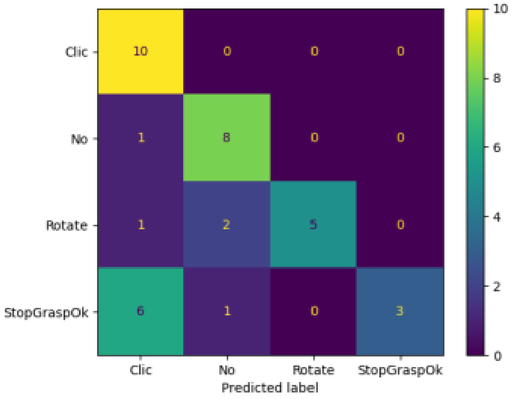
Hand Gesture Recognition Using Motion Energy Images -Dynamic Time Warping (DTW)

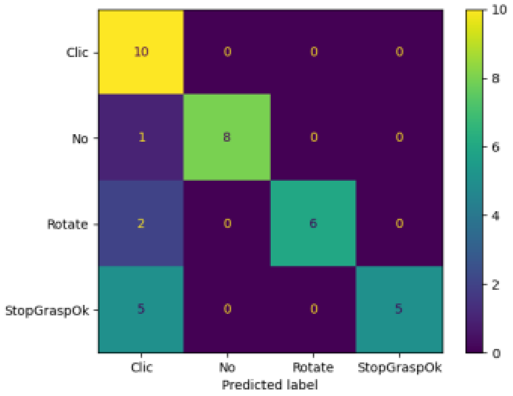
The goal of this project is the visual recognition of four different gestures ('Click', 'No', 'Rotate', 'StopGraspOk'). For each gesture category, we have a set of image sequences that describe the respective gesture. This set of labeled image sequences will be used to train and evaluate a **K-NN classifier**.

First, gesture recognition is executed using a method based on **motion energy images**. The function motion_energy_image_generator() is defined, which takes as input the image files of a sequence of images and produces the motion energy image corresponding to that sequence. Specifically, a motion energy image is a static way of depicting motion (in this case, a hand gesture) and is created from a set of binary images. Each binary image is obtained through the binary thresholding of the absolute difference between two consecutive images (frames) of the sequence. The sum (pixel-wise) of the set of binary images of a sequence (after normalizing to the range [0,255]) produces the corresponding motion energy image.

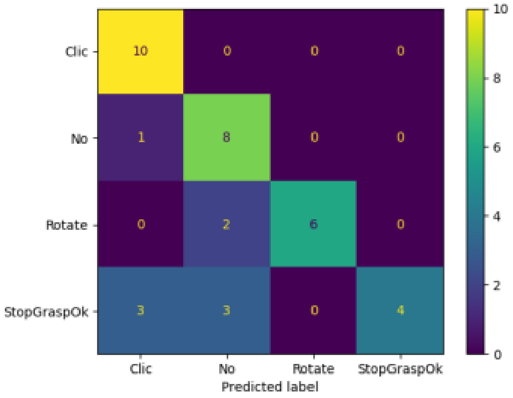
As already mentioned, there are four gestures. The first gesture is characterized by 15 image sequences, the second by 14, the third by 13, and the fourth by 15. In total, we have 57 image sequences, resulting in 57 motion energy images. The motion energy images obtained through the above process will compose the training and test dataset for training and evaluating, a K-NN classifier. Specifically, the features used are the motion energy images themselves and not some descriptor of them. To construct the training set, we use the first 5 motion energy images of each gesture category. The remaining motion energy images form the test set. Four different K-NN models were tested, with the number of nearest neighbors varying from k=2 to k=5. The accuracy of each model and the corresponding confusion matrix of the classification are presented below.



