ANONYMIZING VIDEOS & ACCURACY PREDICTION

CESAR MARRADES CORTES
cmarrades.com

SUPERVISOR:
KEITH QUILLE
ITT Tallaght, AI lecturer

1. ABSTRACT

This document presents the possibility of anonymizing sports videos, and assessing the accuracy of the outputs with Machine Learning. Collaborating with the NCCA, it covers their specific business case where their main objective is to preserve identity of their students before the examination videos presented for grading are visualized by the agents on the organisation, in order to comply with the new GDPR regulation. Combining both, Development and Data Analytics perspective, the project is divided in two different phases.

On a first stage, a Video Anonymizer process has been developed in C# and is delivered as an executable file (.exe). Supported by two different libraries (Accord and Emgu), given one video input, the optimal target is to generate an output video where any visible faces from the students have been recognized and blurred. 100 video inputs have been used for this stage resulting in 200 outputs (1 for each library).

A second component has been developed using WEKA, aiming to predict the accuracy of the outputs from the first component, based on attributes manually extracted from them, such as Camera quality, Camera Angle, Sport, and others that can be found in sections below. This model also hints limitations and boundaries on the libraries used.

This project not only demonstrates the capabilities and limitations of the two libraries used by the first component by stressing them under a multiple set of environments and conditions, such as multiple sports (soccer, karate, boxing, etc.), but also sets a base ground for NCCA research, automation and improvement on a full video anonymization system combined with a Machine Learning component that aims to predict the quality of the blurred videos. The output model provided a 94% accuracy using Naïve Bayes Classifier, with a 95.2% specificity, meaning it may be likely to predict, given a video, how feasible is to detect scenarios with difficulties for automated blurring.

2. LITERATURE REVIEW

Research on this specific business problem has shown how Emgucv has been used on many other face recognition applications by researchers on the field. The following points, extract of the "Master Thesis" of Suad Hajr Ahmed Omar (2016), are a subset of the limitations he discusses about the library:

- The distance of the person from the camera should be in 1 to 3 feet for better result.
- The system will be applied and tested in a fix environment with ordinary illumination.
- The problems such as sunglasses, eyeglasses and other accessories that can partially or fully cover the face are not subject for face recognition.

These are discussed on later sections of the document. Time was also one of the constraints of the project, reducing the available error margin when processing 200 outputs as the basis for the second phase of the project. The "Kodak Facebook Collage Project" by Team, T. B. D., also had similar constraints and bottle neck concerns when developing their project. The team cites on his paper: "*High level language wrappers for OpenCV were evaluated and the team found that EmguCV (), a .NET wrapper, was among the most stable and complete.*"

The paper "RECOGNITION OF FACE DETECTION SYSTEM BASED ON VIDEO" by SARI ABDO ALI MOHAMMED also chose Emgucv (http://www.emgu.com") for his system, although he also remarks limitations on what appears to be one of the best libraries: "*there are many challenges that trouble researchers in this field. A lot of effective factors can limit face detection and recognition. These factors affect the appearance of face such as enlightenment, face poses, occlusion (sunglasses, hairstyle, make up), and face expressions and camera quality.*"

After analysing which properties of the inputs may have an impact on the quality of the produced results, this document set Emgucv as a base for the face recognition system, and offers a new angle on its limitations.

3. INTRODUCTION

The project addresses a specific image processing problem presented by the NCCA. Exams tests are often recorded by students in video format and then evaluated by NCCA. This project aims to develop a face recognition system that shows the possibility of pre-processing videos prior "NCCA evaluation", and anonymize them, by using face recognition libraries that will detect students faces that will be blurred. This is obviously a main concern for the organization as identity preservation is one of the main shifts of the new GDPR regulation coming into play in May 2018.

The final goal of the organization is to have a system that allows to automatize the process of anonymizing videos to preserve the identity of their students and avoid biased marking when evaluating the contents.

This document not also prioritizes findings and recommendations for future development and integration purposes over the big picture of a software engineering

product, but also tackles the problem by isolating responsibilities and addressing them from two different perspectives: software and data analytics. Where the first one oversees video anonymization, the second one tries to find limitations and boundaries on facial recognition libraries used, and provides a prediction on the quality of the outputs processed by this first component.

The project is divided into two main phases.

