Final Presentation - PDO

”Developing Machine Learning Methods to Automatically Geo-Localise Storybooks Artwork”

Anastasia-Valentina Tomita
Registration number: 1801722
Supervisor: Dr. Shoaib Jameel

Project Topic

- Modifying a certain book in order for it to be accessible to different parts of the world is nothing new
- When it comes to illustrations, this process gets more complicated, and it is often overlooked due to its complexity

Project Topic

- This is what my project tries to solve, to bring the first steps towards a **geo-localised image translation algorithm**
- We target children’s books/illustrations, more specifically reading-for-pleasure books

Project Motivation

- A very large proportion of the global population does not have access to **educational material originally created in their mother language**
- This **creates inconsistency** between how children perceive their world and the fictional world a book creates for them
- Thus, they **cannot relate** to the story, making the text understanding much more difficult

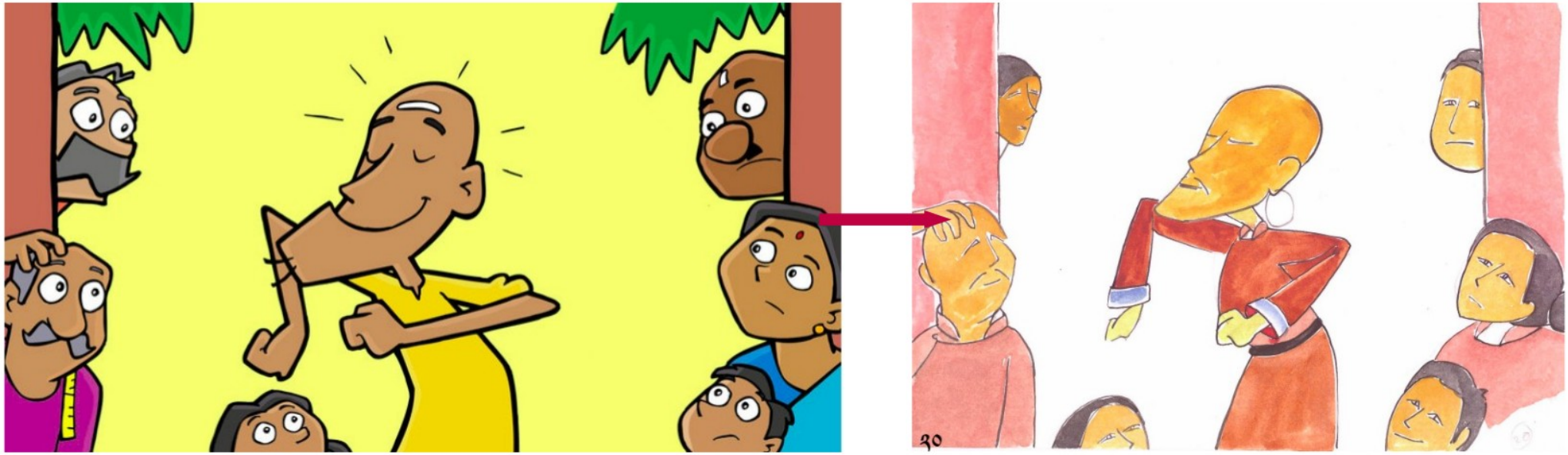
Project Motivation

- Reading constitutes the backbone in each individual's development
- Books expand children's vocabulary early on and help along the process of cognitive development, while also building knowledge about the world and about human morality

Manual Geo-Localisation

from:

to:



- Geo-Localisation is being currently done only manually, which is very time consuming and it is dependant on the skills of a visual artist

Capstone Project Goal

- For the purpose of this project, the emphasis will be put on the **human instances** alone
- They require a more complex process due to them being the principal **emotion transmitters** within any given frame

Capstone Project Goal

- Is to automatically find the best suited replacement for a certain object of class “Human” within a given frame

Image Segmentation

- In order for the process of image translating to be possible, the first step would be **Instance Segmentation**
- Each different object within the inputted frame should first be detected, categorised and delimited through the use of deep learning models.
- For each detected instance, a **segmentation mask** would be generated which would conclude in a pixel-level separation of the given object

Image Instance Segmentation

Replacing Human Instances

- There would be the following factors that would determine if the replacing instance is suitable:
  - Age
  - Gender
  - Facial Emotion
  - Body Position

