# ICT for Health report 2
# Academic Year 2019/20

Lorenzo Cascioli - s275924

December 31, 2019

# 1 Introduction

The work carried out and shown in this report consisted in performing feature extraction from a dataset of images. More precisely, the objects of interest were photos of moles, and the goal would be to analyze several features of a mole in order to help medical doctors in trying to quickly identify a potentially dangerous mole by means of a simple photo.

In a general case, five features are considered by a doctor trying to identify a melanoma (skin tumor); these are commonly expressed with the so-called "ABCDE rule", as they are Asymmetry, Border, Color, Diameter and Evolution. Obviously, a complete study would require all these aspects to be taken into account; however, in this work only the first two features have been considered, for the sake of brevity.

Therefore, the procedure was articulated in three main parts:

- identification of the mole from the image
- analysis of the border indentation
- analysis of the asymmetry

The searched results, of course, were not expected to perfectly portrait clear, universal differences between a melanoma and a normal mole, that would be very difficult even considering all the five features; the intention was instead to perform an accurate extraction and statistical analysis of the two features. This could be certainly useful in a broader study on moles (comprehensive of all the features) with the aim of melanoma identification, as mentioned before.

# 2 Description

## 2.1 Dataset

Figure 1: Example image of a medium risk mole

The set of images at disposal for this study was composed of jpeg photos of moles classified into three different categories: low risk, medium risk and melanoma. More specifically, 11 low risk moles, 16 medium risk moles and 27 melanomas were given. An example of such photos is given in fig. 1.

For each of these samples, values for Asymmetry and Border were to be determined. What follows, therefore, is a detailed exposition of the adopted procedures to identify the mole and extract some meaningful measures for the two features of interest. Then the results are going to be adequately studied.

## 2.2 Clustering and mole identification

### 2.2.1 Basic 3-cluster procedure

The first task to accomplish was to read the image and identify the pixels corresponding to the mole points, on which the further operations about Border and Asymmetry would be executed. Having read the image, clustering was performed on its pixels using the K-Means algorithm. At first, K-Means was applied with three clusters: each pixel of the image was assigned to one of them and was given the color of its corresponding centroid, generating a 3-color picture like the one presented in fig. 2.

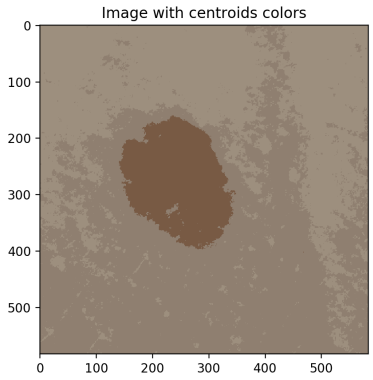


Figure 2: 3-color image of a mole

The idea would be to take all the points of the darker cluster as the mole. However, an operation of "denoising" had to be introduced at this point, because the output darkest cluster of K-Means is likely to have several points included that are not actually part of the true mole, either because there are some other dark parts on the skin of the photo or because of bad light or other noises. Hence, a procedure to try to remove such points as accurately as possible has been implemented.

Having kept only the points of the darkest color, setting the others to white, a three-phase removal took place:

1. **frames removal**: given that in a normal photo the mole is placed at the center of the image, the 10% pixels at the row and column margins were automatically set to white;

2. **whitening**: at this point, the idea was to remove the points which are furthest from the median, which could be reasonably assumed as the center of the mole. Thus, the points' coordinates of the two dimensions were normalized with respect to their median and standard deviation and treated as a Gaussian distribution from which

the points at the tails were set to white. At this stage, the choice was to iterate keeping each time the 95% of the points and "cutting away" the others. The stopping condition was met when the initial variance had been reduced by 33% (a considerable value) or when from one iteration to the other the changes in variance were very limited. If the first iteration instantaneously met the stopping condition, the initial mole was given as output of this phase, as no need of outliers removal was detected (the case in which in a single iteration the variance is reduced by 33% is highly unlikely).

3. **blurring**: because some isolated points might escape being cut during the previous operations, a final phase has been introduced with the precise aim of eliminating single sparse colored pixels. This was called "blurring" as the idea was to examine groups of the first 5 neighbors of a pixel in each direction (going from (0,0) to the end examining all the pixels) and to set the pixel to white if the number of colored neighbors was below a given threshold, which was here set to 3.

