BABELE

Carmine Ippolito

1 CONTEXT OF THE PROJECT

Lips and lip movements are biometrics that have been used for several purposes, such as improving lip reading systems and other applications that fall within the forensic field. Although less developed, lip-based systems are getting better and better as new techniques are researched [2].

In this paper these biometrics will be used for recognition and to identify the language of the subject.

2 GOALS OF THE PROJECT

The main goal of the project is to realize a recognition system that uses videos containing only the lip area, and evaluate whether the knowledge of the subject's language (that needs to be identified) produces better results.

There will also be an initial analysis on some of the non-functional requirements (fairness and explainability).

3 METHODOLOGICAL STEPS CONDUCTED TO ADDRESS THE GOALS

Initially the provided dataset contained 258 videos unevenly split into 6 languages (Italian, English, German, Spanish, Dutch and Russian). For this reason, 32 videos were chosen for each language based on frame rate, duration, size and thumbnail (192 videos in total, as shown in Table 1). Since the videos were downloaded from YouTube by different people without a standardized process, their quality was very variable and generally not the best. For example, there were videos with strange frame rates (27), very short duration (less than a minute) and disproportionate size for their content.

According to the previous methodology each video had to be cut into 3 sub-videos of 15 seconds in which a face was detected by dlib. Subsequently, the area of the labial zone and the Euclidean distances between the landmarks of the inner lip were extracted.

Although the new methodology has remained similar to the previous one (5 10-second sub-videos instead of 3 15-second sub-videos), the provided scripts have been rebuilt from the ground-up to improve their performance (parallelization) and the quality of the videos. Nonetheless, the dataset still needs to be manually rechecked or rebuilt from scratch, even if improved.

3.1 Data acquisition

In the new scripts the distances are no longer rounded to zero but to the nearest integer. Furthermore, the upsample_num_times parameter used by the dlib detector was set from 1 to 0, reducing the execution time by up to 75% with negligible effect on the distances between the landmarks. In general, it should be preferable to use higher resolution videos rather than upsampling.

The first rewritten script was the one responsible for the extraction of the sub-videos. Compared to the previous one, which scaled everything to 360p, its performance varies greatly depending on the resolution (as shown in Figure 1 and Figure 2). Also, the new script may take longer as it encodes with a better codec, looks for a single 10 second sequence instead of merging three 5 second sequences and skips some of the frames to avoid possible transactions, special effects and occlusions.

Language	Men	Women
Italian	15	17
English	12	20
German	14	18
Spanish	15	17
Dutch	23	9
Russian	14	18

Table 1: distribution of the reorganized dataset (192 videos in total).

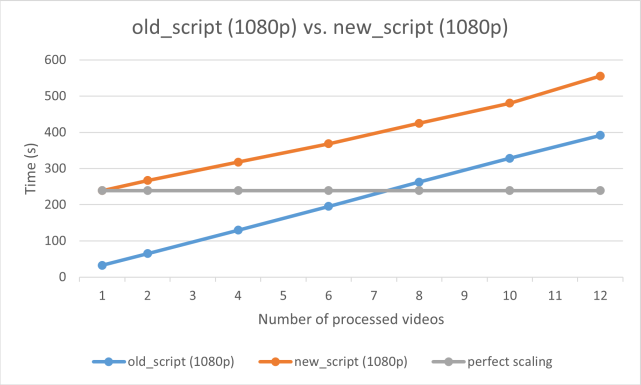

Figure 1: performance of the old and new script at 1080p (worst case scenario) on an AMD Ryzen 5 3600.