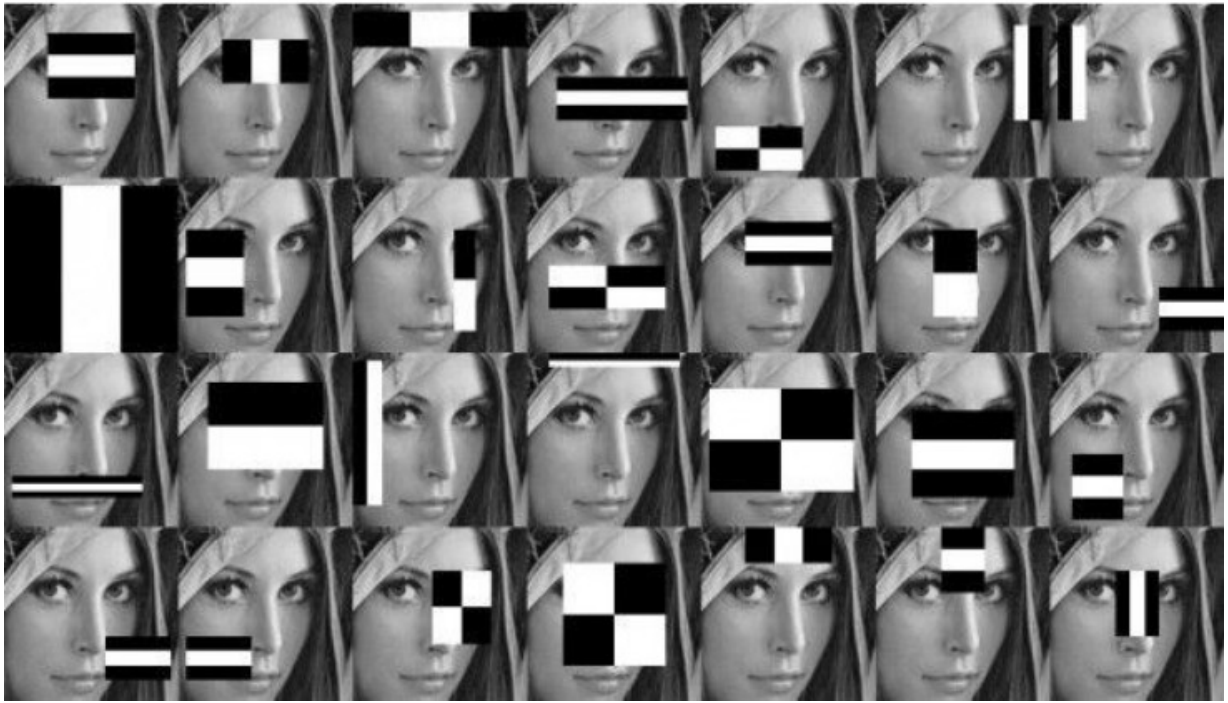
Facial Emotion Detection

- Emotion is a crucial aspect of any animation, as it is the principal method in which children's **emotional awareness** is being strengthened
- It is imperative that the transmitted emotion **is being maintained** as much as possible, as to not confuse the young readers

Facial Emotion Detection

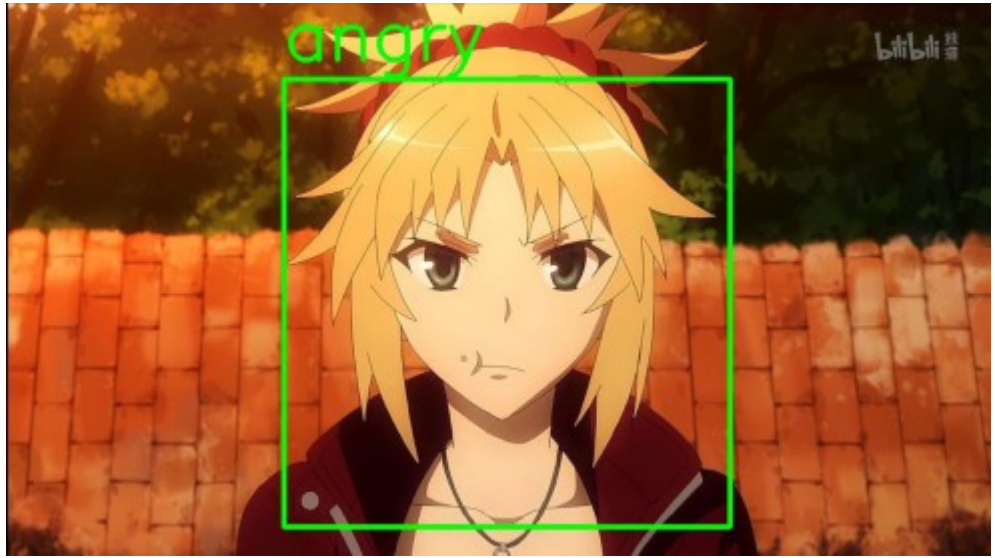
- I have started building the emotion detection module by using Haar Cascades for face detection
- A series of Haar features are being applied, in a manner similar to convolutional kernels, in order to detect certain facial components

