ISSN 1870-4069

Understanding and Tracking of Deformable Objects in the Context of Humanoid Robotics: A Literature Review

Gustavo De Los Ríos-Alatorre, Luis Alberto Muñoz-Ubando

Instituto Tecnológico y de
Estudios Superiores de Monterrey,
Mexico

gustavodlra1999@gmail.com, amunoz@tec.mx

Abstract. The following paper offers a comprehensive literature review of recent advancements and trends in the manipulation of deformable objects by robotics systems focusing as much as possible on the interaction between humanoid robotics and this type of objects. The paper boards different fields such as deep learning, traditional deformable object understanding, representations and handling, as well as the role that context awareness and multi-sensory data plays in the understanding of these objects. The explored works span a period around the last 30 years in order to get the most complete possible understanding.

Keywords: Computer vision, deep learning for visual perception, recognition.

1 Introduction

As a field of artificial intelligence, computer vision empowers computational systems to derive meaningful insights from digital images, videos, and other visual inputs, culminating in informed decision making [18]. This process frequently involves utilization of image processing techniques to aid in the interpretation of intricacies of a 3-D world from time varying 2-D data.

By allowing systems to perceive and understand their environment, computer vision paves the way for crucial applications in diverse sectors such as robotics, medicine, computer graphics and autonomous vehicles. Humanoid robotics consists of a field of study that aims to develop robots that mirror the structure and behavior of the human body, allowing them to function autonomously in environments designed for humans.

Fundamental concepts related to it include bipedal locomotion, balance control and complex motor functions that involve a blend of perception, decision-making, and execution. By integrating technologies from different domains such as mechanics, electronics and computer science, humanoid robots are designed to perform tasks such as walking, climbing stairs, or object manipulation, replicating human agility and dexterity.

Incorporating computer vision significantly enhances the capabilities of humanoid robots, enabling them to perceive and interact with their environment in a more refined way. It is particularly relevant to consider the tracking and manipulation of deformable or flexible objects, which represents an intricate challenge due to their geometries and material properties.

As the complexity of the geometry of an object and material properties increase, traditional computer vision techniques can often fail to deliver to the desired accuracy and robustness. The following text offers a comprehensive literature review focused on how modern computer vision strategies tackle these difficulties, exploring advancements in deep learning techniques that have enhanced tracking performance of flexible objects.

With an emphasis on the integration of object recognition and depth perception algorithms, this literature review intends on delving into the processes that enable precise, real-time tracking of these complex targets. The intersection of depth perception and object recognition forms a robust basis for the analysis of non-rigid objects, despite their inherent complexities.

Context-awareness also emerges as a significant aspect in the development of sophisticated computer vision algorithms. Incorporating contextual understanding allows humanoid robots to improve their interaction with deformable objects, facilitating more reliable tracking and manipulation.

Finally, this review presents an overview of how multi sensory data integration can augment the information gathered by computer vision techniques. Techniques to fuse data from different sensors are explored, such as tactile sensors, with visual data to enhance tracking and manipulation accuracy, particularly in complex environments.

Overall, this literature review aims to provide a thorough insight into the state-of-the-art in computer vision for flexible object tracking, focusing on the latest as well as the most significant advancements, challenges, and promising avenues for future research. The structure of the review will be based on the following research questions:

1. What unique challenges are presented by flexible object tracking for humanoids and how can computer vision effectively address these difficulties?
2. What are the advancements in deep learning techniques and how they can enhance the performance of vision-based tracking of flexible objects for humanoid robots?
3. How can object recognition and depth perception algorithms be integrated for precise real-time tracking of flexible objects?
4. What is the role of context-awareness in computer vision algorithms for improved tracking and manipulation of flexible objects by humanoid robots?
5. How can data from other sensors be integrated with computer vision to increase the accuracy of flexible object tracking and manipulation in complex environments?

2 Background and Historical Development

The first computer vision experiments begin in the 1960s with the experiments done by PhD. Lawrence Roberts, known as the father of computer vision. In his PhD Thesis for MIT many of the bases for computer vision techniques being used today were developed [36]. In this thesis, he managed to make a computer create a 3-D representation of solid objects in a photograph via the use of image transformations. The work done by Roberts establishes the foundations for future 3-D object recognition and tracking.

