CSCI E-25 Final Project Proposal
Instance Segmentation for Fashion Apparel

An Hoang, Mark McDonald, Vivek Bhatia

1. Introduction

Recently, computer vision applications in the fashion domain have attracted significant attention. A historically challenging task has been to build an intelligent recommender system that can suggest items for purchase, based on a semantically rich notion of "personalized style". Despite the seemingly insurmountable complexity of this task, the lucrative end reward has enticed e-commerce behemoths like Amazon [1], Stitch Fix, and Pinterest [2] to participate in a race to develop recommender systems that can understand notions of style.

A recent paper describes a system that can take in "in the wild" scene images to generate complementary items recommendations [4]. However, their model training involves inefficient cropping heuristics that fail on certain clothing items, complicated attention mechanisms to detect each item of clothing in an image, and a manually-curated dataset. The authors remark that constructing appropriate ground truth data to learn the notion of compatibility is a significant challenge due to the amount of human effort to label and categorize.

Being able to segment clothing items and recognize fine-grained attributes would significantly reduce the burden of creating quality training datasets for researchers, resulting in a boost of performance for existing models. The recent introduction of the iMaterialist dataset [6] and Kaggle competition [7] has enabled this ambitious endeavour to materialize. A collaborative effort between the fashion and computer vision communities, these resources allow others to tackle this novel fine-grained segmentation task, which unifies both categorization and segmentation of rich and complete apparel attributes, an important step toward real-world applications.

To make progress towards building an intelligent fashion recommender system, we will first apply state of the art instance segmentation algorithms on this dataset, trying to isolate articles of clothing and their fine-grained attributes from scene images "in the wild". If we can fine tune the model to achieve satisfactory results, we will move on to the next phase of using these segmented and classified representations to train a visual-aware recommender system that can suggest complementary pieces of clothing when presented with a query in the form of an image scene.

2. Implementation Details

We propose to implement the automated fashion recommendation system in a two staged approach. The primary task for the recommendation system is to implement a robust framework for object detection and segmentation geared towards apparel objects from images of any size. The goal of this framework would be to detect and classify main apparel objects (e.g. shirts, pants, jackets, shoes, etc.) and apparel parts (e.g. sleeves, collars, etc.) from input images and to provide a pixel-wise segmentation masks for each instance of detected apparel object.

Many image segmentation techniques are discussed in the literature including region-based segmentation, edge detection, clustering and many others, however, traditional algorithms pose several limitations especially when segmenting images with multiple objects [8]. With the advent of artificial neural networks there have been several new generalized trainable models for object detection which can be retrained for detecting new object classes. We intend to implement our unsupervised segmentation framework by leveraging the existing Mask R-CNN architecture for object instance segmentation [9].

An initial challenge towards implementing the segmentation framework would be to train the existing model for detecting the apparel object categories annotated in the fashion training dataset discussed below. It is well known that training models from scratch is often challenging and time consuming. Our methodology would be to use the transfer learning approach using pre-trained weights for the Mask R-CNN model on the COCO dataset [10] and test the accuracy on test images in the fashion dataset. Since the COCO dataset was not trained on fashion-related items, some combination of transfer learning and fine-tuning of weights will be necessary. We do not anticipate that a full retraining of the model will be required.

Although Mask RCNN is the primary focus, we intend, time permitting, to compare and contrast this with other segmentation architectures (e.g. – FCN, PSPNet, UNet).

2.1. **Dataset: Kaggle Competition Data Source**

The dataset used for training and testing will be sourced from the Kaggle project that presented a fashion segmentation challenge [7]. The dataset includes ~50K training images and ~ 3K test images. Each image has been annotated to include main apparel objects e.g. jackets, shirts, pants,etc. Furthermore, each main object has been annotated with a sub-object e.g. pockets, sleeves, buttons, etc. Overall the dataset includes 45 categories of main and sub objects with a wide distribution of the occurrence rate for each category.