Facial Emotion Detection

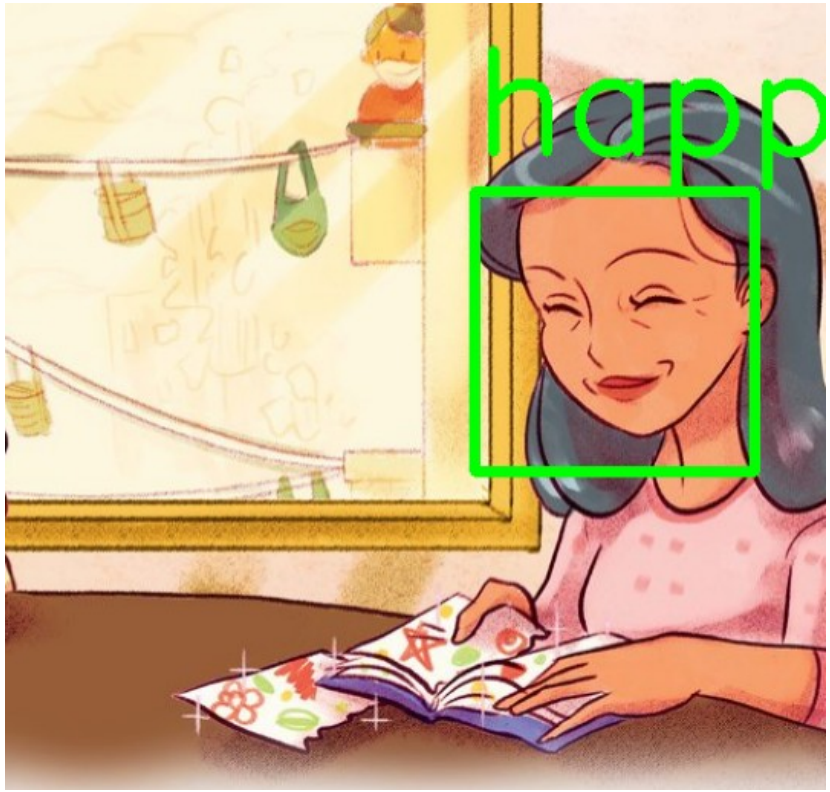


Facial Emotion Detection

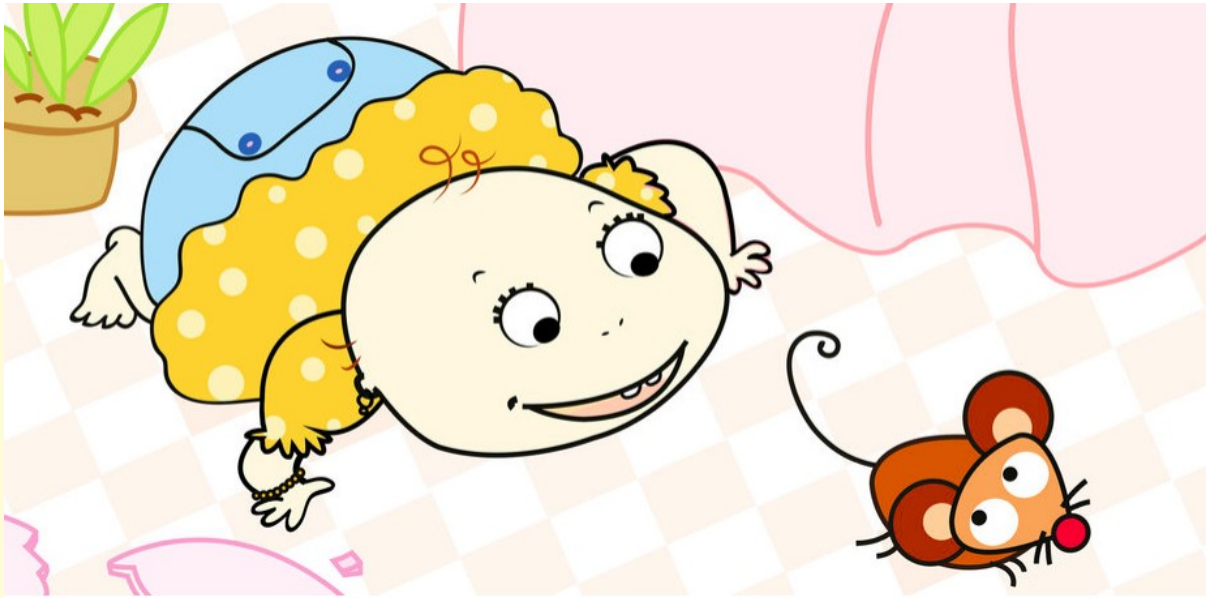
- However, deep learning techniques such as DeepFace [3] proved to have better accuracy on cartooned characters. [3] is a framework created by Facebook which is widely used for emotion detection



Facial Emotion Detection

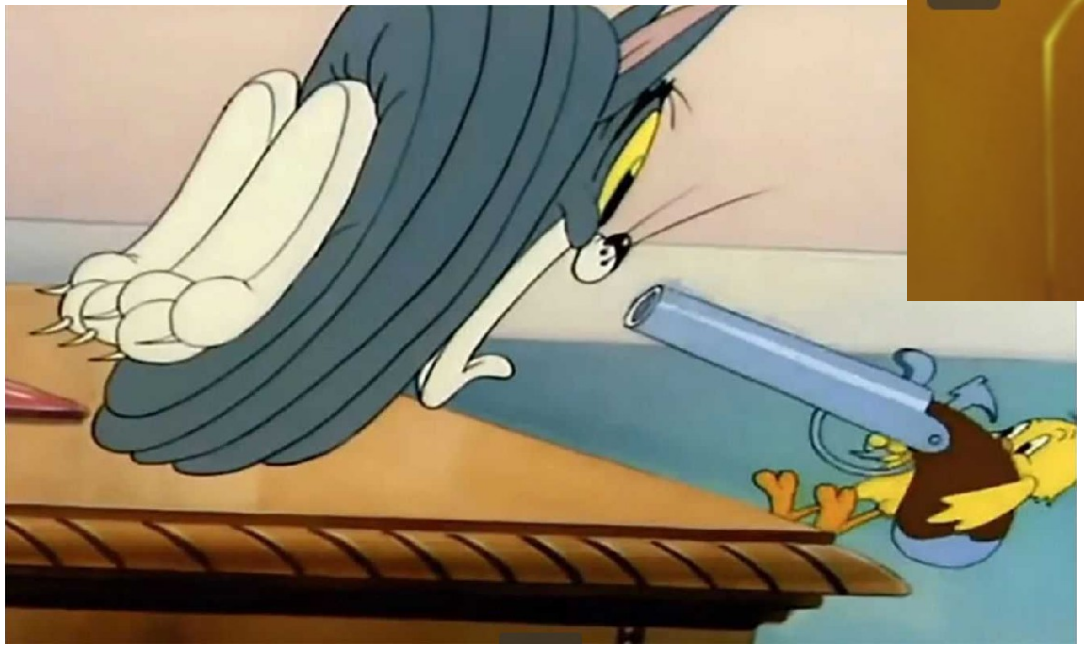


Challenges



```
resizedImage17.png is being processed
Action: emotion:   0%|
No face detected in image: resizedImage17.png
```