Throughout the 1980s and 1990s, deformable models were introduced. This advancement would prove fundamental for tracking deformable objects. Deformable models applicable to 2-D objects were introduced by Kass et al. as explicit deformable contours in 1988. In this early exploration, deformable contours were known as snakes, which are considered as an energy minimizing spline guided by external constraint forces and influenced by image forces that pull it towards features such as lines and edges [19].

Snakes are active contour models, locking into nearby edges and localizing them accurately. These deformable models were later generalized to work in 3-D by Terzuopolos et al. [43], their work on dynamic analytical models initiated the physically based deformable models. These models aimed to represent the physical properties and behavior of materials realistically, using principles from mechanics, physics, and biology.

In 1999, the Scale-Invariant Feature Transform (SIFT) algorithm was introduced by Lowe [28]. The main purpose of this algorithm is to extract information from features that are invariant to image scale and rotation, providing robust matching across a substantial range of distortion, noise or change in viewpoint. Applications of the SIFT algorithm include object tracking and recognition, 3-D reconstruction, and augmented reality.

Variations of the SIFT algorithm have also been utilized in the recognition of deformable objects. In [52], Zickler, S. et al. present an algorithm that utilizes PCA-SIFT in a combination with a clustered voting scheme. This model achieves detection and localization of multiple, highly deformable objects in real-time video footage.

Spanning the 2010s, the rise of deep learning methods saw a wide utilization of these methods in computer vison. Clear examples of these can be seen via the use of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) models being used for object recognition and tracking over time. This technology has been utilized in robotics-related tasks such as classification of slip occurrences by recognizing instances of objects being handled as well as their properties [24, 27, 51].

The main purpose behind this sort of application being object classification. Entering the 2020s, attention mechanisms and transformers which originated in the field of Natural Language Processing have been adapted for computer vision and robotics tasks. Interesting advancements have been made, such as in [40] were the authors proposed a Transformer framework for tabletop tasks encoding language goals and RGB-D voxel observations, outputting discretized 6-DoF actions.

The work done by Yunhai H. et al. [15] is also really promising regarding the utilization of deep learning-related models in tasks related to deformable object manipulation in robotics by utilizing a transformer-based robotic grasping framework for rigid robotic grippers that leverage tactile and visual information for safe grasping. The Transformers involved learn physical feature embeddings with sensor feedback predicting a grasping outcome via the use of a multilayer perceptron (MLP).

3 Flexible Object Tracking Challenges

The understanding and tracking deformable objects in real time applications would broaden the range of applications for robotics systems in areas such as surgery, household robotics, manufacturing, and those mentioned in previous sections.

A complete understanding of these objects is yet to be achieved, mainly due to their complex nature. Thus, representing, understanding, and manipulating deformable objects in robotics remains an open challenge, which has itself given rise to other significant challenges such as detecting the elasticity and plasticity of objects and manipulation planning and control.

Most of the work involving robotic manipulation tends to focus on working with rigid objects [7]. In [5] Billard et al. attribute this to the fact that flexible materials like fruits, vegetables and clothing have varied sizes, weights, and superficial properties, with manipulations that involve deformation being difficult due to the need for an accurate model that represents them. In most interactions, rigid objects are expected to keep their characteristics such as shape intact, with any forces applied to them represented as a series of rigid body transformations.

Conversely, deformable objects experience changes in their shape after a force is applied to them, with the degree of deformation and motion resulting from an interaction being dependent on the material composition of the object, making the physics of the deformation process hard to capture. Dynamics models have been utilized for high-fidelity mechanical modelling, however the information required to make these models is not always present for robotics systems in the real world [3].

All these factors combined with the fact that deformable objects can potentially have an infinite number of degrees of freedom due to their continuous nature [9], have resulted in current attempts of modelling, understanding, and tracking deformable objects being application specific [3]. On the vision front, the understanding of deformable objects is also made difficult because several elements in an image can increase the difficulty of detecting these objects throughout a sequence of frames.