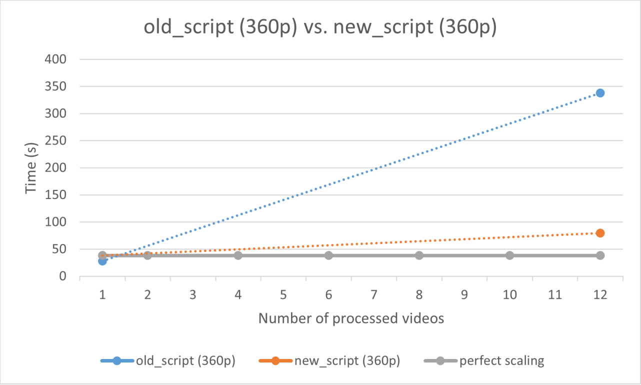

Figure 2: performance of the old and new script at 360p (best case scenario) on an AMD Ryzen 5 3600.

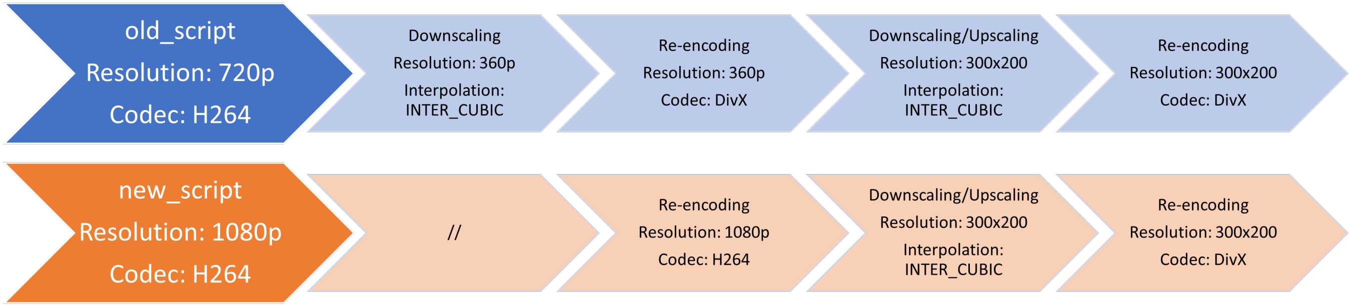

Figure 3: pipeline used to extract the mouth area from the videos.

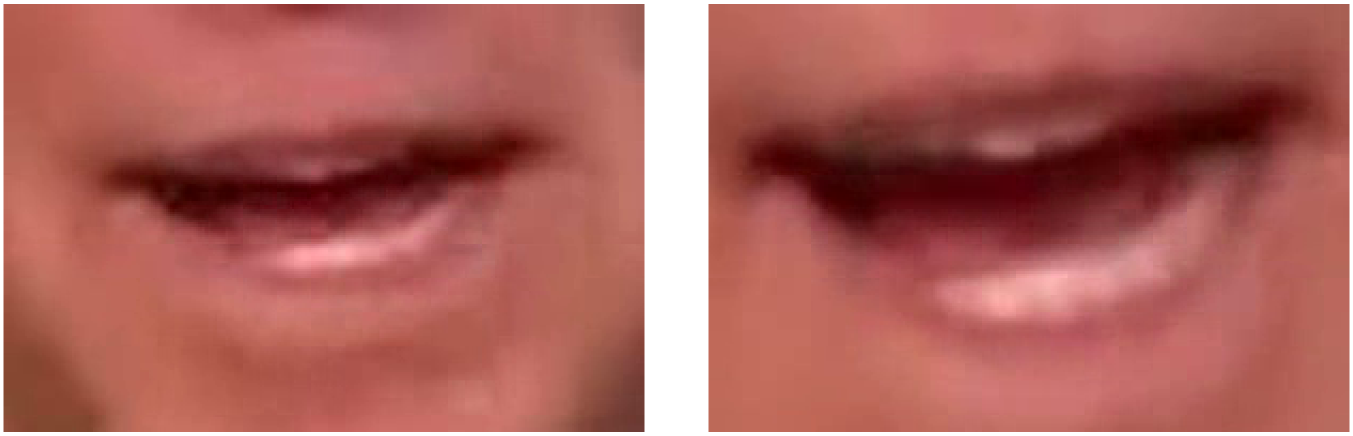

Figure 4: result of the old script on the left and of the new script on the right. The difference in the selected region is caused by the downscaling.

