MultimodalClothes: A Multimodal Real-World Dataset for Robotic Clothes Classification

Niklas Fiedler[1], Ge Gao[2], Fangwei Zhong[3], and Jianwei Zhang[1]

***Abstract*— Classifying pieces of clothing is crucial for robots performing laundry work or dressing assistance tasks. This is a challenging problem since clothes are highly deformable and assume various shapes when being held or manipulated. Deep learning based approaches show promising performance for clothes classification. However, existing datasets are either insufficient in size for deep network training or cover only a few classes. Apart from RGB images, other sensor modalities can also help solve the clothes classification task. In this work, we present *MultimodalClothes*, a multimodal dataset of 90,883 real-world samples of clothes held by a robot arm. The dataset contains the following modalities: RGB and depth images, point clouds, tactile data, and temporal information. Further, we provide a statistical analysis of the dataset as well as classification baselines for the visual and depth modalities. The dataset and our code are publicly available via `https://github.com/TAMS-Group/MultimodalClothes`.**

I. Introduction

Classification of non-rigid objects, such as clothes, is required for robots to accomplish daily tasks, such as laundry work [1] or robot-assisted dressing [2]. Clothes are highly deformable when grasped or manipulated, even compared to other non-rigid objects, such as plush toys or cables, due to their higher flexibility.

In recent years, several deep learning-based frameworks have been proposed to address the challenge of clothing recognition [3], [4]. Frameworks that learn deep features from depth data show promising performance. In order to further aid the deep learning frameworks, sufficient annotated data is needed. Collecting the visual data of clothes manipulated by a robot is very labor-intensive. Most current real-world datasets are insufficient for training deep networks without augmentation with simulated data [5], [6]. On the other hand, some slightly larger datasets contain only a small number of classes, which limits their use in real-world applications [7], [3] as we discuss in the following section. One way to overcome the data shortage problem is to use synthetic data generated in simulators. However, there is a noticeable performance gap between networks trained on real data and those trained on synthetic data only [7]. We provide an overview of related datasets for comparison in Table I.

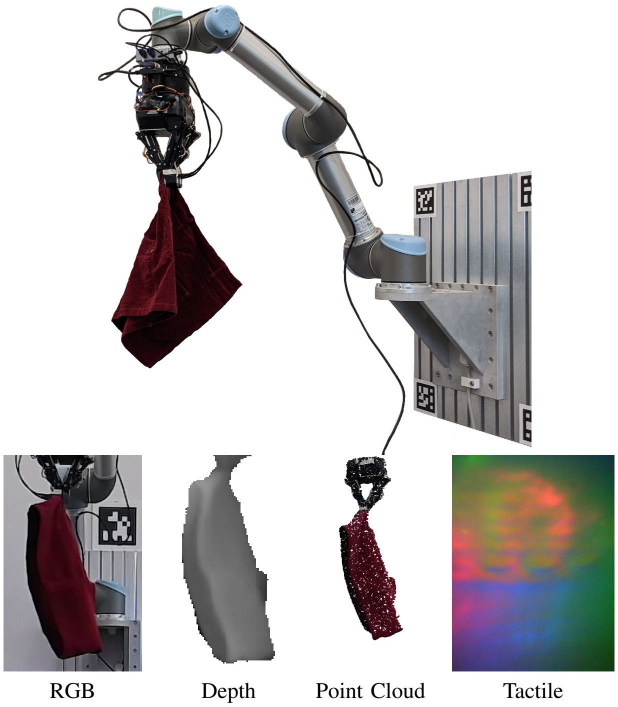

Fig. 1. Data recording setup with exemplary sample of type "*skirt*". The robot arm is holding the piece with tactile sensors making direct contact. While collecting the samples, the piece is rotated by a wrist motion. Below, various modalities included in the dataset are visualized for an exemplary sample.

Furthermore, since recognizing highly deformed clothes is very challenging, relying only on the vision data is potentially insufficient. Since the robot usually picks the garment up before the manipulation task starts, other data modalities can also be perceived.

To pave the way for further research into the topic, we propose *MultimodalClothes*, a large multimodal dataset containing real-world RGB-D data and covering a wide range of clothing classes. An exemplary recording situation is shown in Figure 1. Below, various modalities of a recorded sample are visualized.

Our contributions can be summarized as follows:

- We present a multimodal real-world dataset with 12 classes for clothing recognition.
- We evaluate established baseline methods for the clothing classification task.
- Using the multimodal dataset, we provide the first direct comparison of clothes classification performance using different input modalities and processing methods.