Challenges

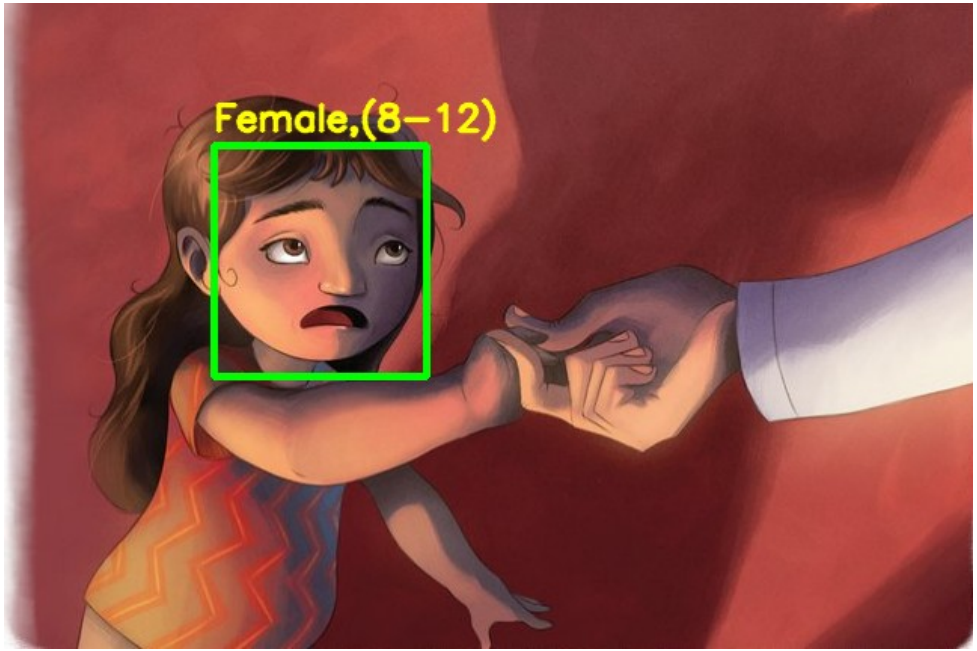


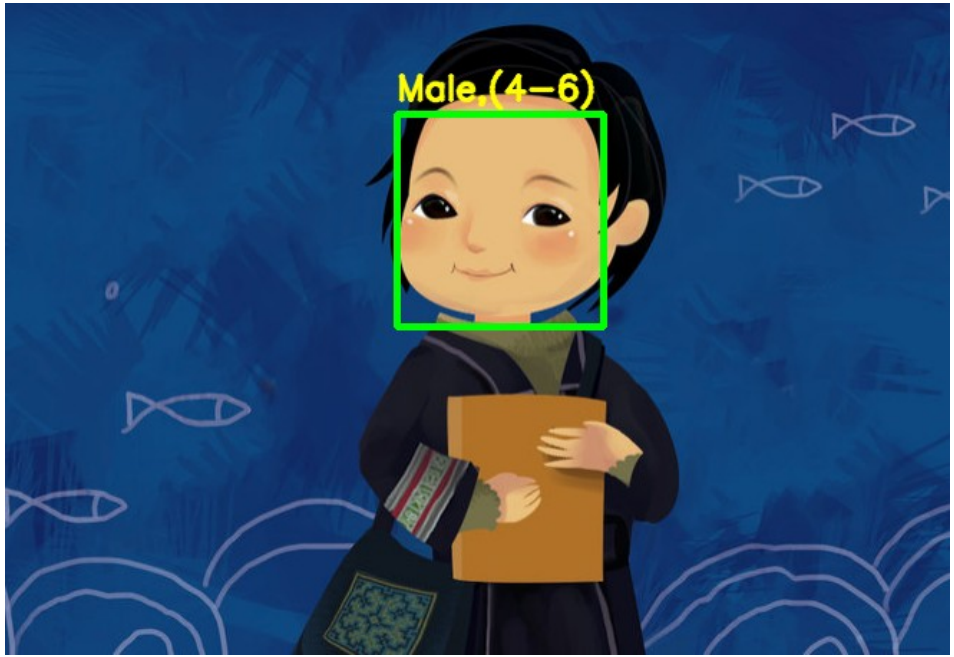
- Emotions are often very dependant on the overall information provided through the narrative thread.
- Facial features are being altered

Age and Gender Analysis

- Geometric modelling techniques are the most widely used for gender detection
- They utilize **taught facial landmarks models** to localise facial elements such as hair, nose or eyes. Such features are then further analysed in order to observe certain patterns which help categorize human faces as being either male or female.
- The same analysis is applied when determining a person's age, as ageing is a natural process which can be observed on any human being

Age and Gender Analysis





Body Position Module

- It is often the case that cartooned characters present very specific body positions in order to anticipate movement, which would otherwise be difficult to express within motionless frames without pose exaggeration

Body Position Module

- I have used OpenCV and a pre-trained Caffe model that won the COCO keypoints challenge in 2016 (shared in a paper entitled *“Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”* by Zhe Cao et. al. [4]) and the work presented on the LearnOpenCV website [5]