3.1. PROJECT PHASES

1ST PHASE - VIDEO ANON YMIZER

On the first phase, the problem was approached from a Development perspective, by isolating the face recognition problem and video anonymization into an executable file. Once the development phase finished, 2 machines were required to bring down the processing time to 8 days. This process does not aim to solve the problem, but to set a base line of accuracy and performance using third party libraries, and provide some insights on their limitations, and recommendations for future research and development.

2ND PHASE - PREDICTION MODEL

On the second phase, the problem was approached from a Data Analytics perspective. Inputs were analysed, and as a result, attributes were manually extracted. Output results were labelled according to scales that can be found in sections below. Having a dataset with all this information, this file was then imported into WEKA tool, used for data analysis, data pre-processing, attribute selection and final model selection. This model would predict the quality of the outputs produced on the first phase, based on a 1-10 scale (see appendix section) and on a binary "*full*" "*partial*" scale.

These two marked phases took an approximate equal amount of time. Processing the videos, marked as 30% of the first phase, took more than 7 days.

3.2. COMPONENTS

VIDEO ANONYMIZER COMPONENT

The executable file has been developed using C# using EmguCv and Accord libraries. The tool processes each video generating two different video outputs (one per each library), where on a perfect scenario, student faces cannot be recognized.

PREDICTION MODEL

WEKA model predicting accuracy of anonymizer process (phase 2).

3.3. INPUTS & OUTPUTS

INPUTS

- 100 YouTube Videos for the anonymizer process (mp4 format).
- Dataset with video attributes and anonymizer result. These attributes are detailed on the next section.

OUTPUTS

- 2 sets of 100 anonymized videos. 2 hundred videos in total (mp4 format).
- Weka prediction on video blurring success (fully or partially blurred).

4. METHODOLOGY

This section breaks up the different phases of the project providing a more granular overview with technical details and specific results. As discussed above, there are to main differentiated phases:

1. Video Anonymizer process
2. Prediction Model

In an attempt to produce the best possible prediction model, the second phase was developed using two different configurations detailed below. The first configuration is based on a 1 to 10 scale, rated from worst to best possible output, and the second one based on a binary scale, assessing whether outputs are "fully" or "partially" blurred.

4.1. 1ST PHASE - VIDEO ANONYMIZER

LIBRARY RESEARCH

Following libraries were evaluated:

- Microsoft Cognitive Services.
- Censor Face (pay per usage)
- Open CV (free)
- Emgu CV (free)
- Accord

Free versions were taken as a start base, and **Emgu** and **Accord** were chosen based on available documentation (accord) and social feedback research (Emgu).

The tool that anonymizes videos is presented as a Console ".exe" file, written on C#. Once the process starts, it processes all the input videos located on the "input" folder (configured on the settings file), generating 2 outputs per each input (1 per library).

As the internals of the program are out of the scope of the project, no further analysis is presented.

INPUTS RESEARCH & DEFINITION

VIDEOS

A set of 100 random sport videos have been extracted from YouTube platform. Initial approach was to provide a variety of different scenarios that would cover the most common cases in NCCA when evaluating their students for exams. Longer videos require longer processing time, therefore shorter videos had a preference over longer ones. On these premises, the whole batch of videos is classified in 10 different sports, indoors and outdoors conditions, a range from 1 to more than 10 persons, recorded on mobile and normal cameras, on static and dynamic movement conditions and on multiple different field types.

SPORTS

The 10 sports have been chosen without knowing the final real business cases, trying to address the problem from a wide variety of possible scenarios and conditions. The following table offers a breakup of the different sports with a minimized overview of generalized tendencies found after inputs visualisation:

Sport	Conditions
Basketball	Indoors, outdoors, dynamic cameras, medium and far distances
Karate, Judo	Indoors, multiple backgrounds, multiple angles, tendency to close to medium distances, head protectors
Boxing	Indoors, multiple angles, head protectors, referee. Close, medium and far ranges
Hurling	Tendency to far distances, high angles, with multiple people
GaelicFootball	Tendency to far distances, medium and high angles, with multiple people.
Tennis	Multiple angles, mixed cam movement types, high angles
TableTennis	Medium and far distances,
Soccer, SoccerInterior	Multiple angles, mixed cam movement types, medium, high angles
HighJump, HighJumpPole	Medium angles from close and medium distances and different camera types
WeightLifting	Static cameras, from close and medium distances and medium angles

Sports table

This information indicates different sports have a tendency of sharing common patterns, although different inputs from same sport have multiple combinations of attributes. These sports were chosen based on the variety of attributes they had.

