Relatable Clothing: Detecting Visual Relationships between People and Clothing

Thomas Truong, Svetlana Yanushkevich
Department of Electrical and Computer Engineering

Abstract

Detecting visual relationships between people and clothing in an image has been a relatively unexplored problem in the field of computer vision and biometrics. The lack of readily available public dataset for ``worn'' and ``unworn'' classification has slowed the development of solutions for this problem. We present the release of the Relatable Clothing Dataset which contains 35,287 person-clothing pairs and segmentation masks for the development of ``worn'' and ``unworn'' classification models. Additionally, we propose a novel soft attention unit for performing ``worn'' and ``unworn'' classification using deep neural networks. The proposed soft attention models have an accuracy of upward 98.55% $\pm$ 0.35% on the Relatable Clothing Dataset and demonstrate high generalizable, allowing us to classify unseen articles of clothing such as high visibility vests as ``worn'' or ``unworn''.

Relatable Clothing Dataset

- There were no datasets containing detailed instance segmentations and visual relationship classes for ``worn'' and ``unworn'' detection of clothing before Relatable Clothing.
- Relatable Clothing is a modified version of DeepFashion2 [1].
  - Mask R-CNN trained with Open Images V6 [2,3] to provide person instance segmentations.
  - ``unworn'' samples added to images with ``worn'' samples.
- 29,852 person-clothing pairs for training
- 5705 person-clothing pairs partitioned into 10 folds for validation and testing.

(a) (b) (c) (d) (e)

Figure 1: Example samples from the Relatable Clothing Dataset. From left to right: a) the person mask, b) the image, c) the first ``unworn'' article of clothing, d) the first ``worn'' article of clothing, e) the second ``worn'' article of clothing

Soft-Attention Unit

- A trainable unit which guides the "attention" of the network to the areas containing masks.
- Input subject-clothing pair masks are added together and then concatenated before being resized.
- The concatenated masks are then passed through a convolutional filter block to produce the soft-attention unit output.
- The soft-attention unit output is added to the output of the 3x3 convolutional layer of each bottleneck unit in ResNet [4].

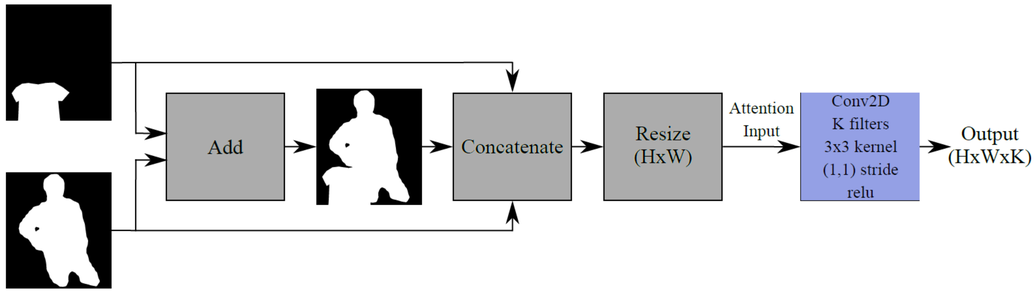


Figure 2: Proposed Soft-Attention unit used for detecting ``worn'' and ``unworn'' articles of clothing given a subject-clothing pair.

Results

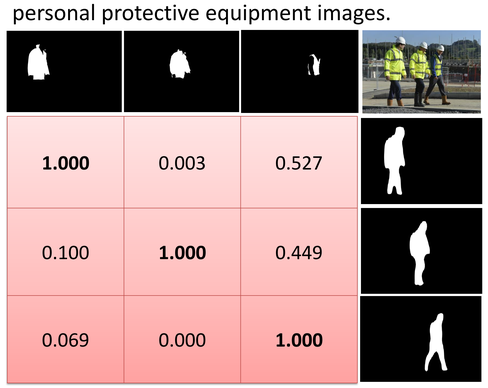
Table 1: Summary of results using the proposed soft-attention unit on the Relatable Clothing dataset validation set.

Backbone	Soft Attention Units	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F_1 (%)
ResNet50V2	1	96.00 ± 1.03	98.79 ± 0.56	94.83 ± 1.41	97.98 ± 0.96	96.76 ± 0.85
ResNet50V2	16	97.74 ± 0.40	97.76 ± 0.61	**98.66 ± 0.46**	96.17 ± 0.87	98.21 ± 0.36
ResNet101V2	1	97.97 ± 0.63	98.96 ± 0.33	97.79 ± 0.98	98.24 ± 0.49	98.37 ± 0.54
ResNet101V2	33	**98.55 ± 0.35**	**99.16 ± 0.40**	98.52 ± 0.50	**98.58 ± 0.65**	**98.84 ± 0.29**

Applications

As an early proof of concept, we test our best detector on unseen personal protective equipment samples from the Personal Protective Equipment dataset from Benedetto et al. [5].

Table 2: ResNet101V2 predictions on unseen personal protective equipment images.



1.000	0.003	0.527	
0.100	**1.000**	0.449	
0.069	0.000	**1.000**	

Conclusions and Future Work

We proposed a novel soft attention unit for detecting worn and unworn clothing given person and clothing mask priors. In addition, we release a dataset for ``worn'' and ``unworn'' clothing detection titled Relatable Clothing. We achieve the best performance using a ResNet101V2 backbone with 33 of our proposed soft attention units. This model achieved 98.55% ± 0.35% accuracy, 98.58% ± 0.65% specificity, and 98.84% ± 0.29% F_1 score on the Relatable Clothing Dataset. We also present some promising use cases in workplace safety through qualitative results on a small high-visibility vest dataset. In our future works we plan on expanding the proposed network to do end-to-end object detection and visual relationship detection. This involves detecting all person-clothing pairs in a given image and classifying ``worn'' and ``unworn'' for each pair.

References

[1] Ge, Yuying, et al. "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images." Proceedings of the IEEE conference on computer vision and pattern recognition. 2019.
[2] Benenson, Rodrigo, Stefan Popov, and Vittorio Ferrari. "Large-scale interactive object segmentation with human annotators." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[3] Kuznetsova, Alina, et al. "The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale." (2020).
[4] He, Kaiming, et al. "Identity mappings in deep residual networks." European conference on computer vision. Springer, Cham, 2016.
[5] Di Benedetto, Marco, et al. "Learning safety equipment detection using virtual worlds." 2019 International Conference on Content-Based Multimedia Indexing (CBMI). IEEE, 2019.

thomas.truong@ucalgary.ca