The result of these three operations should be the mole quite precisely identified and without strange outliers. An example of the mole before and after this "denoising" is presented in fig. 3.

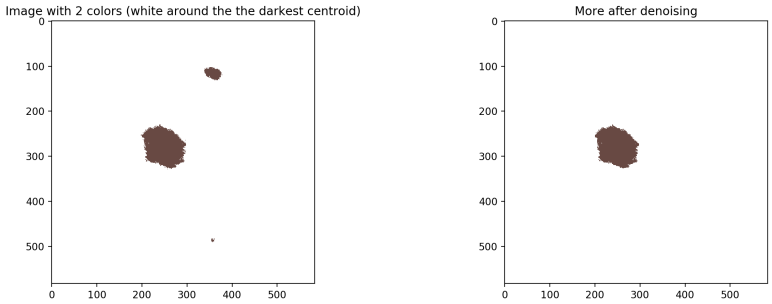


Figure 3: A low risk mole before and after denoising

In this case, a second, little mole had to be removed because it was not the one to analyze. Furthermore, some other points at the bottom of the image were inserted by K-Means in the darkest cluster. By means of the described passages, it is shown how at the end only the true mole was left for the following operations.

To check if this first part went smoothly as planned, a rectangle was built taking as values for its edges the maximum and minimum x and y coordinates between all the points of the final mole. The rectangle was drawn on the initial photo and shown to the user as in fig. 4. Moreover, the median (center) of the mole was highlighted.

At this point, the user was asked whether the rectangle fitted well the mole or not. In case of a positive answer, the analysis of Border and Asymmetry could be done. However, for some moles performing this procedure with only three clusters might not be enough. Therefore, in case of a negative answer, a more complex algorithm was adopted to correctly identify the mole.

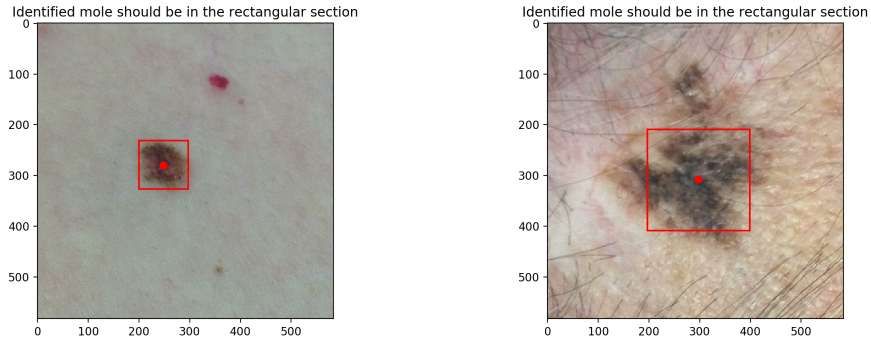


Figure 4: Examples of correct and wrong rectangle obtained with the 3-clusters procedure

### 2.2.2 Procedure with more clusters for difficult moles

The obvious solution to be more precise was to increase the number of clusters when performing K-Means. Hence, for the cases in which the basic procedure was not successful, K-Means was re-applied with 5 clusters. A new problem emerged, though, as for some moles just the darkest cluster was needed for a nice description, while for others two out of five were needed. In the first case, the second darkest cluster had not to be taken into account, or it would have brought a considerable amount of noise. To solve the issue, an output was presented to the user: the initial image was displayed together with another image where the darkest cluster was colored in black, and the second darkest cluster was colored in grey. The user had three possible options: keep only the first cluster, keep both the clusters or repeat the procedure because even the first cluster was not precise enough. An example of this output is presented in fig. 5: in the first case, two clusters were necessary to identify the mole; in the second case, only the darkest one had to be taken into account.

