# Towards Automated Handling and Sorting of Garments combining Visual Language Models and Convolutional Neural Networks

Serkan Ergun[1†], Tobias Mitterer[1†] and Hubert Zangl[1,2]

***Abstract*— Ambitious goals set by the European Union are aiming towards full recycle-ability of garments by 2030. According to the EU, 12 kg of garments are discarded by each citizen per year. In order to process such vast amounts of garments, automation of garment handling and recycling is unavoidable. Automated handling and sorting of such garments is a major challenge in the field of robotics. Current approaches specialize in one part of this challenging task. For sorting, current approaches use cameras and pre-trained networks with a dataset with a pre-defined set of classes. This paper presents an approach of using artificial intelligence (Convolutional Neural Network and Visual Language Models) to locate and separate garments from a pile and identifying and sorting them into dedicated containers. This combines the advantages of both neural network types, where convolutional neural networks are used for grasping (segmentation and corner detection) and visual language models are used for classification of garment types and to help the grasp prediction network in narrowing in on better grasp positions.**

***Index Terms*— Garments, Sorting, Visual Language Models**

## I. Introduction

In recent years, the European Union has set out to combat the huge amount of textiles being discarded every year. To this purpose, a directive has been released stating to achieve full recycle-ability of garments by the year 2030 [1]. To be able to recycle garments, facilities need to sort disposed garment according to their type, material composition and color and detect their state of health regarding faults, unremovable stains or tears. Such facilities currently rely heavily on manual labor to accomplish those tasks. Given that each European Union citizen disposes of approximately 12 kg of clothing annually [1], this results in substantial quantities that require sorting. To be able to better handle this workload, robots in combination with artificial intelligence are a viable alternative. Such a sorting workflow can be split into multiple tasks, starting with retrieving a garment from a pile, inspecting it and sorting it according to pre-defined categories. The robot needs to be able to differentiate between the different garments in the pile to at least be able to pick one textile out of the pile for individual inspection, detect which type the garment belongs, perform an inspection of the garment and sort it into given containers. To be able to retrieve a garment from a pile, a first visual inspection is needed for a preliminary distinction of different textiles and to be able to pick out a single object from the pile. To this purpose, different techniques in artificial intelligence can be used. We propose to use image segmentation and corner detection in conjunction, in a pre-trained Convolutional Neural Network (CNN) to be able to distinguish between different garments and to be able to detect first optimal grasp positions. After the textile has been moved from the pile for individual inspection, a class needs to be assigned to the item. For this purpose, different options are available. One option is to use a pre-trained neural network to try to sort the garment into a given set of classes [2]. This limits the operation, as in such a facility you cannot be sure what types and subtypes of garments are inside the piles that need to be sorted. We propose to use Visual Language Models (VLMs) to match a class to the inspected textile instead. Utilizing a VLM offers benefits, including eliminating the need for further training and providing autonomy from specific subsets of classes. An additional step is needed to semantically match all detected labels given by the VLM to the final list of classes needed for the specific use-case. VLMs can also be used for a first semantic state of health detection of the garment and to find best semantic grasp positions for each given class of garment. This helps the grasp prediction network in the next step to be able to find better grasp positions to place the object into the corresponding container. An overview on the complete

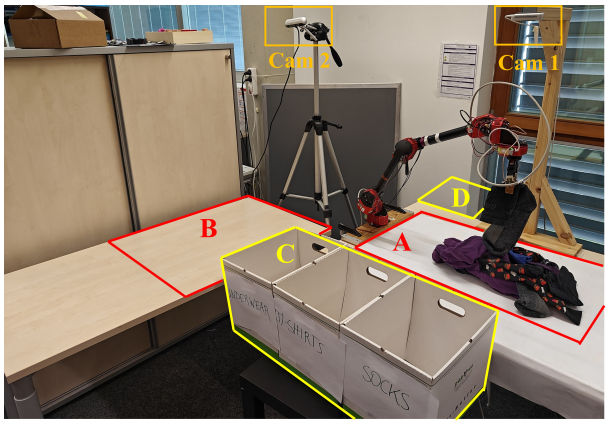


Fig. 1: Illustration of the garment handling and sorting scenario: Garments are picked from a pile (zone A) and manipulated to the inspection desk (zone B). Each garment piece is then sorted in the corresponding container (zone C) or discarded, if it does not meet sorting criteria (zone D). Two RGB-D cameras (Cam 1 and Cam 2) are used for grasp prediction and garment classification, respectively.