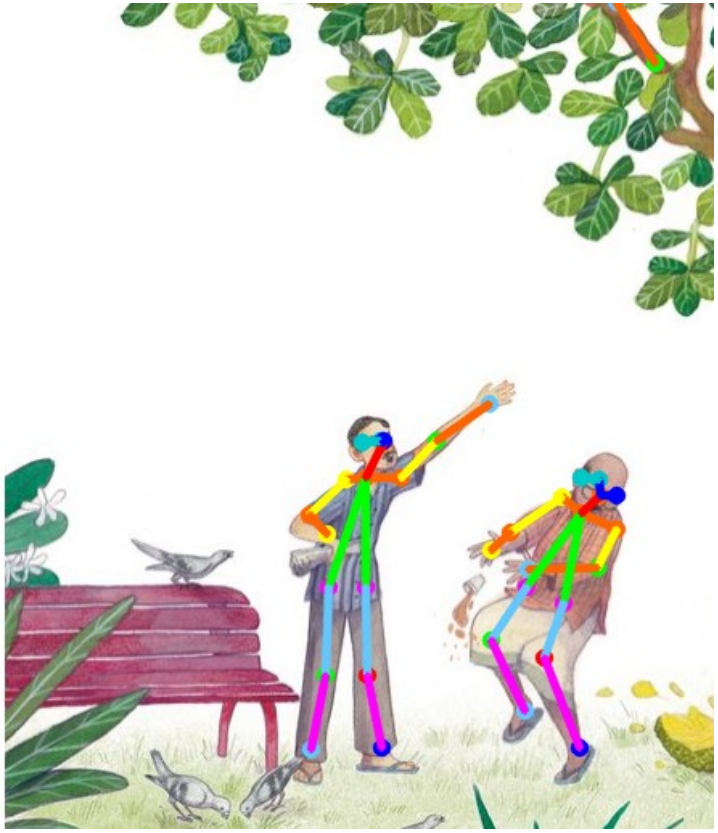
Body Position Module

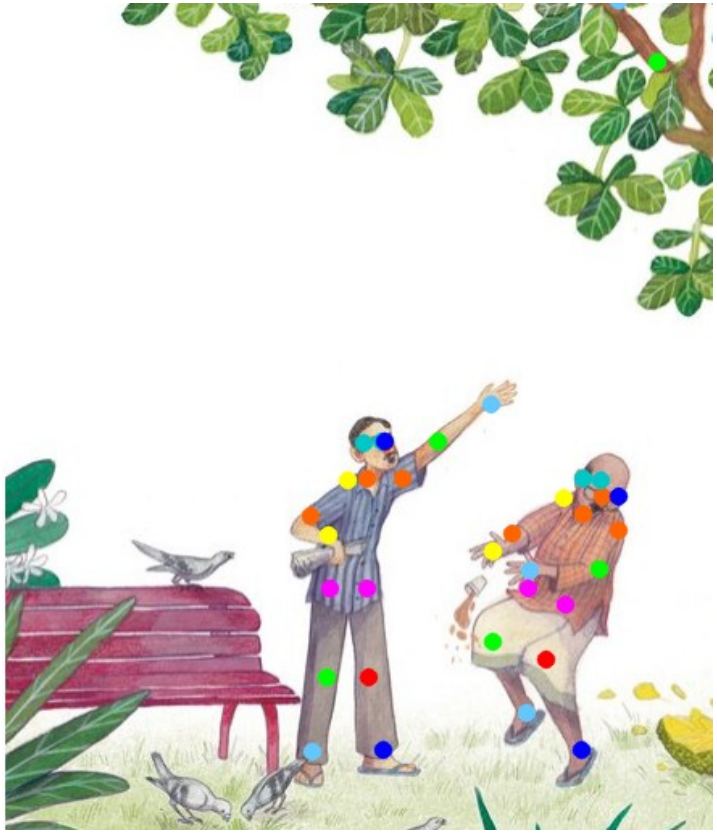
- The COCO model mentioned produces 18 points in the following format:

  Nose – 0, Neck – 1, Right Shoulder – 2, Right Elbow – 3, Right Wrist – 4, Left Shoulder – 5, Left Elbow – 6, Left Wrist – 7, Right Hip – 8, Right Knee – 9, Right Ankle – 10, Left Hip – 11, Left Knee – 12, Left Ankle – 13, Right Eye – 14, Left Eye – 15, Right Ear – 16, Left Ear – 17, Background – 18

Body Position – pose vs key-points

Body Position - pose vs key-points





Body Position Module

- I have not included these outputs when calculating the best-matching replacement as to not further complicate the translation task
- If one would want to compare two human instances through analysing their pose, **the angles** between the lines formed by connecting the 18 detected key-points need to be calculated and matched

Calculating Object Similarity

- After segmenting each human instance from the initial image, they are being analysed and compared with all the instances stored in the database, through calculating their **Euclidean distance**

Calculating Object Similarity

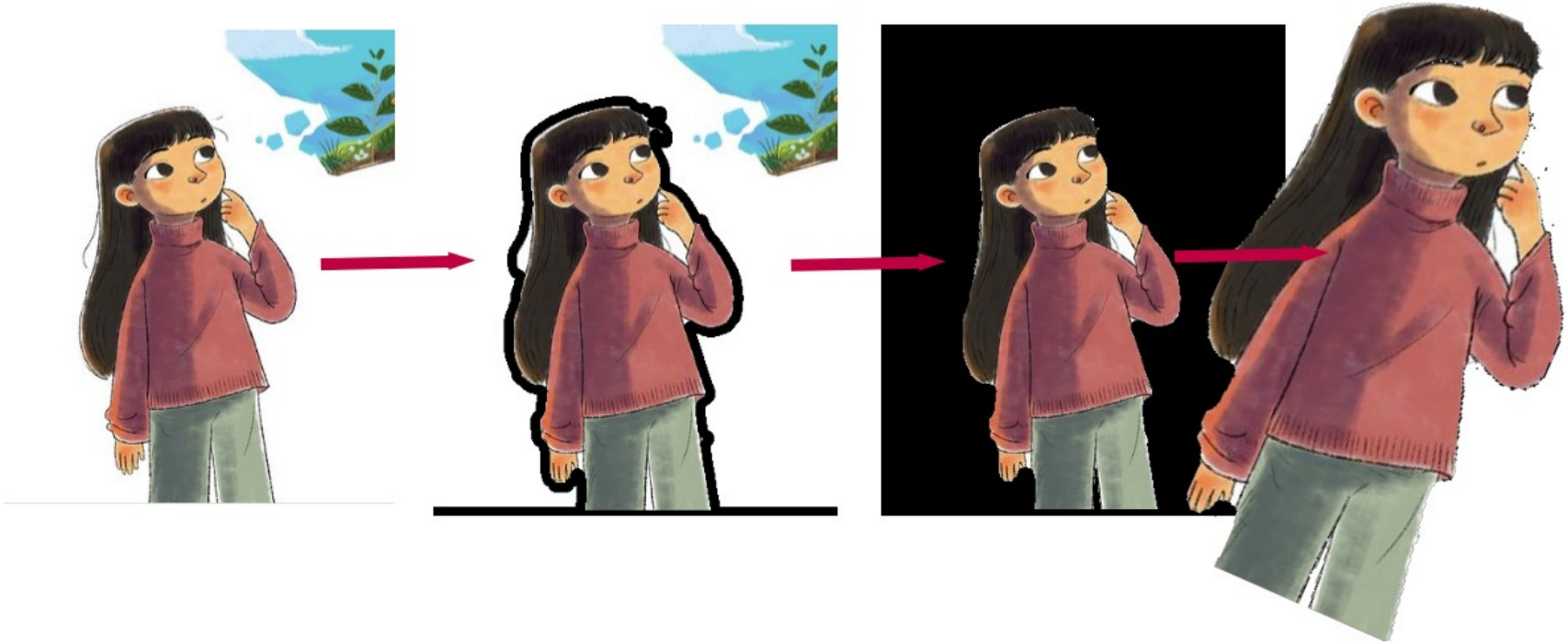
- Between each two instances, a **matching score** is being established, which shows their similarity
- This matching score is defined as:

$$ED = \sqrt{\sum_{e=1}^{7}(P_{e_1} - P_{e_2})^2 + (A_1 - A_2)^2 + (G_1 - G_2)^2}$$

