this reason, a predefined vocabulary was chosen. For each concept, the user can assign an attribute, e.g. for colour: “light”, “dark” and so on.

These concepts are:

1. Colour of lesion [light to dark]
2. Uniformity of colour [uniform to heterogeneous]
3. Asymmetry of lesion [symmetric to asymmetric]
4. Border of lesion [regular to irregular]
5. Texture of lesion [smooth to rough]

In order to collect the data, a web-based questionnaire was created. The questionnaire was developed in PHP and JavaScript, while a PostgreSQL database server was setup to store the user selections. The main page of the questionnaire, where the data collection regarding the image set is done, is shown in Figure 2. Users are presented with a slideshow and they are asked to input how they perceive each image, using five sliders with values ranging from 0 to 10 for each of the five concepts.



The screenshot displays a web-based questionnaire interface. On the left, there is a large image of a brown, circular skin lesion on a light skin background. To the right of the image is a vertical stack of five sliders, each with a numerical value in a small box on the right and three labels below the slider. The sliders are: 1. 'Lesion's Color' with a value of 3.33, labels 'Light', 'Medium', and 'Dark'. 2. 'Color Uniformity' with a value of 3.77, labels 'Uniform', 'Medium', and 'Not_Uniform'. 3. 'Lesion's Asymmetry' with a value of 1.85, labels 'Symmetric', 'Medium', and 'Not_Symmetric'. 4. 'Lesion's Border' with a value of 1.83, labels 'Regular', 'Medium', and 'Irregular'. 5. 'Lesion's Texture' with a value of 1.70, labels 'Smooth', 'Medium', and 'Rough'. Below the image and sliders, there are navigation buttons: 'Previous Image', 'Image 1 / 31', 'Next Image', and 'FINISH !'.

Figure 2. Screenshot of the image description form, where volunteers describe the images through the use of five bars.

The web-based questionnaire consists also of other two parts:

- an introductory page, where the user identification and the IP address check is done. Accepted users are required to provide a nickname (to respect their anonymity) and their medical related qualification (if they are medical doctors or not).
- a final page, which is intended to make an integrity check of the whole data set produced by the user and store it on the database.

Currently, 37 volunteers have submitted their intuitive view of the selected 31 lesions (6 medical doctors and 31 with no medical knowledge). Generally, we noted high intra-class variation in the non-doctors group of volunteers, while slightly lower variation in the group of doctors.

In the Table 1 we report the overall mean and standard deviation of the values gathered by our form for the images shown in Figure 1.

Table 1. Examples of values (mean \pm stdev) of doctor (top lines) and non-doctor (bottom lines) perception for the lesions shown in Figure 1.

	lesion	lesion colour	colour uniformity	lesion asymmetry	lesion border	lesion texture
doctor	(a)	0.59 \pm 0.53	1.45 \pm 1.63	4.13 \pm 3.15	5.78 \pm 3.20	0.78 \pm 0.56
	(b)	8.08 \pm 1.12	6.81 \pm 2.27	6.16 \pm 1.68	3.72 \pm 3.04	7.69 \pm 1.23
	(c)	5.31 \pm 0.48	3.87 \pm 2.34	0.75 \pm 0.63	1.75 \pm 2.09	3.41 \pm 2.05
non-doctor	(a)	0.62 \pm 0.87	2.14 \pm 2.49	7.09 \pm 2.85	6.86 \pm 2.82	1.03 \pm 1.51
	(b)	8.51 \pm 1.18	5.51 \pm 2.05	7.22 \pm 1.97	5.85 \pm 3.06	8.42 \pm 1.14
	(c)	5.25 \pm 1.90	4.51 \pm 1.94	2.34 \pm 1.79	2.72 \pm 1.72	4.01 \pm 2.52

The initial plan was to have two groups of people: doctors and non-doctors and conduct the same experiment on both groups separately. Unfortunately, the number of medical doctors (just 6) was too few to provide reliable results. Moreover, after the analysis of the data of both groups, we can say that the difference of the two groups when it comes to evaluate each lesion is very little. So, it was decided to drop the discrimination between doctors and non-doctors, merging the two groups into a single one. Thus, each image has 37 different evaluations, equal to the number of the volunteers. The standard deviation of these evaluations, averaged over all 31 images, is presented on Table 2. We observe that in most cases, the deviation is more than 2, which, for our domain $\in[0,10]$ is very high.