[1]Serkan Ergun, Tobias Mitterer and Hubert Zangl are affiliated with the Department of Smart Systems Technologies, Sensors, Actuators and Modular Robotics Group, University of Klagenfurt, 9020 Klagenfurt am Wörthersee, Austria, serkan.ergun@aau.at, tobias.mitterer@aau.at, hubert.zangl@aau.at

[2]Hubert Zangl is also affiliated with the Ubiquitous Sensing Lab, University of Klagenfurt, 9020 Klagenfurt am Wörthersee, Austria

[†]*These authors contributed equally.*

https://doi.org/10.34749/3061-0710.2025.4

handling and sorting set-up is shown in Fig. 1.

The main contributions of this paper are an evaluation of using VLMs for classification of textiles, finding optimal grasping positions, comparing between the performance of using a dedicated CNN and the more general VLM and demonstrating a use-case for this evaluation in a robotic gripper textile sorting context.

## II. Related Work

Recent advances in object detection and classification show a shift from pre-trained networks like CNNs, where a specific list of classes is given in the training and detected to a more general, semantic-based approach like in VLMs. One example of a CNN is Suchi et al. [3], where in the recorded dataset, Object Clutter Indoor Dataset (OCID), a given set of objects and their classes are defined and can be detected by the network. While such networks have good performance and require relatively less hardware, there are advantages to use VLMs for object detection and classification. These VLMs use Zero shot object detection to be able to detect classes by using semantics and free-text queries as input. First advances have been made by the Vision Transformer for Open-World Localization (OwlVit) network [4], in which image-text models are transferred to open-vocabulary object detection. Current advances in the field combine advantages of both methods. Examples are GroundVLP [5], which harnesses the visual grounding abilities from pre-trained image-text pair models and open-vocabulary object detection data to better detect and localize objects without dedicated training on those classes or DINO [6], which achieves good performance on the Common Objects in Context (COCO) dataset by incorporating improved de-noising techniques and anchor boxes. For robotics, the next step after detecting objects with a sensor, e.g. a camera, is to identify suitable positions to grasp the detected object. One example of finding such grasping candidates is an extension to the previously mentioned Suchi et al. [3], Ainetter et al. [7], where the OCID dataset has been used as base to annotate grasping positions on labeled data. The CNN combines grasp detection with dense, pixel-wise semantic segmentation and was tested with a parallel-plate gripper in [8] in combination with capacitive sensing in the gripper. A special case of robotic grasping is, if the objects to grasp have a special shape or material such that they can only be grasped properly at dedicated areas, e.g. a cup. One type of object like this are textiles and garments, as due to their size or given constraints by the task (such as robot assisted dressing) they have the need to be grasped at dedicated positions. One such use-case is if a garment needs to be visually inspected for faults or current state and therefore all parts need to be fully visible. Yamazaki [9] presents a CNN trained to detect optimal grasp points for cloth based on shape classification to be able to properly unfold a given textile lying on a table. Another example of textile grasping is Fu et al. [10], where a network is used to detect the state of the textile based on visible corners and decides on grasping points to unfold it. A detailed semantic description of each class of textiles is given by the European Commission in [11]. As already shown for object detection earlier, VLMs can also be used to semantically detect the best location to grasp a specific type of object, thus increasing the number of successful grasps. LanGrasp [12] uses Large Language Models (LLMs) and VLMs to enable semantic one shot object grasping, where the LLM gives the part of the object which should be grasped and the VLM grounds that information in an image. Finally, a grasp planner plans and executes the grasp. Another example on VLMs being used for robot grasping is Huang et al. [13], with a focus on handover tasks of household objects between a human and a robot. The robot uses a combination of VLM and LLM to detect objects of given classes and semantically detect appropriate grasping parts of the objects. As a last step, a grounded VLM is used to segment and detect the grasping parts of the object. Our approach introduces the innovative use of VLMs in combination with a CNN in the environment of sorting textiles from a highly cluttered heap depending on given semantic classes.