Lighting conditions, background noise, and especially occlusion combined with the variety of appearances that a deformable object might have make it a significant challenge. Historically, these issues (especially occlusions) have been faced by complementing vision models with physics simulations, such as in [34] [42]. However, these approaches also suffer from needing a physical model which might not always be available. The uncertainty regarding the shape of the object also makes the task difficult since estimating all possible shape changes of an object is not easy with shape changes often being unpredictable.

Real-time processing of tasks involving deformable object tracking with computer vision remains a major challenge [25]. The complexity comes from the high dimensionality of deformable objects, needing to repetitively compute changes in the state of the object within a short time span. Obtaining results in a quick manner necessitates a trade-off between accuracy and speed. While training data exists for solutions that utilize artificial intelligence models [24] it is still relatively limited, mainly because, it is never possible to consider all deformations for the objects in a set [6].

This compounded by the fact that deformable objects are very diverse with individuals within groups of deformable objects presenting unique deformations. Learning models expected to work with deformable objects require a good capacity for generalization. Manipulation, planning, and control while handling deformable objects is also a significant challenge. Traditional planning and control techniques applied with rigid objects are not applicable [3].

The high dimensionality and non-linear nature of flexible objects requires the development of specialized case specific techniques, such as in [8, 47]. Significant strides continue to be made in the field; however, a long road still lies ahead. The challenges mentioned throughout this section are a reminder of the need for continuous research and development related to tracking and manipulation of these of objects. Upcoming sections will delve into some of the most promising advancements in the field.

4 Deep Learning Techniques for Object Tracking

The recent boom in machine learning that began in the latter part of the 2010s gave way to important developments in the realm of deep learning. Some of the most notable advancements include Convolutional Neural Networks [16, 13], Recurrent Neural Networks [32], Generative Adversarial Networks [10], Deep Reinforcement Learning [46] as well as Transformers and Attention Mechanisms [11].

Although these deep learning technologies were originally intended to be used with certain types of tasks in mind such as computer vision or natural language processing, the field of robotics has managed to adapt part of these technologies for tasks relevant to it. Convolutional Neural Networks have had a major role in expanding tracking capabilities in robotics systems.

For instance, in [29] the research team utilized a hierarchy of deep convolutional neural networks to categorize deformable objects as well as recognizing their pose. The framework used by the team utilizes two distinct CNN layers, the first one is used for classifying a certain garment based on predefined categories.

Meanwhile, the second layer consists of a category specific CNN to perform pose estimation, with the team testing their approach not only in simulations, but also with an actual robotic platform. In their implementation, the deformable garments are hung from a point to have a recognizable starting shape for the deformable object. The task of interacting with garments in a certain state is applicable in the industry while also being a desirable skill for future service robots.

The proposed pipeline achieved a recognition rate of 89.38 percent, outclassing state-of-the-art methods at the time of its release. Deep learning methods are making important contributions to the field of humanoid robotics. One notable study [50] outlines a novel approach for creating a humanoid robot capable of functioning in a manufacturing setting. This approach considers four principles for the robot and the algorithm: the ability to execute tasks, refine performance through repetition, adapt to new situations, and be easily implemented in real-world settings. Data for training is initially gathered through teleoperation.

The study employs a two-stage deep learning model, where a Deep Convolutional Autoencoder (DCAE) is responsible for image feature extraction and reconstruction, while a Time-Delay Neural Network (TDNN) learns the task dynamics based on these features and robot motion data. Since CNNs can handle considerably more input dimensions than fully connected neural networks with less parameters decreasing training time and enhancing the performance for image processing tasks such as feature or edge extraction on images, they present a solid alternative when working with vision tasks.

With a trained DCAE, half of the structure of the model is dedicated to encoding (compressing) the information to small-dimension image features which can represent the state of an input image and provide high-resolution input information in less dimensions, with batch normalization being used in the process to reduce the possibility of overfitting. The second half of the model is used to decode (reconstruct) the information to extracted image features.