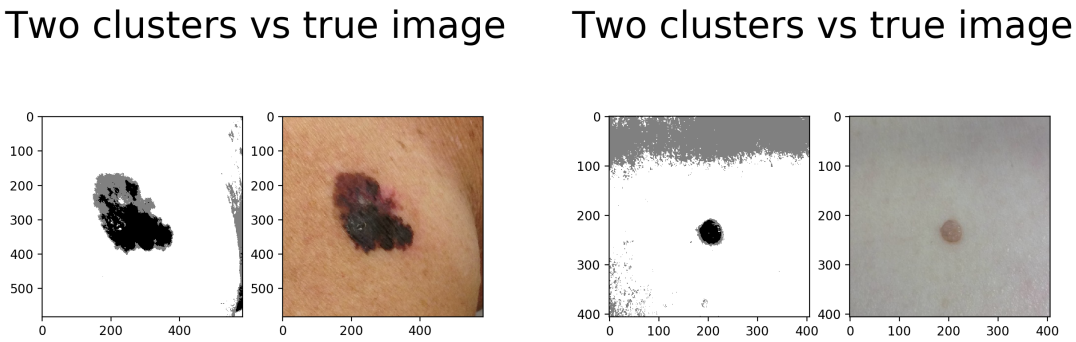


Figure 5: Opposite examples for the selection between one or two clusters

The third option involved applying a third time K-Means with 7 clusters, and having another user choice analogous to the presented one to decide whether to use one or two clusters for the mole identification. This was actually almost never used, and it was needed only for bad photos where the mole was almost of the same color of the rest; in

the used dataset, only one mole needed to be treated this specifically.

In any case, having taken the points of one or two clusters as the mole, a similar sequence of operations to those of the basic procedure was implemented. Frames removal and blurring were initially applied, while the so-called "whitening" has been applied at the end and slightly changed in order to be more precise. Firstly, the iterations removed less than the 5% of the points at a time; then, instead of inserting a stopping condition, after each iteration the rectangle was displayed on the mole and the user was asked whether the result was good or not: doing so, the user defines when a positive result has been reached in way that could be a little longer than before but of certain consistency. Fig. 6 shows a simple case where the basic solution did not work solved with this more complex clustering.

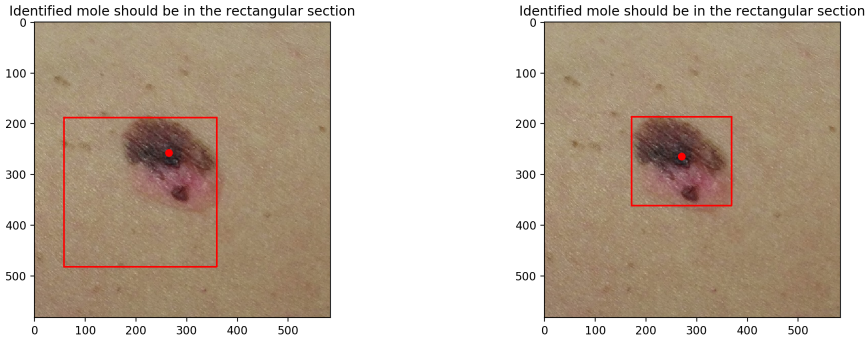


Figure 6: Identification of a difficult mole with 3 and 5 clusters

At this point, every mole in the dataset could be properly identified, thus it was possible to begin working on the mole points to extract information and measurements about the Border and the Asymmetry.

## 2.3 Border

The main idea for the analysis of the Border was that a melanoma would normally have a highly indented, irregular perimeter. Therefore, the objective of this part was to calculate the perimeter of the mole and extract a ratio between this value and the perimeter of a perfect circle having the same area of the mole: the higher this value, the more indented (and thus dangerous) the mole border was supposed to be. In order to achieve so, three main passages might be identified and later explained:

1. "fill-the-holes" procedure
2. perimeter evaluation
3. perimeters ratio calculation

### 2.3.1 "Fill-the-holes" procedure

Since it was very likely that the clustering and mole identification had left some little white areas inside the mole, these spaces had to be filled with the color of the mole in

order to significantly simplify further operations.

