Polytechnic of Turin
Master course in ICT for Smart Societies

ICT for Health

Lab 2 report

Student: Braccio Jacopo (s273999)

Features' extraction

1. Introduction

1.1 Moles and teledermatology

Moles are a common type of skin growth. They often appear as small, dark brown spots and occur when melanocytes, which give skin its natural color, grow in a cluster instead of being spread throughout the skin. Moles generally appear during childhood and adolescence. Most people have 10 to 40 moles, some of which may change in appearance or fade away over time.

An individual mole is unlikely to become malignant (lifetime risk is about 1 in 3,000 to 10,000); however, patients with large numbers of benign moles (> about 50) have an increased risk of developing melanoma, a type of cancer.

Because moles are extremely common and melanomas are uncommon, so precautionary removal is not justifiable. However, biopsy and histologic evaluation should be considered if moles have certain characteristics of concern (known as the ABCDEs of melanoma):

A. Asymmetry: asymmetric appearance;
B. Borders: irregular borders (i.e. not round or oval);
C. Color: color variation within the mole or unusual color;
D. Diameter: should be smaller than 6mm;
E. Evolution: new mole in a patient over 30 years or a changing mole.

Periodical check of moles should be done in order to prevent the transformation in melanoma.

In order to make life easier for people, teledermatology could be of great help.

Teledermatology is a subspecialty of dermatology and among the most popular application of e-health and telemedicine. Teledermatology aims to replace the personal contact between patients and medical doctors with a digital exchange of medical information (such as audios, images or videos). In this way, a patient can consult a dermatologist and seek advice for a skin condition even from his home, which is really useful for people with disabilities and elderlies. Teledermatology is based on two main concepts: *store – and – forward*, the patient transmits a digital image of the skin to his doctor in order to make a diagnosis, and *real time teledermatology*, which consists of a live video communication between patients and medical doctors. Thanks to this, checking moles reduces only on taking pictures of them which will be analyzed.

1.2 Goal of the activity lab

The goal of this lab is analyzing a dataset consisting of images of moles and extract two of the five features used for classifying moles: border and asymmetry.

2. Data preparation and analysis

2.1 Dataset description

The dataset used in this laboratory is made of 54 images (in the 'jpg' format) of moles, already classified in three categories according to the probability of being a melanoma.

The dataset's division in three classes is as follows:

1. 11 images of moles having low probability of being tumors, called "*low_risk_n*" ('n' being the number of the picture);
2. 16 images of moles having medium probability of being tumors, called "*medium_risk_n*";
3. 27 images of moles having high probability of being tumors, called *"high_risk_n".*

2.2 Data preparation

The given images are RGB images, i.e. an additive color model in which red, green and blue are added together in various ways to reproduce a broad array of colors, hence each of them has been read as a 3-dimensional array with shape of 583x583x3 whose elements are unsigned integers with 8 bits.

Then, color quantization is performed. Color quantization is a process that reduces the number of distinct colors used in an image, usually with the intention that the new image should be as visually similar as possible to the original image. This allows the proposed algorithms to work with an image that has just 3 colors instead of the 2^{24} given by the RGB model. The darkest color will correspond to the mole in examination.

In order to quantize the image, the previous 3-dimensional array has been transformed into a 2-dimensional array representing the list of all pixel's colors present in the original image. *K-means* algorithm has been used to cluster together the colors and each point of the 2D array is assigned to the one out of the three clusters according to its color level.

Some complex images in the dataset do not work properly with 3 clusters and many details are lost. Therefore, two different lists were created: *'cluster_3'* containing the name of the moles that work properly with just 3 clusters and *'cluster_5'* containing the name of the moles that work properly with 5 clusters. Those lists are stored in a file called `moles_list.py`.

Taking as example the file '*low_risk_2.jpg'* (Figure 1), the result of the quantization process is shown in Figure 2. The previous file will be used as example for the rest of the report.