The other rewritten script was the one responsible for extracting the mouth area and the distances between the landmarks. It now takes 85% less execution time and is able to calculate the distances based on the Manhattan distance (experiments with Ollivier-Ricci's curvature have had little success and require more in-depth analysis). One of the problems with the old script was the missing extraction of some videos caused by dlib. Although the problem has not been completely resolved, it has been greatly reduced (from 15.3% to 1.8% of missing sub-videos).

Regarding quality, Figure 3 shows the pipeline used to extract the mouth area from the videos, while Figure 4 shows the final results. As can be seen, the increase in quality does not entirely pay off the decrease in performance. It remains desirable to experiment with more advanced codecs (VP9) while using the downscaling to improve performance.

3.2 Data understanding and non-functional requirements

Two different methods were used to understand which distances had the least impact on the results.

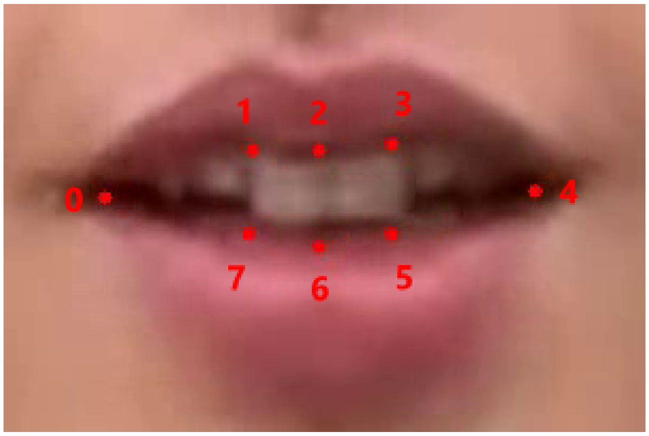

Figure 5: landmarks located by the dlib detector (mouth_intern).

The first method used the standard deviation to find the 4 least variable distances: D8 (1, 2), D14 (2, 3), D26 (5, 6), and D28 (6, 7). See Figure 5 and Figure 6 for more details.

