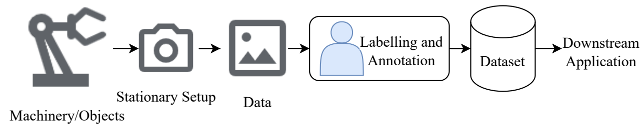
On the Application of Egocentric Computer Vision to Industrial Scenarios

Vivek Chavan*[1] Oliver Heimann[1] Jörg Krüger[1,2]

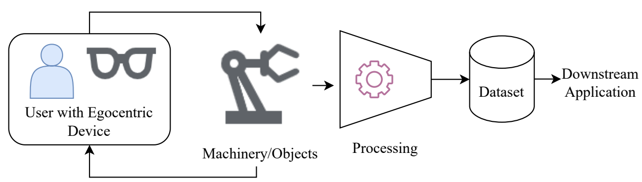
[1] Fraunhofer IPK [2] Technical University of Berlin, Germany

Abstract

Egocentric vision aims to capture and analyse the world from the first-person perspective. We explore the possibilities for egocentric wearable devices to improve and enhance industrial use cases w.r.t. data collection, annotation, labelling and downstream applications. This would contribute to easier data collection and allow the users to provide additional context. We envision that this approach could serve as a supplement to the traditional industrial Machine Vision workflow. Code, Dataset and related resources will be available at: https://github.com/Vivek9Chavan/EgoVis24



(a) The conventional data collection and labelling approach, involving a fixed, stationary setup.



(b) A proposed approach for automated data collection and annotation, where a user describes their observation while interacting with the object. The data is then processed to obtain the labelled dataset.

Figure 1. A comparison of the two approaches. Our work explores the latter.

1. Introduction

The field of Egocentric Computer Vision has seen increased attention in recent years [11, 15]. This has been catalysed due to the increased mainstream focus on wearable Augmented Reality (AR) and Virtual Reality (VR) devices [1, 4]. The Computer Vision community has introduced several novel datasets in recent years, with an aim to unlock and explore new challenges and innovations in this area [2, 9, 10]. These large and diverse datasets capture humans in varying everyday scenarios.

In contrast, we focus on industrial production scenarios. Industry 4.0, or smart manufacturing, focuses on digital transformation of product development, including manufacturing, use, maintenance, and recycling [7]. There is a significant gap between the current state-of-the-art in Artificial Intelligence (AI) and Computer Vision Research, and its integration into traditional production systems [14]. The bottleneck often tends to be the digitisation of workflows and the inability to capture the expertise of the Subject-Matter Experts (SMEs) proficiently.

The conventional exocentric/allocentric data collection in the industry is summarised in Figure 1a, which requires careful labelling, annotation and documentation for training AI models or knowledge transfer. In this ongoing research work, we study the use of lightweight egocentric devices for capturing multimodal egocentric data, which is processed via agentic workflow for adding task relevant labels and contextual information to the tasks. This is shown in Figure 1b and Figure 2.

*Correspondence: vivek.chavan@ipk.fraunhofer.de

2. Related Work

Egocentric Computer Vision. Understanding the world from the first-person perspective is intuitive for humans, but poses several challenges for conventional Machine Vision and AI methods [2, 9]. Several iterations and configurations of wearable devices have been proposed [3, 5, 9, 13]. Such devices enable additional user specific data to be captured alongside visual data, such as eye-gaze, hand pose, voice interaction. Several novel datasets and benchmarks have been released in recent years, which introduce new challenges and research directions for the community.

Industrial Machine Vision. Traditional image processing has led to several important advancements in the in-