This research was funded by the German Research Foundation (DFG) and the National Science Foundation of China in the project Crossmodal Learning, TRR-169.

[1] is with group Technical Aspects of Multimodal Systems (TAMS), Department of Informatics, Universität Hamburg, Germany {`niklas.fiedler`, `jianwei.zhang`}`@uni-hamburg.de`

[2] is with Mech-Mind Robotics Technologies Ltd., China `gao.ge@mech-mind.net`

[3] is with the School of Intelligence Science and Technology, Peking University and Beijing Institute for General Artificial Intelligence (BIGAI), China `zfw@pku.edu.cn`

TABLE I

Comparison of clothes datasets available.

Approach	Year	Recording Scenario	Modalities	# classes	# pieces		# samples		Purpose
					Sim	Real	Sim	Real	
Fashion MNIST [8]	2017	Canonical	Grayscale	10	-	70,000	-	70,000	classification
Deepfashion [9]	2016	Canonical, Worn	RGB	1050	-	?	-	800,000	classification
Gabas et al. [10]	2016	Grasped, Rotated	Depth	4	-	?	-	4,272	classification
Corona et al. [11]	2018	Grasped, Rotated	Depth	4	?	?	60,000	5060	classification, grasp point detection
Saxena et al. [4]	2019	Grasped, Rotated	Depth	3	?	?	94,140	18,305	classification, grasp point detection
CTU Dataset [12]	2013	Canonical	RGB, Depth	9	-	17	-	?	classification, segmentation
Mariolis et al. [7]	2015	Grasped, Rotated	RGB, Depth	3	72	13	643,200	90,077	classification
Sun et al. [3]	2017	Randomly laid on Table	RGB, Depth	5	-	50	-	2100	classification, grasp point detection
Ours	2023	Grasped, Rotated	RGB, Depth, Tactile, Temporal	6	-	18	-	46,031	classification
				12		36		90,883	

'?' : The value is unknown. '-' : There are no such pieces/samples in the dataset.

II. Dataset

Our recording setup consists of a wall-mounted UR5 robot arm equipped with a Robotiq 3-Finger Gripper to grasp the pieces and multiple sensors to capture multimodal data during the manipulation (see Figure 1). We integrated the DIGIT sensors presented by Lambeta et al. [13] by replacing two of the gripper's original fingertips with the sensors.A Microsoft Azure Kinect camera was set up to capture RGB and depth images as well as point clouds from a single position. It was placed at a height of 0.97 m with a distance of 1.5 m from the piece to capture the garment as a whole. Thus, in some cases, large pieces such as jackets or jeans touched the floor.

When choosing the classes for the dataset, we took inspiration from the DeepFashion dataset [9] by selecting their most common garment types. Then, the list was narrowed down by combining multiple types with minor differences into one. For example, the classes *Blouse* and *Button-Down* were combined. Further classes were removed because they were deemed not relevant enough for the scenario and others because we could not source sufficient samples for recording, which also indicates limited relevance.

We propose a set of twelve classes as listed in Table II. Six classes were chosen by maximizing the perceived difference in their general shape for a smaller and easier version of our dataset to accommodate less demanding scenarios. To gather an authentic set of diverse samples, we collected clothes from colleagues and friends in addition to our own for the data collection. For each class, three distinct pieces were selected. Two are used as training data, and one as test data. This helps to identify overfitting on a specific color or pattern. The samples were provided by eleven people (four male, seven female) with varying sizes and color preferences.

In contrast to existing approaches [10], we manually hand over the pieces to the robot to record samples grasped in the distinct grasp areas top, edge, and inside. The impact of the grasp position on the classification accuracy can be investigated by filtering the samples by their grasp area. Thus, it is impossible to let the robot automatically grasp the piece randomly from a surface. A human operator handed the pieces to the arm aiming for a uniform distribution of the grasp points in each grasp area. This also prevents a bias caused by the robot's grasp strategy. After the piece was handed over, the arm moved to a fixed position in front of the Kinect camera. Then, a data collection sequence was started by recording roughly 50 samples while rotating the piece with a speed of 1.5 rad/s in both directions around the vertical axis. Overall, 50 collection sequences were performed for each piece (10×top, 20×edge, and 20×inside). Using a robot arm ensures a consistent rotation speed and holding position and is very similar to the classification scenario.