## III. Experimental Setup

The proposed lab scale experimental setup is shown in Fig. 1. It consists of a single modular series elastic 6-DoF arm with a two-fingered cable gripper (type A-2085-06G) by HEBI Robotics [14]. The finger tips are custom made and flexible. Random garments are placed in a convoluted pile in the initial inspection workspace. The robot is being used to grasp and manipulate a single piece of garment from the pile (zone A in Fig. 1) and place it on a second table for garment type classification: underwear, shirts (including t-shirts and polo shirts) and socks (zone B). As the performance of the VLM is evaluated the garments are just slightly dragged over the table edge to perform an initial unfolding and no further flattening or optimal positioning of the garment is done. The robot then places the garment in the corresponding box in front of the table (zone C). Garments, which do not fall in one of these categories, or which are not identifiable are discarded at the back of the table for manual inspection (zone D).

Two Intel RealSense cameras (models D415i and D455f) are being used for capturing depth images for the grasp prediction from the pile (Cam 1) and the inspection table (Cam 2), respectively. The RGB stream of the D455f was also used for capturing the input for the garment classification.

The grasp prediction algorithm uses a CNN, which is based on the works of [7] and [15]. The CNN has been trained with a modified training set based on [3]. It has furthermore been used effectively in previous works, such as [8].

https://doi.org/10.34749/3061-0710.2025.4


Fig. 2: Procedure of the experiment. A single piece of garment is identified by the grasp prediction algorithm (a), the robot then picks up the garment (b) and places it on the inspection table (c). The VLM identifies the type of garment and its color (d). The grasp prediction algorithm is then run again to identify the optimal grasp position (e). The garment is then picked up (f) and placed in the dedicated container (g). Unrecognized or not categorized are discarded at the back of the table for further manual inspection.

The garment classification is achieved by using VLMs, as they are able to match a semantic text input to a specific object in the picture. Dedicated VLMs like OWL-FIT [4] usually work by giving the network a semantic list of objects to find, which can lead to either a reduced number of objects detected or a long list of possible classes in the prompt. To this purpose a VLM with a broader language part in the model, namely Llama 3.2-Vision 11b from Meta Inc. [16] is used via the Ollama distribution [17]. This enables a more open prompt definition due to its bigger training data-set, where the types of garment to be found are not specified but the network has to match a garment type to the textile presented in the picture. For the proposed use-case, only the output classes of 'sock', 'underwear', 'shirt' and 'unknown' are used. A semantic assignment of the detected sub-types to these four classes, where each garment type which cannot be assigned to the three known classes is assigned to the 'unknown' class is done. This semantic matching process uses the textile classification specification given by the European Commission in [11].

Listing 1: Minimal Python code example for running llama3.2-vision with Ollama

```
import ollama

response = ollama.chat(
    model='llama3.2-vision',
    messages=[{"role": "system",
        "content": "You are an intelligent robotic
    arm."}, {'role': 'user', 'content': 'What
    clothing item can you see? If your confidence
    is below 85 percent, classify the type as
    unknown. Classify them in the classes: Shirt,
    Sock, Underwear or unknown. Just specify type
    of clothing and color. which part of the
    garment makes the most sense to grasp? Name one
     part. Make the answer very short and concise.
    In three words. Your response is garment type,
    color, grasp location. Omit any line breaks or
    newlines.', 'images': [fullPathToImages]
            }]
)
```

The minimal code, necessary to run Llama3.2-vision with Ollama in Python3 is shown in Listing 1,where `fullPathToImages` is the location of the image on the hard drive. The grasp prediction(s) and the VLM are run sequentially on the processing unit,a lab PC with an 11th Gen Intel® Core™ i7-11700KF @ 3.60GHz × 16 CPU with 64 GB of DDR4 RAM. The Graphics Processing Unit (GPU)

https://doi.org/10.34749/3061-0710.2025.4

Fig. 3: Sample pieces for garment classes: shirt, underwear and socks

used is the Nvidia GeForce RTX 3060 with 12GB V-RAM. The entire scenario is shown in Fig. 2. First, an ideal grasp pose is selected by the CNN (a). Then the robot moves to the selected pose and grasps the garment (b) and manipulates it to the inspection table (c). On the inspection table, the VLM is then run to identify the garment type, color and to return a semantic description of the best grasping position on this type of garment. The terminal output is shown alongside the input image in (d). The garment type output of the VLMs is then further refined and categorized semantically into the four given classes. Consequently, the CNN is then run again to identify the ideal grasp pose (e). The robot then picks up the garment once more (f) and places it in the correct bin (g). If the garment does not fall into one of the specified categories, it will be discarded at the back of the table for potential manual inspection.