Table 2. Average standard deviation between users' inputs over all 31 images.

lesion colour	colour uniformity	lesion asymmetry	lesion border	lesion texture
1.419	2.022	2.221	2.341	2.144

2.2 Machine Learning System

In this section we describe the system, based on machine learning algorithms, that imitate the human understanding of lesion images. Figure 3 shows an overview of the system. The system needs to be trained by using low-level image features as input and human description as output.

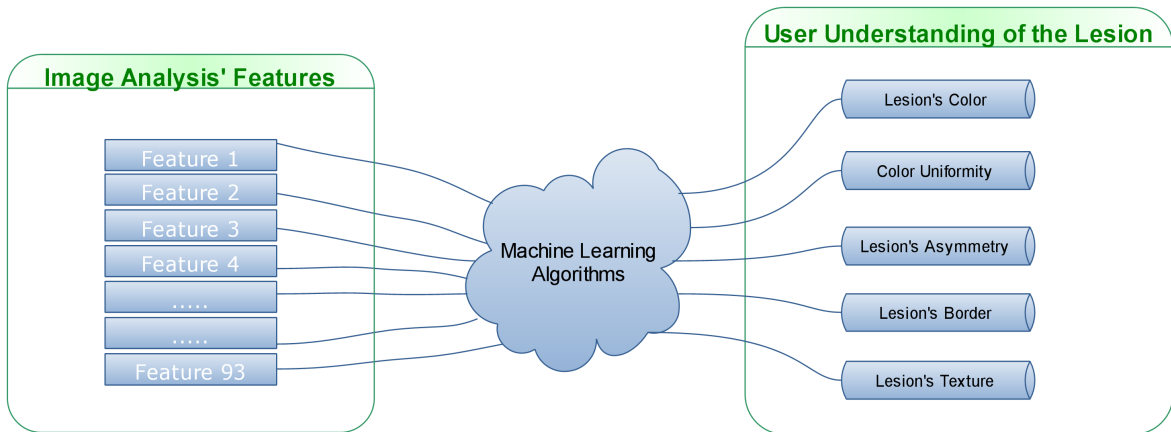


Figure 3. Overview of the machine learning system

2.2.1 Feature Extraction

A simple segmentation algorithm based on the Otsu method¹³ was used to distinguish between background (safe skin) and foreground (lesion). It gave good results in most cases. In the 10% of the cases in which the algorithm fails, a manual segmentation is done.

Ninety-three features were extracted from our images. Extracted features include:

- area, perimeter, compactness of the lesion (3 features)
- central moments of the border of the lesion and of the whole lesion (2 features)
- mean and standard deviation of colour values in RGB (red, green, blue) and HSI (hue, saturation, intensity) colour spaces, calculated for the entire lesion and for 8 subregions (18 features)
- mean and standard deviation of contrast on the border of the lesion (2 features)
- textures based on Fourier transform (4 features)
- textures based on Co-occurrence matrices (36 features)
- textures based on Gabor filters (28 features)

We created 5 different classifiers, each one using a subset of the 93 extracted numerical features as input and one of the human perceived attributes as output. For example, classifier 1 computed the fuzzy description [light, medium, dark] for the lesion colour.

The representative features which were extracted for each lesion are summarised on Table 3. A detailed analysis of the method which was used to extract each one of them is described in another report.¹⁴

The features presented in Table 3 are grouped into the following categories during the system learning process, to train each classifier system:

- Features for “Lesion Colour”: {6-17}
- Features for “Colour Uniformity”: {6-17, 20-25}
- Features for “Lesion Asymmetry”: {1-5, 20-25}
- Features for “Lesion Border”: {3, 4, 5, 18, 19}
- Features for “Lesion Texture”: {26-93}

From the last group, the texture features, we also considered 3 subsets:

- Fourier Features: {26-29}
- Cooccurrence Matrix Features: {30-65}
- Gabor Features: {66-93}

that we used individually to train 3 different classifier systems.

Table 3. The 93 features extracted from the lesions, used as input of the systems.