Image Compositing

- Compositing techniques have been designed in order to combine visual elements from several sources such as if they belonged to the same scene.
- In order to do so, we need to **extract the chosen replacement**

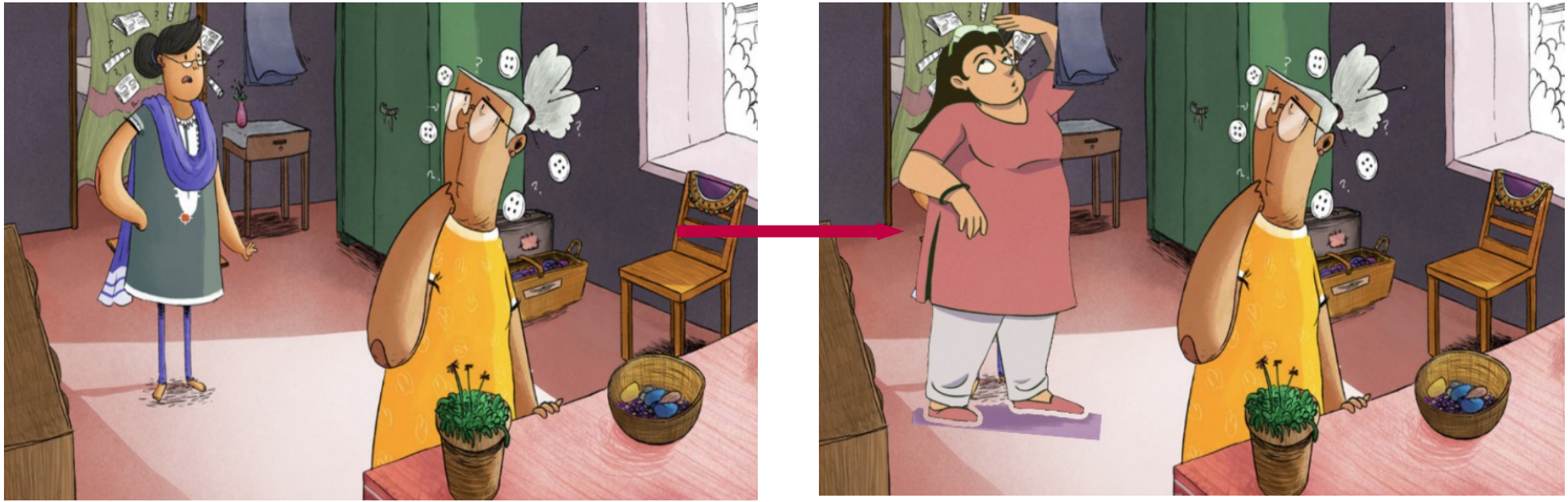
Image Compositing



Placing the object

- In order to accurately place the object at the right location within the frame, the detected position of the segmented mask is being used, alongside with its width and height.
- Its shaping is used to detect the resizing factor that needs to be applied to the replacing object, as to make it better fit within the frame’s context.

Placing the object



Level Of Abstraction

- Children’s illustrations tend to have a **higher level of abstraction**
- Pre-trained deep learning models have only been trained on datasets which depict the real world, without any level of abstraction (such as COCO)
- Because of this, these models don’t have a predictable behaviour
- In order mitigate this issue, image styling has been used

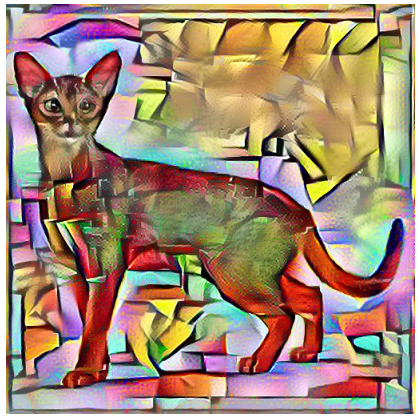
Image Style Transfer

- Style transfer is a technique that takes two images (a ***content image*** and a ***style reference image***) and blends them together so the output image looks like the content image, but having the style of the style reference image