VIDEO ATTRIBUTES

After visualization of input videos,

- Sport (sports defined on *sports table* above)
- Indoors / Outdoors
- Light conditions:
- Whether the video was recorded using a phone or a normal camera.
- Video quality
- Camera Movement. E.g. static, or dynamic.
- Camera Recording angle. E.g. low, medium or high.
- Camera distance. E.g. medium or far.
- Number of people in the video.
- Whether people was wearing head protectors.
- Field type. E.g. court, grass.
- Field Colour.
- Background Colour.
- Whether there was crowd on the background of the video.
- Whether there was a referee.
- Ground position: whether people in the video would spend a considerable amount of time on the floor. E.g. wrestling or combat sports in general.

All this information was gathered on dataset, along with the final "class" attribute for later classification, conforming the basis for the final prediction model.

OUTPUT CLASSIFICATION TABLES

Two different output scales were used to generate the final prediction model. The 10-scale raking system showed underperforming results on the second phase, which forced to reassess of the 10-scale into a binary scale:

10 -Scale Rank	10-Scale Description	2-Scale Rank
1	no accuracy, high noise	Partial
2	no accuracy, noise	Partial
3	extremely low accuracy, noise (any)	Partial
4	low accuracy, high noise	Partial
5	low accuracy, noise	Partial
6	good accuracy, high noise	Partial
7	good accuracy, low noise	Full
8	almost no failures, noise (any)	Full
9	no failures, noise	Full
10	no failures, no noise	Full

Output scale tables

Although evaluation may be subjective to interpretation, the following rules were used when grading outputs with the scales above:

- "Extreme low accuracy" was considered when nearly none of the faces were recognized on the video.
- "Low accuracy" was considered when some of the faces were recognized in some parts of the video.
- "Good accuracy" was considered when faces were recognized many times.
- The 7th mark was used to consider when most of the parts of the video were successfully blurred, but still with failures.
- The 8 mark was used for videos with almost no failures
- The 9 mark, the maximum found on the dataset, was used for a video where nobody could be recognized.

VIDEO OUTPUTS

VIDEO OUTPUTS PROCESSING

Processing 200 hundred outputs (1 per each library) took a considerable amount of time. A micro AWS instance was initially setup processing the 100 inputs located in the "inputs" folder. This instance was later on scaled to a large instance as the process would max out the CPU. After few days processing inputs, a second machine INTEL CORE i7 2.60GHz was added to speed up the process. Processing 200 outputs took around 8 days following these steps. No performance measurements were taken. There is however room for performance improvements, detailed in sections below.

VIDEO OUTPUTS CLASSIFICATION

Visualization of output videos was required to assess the blurring quality based on the scales detailed on the "Output Classification Tables" section. This step completed the final dataset used as a basis for the prediction model phase.

4.2. 2ND PHASE - PREDICTION MODEL

TWO SCALES, ONE MODEL

All the steps required to elaborate the prediction model (data pre-processing, attribute selection, testing, etc) were carried out using WEKA tool. A first intend using the 10-scale output class showed underperforming results with a maximum accuracy of 56%, therefore it was discarded. After revisiting inputs, a binary threshold (fully or partially blurred) was defined based on "***Output scale tables***" above, with only 2 inputs ranked as 7. All the steps were then followed using the binary output scale performing with a 94% accuracy. Although no further tests were done, final results indicate increasing the "*Full*" threshold at rank 8 would still have high probability of positive results.

DATA INSIGHTS

Data inspection and observations revealed following findings:

Videos are nearly 50% divided between indoors and outdoors. Only 5% of the videos are recorded under bad light conditions, and there is a high tendency to record these videos on cameras rather than on mobile phones, as only 5% of them are recorded using a phone. There are very few of the videos where persons are wearing head protections and there is a tendency to keep the same cam distance along all the duration of the video. Recordings are evenly divided with and without referee, and either with or without crowd. It is remarkable the very few high scored videos found on the dataset. No obvious correlations are shown after visual correlation inspection.