The other neural architecture is the Time-Delay Neural Network (TDNN) [23]. TDNNs are designed to work with data that has a temporal component, making them ideal for tasks like sequence reconstruction. In [50], researchers employed a TDNN with multiple layers to generate continuous sequences by dynamically adjusting the input window over time.

This real-time adjustment is made possible by continuously feeding the network with new sensory and motor data, such as images captured by a camera and robot arm movements. The TDNN model in this study was capable of learning from multiple sensory-motor signal inputs, with the features extracted from the DCAE and robot motions being used by it. The experimental setup included a robot equipped with two 6-DOF non-back-drivable arms and a camera for precise manipulation.

Non-supervised learning has also been utilized to a great extent in relation to robotics and deformable object manipulation. For instance, in [31], the research team utilized a reinforcement learning approach to attempt to instruct a robot arm on how to interact with a deformable object in the form of cloth, considering actions such as diagonal folding, hanging, and folding the cloth up to a mark.

This was done utilizing a reinforcement learning technique known as Deep Deterministic Policy Gradients from Demonstrations (DDPGfD) [44], which is an extension of a prior algorithm known as Deep Deterministic Policy Gradients (DDPG) [26]. DDPG is a popular deep reinforcement learning technique used for solving continuous action space problems, meaning that the possible actions the agent can take exist in a continuous domain.

DDPGfD improves upon the original method by encouraging behavioral cloning, meaning that the agent that undergoes the reinforcement learning is rewarded by following a set of predetermined observations provided by an expert system. This helps mitigate usual reinforcement learning hassles such as slow convergence or having vast differences throughout iterations due to each of these iterations being conformed by random actions. The research team in charge of this study utilized the approach in both simulated and real-world environments, having a success rate of 90 percent when dealing with diagonal cloth folding in simulations while scoring a 66.6 percent grasp accuracy in the real world.

Transformer models carry with them great amounts of potential being capable of learning the data they are given in an intricate manner thanks to their attention mechanisms, which allow for models to focus on specific features of the input data that have a high relevance for the task at hand. In [15], the research team utilized the robotics specific transformer models TimeSformer [4] and ViViT [2] for the task of predicting slip detection when robotics systems handle fruits (deformable objects) which is relevant to factory and service robotics, outperforming models that made use of a CNN + LSTM pipeline that utilized Resnet18 [17].

The TimeSformer model works by processing spatial-temporal dimensions sequentially. The attention is applied first in the temporal dimension of the inputs at the same spatial position. Since these transformers work with computer vision tasks, the input image is dismantled into patches which are later flattened then linearly embedded to vectors of a certain size with positional embedding being added to each of them with a classifying token designed to extract task-level representations by tending to all other vectors. From there an input matrix is obtained and fed into a series of transformer layers.

The output of the classifying token is then used by a variety of tasks. The ViViT transformer works similarly to the TimeSformer model with a few key differences. For instance, ViViT processes the involved dimensions in parallel with half of the heads attending the spatial dimension and the other half the temporal dimension. The outputs are later combined via concatenation adding a linear transformation to half the size. It also has no need for a classification token, instead, the average of all output patches from the last Transformer layer is obtained passing it to a Multi-Layer Perceptron network (deep learning network used for classification) to predict a slip detection.

5 Integration of Object Recognition and Depth Perception Algorithms

Having a robotics system be capable of performing object recognition as well as having a certain ability to understand its environment via depth perception can be extremely beneficial towards working with deformable objects. Earlier research regarding the topic would utilize the extraction of visual features, such as silhouette features [22, 48, 21], however new more diverse approaches have been implemented partly thanks to the existence of low-cost RGB-D cameras.

These cameras provide both color (RGB) and depth information (D), enabling robots to better understand their environment in three dimensions. The two main currents regarding deformable object recognition tend to identify the object either when it is on a table [49, 41, 35] and those that recognize it when the object in question is hanging from a gripper [22, 12].