During the recording process, the visual and depth modality are captured in the form of RGB images, depth images, and point clouds with a color channel. The point cloud is cropped in Cartesian space before it is saved as a `pcd`-file to save disc space. The resulting point clouds only contain the piece and a small part of the robot gripper. Two RGB images make up the tactile measurements of the two DIGIT sensors. Further, due to the sequential recording style with a fixed rotation speed and data caption frequency, temporal information is available when all individual samples recorded in a single collection sequence are considered as a single time series sample.

III. Evaluation

We provide statistical information about the number and distribution of samples across the subsets of the dataset. Further, we trained multiple state-of-the-art classification approaches to classify the samples in the set using various input types.

For the benchmarking tests, we define fixed train, validation, and test sets of the data. Table II gives an overview of the number of samples in each part of our dataset in its two versions. The classes used in both the 12-class set and the 6-class subset are highlighted in bold font. As shown, the whole dataset consists of 90,883 samples. Thus, regarding the overall dataset size, our real-world dataset is larger than most of the sets presented in Table I. Therefore, we believe that the dataset is sufficiently dimensioned to evaluate state-of-the-art approaches which require a large amount of training

TABLE II

Sample count by class for the train, validation, and test sets of the dataset in both the 6- class and 12-class versions.

Class	MultimodalClothes-12 Train Set	MultimodalClothes-12 Val. Set	MultimodalClothes-12 Test Set	MultimodalClothes-6 Train Set	MultimodalClothes-6 Val. Set	MultimodalClothes-6 Test Set
Jacket	3966	1034	2424	4017	983	2424
Button-Down	3894	1019	2500	-	-	-
Jersey	4015	985	2395	-	-	-
Hoodie	4004	996	2500	3978	1022	2500
Sweater	4016	984	2500	-	-	-
Tee	4000	1050	2500	4058	992	2500
Jeans	4892	1207	2500	4871	1228	2500
Sweatpants	3986	1013	2550	-	-	-
Shorts	3993	939	2500	3924	1008	2500
Dress	4006	1043	2477	4056	993	2477
Skirt	4017	981	2500	-	-	-
Top	4041	956	2500	-	-	-
Sum	48,830	12,207	29,846	24,904	6226	14,901
	90,883			46,031		

TABLE III

Achieved classification accuracies of established approaches on the 12- and 6-class versions of the dataset using various input types.

Method	Input Type	MultimodalClothes-6		MultimodalClothes-12	
		Validation	Test	Validation	Test
PointNet	Point Cloud	93.1%	34%	76.3%	33.6%
PointNet++		99.9%	48.1%	99.7%	**47.5%**
DGCNN		99.1%	46.8%	96.0%	**47.5%**
ResNet-18	RGB	100%	64.9%	100%	8.7%
	Depth	99.6%	**69.7%**	99%	46.6%

data. This is especially relevant as, at the time of writing, some modalities (especially the tactile recordings) included in the dataset cannot be simulated sufficiently well for real-world transfers. It is evident that we include more modalities than other sets listed in Table I and provide a relatively high number of classes compared to other sets with a robot grasping scenario.

To provide a performance baseline for the dataset in both versions (12 and 6 classes), we use it to train well-known and established architectures. First, we compare the performance of the point cloud based classification networks PointNet [14], PointNet++ [15] (in the MSG configuration), and DGCNN [16]. Second, we apply ResNet-18 [17] onto the RGB images in the dataset because it is a well-established and proven tool for image classification. We use the pre-trained weights provided by PyTorch to initialize the whole network. Finally, similar to most approaches related to ours, we train a convolutional neural network on depth images. For this task, we also make use of a ResNet-18. To process the single-channel depth input, we adapted the first convolutional layer of the network.

The results of the benchmark runs are listed in Table III. The significant difference in classification accuracy between validation- and test data indicates a generally bad generalization. Point cloud based methods perform similarly well both on the 6- and the 12-class version of the dataset. This indicates that the differences in general shapes of the garments are less relevant for the architectures. While the approach operating on RGB input yields competitive results on the 6-class subset, clear indications of overfitting are apparent on the 12-class test set with a classification performance insignificantly better than a random choice. The depth image based approach shows a better generalization performance than the RGB based one. While it outperforms all other approaches on the small subset, point cloud based approaches with local feature processing capabilities are more accurate on the 12-class set.

Saxena et al. present confusion matrices for both the classification of pieces that were used during training and of ones that were not. They achieve a classification accuracy of 71% for unseen garments, with a convolutional neural network processing depth image input. However, they only distinguish between three visually very different classes, which makes the task easier than our 6-class dataset. Nevertheless, this is very close to the classification accuracy of 69.7% achieved by our ResNet distinguishing between six classes using depth images as input. We cannot directly compare the ResNet used for our benchmarking tests as the dataset is not available online anymore.