The distributions show "fully" blurred outputs have very specific characteristics, where the "partially" classes occupy

most of the dataset. The plots indicate that all the "fully" outputs are taken from close distances. They also tend to be recorded from static or "steady" cameras, from a medium angle, with no crowd. These attributes will be important to determine whether a video will be successfully anonymized.

Only 10% of the outputs rated above 6. Figures below show distributions of 10-output class (left) vs 2-output class (right).

10-output scale (left) vs 2- output scale distributions (right)
10-output scale: Left to right, 1 to 10
2-output scale: blue = fully blurred, red = partially blurred

The 10-scale distribution shows how most outputs are ranked below 5, where the binary shows only 10 out of 100 outputs are ranked as "fully" blurred.

MOST RELEVANT ATTRIBUTES

Attribute	Rank
Crowd	0.3242
CamDistance	0.3203
In-Out	0.3203
CamMovement	0.2958
CamAngle	0.2695
Referee	0.261

Correlation Attribute Evaluation scored Crowd, CamDistance, In-Out and CamMovement as the most relevant attributes affecting video blurring accuracy.

5. MODEL TESTS AND RESULTS

Based on the previous list of attributes extracted using the *Correlation Attribute Evaluation* for the dataset using the 2-scale output class, different tests were run **with** (*attribute set 1*) and **without** the "*Referee*" (attribute set 2) attribute. These tests results are presented on the table below, with accuracy, sensitivity and specificity columns.

Attribute Set	Algorithm	Accuracy	Sensitivity	Specificity
1	Naïve Bayes	91%	**100%**	90%
1	Trees.J48	89%	30%	**95.5%**
2	Naïve Bayes	**94%**	90%	94.4%
2	Trees.J48	89%	30%	**95.5%**

There are three main components on the table:

ACCURACY

Shows an average percentage of how the model would be able to predict "fully" and "partially" blurred videos. On this context, the third configuration has the highest accuracy with a 94% which is an excellent indicator on how this model would perform predicting blurring accuracy based on video properties.

SENSITIVITY

Focuses on predicting the "**fully**" blurred videos. Based on this value, the first configuration would recognize 100% of the "fully" blurred videos.

SPECIFICITY

This value indicates how well the model performs when predicting "**partially**" blurred videos. J48 three algorithm shows 95.5% of Specificity, meaning with this, the model would predict 95.5% of the times when a video would be "**partially**" blurred, therefore the second option is better when looking for negative patterns.

6. COMMENTS

A 10-scale output class showed underperforming results. These were discarded with a maximum accuracy of 56%. After rescaling the 10-scale table into a binary one, the results of the final model show how different algorithms can successfully perform with 2 different sets of attributes.

Final results show how accuracy varies from 91% to 94% with two different attribute sets. It is very important to remark on this case that more accuracy does not mean a better model, as sensitivity and specificity are different both cases meaning with this, each of the algorithms is better depending on what is looking for.

This investigation indicates it is more likely to recognize patterns and detect "successfully" or "unsuccessfully" blurred videos using a "yes", "no" binary model, than a 1-10 ranking method.

7. DISCUSSION AND CONCLUSION

7.1. OVERVIEW AND RECOMMENDATIONS

The system described in the sections above tried to solve two specific problems:

- Anonymizing sport videos.
  - In this phase many limitations arose when using EmguCv and Accord Libraries. The initial guess was, close ranges and static videos had a better success rate.

- Predicting whether these anonymizations were "fully" or "partially" blurred.
  - On a second phase, outputs were tackled from a Data Analytics perspective aiming to provide a prediction of the quality of the results based on properties from the inputs. Research and tests on this area confirmed the first assumptions and highlighted the impact that some attributes had on the produced outputs.

The points below overlay the steps followed with a production scenario, in order to identify areas open for improvement and automation:

AREAS FOR IMPROVEMENT

INPUTS GATHERING FACE

Manual process followed in this project would be easily replaced by an agent uploading them to the system. System space could be a constraint on a production scenario depending on the amount and frequency of inputs.

INPUT PROPERTIES-ATTRIBUTES EXTRACTION

This step was carried out manually. For a full automated system, this could be one of the most challenging problems to solve, although no more than a quick video inspection is required to fill this information.

On a full integrated system, Microsoft Cognitive Services can provide this information for specific frames, although it is

not granted any snapshot data from the video will remain constant.