When the classification is done with the object on a table, it can be done with a single image, which is known as single shot perception. However, the scores obtained by single shot perception are usually outmatched by approaches where the robotics system interacts with the deformable object, gaining more information about the topology and deformations present in that object. In some instances, a 3-D model that comes because of the volumetric features in 3-D images.

Table 1. Data in the table is sourced from the original articles. Entries marked with "NA" indicate that the information was not reported in the original study. The abbreviation "FF" stands for Feed-Forward, used in the context of neural network-based methods.

Category	Method	Year	Accuracy	Inference Time
Deep Learning	Inception-V3 CNN	2018	88.03 %	Real-time
	TimeSformer	2021	85.0 %	2.46 s (FF)
	ViViT	2021	83.9 %	2.43 s (FF)
	Pose-recognition CNN	2015	89.4 %	1.8 s
	2-phase DL Model	2016	77.8 %	Real-time
	RL for manipulation	2018	90 % (folding)	24h
Deformation Models	Volumetric Approach	2014	90 % (Shorts)	0.22 s
	3D DLO Shape detection	2021	2.8 pixel error	0.2 s
	Clothes state recognition	2009	81.5 %	5-20 s
	Model-driven clothes state estimation	2002	72.41%	NA

Considering a continuous perception of dynamic interactions focusing on the material and shape of a deformable object as it is being interacted with can be useful when trying to understand said object for classification purposes [30]. In [30] the research team adopts a framework inspired by [41] and [22].

The work in question presents and demonstrates a continuous visual perception approach for deformable object classification while a robot picks and observes how the object in question changes over time. The team extracts visual features from 2.5 D images in consecutive frames to learn a temporal-consistent representation of the dynamic attributes present in a particular piece of clothing.

First, the object is placed in a random configuration on a flat surface where a robot grasps it to observe its physical deformation. Different views of the object are utilized (egocentric and exocentric). This implementation was evaluated using two clothing databases with it working well for highly deformed garments. It obtained an accuracy of 66.7 percent among five categories, having an increase of 39.4 percent of classification score when considering other current approaches to clothing perception and recognition.

6 Context-Awareness and Multi Sensor Data Integration

Context-awareness in robotics and computer vision systems indicates that the systems that have this characteristic have a certain degree of knowledge about their environment. This knowledge is frequently obtained via a particular sensor such as a camera, a proximity sensor, an RGB-D sensor among others. Typical parts of a context may include the location, identity, activity and state of people, groups, and objects [38].

Physical variables can also be considered such as temperature and lighting. The field of collaborative robotics is in fact reliant in context-awareness. For operations that include collaborative robots and humans, context awareness must be timely in order to keep a safe and efficient working environment.

In [45], the research team suggests that visual observation of the motion of human workers may provide useful clues about tasks that need to be performed, with these having solid potential of being explored. Context awareness is also prevalent in autonomous vehicles, where it plays a central role in maintaining not only the safety of the autonomous vehicle and its occupants but also the safety of everything and everyone outside.

In [37] an advanced driving assistance system (ADAS) [1] is proposed and implemented with it being based on context-awareness of the environment that surrounds the vehicle. Ontological context awareness ADAS represent a significant advancement when compared to current systems, since these can use their context regardless of having to meet certain conditions, such as for the street that the driver is transiting to have lines.

Ontological ADAS also consider Mobile entities, static entities, and context parameters [1]. In many cases additional sensors or functionalities are suggested as would be the case with improving low-level object detection or better estimating road conditions [14], as well as more safety centered approaches [39]. A great way in which humanoid robots could acquire information about both their context and the state of the deformable object is via multisensory input.

Tactile sensor implementations alongside visual sensors are some of the most promising tactics regarding the understanding of deformable objects. For instance, in [15] visual data is combined with tactile data coming from a GelSight sensor to estimate when a slip may happen to a robot arm manipulating deformable objects such as fruits. Similar ideas such as visual servoing complemented with additional sensor data are explored in [20].

