EPITECH. TECHNOLOGY

NOSTRADAMOVIES

BOOTSTRAP

NOSTRADAMOVIES

Features

If you are not convinced by poster uniformity, search for examples on the Internet.
You should find lots of informative videos and articles.

You may trust serendipity from the following link

Try to anticipate the most valuable features to be extracted from posters.

There are lots of possible features, such as color, histogram of oriented gradient (HOG, title position, fonts…)

Face detection

The number of faces, or their positions, are important features.
Whether it's full face or profil, it can also have an impact.

Some algorithms are already pre-trained to detect faces, but if they are only trained to detect full faces and not a profil one, you will perform badly on some movie posters.

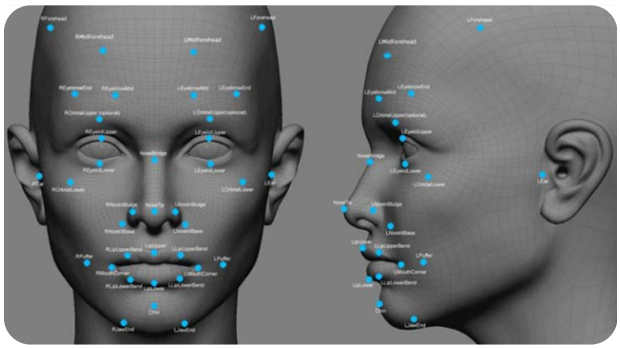
Adaboost is a basic algorithm to do face detection, it's used in the **OpenCV** library.
Go and find out more into how it works.

Using an unsupervised framework requires being confident in your algorithm, since one can not automatically test the final performance.

Pick about 10 images and test with your bare hands to see if you correctly detect faces in the posters, manually applying your algorithm.

For each poster, you should be able to compute and display the number of faces, but errors are sometimes inevitable (well, almost inevitable).

Apart from getting the number of faces, you can also locate them on the poster.
If you want to detect faces you need to start looking for pattern (such as the line of the eyes).



In a gray scale image, patterns can be detected thanks to large variations in the intensity.

Online, and depending on the computer vision library, you will find more details on how it can be done.

Write a program that displays detected faces in a poster.

You may see that sometimes non faces are detected as faces. If you understand how your alogirthm works, you might also understand what caused the mistakes. It's the first step to do if you want to correct your errors.

Once again, test your algorithm on a few images to see if it works correctly.

Colorimetry

Colorimetry may be rich in information.
But it does not restrict to mean color only.
It encompasses color emotion, color harmony, color variance…

OCR and text analysis

Text-based features are probably the most important ones. It includes font, text size, name of the movie maker and actors, composition, text, color, etc. Extracting these features requires two steps:

- ✓ OCR, which is getting the text from the poster ;
- ✓ text analysis.

OCR is a complex data, and you are not expected to develop your own algorithm.
Good luck if you want to go down this road! Find and sandbox OCR libraries.

OpenCV, pyTessaract for instance

Once you have extracted the text along with the composition (size, location, colorimetry), you need to analyse it. Look for semantic tagging and content classification. You can look for various things such as locations, sentiment analysis, title length, names, specific vocabulary…

Grammatical analysis will require Natural Language Processing (NLP). There are many libraries for NLP to help you analyse the text.
Those new features can be desicive to find correlations and to increase the quality of your predictions.

SHapley Additive exPlanations

The *black box* problem refers to the lack of understanding you may have about predictions.
It sometimes seems that data turns into prediction through a foggy mechanism.

Therefore, features explainability is crucial. Most algorithms like Random Forest have a built in method for features importance but it can be sometimes misleading and depends on the algorithm choosen.

SHapley Additive exPlanations will give you a global model explainability that can be apply to every tree based algorithms. Besides, it can look at one specific prediction in the dataset to understand the decision made. You can also add visualisation showing the positive or negative impact for each features.

Check out this tutorial if you feel lost.

v 1.5