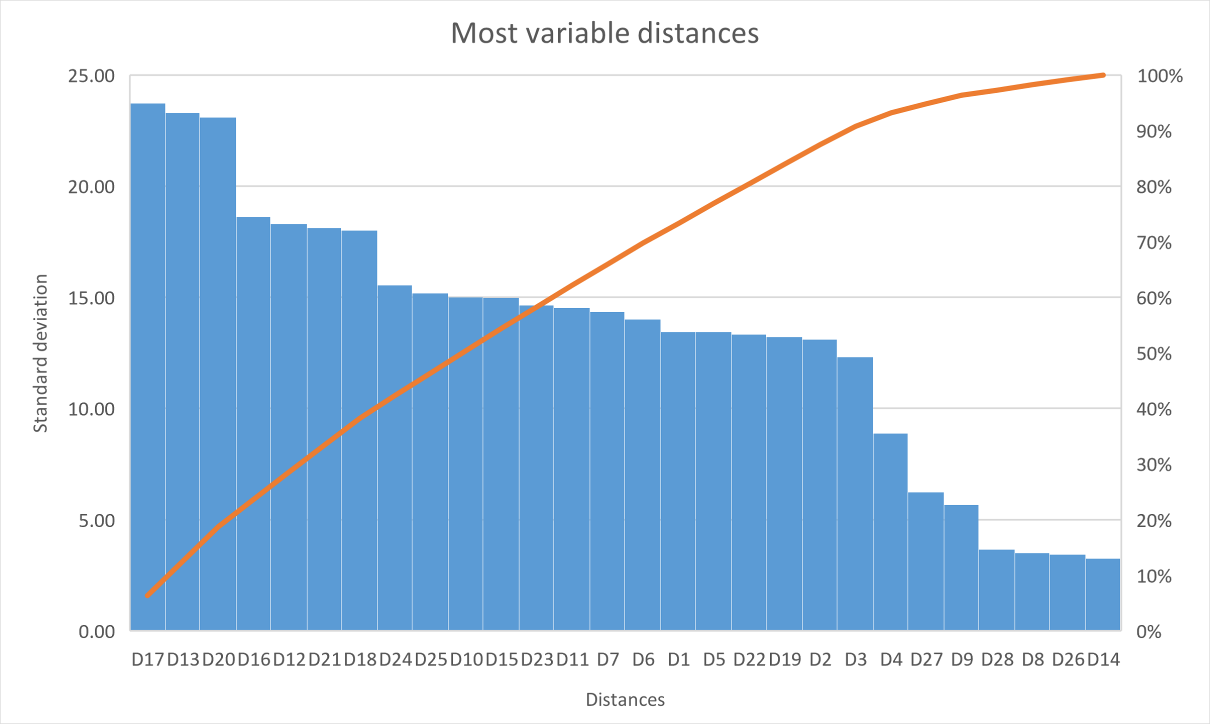

Figure 6: distances ordered by their variability.

The second method used one of the trained classifiers (random forest) to find the importance of the features based on their permutation. While the results for the Manhattan distances were fine, the results for the Euclidean distances could not be considered reliable (probably due to the lower accuracy of the model). Because of this problem, only the first method was used.

Regarding fairness, a problem was found in some of the videos featuring men with mustaches, whose presence caused an incorrect localization of the landmarks (as shown in Figure 7). Unfortunately, no workaround has been found (a better detector than dlib should be used to fix the problem).

3.3 Modeling

To create the models for the identification of the language, 4 datasets were created by varying the type (Manhattan or Euclidean) and the number (28 or 24) of the distances. The data was divided into an 80-10-10 split, ensuring that videos featuring the same subject were grouped together. This was done to allow the models to learn the differences between the languages and not between the subjects. After being standardized with scikit-learn's StandardScaler, the data was optionally passed through the PCA.

The classifiers tested were the k-nearest neighbors, the support vector machine and the random forest, but due to a problem with the predict_proba function the SVM was not used. The parameters were selected and saved with a grid search and the MLFlow Tracking API, respectively. The most accurate classifier was the random forest trained with 300 estimators on the Manhattan28 dataset passed through the PCA with 0.95 components (42.2% accuracy).

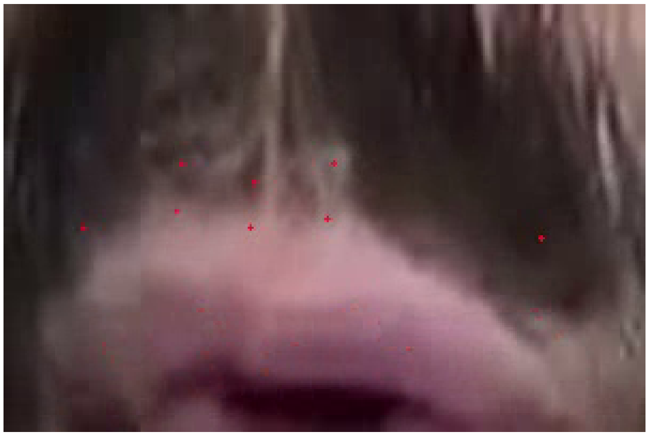

Figure 7: example of an incorrect localization of the landmarks caused by a mustache.

Two different strategies were used to create the models for the recognition: one based on machine learning (k-nearest neighbors and random forest) and one based on deep learning (CNN).

For the one based on machine learning, the dataset was created by merging the validation and test sets of the datasets with 28 distances and dividing them into a 60-20-20 split. The best classifier was the random forest trained with 300 estimators on the Manhattan28 dataset (61.1% accuracy).

%-% Weights	ML accuracy	DL accuracy
100%-0%	61.1%	81.9%
80%-20%	66.7%	81.9%
60%-40%	55.6%	81.9%
40%-60%	41.7%	81.9%
20%-80%	30.6%	76.4%

Table 2: accuracy of the models combined through a weighted arithmetic mean of the probabilities (recognition on the left and language identification on the right).