7 Conclusion

Advancements in the handling of deformable objects are being driven by a variety of evolving approaches. The integration of deep learning techniques, such as reinforcement learning, and the use of transformer models lay the foundation for even better models that may come while already delivering solid results. Also, the inclusion of multi-sensory inputs and context awareness will certainly improve the results that humanoid robots will have in terms of managing and interacting with deformable objects. There are still many challenges in the way but the stride towards the understanding of deformable objects is not stopping.

The tracking and understanding of deformable objects is particularly important in the development of robotics systems capable of working in unstructured industrial environments. Expanding the capacity of robots to understand deformable objects in this type of environments will allow for them to be utilized in more ample, novel applications, while working in environments with non-ideal conditions [33]. Some of the most significant models discussed in the review, based on criteria such as performance, novelty, and impact are present in Table 1.

Acknowledgments. Supported by ITESM.

References

1. Armand, A., Filliat, D., Ibañez-Guzman, J.: Ontology-based context awareness for driving assistance systems. In: Proceedings of the IEEE Intelligent Vehicles Symposium Proceedings, pp. 227–233 (2014) doi: 10.1109/IVS.2014.6856509
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: ViViT: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6836–6846 (2021) doi: 10.1109/iccv48922.2021.00676
3. Arriola-Rios, V. E., Guler, P., Ficuciello, F., Kragic, D., Siciliano, B., Wyatt, J. L.: Modeling of deformable objects for robotic manipulation: A tutorial and review. Frontiers in Robotics and AI, vol. 7, pp. 82 (2020) doi: 10.3389/frobt.2020.00082
4. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Proceedings of the 38th International Conference on Machine Learning, vol. 139, pp. 813–821 (2021) doi: 10.48550/ARXIV.2102.05095
5. Billard, A., Kragic, D.: Trends and challenges in robot manipulation. Science American Association for the Advancement of Science, vol. 364, no. 6446 (2019) doi: 10.1126/science.aat8414
6. Blanes, C., Mellado, M., Ortiz, C., Valera, A.: Review. Technologies for robot grippers in pick and place operations for fresh fruits and vegetables. Spanish Journal of Agricultural Research, vol. 9, no. 4, pp. 1130 (2011) doi: 10.5424/sjar/20110904-501-10
7. Bohg, J., Morales, A., Asfour, T., Kragic, D.: Data-driven grasp synthesis—A survey. IEEE Transactions on Robotics, vol. 30, no. 2, pp. 289–309 (2014) doi: 10.1109/tro.2013.2289018
8. Caporali, A., Galassi, K., Palli, G.: 3D DLO shape detection and grasp planning from multiple 2D views. In: Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics, pp. 424–429 (2021) doi: 10.1109/AIM46487.2021.9517655
9. Chi, C., Berenson, D.: Occlusion-robust deformable object tracking without physics simulation. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 6443–6450 (2019) doi: 10.1109/IROS40897.2019.8967827
10. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A. A.: Generative adversarial networks: An overview. IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 53–65 (2018) doi: 10.1109/msp.2017.2765202
11. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186 (2019)
12. Doumanoglou, A., Kim, T. K., Zhao, X., Malassiotis, S.: Active random forests: An application to autonomous unfolding of clothes. In: Proceedings of the European Conference on Computer Vision, pp. 644–658 (2014) doi: 10.1007/978-3-319-10602-1_42
13. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015) doi: 10.1109/iccv.2015.169
14. Guo, C., Meguro, J., Kojima, Y., Naito, T.: A multimodal ADAS system for unmarked urban scenarios based on road context understanding. IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 4, pp. 1690–1704 (2015) doi: 10.1109/tits.2014.2368980
15. Han, Y., Batra, R., Boyd, N., Zhao, T., She, Y., Hutchinson, S., Zhao, Y.: Learning generalizable vision-tactile robotic grasping strategy for deformable objects via transformer. arXiv, pp. 1–11 (2021) doi: 10.48550/ARXIV.2112.06374
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017) doi: 10.48550/ARXIV.1703.06870