#	Feature	Additional Info
1	Perimeter of the lesion	
2	Area of the lesion	
3	Compactness Index	
4	Normalised Central Moment of the border	
5	Normalised Central Moment of the whole lesion	
6-8	Mean {red,green,blue} of lesion	
9-11	Std {red,green,blue} of lesion	
12-14	Mean {hue,saturation,intensity} of lesion	
15-17	Std {hue,saturation,intensity} of lesion	
18	Mean contrast on the border	over 8 points
19	Std contrast on the border	over 8 points
20-22	Std of {red,green,blue}	over 8 regions of the lesion
23-25	Std of {hue,saturation,intensity}	over 8 regions of the lesion
26	Entropy of Fourier transform	over the lesion's bounding box
27	Inertia of Fourier transform	over the lesion's bounding box
28	Energy of Fourier transform	over the lesion's bounding box
29	Weighted Distance of Fourier transform	over the lesion's bounding box
30-35	Contrast of Co-occurrence matrix	for 6 offsets {1,2,3,4,5,6}, averaged over 4 directions {0, $\pi/4$, $\pi/2$, $3\pi/4$ }
36-41	Dissimilarity of Co-occurrence matrix	
42-47	Homogeneity of Co-occurrence matrix	
48-53	Energy of Co-occurrence matrix	
54-59	Entropy of Co-occurrence matrix	
60-65	Correlation of Co-occurrence matrix	
66-69	1 st Gray scale Hu invariant of Gabor filtered image	for 4 rotations {0, $\pi/4$, $\pi/2$, $3\pi/4$ } of the Gabor kernel
70-73	2 nd Gray scale Hu invariant of Gabor filtered image	
74-77	3 rd Gray scale Hu invariant of Gabor filtered image	
78-81	4 th Gray scale Hu invariant of Gabor filtered image	
82-85	5 th Gray scale Hu invariant of Gabor filtered image	
86-89	6 th Gray scale Hu invariant of Gabor filtered image	
90-93	7 th Gray scale Hu invariant of Gabor filtered image	

2.2.2 User Data Interpretation

The output of the system has been created from the data collected from the questionnaire as follows.

For each image attribute, we:

- a) create a histogram, discretizing into 3 bins of answers,
- b) create a probabilistic model of whether a specific evaluation is in one class (bin) or another.

For example, assuming we have a set of 37 "lesion colour" evaluations for image (c) in Figure 1:

$$\{7.51, 7.29, 4.41, 1.70, 4.67, 7.26, 2.44, 6.89, 7.52 \dots, 4.81\}.$$

We create 3 bins: $[0 \leftarrow 3.333)$, $[3.333 \leftarrow 6.667)$, $[6.667 \leftarrow 10]$.

Thus the histogram is: $[7, 18, 12]$.

The resulting probabilistic model is: image (c) has a probability of $7/37 = 0.19$ to belong to light, a probability of $18/37 = 0.49$ to belong to medium and a probability of $12/37 = 0.32$ to belong to dark.

Figure 4 sketches our system.

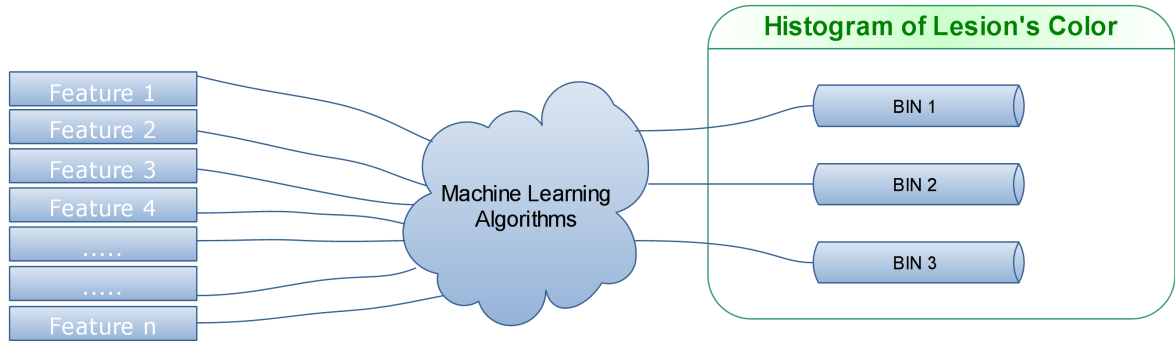


Figure 4. Overview of our proposal to use histogram instead of a single value for each attribute

To evaluate the correctness of each system we propose an evaluation algorithm based on the comparison of the two histograms, not in terms of error (average of the absolute difference on each bin) but in terms of bin ranking.