Image Style Transfer

+

results



+

results

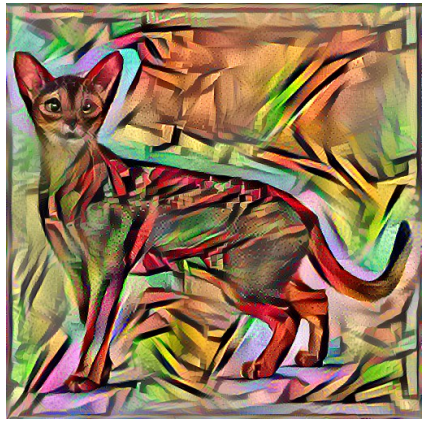
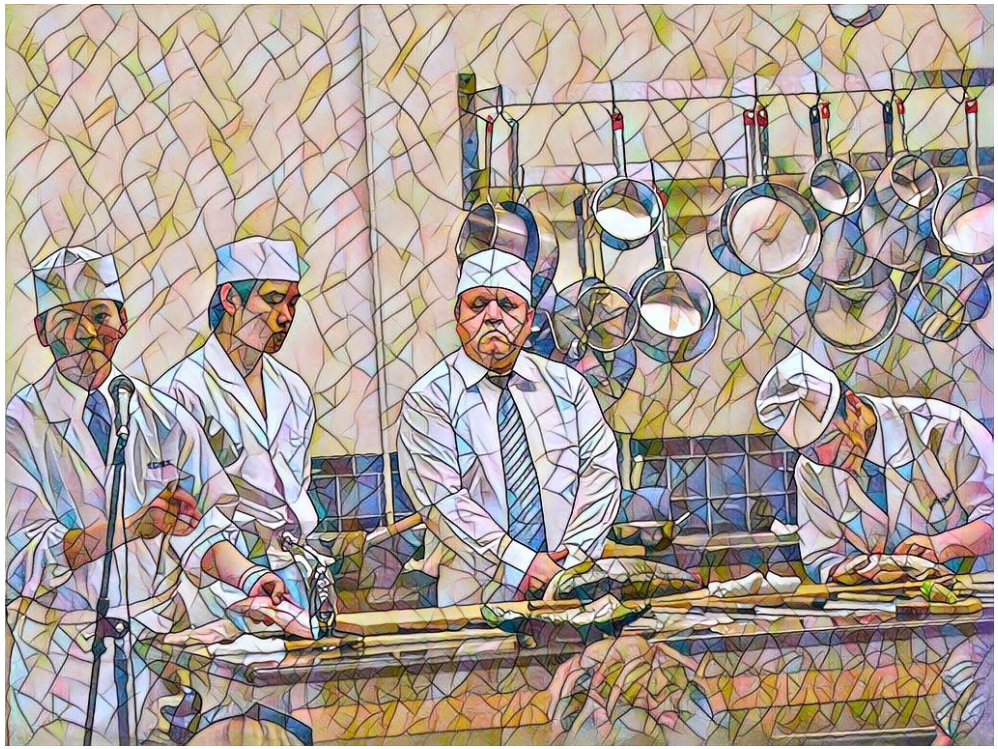


Image Styles

- The two styles I have chosen are as follow:
  - One style that introduces more abstract shaping and more rough edging to the image
  - Another style that makes the image appear drawing-like

