Advanced Computer Vision Summative Assignment 21-22

rtjm84

Note: The following study used Google Colab for implementation and experimentation. As a result, memory and usage was limited. This primarily affected the number of training epochs, models were limited to 12 epochs in this study with the learning rate decaying halfway through, and the resolution of the model outputs.

1 Human Feature Analysis

1.1 Human Patch Extraction

A pre-trained Mask R-CNN [6], Fig. 1, has been implemented for patch extraction where "person" items detected with a confidence greater than 0.965 are retained. Originally, the mask of the person was saved in the dimension of the full scene, this was later changed so that the patch extracted instead had the dimension of the bounding box surrounding the item, cut from the entire frame. This new extraction technique was reasonably accurate and allowed for an easier combination with other model outputs in the full pipeline. The disadvantage of this approach is that overlapping bounding boxes will obscure each other.

1.2 Classification

Pose estimation, using OpenPose [1], involves mapping detected keypoints to body parts. If facial features aren't detected, an *other* classification is given, otherwise, if knees are detected then classify as *full body*. For these instances, the knee-hip gradient estimates membership to *full-body sitting* or *standing*. Alternatively, if hips are detected then classify as *half-body* and *head-only* otherwise. This logic allows for fast computation and an accuracy of 45%, as seen in Table 1. Misclassifications occur from baggy or dark clothing affecting keypoint detection, resulting in only 4 Movie *full-body standing* classifications. An ML approach could tune the confidence parameters used.

Figure 1: Mask R-CNN Framework for Instance Segmentation [3]

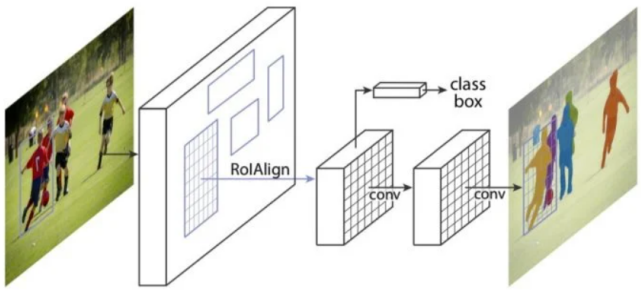


Table 1: Pose Classification Accuracy over 100 Sample (analysis was conducted manually over 50 samples in each of the 2 domains)

		Actual				
		Full-body Standing	Full-body Sitting	Half-Body	Head-only (profile)	Others
Predicted	Full-body Standing	13	1	0	0	0
	Full-body Sitting	16	4	0	0	0
	Half-Body	5	2	13	0	0
	Head-only (profile)	4	0	8	8	0
	Others	3	2	7	1	7

1.3 Training Data Selection

The average RGB values across all patches is found, and then differences between each patch's individual average RGB values and the global average are used to strategically sample 50% of the patches. This means that similar patches are less likely to be selected whereas patches with more variation from the mean have a higher probability as they can provide new data. It is believed that this is a novel approach. Additionally, images are eliminated from the training data if they are smaller than 128 in either dimension as other images will have more data.

1.4 Training Data Augmentation

First, a 50% probability of a horizontal flip reduces the likelihood of similar images. Next, a 20% probability of rotation by a maximum of 10 degrees increases variation. Additionally, a 50% probability of a contract limited adaptive histogram equalization (CLAHE) is implemented. CLAHE affects the image contrast and aims to benefit the model's capability of recognising textures instead of just colours. Colour altering augmentations were not implemented so as to not affect the styles learned. A probability of random cropping was initially implemented but then removed as it reduced the likelihood of full body classifications when combined with section 1.2.

2 Style Transfer

2.1 Model Deployment

A pre-trained horse2zebra CycleGAN [8] style transfers using the entire frames. The loss used in the cycleGAN is seen in Eq. 1-3 [7]. Results are seen in Fig. 2. The game-to-movie transfer performed well but introduced a few artefacts. The movie-to-game results are less accurate, potentially due to there being more movie styles than game style. It should be easier to map a smaller source distribution across a variety of target styles than it will be to condense a large source distribution to only a few. The horse2zebra architecture may be better suited to smaller distributions indicating the domain-shifting problem.

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda \mathcal{L}_{cyc}(G, F) \tag{1}$$

$$\mathcal{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[log(D_Y(y))] + \mathbb{E}_{x \sim p_{data}(x)}[log(1 - D_Y(G(x)))] \tag{2}$$

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{y \sim p_{data}(y)}[||G(F(y)) - y||_1] + \mathbb{E}_{x \sim p_{data}(x)}[||F(G(x)) - x||_1] \tag{3}$$

2.2 Model Fine-tuning

This solution's results are frames whose backgrounds have been acted upon by the 2.1 model, and whose people have been developed by a separate general human model trained on the patches extracted in section 1.1. Example results can be found in Fig. 3. The model performs well with a clear attempt at style transfer, for example, the blue shirt in Fig. 3(c)(d). The addition of the human patches does introduce more artefacts, predominantly at the bounding box perimeter. This could potentially be resolved by introducing a slight Gaussian blur along this border [5]. There is also, as expected, some issue with colliding patches which would be removed if masks were extracted instead of bounding boxes. In general, the consistently reasonable outputs of this approach indicates that it is more suitable than the one proposed in section 2.1. The quality is noticeably improved with less variation between frames.