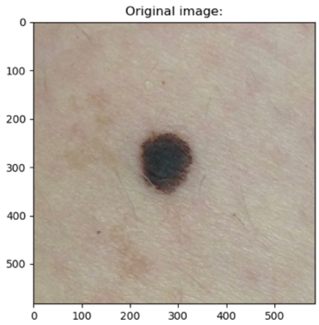


Figure 1. Original 'low_risk_2'.

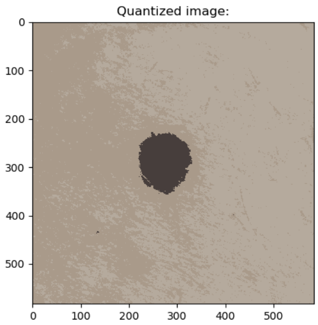


Figure 2. Quantized 'low_risk_2', 3 clusters.

3. The contour

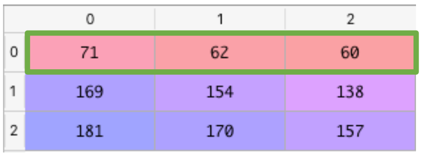
This section is focused on the research of the contour of the mole so its perimeter can be evaluated. By confronting the found perimeter with the ideal perimeter of a circle having the same area as the mole, a measure of its indentation is given: if the ratio between the measured perimeter and the ideal perimeter is large, it means the mole is very indented, otherwise is almost circular.

The following subtask are performed to meet this quest.

3.1 Darkest color

The darkest color is the color representing the mole. In RGB notation, [0,0,0] corresponds to black and [255,255,255] corresponds to white, hence the darkest cluster is the one whose elements are the smallest among all the centroids.

The values of the 3 centroids of the image 'low_risk_2' are shown in Fig. 3. They have been sorted in ascendant order, meaning that the first row corresponds to the darkest cluster.



	0	1	2
0	71	62	60
1	169	154	138
2	181	170	157

Figure 3 Centroids of 'low_risk_2'; first row is the darkest colour.

For images belonging to list '*cluster_5*' the palette representing the centroids has 5 rows, the first two (the rows with lowest values) are chosen and considered together as the darkest colours.

3.2 Binarization of the image

In order to work with even easier images, the images are transformed into binary images by replacing pixels equal to the darkest colour (or colours, case number of clusters = 5) with black, and all the others with white, as shown in Fig.4.

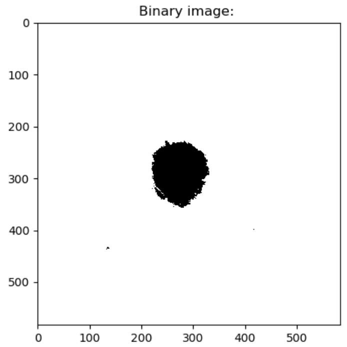


Figure 4. Binary representation of 'low_risk_2'.

3.3 Cropping

File `crop_image.py` contains a function used to find the rectangular region surrounding the mole. The algorithm aims at finding the four outer points of the mole:

- *top_left*: the leftmost pixel of the mole;
- *top_right:* the rightmost pixel of the mole;
- *top_up:* the northernmost pixel of the mole;
- *top_down:* the most south pixel of the mole

and then cropping the mole according to the straight lines that pass through these points and perpendicular to the main axes.

The starting point of the algorithm is the center of the image C_{im} whose coordinates are:

$$C_{im} = \left(\frac{width}{2}, \frac{height}{2}\right).$$

The algorithm requires that the center of the image corresponds to a pixel of the mole, so the pictures in the dataset that did not satisfy this request where manually fixed.