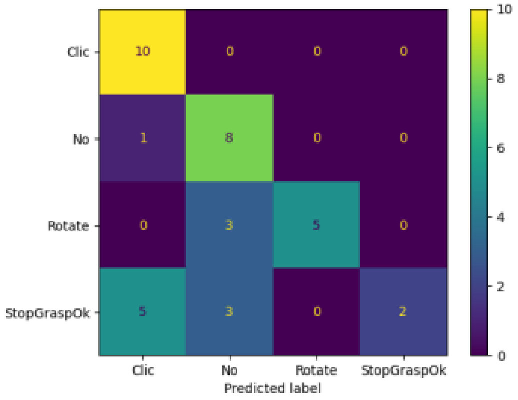
Confusion Matrix of K-NN classification(k=2)
Accuracy: 0.7027



Confusion Matrix of K-NN classification(k=3)
Accuracy: 0.7838



Confusion Matrix of K-NN classification(k=4)
Accuracy: 0.7568



Confusion Matrix of K-NN classification(k=5)
Accuracy: 0.6757

Generally, motion energy images are not the most robust representation of motion as they are significantly affected by factors such as noise, brightness differences, and unwanted (and unrelated to the gesture) movement within the scene. They also present difficulties in accurately capturing information and representing more complex motions, especially when they consist of simpler overlapping movements.

Despite the negative characteristics and limited capabilities of motion energy images as a means of representing movement, their use as features for training and evaluating A K-NN classifier produces relatively good results, where in the best case (k=3), 78.38% of the gestures in the test set are correctly classified, especially considering the small number of images (57 images) available. The classification accuracy could potentially be increased by using different features (using a descriptor for the motion energy images rather than the images themselves) and by using a different threshold value during the thresholding process for producing the binary images that compose a motion energy image.

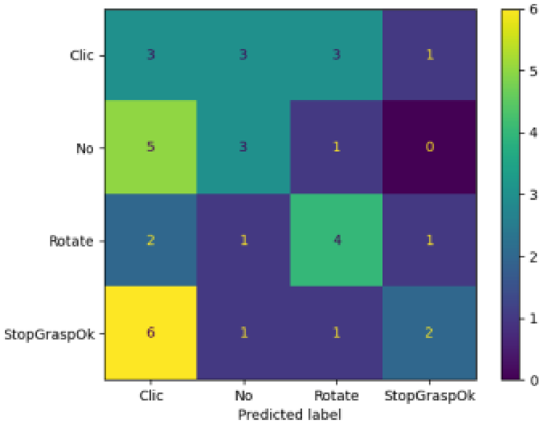
In the second part of the project, hand gesture classification is executed by combining a K-NN classifier with the **DTW algorithm**. For each gesture - image sequence in the set, we extract a vector that describes it, and by comparing the vector of a query gesture with vectors of known gestures, we decide which category it belongs to. Specifically, the vector describing a gesture - image sequence is a set of representations of the hand in the set of images of the particular sequence. The hand can be represented in various ways, but in the context of this project I examined two specific ways of representation. Initially, for each image in a sequence, the hand is represented by a point - pixel (i,j) of the image. This point can be perceived as the "center of mass" of the hand in each frame.

We are interested in obtaining a representative point for the position of the hand. This point was chosen to be the following: We calculate the external contours of the image, and considering all the pixels of all contours, we compute an average value for the position on the vertical and horizontal axes, say y_avg and x_avg, respectively. The list [x_avg, y_avg] constitutes the vector representing the hand in the case of an image in the sequence. However, a sequence (which defines a gesture) consists of a set of such images; thus, the final vector representing it will be a set of [x_avg, y_avg] representations. The function gesture_vector_generator() produces the vector representing a gesture. The set of vectors generated defines the dataset, which is divided into train and test sets as before.

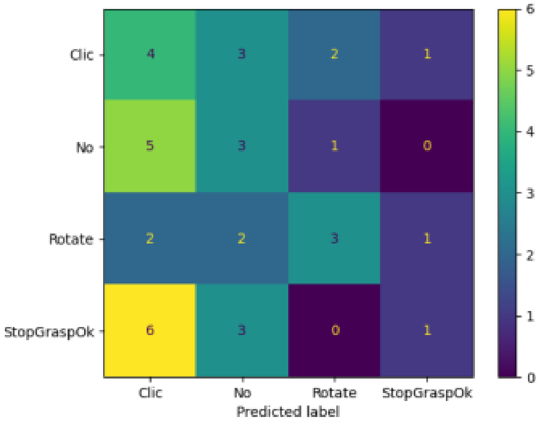
To perform the classification, a new class is created called Classifier(). Objects of this class have only one attribute, the number of nearest neighbors k, in other words, they are K-NN classifiers. The reason for not using a pre-existing model like in the case of motion energy images is the requirement to integrate the DTW algorithm into the process. The fit_predict() method of the class receives the data and the labels of the train set as well as the data of the test set (query gestures) as input, and returns the predicted categories that the query gestures belong to. The **detailed classification process** is as follows:

For each vector-gesture in the test set, we calculate its distance from each vector-gesture in the train set. This distance calculation is implemented using the DTW algorithm (dtw() function from the tslearn library) with the distance metric being the Euclidean distance. The necessity of using DTW for distance calculation arises from the fact that, generally, different gesture samples are characterized by vectors of different lengths; hence, using a more classical methodology —e.g., the Euclidean distance— which requires the vectors under comparison to be of the same length, is not feasible. Additionally, DTW is a very robust methodology for matching the elements of two vectors due to the time warping that occurs. This way, the speed at which the hand moves during a gesture does not affect the matching result. For example, if a user performs the same gesture twice, once slowly and once much faster, calculating the Euclidean distance between the vectors of these two gestures (assuming the vectors are of the same length) will result in a large distance value, indicating that the two gestures are not the same, whereas in reality, they are. On the other hand, DTW will return a relatively small distance value, as expected since it is the same gesture.

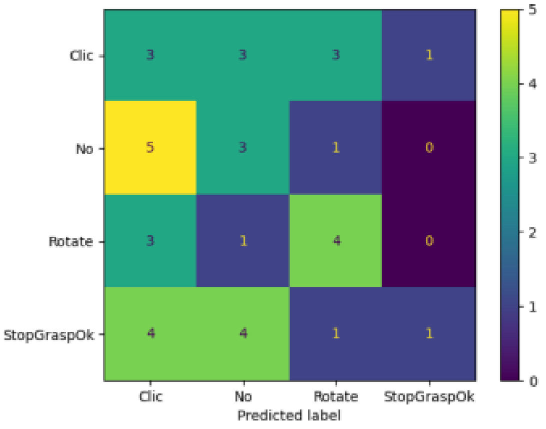
Having obtained the distances of a query gesture from each gesture of the train set, we identify its k closest neighbors, i.e., the k train set gestures from which it has the smallest distance. The category to which the query gesture is classified to belong to, is the most frequently occurring category among the k closest neighbors. Each neighbor votes with its label, and the label with the most votes will be the label of the query gesture. In case of a tie, when the number of neighbors is even, the label that appears first in the set of labels is chosen. Four different K-NN models were examined, each time varying the value of k. The classification results are as follows.



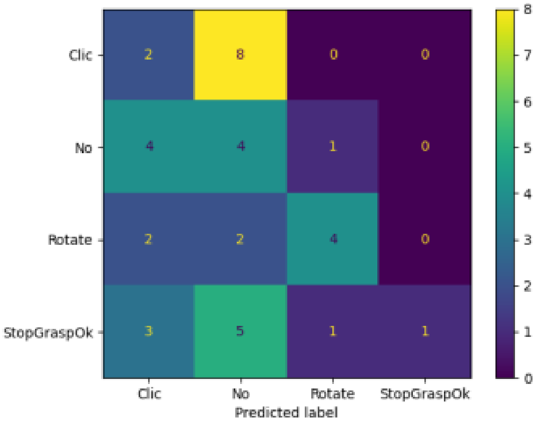
Confusion Matrix of K-NN classification(k=2)
Accuracy: 0.3243