For example:

1. Let the “Lesion Colour” histogram have 3 bins: {Bin_1 , Bin_2 , Bin_3}.
2. Assume that a specific image has actual probabilities for each bin {0.1 , 0.3 , 0.6}.
3. We order the bins according to their probabilities from the most probable to the least probable: {Bin_3 , Bin_2 , Bin_1}.
4. We order in the same way the bin value estimation of the machine learning algorithm for this image and compare the two ordered lists.
5. If the two lists contain the bins in the same order, then the classification is correct.

So, an output of {0, 0.1, 0.9} which has an average error of 0.2 on each bin is considered correct, while an output of {0.25 , 0.4 , 0.35} which has a lower average error (0.16666) is considered wrong.

However, this might have some flaws: in case that the distribution of the histogram is near-uniform, where all bins have almost the same number (i.e. {0.31, 0.33, 0.36}), it would be a problematic scenario for our method. But this happens rarely, because usually the histogram distribution approximates the Gaussian.

3. RESULTS

In this part, almost all the machine learning algorithms of Weka software¹⁵ have been utilised to train each system by using the corresponding extracted features and the best results achieved in each case are presented. Each figure reports the highest percentages of correct classification achieved for each system, and the feature set used to get that result. The feature numbering refers to Table 3. A good description for each algorithm included in the Weka toolbox can be found in the book of Witten and Frank.¹⁶

In each bar diagram, the first three (as they appear from top) classifiers used are the same (so the reader can compare the effectiveness of the same classification method on different attributes), and in some cases, we report additional ones which might perform slightly better. All the settings used for the classification are the default for each classifier, as set by the authors of Weka, except where it is explicitly mentioned. The leave-one-out cross-validation algorithm is used.

As can be seen in Figure 5, the best results (80.64%) are obtained by system 1 trained for the colour attribute. This system correctly classifies 25 lesions as the most common selected labels. The classifier used in this case was a Multilayer Perceptron. The other 4 systems gave results varying between 52% and 68%, depending on the classifier used and the attribute considered.

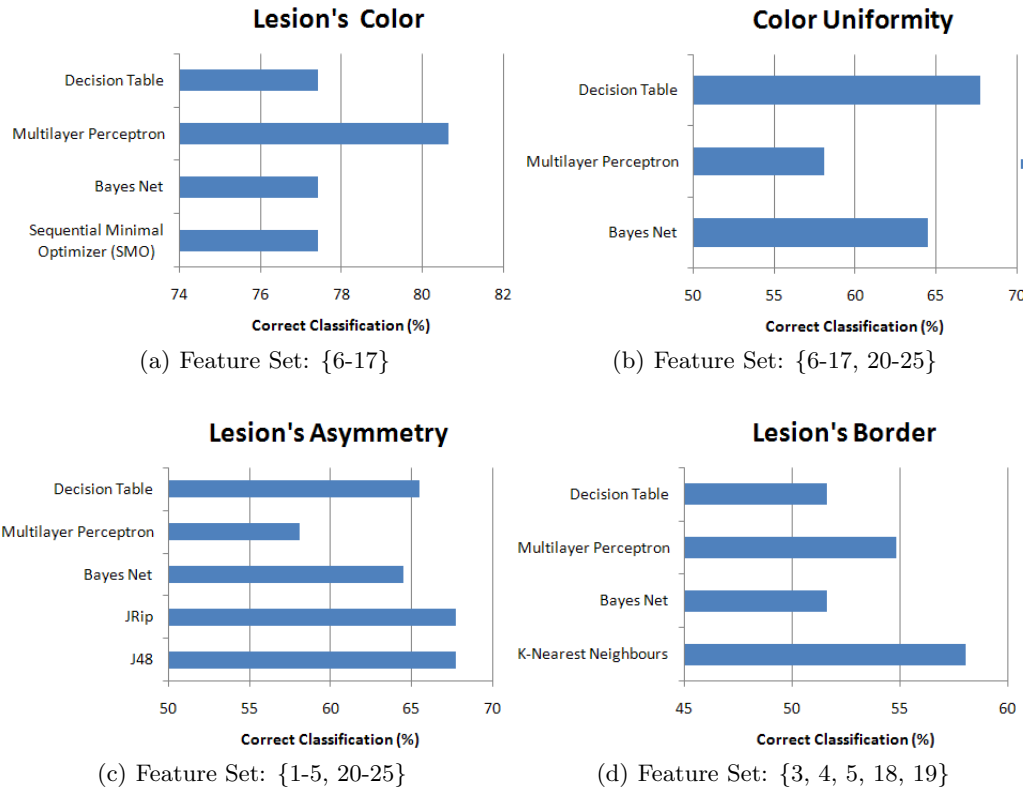


Figure 5. Best classification results (%) for the first four attributes using 3 discrete classes. For each diagram, Feature Set refers to Table 3.