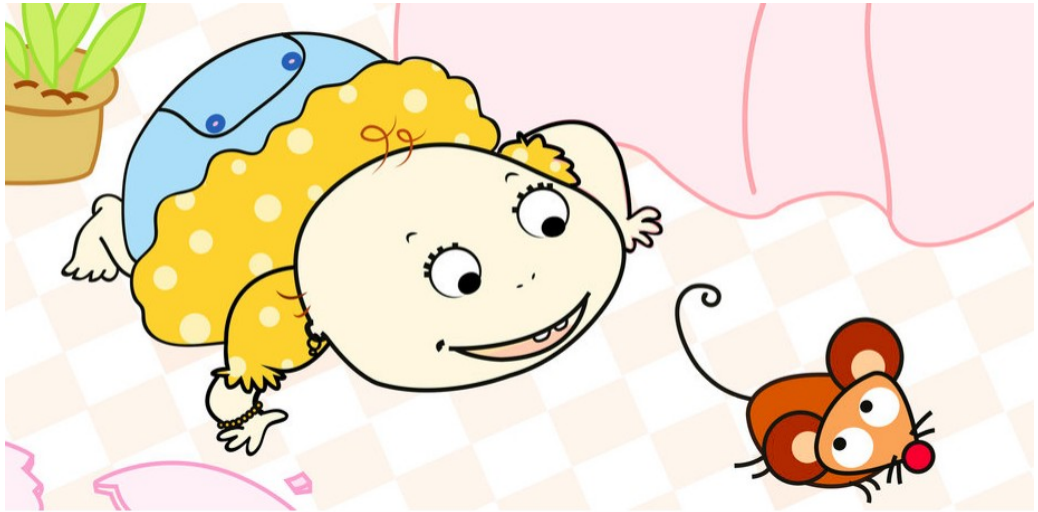
Style 1 VS Style 2

Style 1 VS Style 2

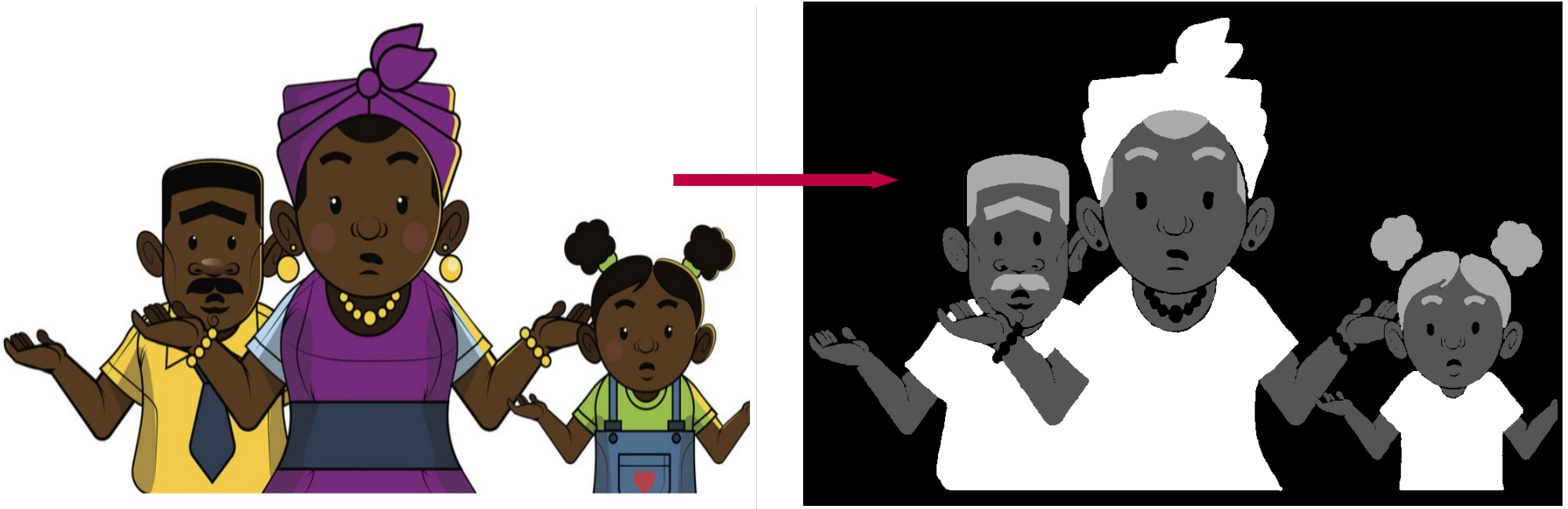


Weakness of the Styling Process

- Through styling we maintain the shaping of the objects/people within the content image



Fine Tuning



Obs: masks done for semantic segmentation - will be changed for instance segmentation

Fine tuning

Future Work

- Creation of a **cartooned dataset** containing various image styles, alongside with their annotations
- Sequential image handling – maintaining object identity throughout consecutive frames

Project Risks

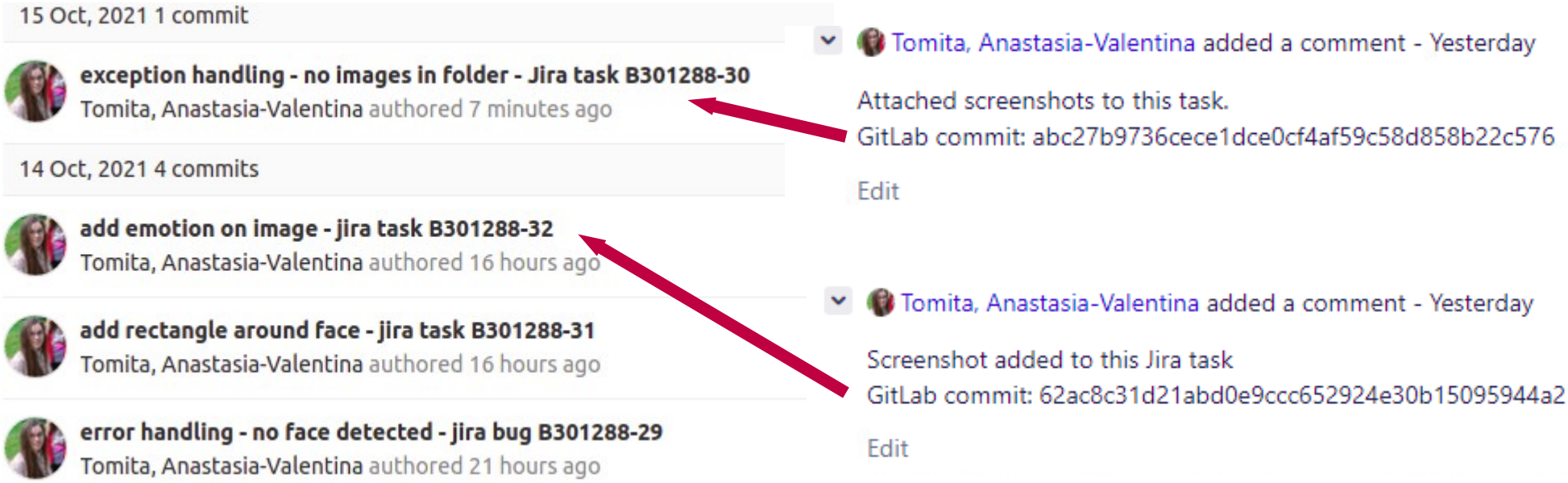
- The biggest risk is related to the level of abstraction that the illustrations contain (Jira risk B301288-27). All models are trained on real scenarios/images, thus not always predicting accurately what an abstract drawing represents

Project Risks

- Not being able to find an already designed **dataset containing children's illustration** and not having the resources and time to build a dataset from scratch (Jira risk B301288-38)
- As believed in the AI community: *"Every problem is just a good dataset away from being solved"*
- In order to mitigate this issue: the use of **Neural Style Transfer** [2] to give standard datasets a more artistic style

Project Management

- Tools that have been used throughout the project: **Jira & GitLab**
- Jira tasks and GitLab commits are being linked together



Project Objectives

- The project lies within the third category below: *the use of software packages to deliver a product*

Project objectives Clear formulation of the Project Objectives, Implementation plan, risks, context. Specific deliverables to be achieved by the oral interview in Week 11 must be agreed with your supervisor. Suitable evidence must be uploaded in Jira/Gitlab. The product may be hardware (e.g. electronic device), software (e.g. game, app, website, database), or the use of software packages to deliver a product (e.g. virtual network with security, device simulation). It is important that you describe the goals in a manner that gives confidence to their being achieved. All code for a project and the technical documentation must be held in the CSEE GitLab repository. Do not worry about how to develop the technical documentation within GitLab. This will be explained once the Autumn term has started. This week 11 interim oral will be undertaken by the second assessor. S/he will need access to your

Source: Moodle - CE301 Project Handbook

References

- [1] K. He, G. Gkioxari, P. Dollar, and R. Girshick: “Mask R-CNN,” in Proc. IEEE ICCV, 2017, pp. 2980–2988
- [2] L. Gatys, A. Ecker, and M. Bethge: "A neural algorithm of artistic style". Nature Communications, 2015
- [3] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf: “Deep-Face: Closing the gap to human-level performance in face verification”. In Proc. CVPR, 2014.

References 2

- [4] Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7291-7299).
- [5] *https://learnopencv.com/deep-learning-based-human-pose-estimation-using-opencv-cpp-python/*
- Presented images are selected from https://storyweaver.org.in/, a digital repository of children's stories designed by Pratham Books

Thank you for your time!