17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016) doi: 10.1109/cvpr.2016.90
18. IBM: What is computer vision? (2023) www.ibm.com/topics/computer-vision
19. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. International Journal of Computer Vision, vol. 1, no. 4, pp. 321–331 (1988) doi: 10.1007/bf00133570
20. Khalil, F. F., Payeur, P.: Dexterous robotic manipulation of deformable objects with multi-sensory feedback - A review. Robot Manipulators Trends and Development, pp. 587–619 (2010) doi: 10.5772/9183
21. Kita, Y., Kita, N.: A model-driven method of estimating the state of clothes for manipulating it. In: Proceedings of the 6th IEEE Workshop on Applications of Computer Vision, pp. 63–69 (2002) doi: 10.1109/acv.2002.1182158
22. Kita, Y., Ueshiba, T., Kanehiro, F., Kita, N.: Recognizing clothing states using 3D data observed from multiple directions. In: Proceedings of the 13th IEEE-RAS International Conference on Humanoid Robots, pp. 227–233 (2013) doi: 10.1109/humanoids.2013.7029980
23. Lang, K. J., Waibel, A. H., Hinton, G. E.: A time-delay neural network architecture for isolated word recognition. Neural Networks, vol. 3, no. 1, pp. 23–43 (1990) doi: 10.1016/0893-6080(90)90044-l
24. Li, J., Dong, S., Adelson, E.: Slip detection with combined tactile and visual information. In: Proceedings of the International Conference on Robotics and Automation, pp. 7772–7777 doi: 10.48550/ARXIV.1802.10153
25. Li, Y., Wang, Y., Case, M., Chang, S. F., Allen, P. K.: Real-time pose estimation of deformable objects using a volumetric approach. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1046–1052 (2014) doi: 10.1109/iros.2014.6942687
26. Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning. In: Proceedings of the International Conference on Learning Representations, pp. 1–14 (2016)
27. Lin, J., Calandra, R., Levine, S.: Learning to identify object instances by touch: Tactile recognition via multimodal matching. In: Proceedings of the International Conference on Robotics and Automation, pp. 3644–3650 (2019) doi: 10.1109/ICRA.2019.8793885
28. Lowe, D. G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110 (2004) doi: 10.1023/b:visi.0000029664.99615.94
29. Mariolis, I., Peleka, G., Kargakos, A., Malassiotis, S.: Pose and category recognition of highly deformable objects using deep learning. In: Proceedings of the International Conference on Advanced Robotics, pp. 655–662 (2015) doi: 10.1109/icar.2015.7251526
30. Martínez, L., Ruiz-del Solar, J., Sun, L., Siebert, J. P., Aragon-Camarasa, G.: Continuous perception for deformable objects understanding. Robotics and Autonomous Systems, vol. 118, pp. 220–230 (2019) doi: 10.1016/j.robot.2019.05.010
31. Matas, J., James, S., Davison, A. J.: Sim-to-real reinforcement learning for deformable object manipulation. In: Proceedings of The 2nd Conference on Robot Learning. Proceedings of Machine Learning Research, pp. 734–743 (2018)
32. Medsker, L. R., Jain, L.: Recurrent neural networks: Design and Applications. Chemical Rubber Co Press, Inc (1999)
33. Mitrano, P., MConachie, D., Berenson, D.: Learning where to trust unreliable models in an unstructured world for deformable object manipulation. Science Robotics, vol. 6, no. 54 (2021) doi: 10.1126/scirobotics.abd8170