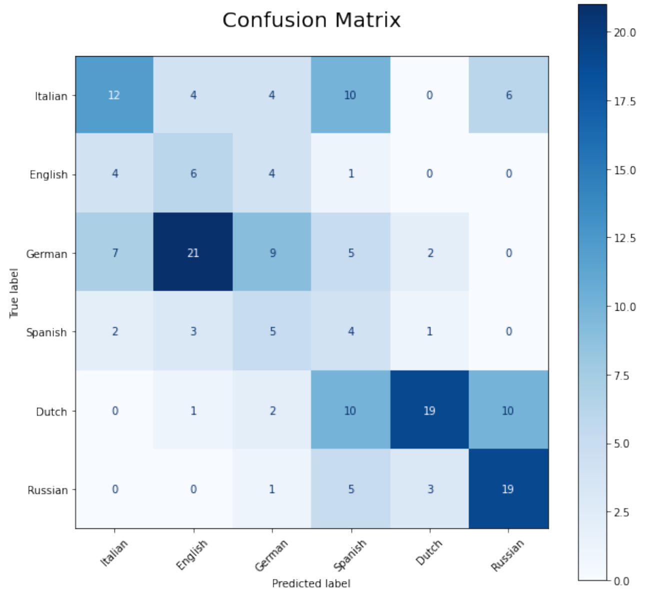

Figure 8: confusion matrix of the model for the language identification (there are 32 videos for each language).

For the one based on deep learning, the dataset was created by extracting 10 frames for each subject present in the aforementioned dataset (the data was preprocessed so that each pixel had zero mean and unit standard deviation).

Because of the small dataset, only the simple CNN from Keiron and Ryan's paper [1] was used (accuracy 81.9%). The network structure is composed of a convolutional layer, a maxpooling layer, a dropout layer, a flatten layer and two fully connected layers (the dropout layer was added to improve generalization).

Finally, the two models for the recognition were combined with the model for the language identification through a weighted arithmetic mean of the probabilities.

4 PRELIMINARY RESULTS AND FINDINGS

Figure 8 shows the confusion matrix of the language identification model (random forest). Dutch and Russian were the most correctly classified languages, while English and Spanish were the worst. English, Spanish and Russian were also the languages with the greatest number of predictions.

Table 2 shows the accuracy of the models combined through a weighted arithmetic mean of the probabilities. The deep learning model was not influenced by the language of the subject as its results were already very decisive. On the other hand, the machine learning model performed better due to its uncertainty and lower accuracy. It is safe to assume that with a larger dataset even the deep learning model could perform better as only 36 of the 192 subjects were included.

5 IMPLICATIONS OF THE RESULTS

The results show how the knowledge of the language can improve the recognition of a subject. Although the improvements are small, there is still room for further progress due to the various limitations of the dataset (the mustache problem in particular). Classifiers that use temporal information could also be considered.

Regarding explainability, the dataset with 24 distances always performed the same or better than the dataset with 28 distances, with the exception of the random forest trained on the Manhattan28 dataset. Therefore, it is safe to assume that further analysis on explainability could lead to better accuracy.

6 CONCLUSION

Lips and lip movements are biometrics that have been used for several purposes, especially in the forensic field. The obtained results show how these can be used to identify a subject's language and improve the models for the recognition. While the improvements are small, there is still room for further progress due to dataset limitations and the availability of more powerful classifiers (especially those that use temporal information).

A more in-depth analysis of explainability remains desirable to gain a better understanding of the model and improve performance. Regarding fairness, the mustache problem needs to be solved as soon as possible by using a better detector than dlib.

REFERENCES

[1] Keiron O'Shea and Ryan Nash. 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458* (2015).

[2] Krzysztof Wrobel, Rafal Doroz, Piotr Porwik, Jacek Naruniec, and Marek Kowalski. 2017. Using a probabilistic neural network for lip-based biometric verification. *Engineering Applications of Artificial Intelligence* 64 (2017), 112–127.