2.3 Part-based Style Transfer

Unlike section 2.2, patches here are filtered and augmented using the approaches in sections 1.3 and 1.4. They are classified, see section 1.2, and used to train a model for each of the 5 poses. When testing, the appropriate pose model is chosen based on the patch and overlays the results with the background results from section 2.1. This approach performed well at times (see Fig. 4 (e-h)), but also performed very differently depending on the pose-specific model implemented, leading to varying patch quality within multi-person frames (see Fig. 4 (a-d)). This is due to the pose models having different, and reduced, amounts of training data.

An implementation bug affected classifications given and so, for some patches, section 2.2's general model was used. With this and the quality of results in consideration, it is therefore reasoned that this approach is less suitable than the one proposed in section 2.2.

3 Real-world Application

3.1 Results Analysis

The general human model was used to create a video with style-transferred humans. There are clear boundaries around patches which could be blended using a Gaussian blur. The mask R-CNN error also results in a flickering effect which would be less noticeable with the initial mask extraction. Colliding patches fragment multi-person frames.

The approach generally performs well, particularly in frames where human patches are smaller and so resolution differences are less clear. Visual artefacts are minimal, only typically appearing on patch borders, and the style seen is similar to the beige and blue commonly seen in the movie scenes.

3.2 Pipeline

If the game could be rendered without lighting there wouldn't be conflicting light sources between domains and so may allow for a more accurate transfer of style and colours. Game meshes could improve pose classification accuracy by either using section 2.2's logic or feeding keypoints into a neural network. Game textures could also be utilised in a game-movie texture transfer model [2]. A human patch pipeline would use meshes to classify pose and accurately extract the patch without lighting before conducting a texture transfer. Following this, a random CNN with short-time fourier transforms could perform an audio style transfer [4].

References

[1] Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." Proceedings of the IEEE conference on computer vision and pattern recognition. (2017)

[2] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. "Texture and art with deep neural networks." Current opinion in neurobiology 46 (2017): 178-186.

[3] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision (2017)

[4] Huang, Zhixian, Shaotian Chen, and Bingjing Zhu. "Deep Learning for Audio Style Transfer."

[5] OpenCV "Image Processing in OpenCV - Smoothing Images" *docs.opencv.org*, Available online at: `https://docs.opencv.org/4.x/d4/d13/tutorial_py_filtering.html` (Accessed: May 2022)

[6] TorchVision Implementation "Semantic Segmentation Mask R-CNN.ipynb" Available online at: `https://colab.research.google.com/drive/1bWLB3tmWv4XyJSu4DHtZO-b-n-DACSKx?usp=sharing#scrollTo=ZPChK3HBLdQq` (Accessed: May 2022)

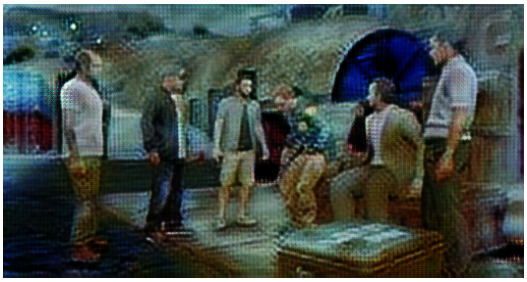
[7] Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." Proceedings of the IEEE international conference on computer vision (2017)

[8] Zhu, Jun-Yan, et al. "junyanz/pytorch-CycleGAN-and-pix2pix" *Github* Available online at: `https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix` (Accessed: May 2022)

(a) G2M Ex. 1 - Real

(b) G2M Ex. 1 - Fake

(c) G2M Ex. 2 - Real

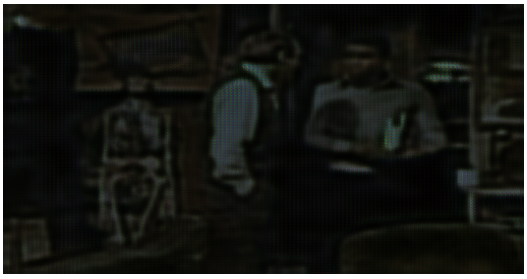


(d) G2M Ex. 2 - Fake

(e) M2G Ex. 1 - Real

(f) M2G Ex. 1 - Fake

(g) M2G Ex. 2 - Real



(h) M2G Ex. 1 - Fake

Figure 2: 2.1 Results (General Frame Model, G2M - Game-to-Movie, M2G - Movie-to-Game)

(a) Ex. 1 - Real

(b) Ex. 1 - Fake

(c) Ex. 2 - Real

(d) Ex. 2 - Fake

(e) Ex. 3 - Real

(f) Ex. 3 - Fake

(g) Ex. 4 - Real

(h) Ex. 4 - Fake

Figure 3: 2.2 Results (Combined Background and General Human Patch Models)

(a) Ex. 1 - Real

(b) Ex. 1 - Fake

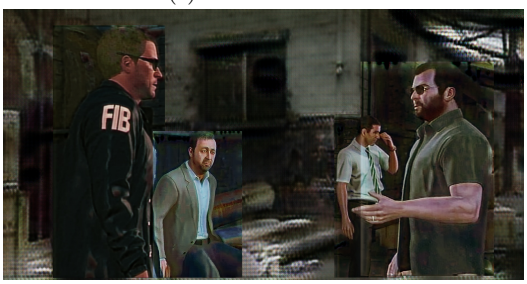
(c) Ex. 2 - Real

(d) Ex. 2 - Fake

(e) Ex. 3 - Real

(f) Ex. 3 - Fake

(g) Ex. 4 - Real



(h) Ex. 4 - Fake

Figure 4: 2.3 Results (Combined Background and Specific Human Pose Models)