The pseudocode of how the search for the northernmost pixel of the mole works is given below:

```
1. Start from the center of the image $C_{im} = (C_x, C_y)$;
2. Set top_down = top_up = $C_y$ and top_left = top_right = $C_x$;
3. Check if the color of the pixel [$C_x$, top_right] is black; if the check
   is positive set top_right = top_right + 1 (the position moves to the
   right); continue until the condition is false;
4. Check if the color of the pixel [$C_x$, top_left] is black; if the check
   is positive set top_left = top_left - 1 (the position moves to the
   left); continue until the condition is false;
5. Check if the color of the pixel [top_up, $C_y$] is black; if the check is
   positive set top_up = top_up - 1 (the position moves to the north);
   continue until the condition is false;
6. Check if the color of the pixel [top_down, $C_y$] is black; if the check
   is positive set top_up = top_down + 1 (the position moves to the
   north); continue until the condition is false;
```

7. Loop through the columns between *top_left* and *top_right* and through the row between top_up and 0 (inverse loop) to check the presence of a black pixel in one of the upper rows. If so, top_up will be updated to that value.

The other 3 points can be found by changing the index values to this algorithm. The results of the crop is shown in Figure 5.

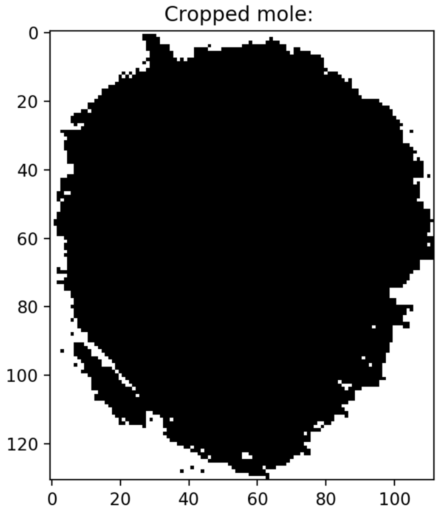


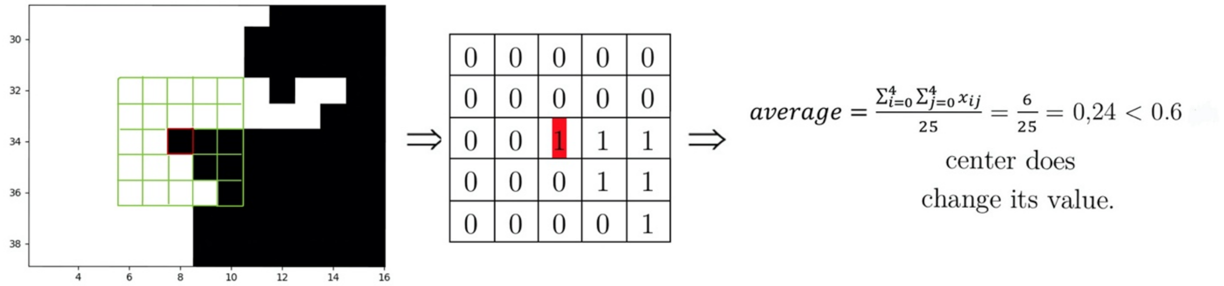
Figure 5. *Cropped image 'low_risk_2'.*

3.4 Smoothing

Quantization introduces noise inside the pictures: many moles shows empty holes inside or isolated pixels around the border that may interfere with the calculus of the perimeter. In order to have a more accurate measurement, the images have been cleaned and smoothed. This operation is done by checking the value of each pixel and confronting it with its neighbors.

A 2D array, having same dimensions as the image is created. The value of each cells is 0 if the corresponding pixel in the image is white, otherwise 1. Then a 5x5 kernel, whose center is the pixel in examination, is used. Starting from the first pixel, the kernel is used to understand the average color value of each area surrounding it i.e. its 24 neighbors. The average is then compared with a threshold (set by trial and error) in order to check if the pixel color has to be changed or not: if the average is bigger than 0.6, the pixel is set equal to black, otherwise white. Once the check is completed for one pixel, kernel's center is shifted by 1 (stride = 1) and the operation starts again until the last pixel of the image is reached.

An example of the process is illustrated in Figure 5 representing a small portion of *'low_risk_2'*.