Figure 6 reports results obtained for the Texture attribute using all texture features and the 3 subsets.

The classification results for most systems seem to be mediocre, except for the colour attribute. In fact, the classification accuracy is strongly affected by the variance of the user input, being inversely proportional to the variability (the more the variability is, the less the ideal accuracy gets). As a matter of fact, this is reflected in our results. By observing Figure 5, it can be seen that the “Lesion’s Colour” is in general classified better than the “Colour Uniformity”. Similarly, the “Colour Uniformity” and the “Lesion’s Asymmetry” are better classified than the “Lesion’s Border”. Table 2 contains the average standard deviation of users’ input for each attribute. We see that the attribute with the lowest deviation enables the classification algorithms to reach higher levels of accuracy, while as the deviation increases, the algorithms are restricted to lower accuracy.

4. CONCLUSIONS

We have proposed a model of human perception of skin lesion images that classifies lesions into the same semantic classes as humans. In this system we focus on some attributes (colour, shape, texture). Generalisation to other semantic properties are possible. Future work would be to automatically generate fuzzy rather than discrete descriptions.

We noticed the the labels assigned to each image have a very high deviation, meaning different people describe the same image in different way. This high variation of user description of lesions using attributes made the resulting classification accuracy quite low in some cases. One future project will be develop a tool that will show several images to the user and will allow him to assign each image to the most visually similar one, instead of assigning a conceptual label.