Additionally, each object has also been annotated with a fine-grained attribute e.g. material, style, etc. about 91 in total. The task of detecting and classifying objects based on fine-grained attributes is required for the Kaggle challenge but will be **outside the scope** of our project implementation. We will be limiting the segmentation task to main objects and sub objects only.

2.2. **Proposed Timeline and Effort matrix**

ID	Task Description	Timeline	Ownership
1	Dataset exploration and scrubbing	10/19 - 11/6	Team
	MILESTONE: FINAL PROPOSAL	11/1 - 11/6	Team
2	**Training Mask R-CNN model**	**11/4 - 11/17**	

2.a	Using pre-trained weights and fine-tuning for segmentation and testing	11/4 - 11/17	Vivek
2.b	Final Model Selection	11/17	Team
3	**Using trained model for segmenting image**	**11/10 - 11/30**	
3. a	Product based segmentation (single objects in images)	11/10 - 11/24	An
3.b	Scene based segmentation (multiple objects in images)	11/17 - 11/30	Vivek
4	**Suggestion algorithm for product-based query system**	**11/20 - 12/7**	
4.a	Develop algorithm and process	11/20 - 12/7	Mark
5	Extension of segmentation model to live images and videos	**11/20 - 12/7**	
5.a	Use segmentation algorithm for real-time processing	11/20 - 12/7	An
6	**Final Presentation and Documentation**	**12/10 - 12/17**	
6.a	Presentation	12/10 - 12/17	Mark
6.b	Documentation	12/10 - 12/17	An/Vivek
6.c	Final Presentation and Submittal	12/18	Team

3. Risk Management

The fashion segmentation challenge is a relatively new problem to be solved using deep learning. Based on the statistics provided in the Kaggle challenge, the accuracy for the currently implemented models remains low and an efficient solution is yet to be presented. Given the complexity of the problem and the limitation in time and resources for our project completion, we intend to tackle the project in phased approach.

The primary endpoint of the project would be to successfully detect and segment main apparel objects from product based still images e.g. catalog images of shirts, pants, jackets, etc. Once this is achieved, we will attempt to use the same model towards segmenting scene-based images e.g model wearing a complete ensemble of shirt, pants, shoes and hat.

4. References

[1] Krishnan, A. (2019, June 19). StyleSnap will change the way you shop, forever. Retrieved from aboutamazon.com

[2] Le, J. (2018, January 14). Pinterest's Visual Lens: How computer vision explores your taste. Retrieved from . *Medium.com*.

[3] Kang, W.-C., Fang, C., Wang, Z., & Mcauley, J. (2017). Visually-Aware Fashion Recommendation and Design with Generative Image Models. *2017 IEEE International Conference on Data Mining (ICDM)*. doi: 10.1109/icdm.2017.30

[4] Kang, W.-C., Kim E., Leskovec J., Rosenberg C., McAuley J.. Complete the Look: Scene-based Complementary Product Recommendation.

[5] Veit, A., Kovacs, B., Bell, S., Mcauley, J., Bala, K., & Belongie, S. (2015). Learning Visual Clothing Style with Heterogeneous Dyadic Co-Occurrences. 2015 IEEE International Conference on Computer Vision (ICCV). doi: 10.1109/iccv.2015.527

[6] Visipedia. (2019, June 26). visipedia/imat_comp. Retrieved from https://github.com/visipedia/imat_comp.

[7] iMaterialist (Fashion) 2019 at FGVC6. (n.d.). Retrieved from https://www.kaggle.com/c/imaterialist-fashion-2019-FGVC6.

[8] Kang W., Yang Q., Liang R. (2009). The Comparative Research on Image Segmentation Algorithms. IEEE Conference on ETCS. doi: 10.1109/ETCS.2009.417

[9] He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV). doi: 10.1109/iccv.2017.322

[10] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. *Computer Vision – ECCV 2014 Lecture Notes in Computer Science*, 740–755. doi: 10.1007/978-3-319-10602-1_48

[11]iMaterialist (Fashion) 2019 at FGVC6. (n.d.). Retrieved from https://www.kaggle.com/c/imaterialist-fashion-2019-FGVC6/overview/evaluation.