0	0	0	0	0
0	0	0	0	0
0	0	1	1	1
0	0	0	1	1
0	0	0	0	1

$$average = \frac{\sum_{i=0}^{4}\sum_{j=0}^{4} x_{ij}}{25} = \frac{6}{25} = 0{,}24 < 0.6$$

center does
change its value.

Figure 6. *Smoothing of the image: the zero-one kernel is averaged and compared with the threshold of 0.6. Being the result smaller than the threshold, the center will be changed into white.*

Figure 6 and Figure 7 show *'low_risk_2'* before and after the smoothing process. Depending on the image in examination, the clean-up of the mole removes from 5% to 25% of pixels allowing a more precise evaluation of the perimeter.

The total number of black pixels represents the area of the mole.

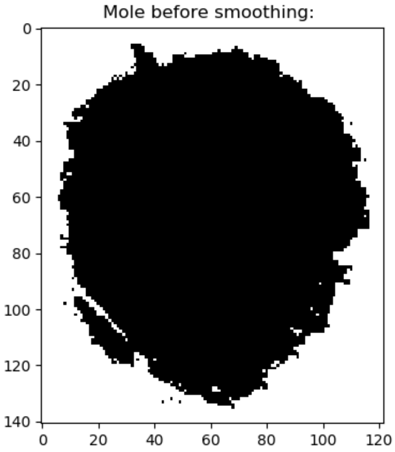


Figure 7. *'low_risk_2' before smoothing*

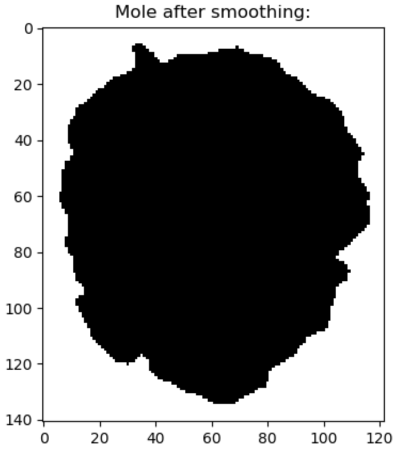


Figure 7. *'low_risk_2' after smoothing.*

3.5 Border

Since the image is binary, the border was found by looking at the possible variation of color between two subsequent pixels. Along the columns, if a pixel is white and the next one is black (and vice versa), it will belong to the border of the mole. Along the rows, if a pixel is white and the one below is black (and vice versa), it will belong to the border of the mole. The border is highlighted in blue, and the number of pixels making it up represents the perimeter of the mole.

The algorithm developed is the following:

```
1. Set perimeter = 0, i = 0, j = 0;
2. Check if the color of the pixel [i,j] is the same of the pixel
   [i,j+1] or the same of pixel [i+1,j]. if the check is true, set
   perimeter = perimeter +1;
3. Set j = j+1; go back to step 2 until the width of the image is
   reached;
```

```
4. Set i = i+1; go back to step 2 until the height of the image is
   reached.
```

Figure 8 shows the border of *'low_risk_2'*.

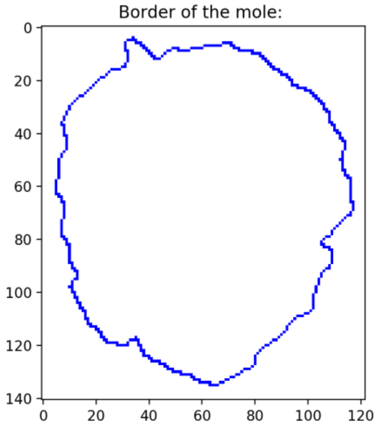


Figure 8. *Border of 'low_image_2'.*

3.6 Calculating indentation

Having the area of the mole, it is possible to find the ideal perimeter, i.e. the perimeter of a circle having same area as the mole:

$$A_{mole} = \pi\, r_{ideal}^2 \;\rightarrow\; p_{ideal} = 2\pi r_{ideal} = 2\pi \sqrt{\frac{A_{mole}}{\pi}}$$

It is now possible to evaluate the ratio between the perimeter of the mole and the perimeter of the equivalent circle:

$$ratio = \frac{p_{mole}}{p_{ideal}}$$

4. Asymmetry

Focus of this section is to find a way of expressing the concept of symmetry as a number, Asymmetry in a mole can be sign of its malevolent nature.