## IV. Results

Sample pieces for each considered garment class (underwear, shirt, sock) are shown in Fig. 3. Further garments are shown in Table I. A subset of the garment samples (five to ten pieces) were placed untidily in zone "A" (see Figure 2c). The experiment described in section III was repeated until zone "A" was cleared. In total, more than 100 individual grasps were taken.

### A. Validation of the approach

The grasp prediction algorithm was previously tested and validated in [8]. In some cases, the grasp prediction algorithm does not return an ideal grasp pose on every attempt. In such cases, another image is taken automatically and the grasp prediction is repeated until a valid grasp pose is found.
During our experiments, no loss of garments during manipulation occurred. However, due to the small size ("arm reach") of the robot, only smaller garment pieces were considered in our experiments. Large and heavier pieces were only used for the validation of the VLM.

### B. Validation of the VLM

To validate the accuracy of the VLM a larger set of garments were analyzed (zone "B"). In total, 122 images from trousers, jeans, jackets, sweatshirts were investigated alongside our pre-determined classes shirts, underwear and socks. Additionally, at random instances, non-garment type objects were presented to the camera (e.g. bottom entry in Table I).
Ten exemplary results of the garment classification algorithm are shown in Table I. Out of these 122 images, the garment type was correctly classified in 118 cases. The color of the garments were correctly identified in 112 cases, caused by inconsistent lighting conditions during the experiments. The color information was not used in our experiments but recorded alongside for future use, if the need arises to sort the garments not only by type but also color. An additional analysis was done on identifying semantic grasp positions for each type of garment, to optionally help the grasp prediction in narrowing in on better grasp positions for specific type of garments. In the testing data-set, non-textiles were included to observe the output if the prompt tells the network to identify a given textile, when no garment is presented. As can be seen in Table II for most cases the precision is greater than 80 %. For the non-garment objects, they were classified correctly as 'unknown', but a non-ideal grasping position was returned.

## V. Summary and Outlook

This paper presented a method to utilize a combination of pre-trained CNN and VLM for automated handling and sorting of garments. Overall, the VLM is able to identify the garment type with a precision of 96.72 %. The color of the garments is correctly identified in 91.80 % of our experiments. The proposed sorting setup has the capability to scale in size and thus, productivity.
By using a serial arm manipulator with higher reach, larger garments can be handled. Furthermore, a larger arm reach allows more containers for sorting to be placed in the scene. The usage of a second arm manipulator for picking up inspected garments while the other one is grasping garments from the pile increases productivity. Alternatively, a conveyor belt type setup may also be considered using pushers to slide garments in the corresponding containers. A second processing unit with a dedicated GPU allows to run both grasp prediction algorithms in parallel together with the VLM to further increase productivity. The current work can be used to include multi-point grasp for improving the handling of bigger textiles like shirts or trousers and allowing all-round inspection of garments for visual defects.

## Acknowledgment

This work has received funding from the "Austrian Research Promotion Agency" (FFG) within the AdapTex project under grant number 899044.

TABLE I: Exemplary results of the garment classification algorithm. The image on left is the input of the VLM. Type, Color, Suggested Grasp Pose and Class are returned by the VLM.

| Image | Type | Color | Suggested Grasp Pose | Class |
|---|---|---|---|---|
|  | Sock | Black | Heel | Sock |
|  | Shirt | Green | Sleeve | Shirt |
|  | Shirt | Blue | Collar | Shirt |
|  | Boxers | Blue | Waistband | Underwear |
|  | Shirt | Purple | Collar | Shirt |
|  | Jeans | Blue | Waistband | Unknown |
|  | Shirt | Blue | Cuff | Shirt |
|  | Boxers | Blue | Waistband | Underwear |
|  | Trousers | Grey | Waistband | Unknown |
|  | Unknown | Blue | Heel | Unknown |