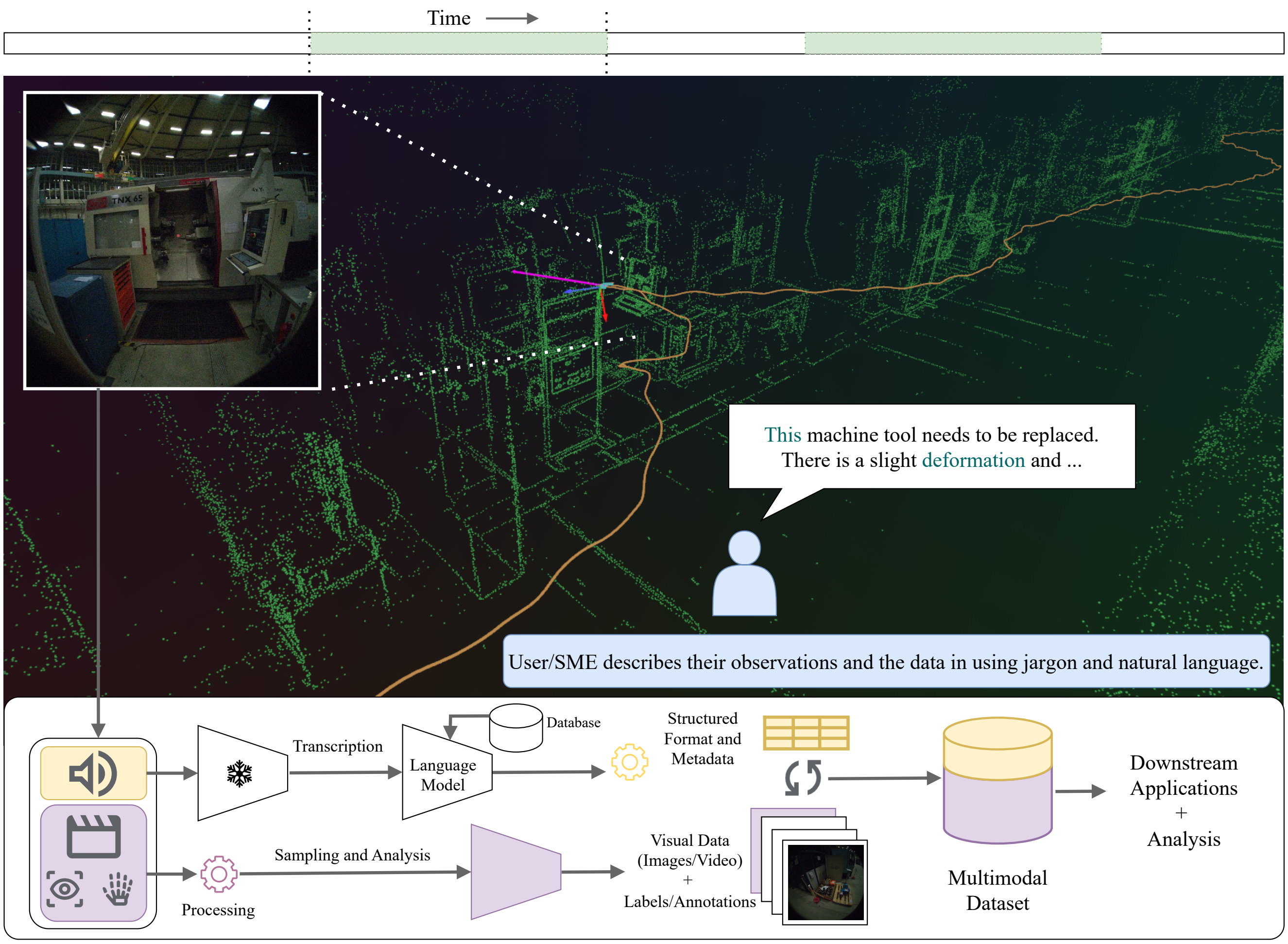


Figure 2. A summary of the proposed pipeline. The User/SME wearing the egocentric device interacts with the object/machinery and documents their observation in natural language. The multimodal dataset is then processed to obtain image/video data, and the transcription, eye-gaze, hand interaction provides the labels and annotations, along with metadata. **Top:** Point cloud reconstruction example from a use case. **Bottom:** A conceptualisation of the data processing.

dustry, which is further accelerated by deep learning based approaches [14]. The most important areas of application tend to be classification, object detection, segmentation, anomaly/defect detection.

3. Methods

Proposed Pipeline. Figure 2 shows the planned implementation in an industrial setting. We use the Meta Aria glasses [3] as the data capturing device. The multimodal data captured by the user is then processed to extract the most meaningful information about the process, or the machinery. The user guidance via voice serves as the lead indicator for understanding which portion of the continuous stream of the data should be processed. The audio data is processed via a custom language model setup, to obtain structured metadata and labels about the given portion of the stream. The camera stream data, augmented by user interaction (eye-gaze or hands), is processed and synchronised with the audio description data to add annotation and context (e.g. object labels, defects, miscellaneous observations). Additional processed data, such as user trajectory, location and other modalities would also be valuable for adding more context to the captured data.

Current egocentric datasets primarily encompass everyday and outdoor activities, with sparse representation of industry-specific scenarios, necessitating domain-specific data collection. Fine-grained activities like screwing and unscrewing bolts require high-resolution classification due to their visually similar but functionally distinct nature. We plan to open-source our data to encourage the broader research community to address such problems.

Challenges. Industrial Machine Vision often requires controlled settings and high precision image processing.

Egocentric data capture cannot fully replace standard digitisation stations and setups. In such cases, egocentric data would augment and assist the user in understanding the workflows and operations. Capturing user guidance via voice may be challenging due to noise, presence of other loud voices or perceived discomfort. In such cases, controlling parts of the user input via hand gestures or other means may be valuable. Additionally, capturing user eye gaze, hand gestures and other personal data poses inherent challenges in such cases. Egocentric AI systems produce large volumes of data, challenging the processing capabilities of on-device hardware. This necessitates adaptive, privacy-centric continual learning strategies and optimisation of data transfer to mitigate compute and bandwidth bottlenecks.