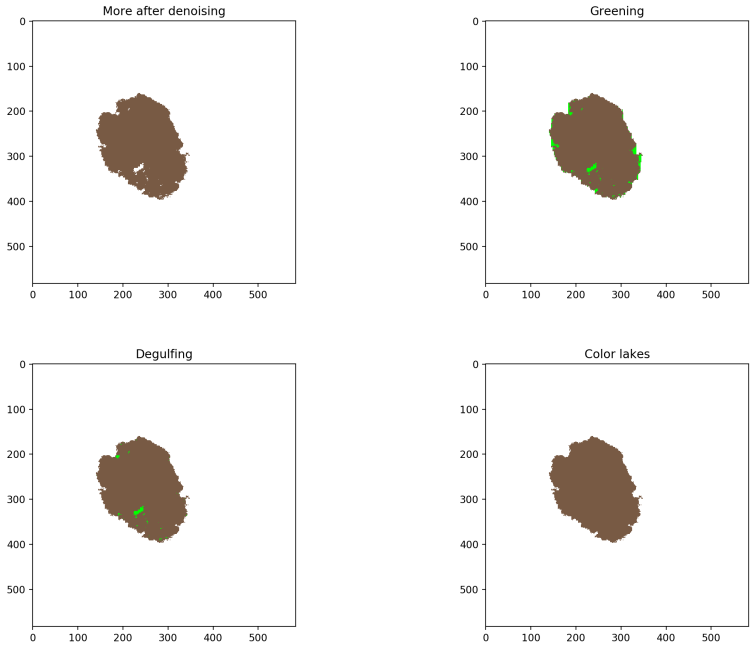


Figure 7: Subsequent greening, degulfing and coloring to eliminate white holes

To do so, a three-point algorithm has been implemented:

1. **greening**: for each row where at least one colored point was present, the colored points at the maximum right and left were found, and all the white points between them were set to green. A deeper look at this situation could be useful at this point; green points might be of two types:
    - points part of a "lake" inside the mole, which should then be colored
    - points in a "gulf", thus in contact with the external white area, that were not to be colored, because they were not part of the mole
2. **degulf**: this passage's aim was to properly eliminate from the green pixels those that were not to be colored because part of "gulfs" and not "lakes". This was achieved with an iterative procedure: each time, the four neighbors of each green pixel were checked; if at least one was white, it meant that the position of the pixel was outside the mole (otherwise it could have only dark or green neighbors). Such identified pixels were set to white again. The procedure was repeated until for one iteration no new green pixels were set to white;
3. **coloring**: only green points that were inside the mole were now left, so they could easily be set of the color of rest of the mole (darkest centroid color).

A working example is presented in fig. 7.

### 2.3.2 Perimeter evaluation and analysis

Having obtained a well-defined mole, it was quite straightforward to pick the points of the perimeter: it was sufficient to take only the points that among their four neighbors had a white pixel. The perimeter was then easily computed and highlighted in red as in fig. 8.

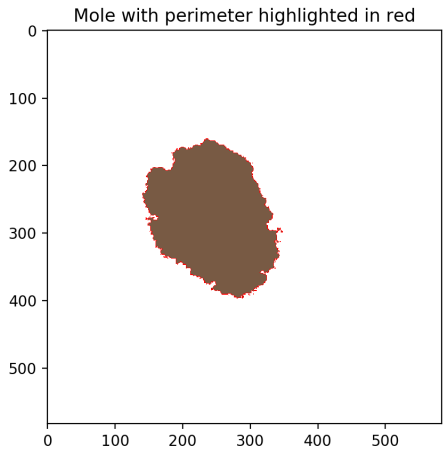


Figure 8: Perimeter identification

As the goal was to evaluate the mentioned ratio, two quick calculations had to be done; having counted the total number of colored pixels, the area of the mole was retrieved. Treating it as the area of a circle, what was next was the evaluation of the circumference of such circle. Finally, the searched ratio could be computed dividing the mole perimeter by the circumference value. The results are going to be displayed later on.

## 2.4 Asymmetry

The second feature to be analyzed was the Asymmetry. There are of course several different possible ways to assign a value to the Asymmetry of an object; what follows here is one of the possible ideas adoptable in this case.
The general workflow here was articulated once again in three main steps:

1. center and rotate the mole along its principal direction
2. compute the median of each row and see how far it is from the zero
3. compute the average of this values to get a numerical measure of the Asymmetry

### 2.4.1 Mole rotation

After having centered the x and y coordinates of all the mole points by subtracting the respective means, the idea was to rotate the mole setting its principal axis in vertical position. To do so, the procedure was the following:

- the two arrays containing the x and y coordinates of the mole points were put together in a matrix, whose covariance matrix was evaluated;

- the eigenvalues and eigenvectors of such matrix were calculated to obtain the rotation angle $\theta$ from the x and y values of the eigenvector corresponding to the largest eigenvalue (dividing the x by the y);
- having calculated $\theta$, a rotation matrix was built and a product with the coordinates matrix was performed, obtaining the new coordinates of the rotated matrix.