4.1 Centre the mole

Symmetry is the quality of being made up of exactly similar parts facing each other or around an axis. The vertical line cutting the image in half is chosen as axis of symmetry. Therefore, mole's center must coincide with the center of the image.

The same function used in Section 3.3 is used to find the extreme pixels of the filtered image and the center of the image is evaluated as:

$$C_{img} = \left(\frac{top_{left} + top_{right}}{2}, \frac{top_{up} + top_{down}}{2}\right)$$

where:

- top_{left} represents the first column with a black pixel (first abscissa);
- top_{right} represents the last column with a black pixel (second abscissa);
- top_{up} represents the first row with a black pixel (first ordinate);
- top_{down} represents the last row with a black pixel (second ordinate);

The axis of symmetry will be the vertical line going through the center C_{img}.
Figure 7 represents the found axis for *low_risk_2*.

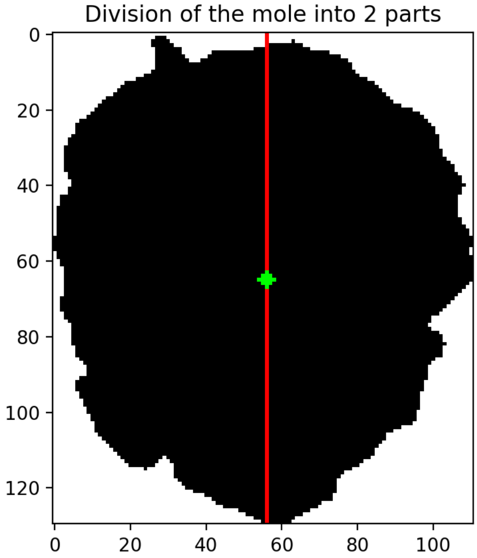


Figure 8. *Vertical line cutting 'low_risk_2' in half. It will be used as axis of symmetry.*

4.2 Measuring asymmetry

In order to give a numeric value to symmetry, the percentage of white pixel of the two halves of the image is calculated as the ratio between the number of white pixels in each of the two halves and the total number of pixels in the images divided by 2.
The asymmetry is then evaluated as follow:

$$asymmetry = \begin{cases} \dfrac{white_{left}}{white_{right}}, if\ white_{left} < white_{right} \\ 1, if\ white_{left} = white_{right} \\ \dfrac{white_{right}}{white_{left}}, if\ white_{right} < white_{left} \end{cases}$$

where: $white_{left}$ is the percentage of withe in the left half and $white_{right}$ is the percentage of white in the right half. Even though this method is working, it is just an estimate of the real symmetry. For a more precise measurement, the axis of symmetry to be used should be the main axis of the mole, found using Principal Component Analysis.

5. Conclusions

Table 1 shows the values of the founded features, ratios between ideal perimeter and measured perimeter and symmetry, for all the moles in the dataset.