34. Navarro-Alarcon, D., Yip, H. M., Wang, Z., Liu, Y. H., Zhong, F., Zhang, T., Li, P.: Automatic 3-D manipulation of soft objects by robotic arms with an adaptive deformation model. IEEE Transactions on Robotics, vol. 32, no. 2, pp. 429–441 (2016) doi: 10.1109/tro.2016.2533639
35. Ramisa, A., Alenya, G., Moreno-Noguer, F., Torras, C.: Using depth and appearance features for informed robot grasping of highly wrinkled clothes. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 1703–1708 (2012) doi: 10.1109/ICRA.2012.6225045
36. Roberts, L.: Machine perception of three-dimensional solids. Massachusetts Institute of Technology (1963)
37. Ryu, M., Cha, S. H.: Context-awareness based driving assistance system for autonomous vehicles. International Journal of Control and Automation, vol. 11, no. 1, pp. 153–162 (2018) doi: 10.14257/ijca.2018.11.1.14
38. Salber, D., Dey, A. K., Abowd, G. D.: The context toolkit: Aiding the development of context-enabled applications. In: Proceedings of the Special Interest Group on Computer-Human Interaction Conference on Human Factors in Computing Systems, pp. 434–441 (1999) doi: 10.1145/302979.303126
39. Shen, Y., Jeong, J., Oh, T., Son, S. H.: CASD: A framework of context-awareness safety driving in vehicular networks. In: Proceedings of the 30th International Conference on Advanced Information Networking and Applications Workshops, pp. 252–257 (2016) doi: 10.1109/WAINA.2016.74
40. Shridhar, M., Manuelli, L., Fox, D.: Perceiver-actor: A multi-task transformer for robotic manipulation. In: Proceedings of the 6th Conference on Robot Learning, pp. 1–28 (2022) doi: 10.48550/ARXIV.2209.05451
41. Sun, L., Aragon-Camarasa, G., Rogers, S., Stolkin, R., Siebert, J. P.: Single-shot clothing category recognition in free-configurations with application to autonomous clothes sorting. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 6699–6706 (2017) doi: 10.1109/IROS.2017.8206586
42. Tang, T., Fan, Y., Lin, H. C., Tomizuka, M.: State estimation for deformable objects by point registration and dynamic simulation. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2427–2433 (2017) doi: 10.1109/IROS.2017.8206058
43. Terzopoulos, D., Witkin, A., Kass, M.: Constraints on deformable models: Recovering 3D shape and nonrigid motion. Artificial Intelligence, vol. 36, no. 1, pp. 91–123 (1988) doi: 10.1016/0004-3702(88)90080-x
44. Vecerik, M., Hester, T., Scholz, J., Wang, F., Pietquin, O., Piot, B., Heess, N., Rothörl, T., Lampe, T., Riedmiller, M.: Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. arXiv, pp. 1–10 (2017) doi: 10.48550/ARXIV.1707.08817
45. Wang, P., Liu, H., Wang, L., Gao, R. X.: Deep learning-based human motion recognition for predictive context-aware human-robot collaboration. CIRP Annals, vol. 67, no. 1, pp. 17–20 (2018) doi: 10.1016/j.cirp.2018.04.066
46. Watkins, C. J. C. H., Dayan, P.: Q-learning. Machine Learning, vol. 8, no. 3–4, pp. 279–292 (1992) doi: 10.1007/bf00992698
47. Wi, Y., Florence, P., Zeng, A., Fazeli, N.: VIRDO: Visio-tactile implicit representations of deformable objects. In: Proceedings of the International Conference on Robotics and Automation, pp. 3583–3590 (2022) doi: 10.1109/icra46639.2022.9812097
48. Willimon, B., Birchfield, S., Walker, I.: Model for unfolding laundry using interactive perception. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4871–4876 (2011) doi: 10.1109/IROS.2011.6095066
49. Willimon, B., Walker, I., Birchfield, S.: Classification of clothing using midlevel layers. ISRN Robotics, vol. 2013, pp. 1–17 (2013) doi: 10.5402/2013/630579

50. Yang, P. C., Sasaki, K., Suzuki, K., Kase, K., Sugano, S., Ogata, T.: Repeatable folding task by humanoid robot worker using deep learning. IEEE Robotics and Automation Letters, vol. 2, no. 2, pp. 397–403 (2017) doi: 10.1109/lra.2016.2633383
51. Yuan, W., Wang, S., Dong, S., Adelson, E.: Connecting look and feel: Associating the visual and tactile properties of physical materials. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4494–4502 (2017) doi: 10.1109/cvpr.2017.478
52. Zickler, S., Efros, A.: Detection of multiple deformable objects using PCA-SIFT. In: Proceedings of the 22nd National Conference on Artificial Intelligence, vol. 2, pp. 1127–1132 (2007)