### 2.4.2 Asymmetry evaluation

From the obtained vertically-aligned mole, a measure of the Asymmetry could be implemented. For each row, the median of the x coordinates of its points was found and plotted; this globally resulted in a graphical visualization of a pseudo-axis of symmetry: the more close to zero its points were, the more symmetrical the mole was. To get a numerical value, the absolute values of each median were added to a summation: being the mole centered, a normal median for a perfectly symmetric row would have x coordinate zero, thus the x coordinates of the medians could be seen as their distances from the optimal situation. Dividing the summation by the total number of rows of the mole, an average value for a row's median was calculated and taken as symmetry indicator: the closer to zero, the more symmetrical the mole was.

The operation was repeated with respect to the horizontal axis and a mean value of the two found indicators was taken as global measure of the Asymmetry. An example of this operations is presented in fig. 9, with the principal axis in red and the orthogonal one in blue.

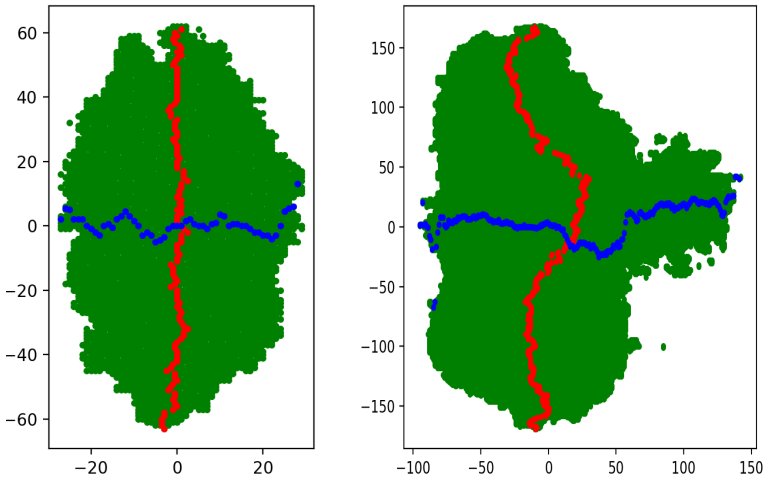


Figure 9: Moles rotated along principal direction with symmetry analysis

# 3 Conclusions

## 3.1 Results

Having performed the planned procedures to extract a value for the Border and one for the Asymmetry for each mole, the obtained results are given in table 1, while a brief statistical analysis for the three classes of moles is presented in table 2.