***Table 1**. Results of border indentation (ratio) and symmetry for each mole in the dataset.*

Low_risk_n			***medium_risk_n***			***melanoma_n***		
n	**ratio**	**asymmetry**	**n**	**ratio**	**asymmetry**	**n**	**ratio**	**asymmetry**
1	1.3594	0.80	1	1.1313	0.9794	1	1.2815	0.8498
2	1.2395	0.821	2	1.2376	0.7836	2	1.4655	0.6424
3	1.28	0.9567	3	1.238	0.8983	3	1.4911	0.913
4	1.5605	0.9683	4	1.0779	0.707	4	1.2765	0.7371
5	1.1734	0.8027	5	1.5836	0.7538	5	1.5345	0.5238
6	1.2981	0.933	6	2.1871	0.8322	6	1.6155	0.7729
7	1.5581	0.837	7	1.4306	0.8821	7	1.3931	0.9716
8	1.2622	0.8431	8	1.3381	0.8149	8	2.168	0.9
9	1.0713	0.9083	9	1.6658	0.8012	9	1.5133	0.907
10	1.1104	0.8476	10	1.5076	0.8436	10	1.7876	0.8066
11	1.1637	0.86	11	1.4722	0.8716	11	1.5421	0.8732
			12	1.2474	0.8384	12	1.383	0.6546
			13	1.4482	0.6367	13	1.2887	0.4708
			14	1.3021	0.8734	14	1.506	0.3381
			15	1.2979	0.973	15	2.0304	0.758
			16	1.312	0.9695	16	1.5832	0.5415
						17	2.7616	0.8499
						18	1.252	0.852
						19	1.5799	0.9424
						20	1.4273	0.8605
						21	1.9822	0.918
						22	1.4273	0.8605
						23	2.4839	0.775
						24	1.461	0.5501
						25	1.278	0.5858
						26	1.9314	0.7993
						27	1.7579	0.5612

Table 2 Mean value and standard deviation for each class:

Table 2. Statistical values for each class

Class	*Features*	*Mean Value*	*standard deviation*
low risk	ratio	1,253	0,127
	asymmetry	0,878	0,081
medium risk	ratio	1,347	0,356
	asymmetry	0,824	0,115
melanoma	ratio	1,623	0,355
	asymmetry	0,749	0,180

Both *low* and *medium risk* classes have similar mean values for both of the two features. Class *'melanoma'*, on the other hand, has the greatest mean value of border indentation, and the lowest mean value of symmetry, reflecting melanoma's trend of being irregular.

However, in the overall there is not a huge difference in the found parameters, so it is impossible to use them as classification criteria.

In order to prove the inefficiency of a classification based on these two features, a test classifier was created and stored in `classifier.py`.

A training set, made of all the images in the dataset with their features and class label (0 = low risk, 1 = medium_risk and 2 =melanoma), is stored in a .csv file. As test vector '*low_risk_10'* is chosen. Since the number of features of each mole is two, they can be plotted into a 2dimensional space having as abscissa the asymmetry parameter and as ordinate the ratio between the ideal perimeter and the measured one, as in Fig. 8.

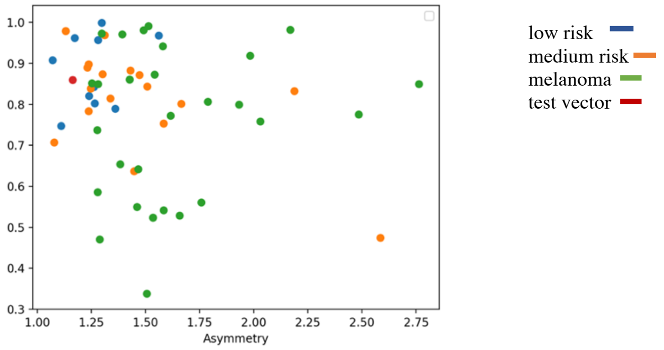


Figure 9. *Visual representation of the dataset. Due to the values being very close, not every single mole is shown.*

The classifier performs K-NN (k-nearest neighbors, K = 5) on the test vector and print the probability of it belonging to one of the three classes. The resulting probabilities for the test vector are: 20% of being low risk, 80% of being medium risk and 0% of being melanoma, in contrast with the real class that is 'low risk'.

Therefore, considering just these two features **is not** sufficient to diagnose melanoma. The results obtained in this laboratory should be part of a more complex study, including also the other 3 features of the ABCDE rule and a medical response on the measured values.