Confusion Matrix of K-NN classification(k=3)
Accuracy: 0.2973

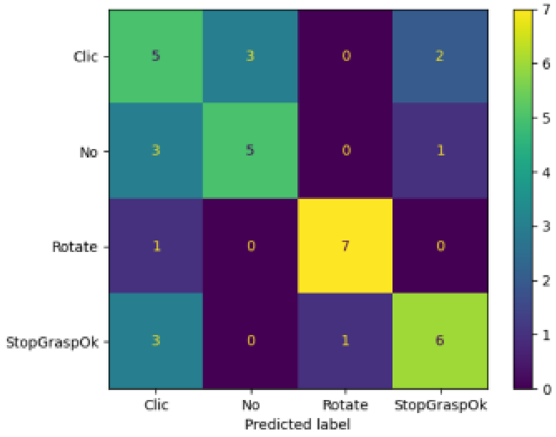


Confusion Matrix of K-NN classification(k=4)
Accuracy: 0.2973

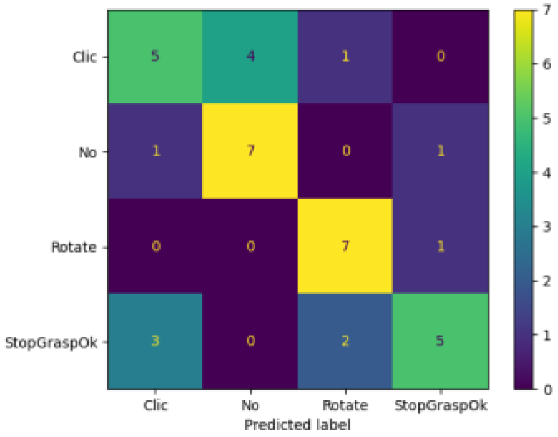


Confusion Matrix of K-NN classification(k=5)
Accuracy: 0.2973

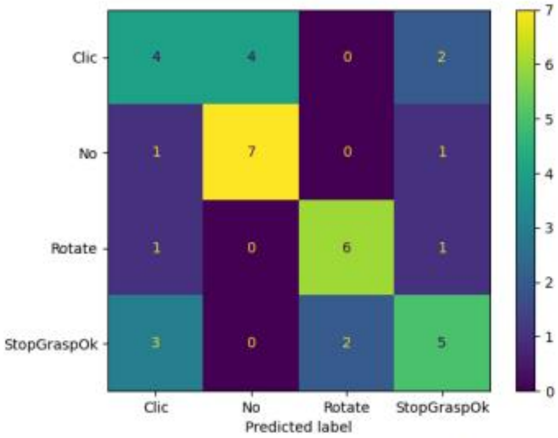
We observe that the model achieving the best classification is the one where k=2. The accuracy of this model is 0.3243, meaning it correctly classified only 32.43% of the cases. This percentage is quite low, especially compared to the best percentage achieved using motion energy images, which was 78.38%. The failure of the classification using DTW is likely due to the simplistic representation we have chosen for the position of the hand. To examine if this is indeed the case, we will re-implement the classification as previously described, but now the hand will not be represented by the pair [x_avg, y_avg]. Instead, we will extend the representation to include the width and height of the minimum bounding rectangle (MBR) that encloses the hand. This way, we obtain a more complex representation of the hand, which is [x_avg, y_avg, w, h] (where w is the width and h is the height of the MBR). The classification results using this new, more complex representation are as follows.



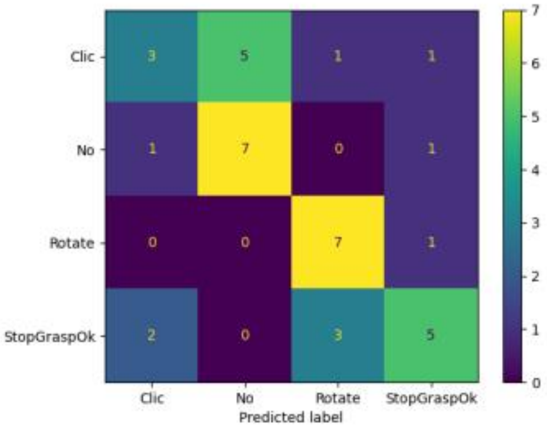
Confusion Matrix of K-NN classification(k=2)
Accuracy: 0.6216



Confusion Matrix of K-NN classification(k=3)
Accuracy: 0.6486



Confusion Matrix of K-NN classification(k=4)
Accuracy: 0.5946



Confusion Matrix of K-NN classification(k=5)
Accuracy: 0.5946

We observe that in all four model cases, the percentage of correctly classified gestures has significantly increased compared to the previous models where the more simplistic representation was used. We thus confirm that the [x_avg, y_avg] representation of the hand is considerably insufficient, resulting in limited capabilities of the classifiers. In this case, the best performance is achieved by the k=3 model, which correctly classified 64.86% of the query hand gestures.

In conclusion, we observed that using a more complex representation for the hand led to a notable increase in the models' performance. This underscores the importance of the choice of hand representation and the need to explore more representation methods to further improve performance. Comparing the set of methodologies applied, the best results were obtained in the initial part of the project where motion energy images were used. However, this methodology is quite limited for the reasons mentioned earlier. In the general case of gesture recognition, we expect to achieve better and more robust results using DTW, provided that the best possible representation of the hand is chosen.