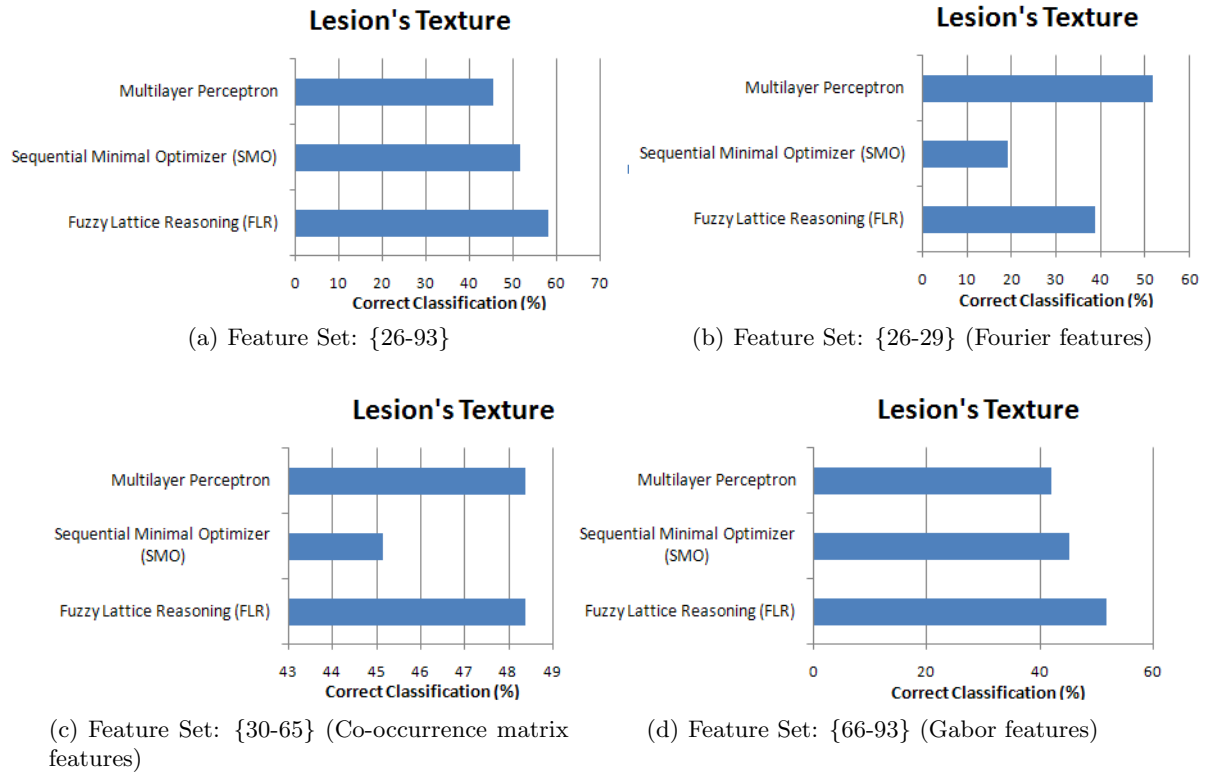


Figure 6. Best classification results (%) for the “Lesion’s Texture” attribute using 3 discrete classes. For each diagram, Feature Set refers to Table 3.

ACKNOWLEDGMENTS

We thank the Wellcome Trust for funding this project.

REFERENCES

- [1] Liu, Y., Zhang, D., Lu, G., and Ma, W.-Y., “A survey of content-based image retrieval with high-level semantics,” *Pattern Recognition* **40**, 262–282 (2007).
- [2] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R., “Content-based image retrieval at the end of the early years,” *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(12), 1349–1380 (2000).
- [3] Carneiro, G., Chan, A. B., Moreno, P. J., and Vasconcelos, N., “Supervised learning of semantic classes for image annotation and retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**(3), 394–410 (2007).
- [4] Vasconcelos, N., “From pixels to semantic spaces: Advances in content-based image retrieval,” *IEEE Computer* **40**, 20–26 (2007).
- [5] Celebi, M. E., Iyatomi, H., Schaefer, G., and Stoecker, W. V., “Lesion border detection in dermoscopy images,” *Computerized Medical Imaging and Graphics* **33**(2), 148–153 (2009).
- [6] Wollina, U., Burroni, M., Torricelli, R., Gilardi, S., Dell’Eva, G., Helm, C., and Bardey, W., “Digital dermoscopy in clinical practise: a three-centre analysis,” *Skin Research and Technology* **13**, 133–142(10) (May 2007).
- [7] Seidenari, S., Pellacani, G., and Pepe, P., “Digital videomicroscopy improves diagnostic accuracy for melanoma,” *Journal of the American Academy of Dermatology* **39**(2), 175–181 (1998).
- [8] Lee, T. K. and Claridge, E., “Predictive power of irregular border shapes for malignant melanomas,” *Skin Research and Technology* **11**(1), 1–8 (2005).

- [9] Schmid-Saugeons, P., Guillod, J., and Thiran, J.-P., “Towards a computer-aided diagnosis system for pigmented skin lesions,” *Computerized Medical Imaging and Graphics* **27**, 65–78 (2003).
- [10] Maglogiannis, I., Pavlopoulos, S., and Koutsouris, D., “An integrated computer supported acquisition, handling, and characterization system for pigmented skin lesions in dermatological images,” *IEEE Transactions on Information Technology in Biomedicine* **9**(1), 86–98 (2005).
- [11] Celebi, M. E., Kingravi, H. A., Uddin, B., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V., and Moss, R. H., “A methodological approach to the classification of dermoscopy images,” *Computerized Medical Imaging and Graphics* **31**(6), 362 – 373 (2007).
- [12] Celebi, M. E. and Aslandogan, Y. A., “Content-based image retrieval incorporating models of human perception,” in [*International Conference on Information Technology: Coding and Computing (ITCC 2004)*], **2**, 241–245, IEEE Computer Society, Los Alamitos, CA, USA (2004).
- [13] Otsu, N., “A threshold selection method from gray-level histograms,” *IEEE Transactions on Systems, Man and Cybernetics* **9**, 62–66 (1979).
- [14] Laskaris, N., *Fuzzy Description of Skin Lesion Images*, Master’s thesis, School of Informatics, University of Edinburgh, UK (2009).
- [15] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., “The WEKA data mining software: An update,” *SIGKDD Explorations* **11**(1), 10–18 (2009). available at: <http://www.cs.waikato.ac.nz/ml/weka/>.
- [16] Witten, I. H. and Frank, E., [*Data Mining: Practical Machine Learning Tools and Techniques*], Morgan Kaufmann, Elsevier (2005). Second Edition.