VIDEO INPUTS PROCESSING

SYSTEM DESIGN

The component in charge of this has been presented as an executable file as briefly detailed in first sections. Recommendation here is a system that will allow uploading these inputs to the server file system. Basic UI is recommended for this step. Once the file has been successfully uploaded, a service "*listening*" to this folder should automatically process the video and produce the output on a different folder.

LIBRARY LIMITATIONS

The points below elaborate a set of limitations detected when processing the videos on the first phase of the project:

- It has been observed height camera angle has a high impact on accuracy, as higher angles present lowest accuracy when detecting faces. This can be observed on video 26. The video is recorded using different heights. E.g. second 2:09 (higher) vs 2:27(medium). Both sequences show how the referee has a lower accuracy ratio when the camera is higher than when the camera is recording at the same height than the head.

- Faces close to 45 degrees angle or close to this angle show high rates of error. Therefore, another technique is recommended for these scenarios, such as full body recognition and partial anonymization.

- The library also shows difficulty to recognize faces when arms or other elements are partially interfering with the head. This can be observed on tennis videos when players are serving. This indicates this specific problem should be addressed using a different approach.

- The system also shows difficulty to recognize faces on some scenarios where people are wearing head protections.

- A face looking down won't be detected depending on how tilted the head is. E.g. video 48, second 00:39. When played at slow motion it can be appreciated the face of this player is only recognized when he raises his head up facing the camera.

- Focus is obviously another element impacting the accuracy. Faces stop being recognized when faces are shown as unfocussed or blurred.

Main recommendation here is adding body recognition that will allow the system to partially anonymize faces based on size calculations. This would positively affect the accuracy of the outputs and minimize the scenarios above happening, although would also have a negative impact on the time required to process one single file.

PERFORMANCE

The presented tool iterates each of the frames and process them one by one. On these premises, there is also potential room for speed optimization by parallelizing frames computation.

PREDICTING OUTPUTS ACCURACY

This step is done through WEKA, although there is wide range of options for full integration. Python or C# are recommended options to build a component that would complete the automated pipeline.

7.2. CONCLUSIONS

Given the initial problem, this document sets the base ground for its resolution, and provides an architectural option while analyses the outputs generated. It also identifies areas with limitations and other points open for improvement and research. It is important to remark this project does not seek to develop a software system that anonymizes faces, but to demonstrate the feasibility to develop one based on the provided specifications, limitations and time constraints.

It also has been detected room to improve the accuracy obtained on the anonymization process. Recommendations section above provides guidelines and specific instructions in this area.

The model tries to predict accuracy on video outputs based on properties extracted from them. The library however, has defined limitations (exposed on the "Library Limitations" sub-section) that are not factored into these attributes. Therefore, whether there is any pattern affecting the accuracy of face recognition algorithms can get diffused with boundaries and limitations of this library. When overlaying this point with a real production case, a business analyst should draw a line at some point between "library limitations" and patterns that may have an impact on accuracy. Further open discussion on this topic is suggested.

High accuracy, sensitivity and accuracy obtained from the model generated may suggest there is a clear problem to be tackled and analysed in depth, but also confirms the importance that basic properties, such as closeness or movement of the camera have when working with face recognition algorithms.

7.3. ETHICS

Face recognition is a global field where word wide governments invest millions in an attempt to isolate the perfect model, and in many cases aiming to have full control of the data. Although this project tackles a problem from an identity preservation point of view, it has a wide range of applications out of the NCCA´s specific business case, as nowadays identity is one of our the most valuable assets we have as members of our society.

8. ACKNOWLEDGEMENTS

The present project is elaborated as a final year research project of Data Analytics on ITT Tallaght, Dublin, Ireland, under the supervision of Keith Quille, AI lecturer.

9. REFERENCES

TEAM, T. B. D. 'Kodak Facebook Collage Project'.

MOHAMMED, S. A. A. (2015) 'RECOGNITION OF FACE DETECTION SYSTEM BASED ON VIDEO'.

Omar, S. H. A. (2016) 'BIOMETRIC SYSTEM BASED ON FACE RECOGNITION SYSTEM'.

Emgu CV: OpenCV in .NET (C#, VB, C++ and more). Available at: http://www.emgu.com.

Weka - Data Mining with Open Source Machine Learning Software in Java (no date). Available at: https://www.cs.waikato.ac.nz/ml/weka.