Depending on the classification method, the two dataset versions pose a dissimilar challenge. Also, it is a generally hard challenge for state-of-the-art approaches and inspires further research on the topic.

IV. Conclusion

We present a multimodal real-world dataset of 12 classes for clothing classification. By conducting experiments using state-of-the-art approaches to the visual modalities of both versions, we show insights into the clothes classification problem. The depth and point cloud based approaches show a better generalization ability compared to the RGB based approach. Our dataset allows researchers to directly compare the applicability of modalities and combinations of them in their problem cases. Novel approaches to uni- and multimodal clothes classification can be effectively evaluated using this dataset. Even if it is not used for benchmarking, a portion or even the full dataset can be utilized for pre-training as a replacement or addition to simulated data.

In future work, the multimodal samples allow researchers to directly compare the performance of various input modalities and develop multimodal classification approaches. This is especially interesting, considering the hard challenge posed by the benchmarking set.

Acknowledgment

Thanks to Sebastian Stelter for helping in the data collection process and to Yannick Jonetzko and Jonas Hagge for helpful insights and discussions. We thank the Hamburg Bit-Bots for providing processing hardware. Thanks to friends and colleagues for providing their clothes.

References

[1] C. Xiong, N. Shukla, W. Xiong, and S.-C. Zhu, “Robot learning with a spatial, temporal, and causal and-or graph,” in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 2144–2151.

[2] F. Zhang and Y. Demiris, “Learning garment manipulation policies toward robot-assisted dressing,” *Science Robotics*, vol. 7, no. 65, p. eabm6010, 2022.

[3] L. Sun, G. Aragon-Camarasa, S. Rogers, R. Stolkin, and J. P. Siebert, “Single-shot clothing category recognition in free-configurations with application to autonomous clothes sorting,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2017, pp. 6699–6706, iSSN: 2153-0866.

[4] K. Saxena and T. Shibata, “Garment Recognition and Grasping Point Detection for Clothing Assistance Task using Deep Learning,” in *2019 IEEE/SICE International Symposium on System Integration (SII)*, Jan. 2019, pp. 632–637, iSSN: 2474-2325.

[5] I. Mariolis and S. Malassiotis, “Matching Folded Garments to Unfolded Templates Using Robust Shape Analysis Techniques,” in *Computer Analysis of Images and Patterns*, ser. Lecture Notes in Computer Science, R. Wilson, E. Hancock, A. Bors, and W. Smith, Eds. Berlin, Heidelberg: Springer, 2013, pp. 193–200.

[6] A. Doumanoglou, A. Kargakos, T. Kim, and S. Malassiotis, “Autonomous active recognition and unfolding of clothes using random decision forests and probabilistic planning,” in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 987–993, iSSN: 1050-4729.

[7] I. Mariolis, G. Peleka, A. Kargakos, and S. Malassiotis, “Pose and category recognition of highly deformable objects using deep learning,” in *2015 International Conference on Advanced Robotics (ICAR)*, Jul. 2015, pp. 655–662.

[8] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms,” *arXiv:1708.07747 [cs, stat]*, Sep. 2017.

[9] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1096–1104.

[10] A. Gabas, E. Corona, G. Alenyà, and C. Torras, “Robot-Aided Cloth Classification Using Depth Information and CNNs,” in *Articulated Motion and Deformable Objects*, ser. Lecture Notes in Computer Science, F. J. Perales and J. Kittler, Eds. Cham: Springer International Publishing, 2016, pp. 16–23.

[11] E. Corona, G. Alenyà, A. Gabas, and C. Torras, “Active garment recognition and target grasping point detection using deep learning,” *Pattern Recognition*, vol. 74, pp. 629–641, Feb. 2018.

[12] L. Wagner, D. Krejčová, and V. Smutný, “CTU Color and Depth Image Dataset of Spread Garments,” p. 13, center for Machine Perception, Czech Technical University, Tech. Rep. CTU-CMP-2013-25, 2013.

[13] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer *et al.*, “DIGIT: A Novel Design for a Low-Cost Compact High-Resolution Tactile Sensor With Application to In-Hand Manipulation,” *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[14] C. R. Qi, H. Su, M. Kaichun, and L. J. Guibas, “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, Jul. 2017, pp. 77–85.

[15] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[16] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, “Dynamic Graph CNN for Learning on Point Clouds,” *ACM Transactions on Graphics*, vol. 38, Jun. 2019.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.