| Mole | Border | Asymmetry | Mole | Border | Asymmetry |
|---|---|---|---|---|---|
| low risk 1 | 1.346 | 2.08 | melanoma 1 | 1.266 | 2.53 |
| low risk 2 | 1.356 | 2.00 | melanoma 2 | 1.675 | 6.09 |
| low risk 3 | 1.365 | 1.46 | melanoma 3 | 1.938 | 3.58 |
| low risk 4 | 1.380 | 2.27 | melanoma 4 | 2.858 | 6.62 |
| low risk 5 | 1.331 | 3.02 | melanoma 5 | 1.854 | 4.65 |
| low risk 6 | 1.315 | 2.02 | melanoma 6 | 2.119 | 13.50 |
| low risk 7 | 1.756 | 3.50 | melanoma 7 | 1.641 | 7.37 |
| low risk 8 | 1.408 | 2.63 | melanoma 8 | 2.247 | 3.80 |
| low risk 9 | 1.054 | 0.79 | melanoma 9 | 1.508 | 7.08 |
| low risk 10 | 1.176 | 0.66 | melanoma 10 | 2.002 | 8.60 |
| low risk 11 | 1.085 | 1.52 | melanoma 11 | 1.672 | 3.61 |
| medium risk 1 | 1.127 | 0.94 | melanoma 12 | 1.442 | 5.32 |
| medium risk 2 | 1.193 | 2.86 | melanoma 13 | 1.369 | 4.81 |
| medium risk 3 | 1.071 | 0.91 | melanoma 14 | 1.450 | 5.44 |
| medium risk 4 | 1.103 | 2.82 | melanoma 15 | 1.667 | 10.54 |
| medium risk 5 | 1.664 | 4.46 | melanoma 16 | 1.961 | 7.44 |
| medium risk 6 | 1.460 | 5.24 | melanoma 17 | 2.483 | 7.44 |
| medium risk 7 | 1.487 | 5.88 | melanoma 18 | 1.076 | 6.78 |
| medium risk 8 | 1.337 | 2.92 | melanoma 19 | 1.679 | 8.85 |
| medium risk 9 | 2.081 | 1.34 | melanoma 20 | 1.464 | 9.54 |
| medium risk 10 | 1.909 | 3.61 | melanoma 21 | 2.161 | 4.59 |
| medium risk 11 | 1.828 | 8.26 | melanoma 22 | 1.466 | 4.28 |
| medium risk 12 | 1.155 | 3.58 | melanoma 23 | 4.339 | 22.18 |
| medium risk 13 | 1.456 | 7.71 | melanoma 24 | 1.618 | 12.26 |
| medium risk 14 | 1.226 | 3.68 | melanoma 25 | 1.288 | 3.75 |
| medium risk 15 | 1.231 | 2.70 | melanoma 26 | 1.595 | 8.57 |
| medium risk 16 | 1.372 | 2.08 | melanoma 27 | 3.978 | 10.52 |

Table 1: Evaluated values for the two features for each of the considered moles

## 3.2 Final considerations

Some comments on the obtained results may be expressed in order to discuss the outcome of this work. Fist of all, it must be pointed out that while the Border indentation value is

| Mole class | Border | | Asymmetry | |
|---|---|---|---|---|
| | Mean | StDev | Mean | StDev |
| low risk | 1.325 | 0.179 | 1.995 | 0.828 |
| medium risk | 1.419 | 0.298 | 3.687 | 2.111 |
| melanoma | 1.919 | 0.743 | 7.398 | 3.992 |

Table 2: Statistical analysis of the features for the three classes of moles

quite objective, several ways could have been used to quantify the Asymmetry of a mole. The adopted one uses an average of the axial symmetry with respect to the principal direction of the mole and the orthogonal one, but a central symmetry or many other types of axial measurements could have been adopted equally validly. The choice made, however, appeared to give quite consistent results. While the precision of the measurements is certainly good (the same result is obtained at every try on one mole), the accuracy of the procedures cannot be perfectly valid; the identification of the mole is performed through the described procedure that does not guarantee a 100% correct individuation of the mole's pixels, as a few of them may be left out and a few wrong ones may be counted in. The results seem however to have acceptable accuracy. From the first sight, it is rather clear that the statistical analysis of the two features for the three classes of moles makes sense with what had been supposed at the beginning; the mean values of both Border and Asymmetry grow going towards potentially bad or bad moles, meaning that the assumptions that a melanoma has a more indented perimeter and is less symmetrical than a normal mole were undoubtedly true.

Melanomas, however, can be very different one from the other, as it is pictured by the variances of the two features, consistently higher than in the other cases. What is more, the results are quite heavily influenced by a few very peculiar and strangely-shaped moles, especially for melanomas and especially for what concerns the symmetry analysis. In these cases, it is clear from the beginning that the object of the image is a skin tumor, but for many other melanomas it is not that obvious. In fact, the distinction between a normal mole and a melanoma is not always easy; there are indeed several cases in which a melanoma has values for the two features very similar to a medium risk or even (but rarely) to a low risk mole. It is natural and it is the reason why two features are too few to conduct a reasonably certain identification of a mole's class.
Nevertheless, these results show that interesting differences between classes could be observed even with a few data, and that the procedures for the mole identification and for the two selected features extraction have brought reasonable results that can certainly be of help.

The applications of this study are of course in the direction of performing a mole classification from its photo. This work could be of sure help in a field that could significantly simplify the life of the doctors, allowing a first remote evaluation of moles. This would urge only people with a potentially dangerous mole already identified from a photo to personally go to the doctor, reducing waiting times and bringing a powerful help to the sanitary system.