TABLE II: Overall results of the garment classification. In total, 122 images of garments were captured. The VLM output was then manually examined. Accuracies for determining the correct class, color and potential grasp position are given for each garment type were recorded.

| Type | Count | Class | Color | Grasp Position |
|---|---|---|---|---|
| Underwear | 14 | 100 % | 85.71 % | 100 % |
| Shirt | 46 | 97.83 % | 95.65 % | 97.83 % |
| Sock | 36 | 97.22 % | 91.67 % | 97.22 % |
| Unknown | 26 | 92.31 % | 88.46 % | 53.85 % |
| Total | 122 | 96.72 % | 91.80 % | 88.52 % |

## References


[1] Directorate-General for Environment, "COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS - EU Strategy for Sustainable and Circular Textiles," 2022. [Online]. Available: https://environment.ec.europa.eu/publications/textiles-strategy_en

[2] R. Tian, Z. Lv, Y. Fan, T. Wang, M. Sun, and Z. Xu, "Qualitative classification of waste garments for textile recycling based on machine vision and attention mechanisms," *Waste Management*, vol. 183, pp. 74–86, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0956053X24002629

[3] M. Suchi, T. Patten, D. Fischinger, and M. Vincze, "Easylabel: a semi-automatic pixel-wise object annotation tool for creating robotic rgb-d datasets," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6678–6684.

[4] Minderer, Matthias and Gritsenko, Alexey and Stone, Austin and Neumann, Maxim and Weissenborn, Dirk and Dosovitskiy, Alexey and Mahendran, Aravindh and Arnab, Anurag and Dehghani, Mostafa and Shen, Zhuoran and Wang, Xiao and Zhai, Xiaohua and Kipf, Thomas and Houlsby, Neil, "Simple Open-Vocabulary Object Detection," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 728–755. [Online]. Available: https://doi.org/10.1007/978-3-031-20080-9_42

[5] H. Shen, T. Zhao, M. Zhu, and J. Yin, "Groundvlp: harnessing zero-shot visual grounding from vision-language pre-training and open-vocabulary object detection," in *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. [Online]. Available: https://doi.org/10.1609/aaai.v38i5.28278

[6] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H.-Y. Shum, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=3mRwyG5one

[7] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 452–13 458.

[8] S. Ergun, T. Mitterer, S. Khan, N. Anandan, R. B. Mishra, J. Kosel, and H. Zangl, "Wireless capacitive tactile sensor arrays for sensitive/delicate robot grasping," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 10 777–10 784.

[9] K. Yamazaki, "Selection of grasp points of cloth product on a table based on shape classification feature," in *2017 IEEE International*

*Conference on Information and Automation (ICIA)*, 2017, pp. 136–141.

[10] T. Fu, C. Li, J. Liu, F. Li, C. Wang, and R. Song, "FlingFlow: LLM-Driven Dynamic Strategies for Efficient Cloth Flattening," *IEEE Robotics and Automation Letters*, vol. 9, no. 10, pp. 8714–8721, 2024.

[11] European Commission. (2025) Classifying textiles. [Online]. Available: https://trade.ec.europa.eu/access-to-markets/en/content/classifying-textiles

[12] R. Mirjalili, M. Krawez, S. Silenzi, Y. Blei, and W. Burgard, "LAN-grasp: An effective approach to semantic object grasping using large language models," in *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024. [Online]. Available: https://openreview.net/forum?id=SfHjWbfW02

[13] J. Huang, C. Limberg, S. M. N. Arshad, Q. Zhang, and Q. Li, "Combining vlm and llm for enhanced semantic object perception in robotic handover tasks," in *2024 WRC Symposium on Advanced Robotics and Automation (WRC SARA)*, 2024, pp. 135–140.

[14] Hebi Robotics Inc., "Hebi A20856G Data Sheet," February 2025. [Online]. Available: http://docs.hebi.us/resources/kits/datasheets/x-series/A-2085-06G_Datasheet.pdf

[15] L. Porzi, S. Rota Bulò, A. Colovic, and P. Kontschieder, "Seamless scene segmentation," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[16] Meta Inc. (2025) Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. [Online]. Available: https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

[17] Ollama. (2025) Ollama Website. [Online]. Available: https://ollama.com/

https://doi.org/10.34749/3061-0710.2025.4