Downstream Applications. We explore diverse applications of egocentric computer vision in industrial settings, aimed at enhancing operational efficiency and accuracy [15]. Key applications include improving data collection and annotation through automated processes, enabling precise part recognition and analysis, and facilitating defect and anomaly labelling. Additionally, we investigate scene understanding and action recognition within these environments, offering substantial support for operators through real-time assistance and guidance. The system would also play a crucial role in training and knowledge transfer, ensuring that new and existing employees quickly adapt to evolving industrial demands. Each of these applications underscores the potential of egocentric computer vision to augment traditional industrial operations, making processes more intuitive and interactive.

Continual Learning. NN models often suffer due to catastrophic forgetting when they are retrained on newer data [6, 8]. It is necessary to develop systems that learn continuously and perform increasingly better on the most important tasks, while still retaining the knowledge from previous broad scale training. We believe such adaptive and intelligent systems would provide users with a more meaningful and helpful experience over time. For our work, we focus on a limited number of use cases and working environments. The participants perform similar tasks on different industrial objects multiple times over several weeks. These include tasks involving similarities w.r.t. the objects, working environments, and the actions being performed.

One of the key challenges with training and continually improving ML models based on egocentric inputs is handling the large amount of multimodal data available to the sensors every second. Moreover, personal data (such as the user eye-gaze and hand pose) or sensitive data (such as confidential office work, conversations at home or work) may not be suitable to be saved and sent for training. Hence, a federated or distributed learning paradigm would be required for handling the data and continually training the

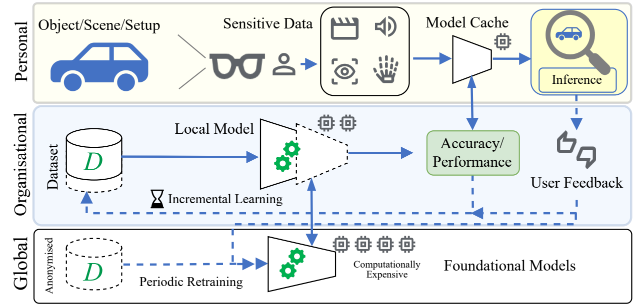


Figure 3. A summary of a distributed Continual Learning framework for egocentric applications. The three layers of application include personal (top), organisational (middle) and global (bottom). The most sensitive information is stored and processed by the personal computing setup with limited compute. The organisational layer trains the local models incrementally, which receive user feedback and related data from the egocentric device. The global foundational models require large amounts of data, which could be periodically shared by the organisation (after anonymization and review).

models [12, 16], as shown in Figure 3.

4. Summary

In this extended abstract, we propose an approach for automated data collection and labelling for industrial use cases. The methods and challenges were briefly discussed. This undertaking brings several eccentric benchmarks and tasks, including scene understanding, object detection and tracking, diarisation, action recognition, hand, and eye tracking, among others. We believe such workflows could significantly reduce the efforts required for digitisation and automation, and would improve knowledge transfer between SMEs and trainees, and aid the development of context aware models.

Acknowledgments

We thank the Meta AI team and Reality Labs for the Project Aria initiative, including the research kit, the open source tools and related services. The data collection for this study is carried out at the IWF research labs of TU Berlin.

References

[1] Apple. Apple vision pro, 2024. Accessed: 2024-05-10. 1

[2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[3] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talat-

tof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research, 2023. 1, 2

[4] Facebook. Facebook to acquire oculus, 2014. Accessed: 2024-05-10. 1

[5] Alireza Fathi, Ali Farhadi, and James M. Rehg. Understanding egocentric activities. In *2011 International Conference on Computer Vision*, pages 407–414, 2011. 1

[6] Robert French. Catastrophic interference in connectionist networks: Can it be predicted, can it be prevented? In *Advances in Neural Information Processing Systems*. Morgan-Kaufmann, 1993. 3

[7] German Federal Ministry of Economics and Climate Action (BMWK). Industry 4.0, 2024. Accessed: 2024-05-10. 1

[8] Ian J. Goodfellow, Mehdi Mirza, Xia Da, Aaron C. Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgeting in gradient-based neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 3

[9] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, 2022. 1

[10] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives, 2024. 1

[11] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1346–1353, 2012. 1

[12] Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020. 3

[13] S. Mann. Humanistic computing: "wearcomp" as a new framework and application for intelligent signal processing. *Proceedings of the IEEE*, 86(11):2123–2151, 1998. 1

[14] Daniele Mazzei and Reshawn Ramjattan. Machine learning for industry 4.0: A systematic review using deep learning-based topic modelling. *Sensors*, 22(22), 2022. 1, 2

[15] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision, 2024. 1, 3

[16] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020. 3