Mobilizing Visual Perception: A Strategy for Enhanced Human-Robot Interaction

Taehyeon Kim†
Contents Convergence Research Center
Korea Electronics Technology Institute
Seoul, Korea
taehyeon.kim@keti.re.kr

Seho Park
Contents Convergence Research Center
Korea Electronics Technology Institute
Seoul, Korea
sehopark@keti.re.kr

Junho Kwon
Contents Convergence Research Center
Korea Electronics Technology Institute
Seoul, Korea
kwonkai@keti.re.kr

***Abstract*—As robotic technologies continue to advance, their integration into daily human lives becomes ever more paramount. At the heart of this integration lies the challenge of fostering seamless human-robot interactions. This paper introduces a robust visual perception framework tailored for real-time operation on mainstream mobile devices. By harnessing the power of deep neural networks, our approach not only detects but also interprets human activities from visual cues. Such interpretation aids robots in making informed decisions during interactions. We demonstrate the efficiency of our framework on the Samsung Galaxy A24, a mid-tier smartphone, achieving impressive real-time performance. The results highlight the framework's potential to revolutionize human-robot interactions across diverse real-world scenarios.**

***Index Terms*—human-robot interaction, computer vision, mobile computing, deep learning, system integration.**

I. Introduction

In the realm of robotics, mobilizing visual perception stands as a transformative approach to revolutionizing human-robot interactions. With advancements in deep neural networks, robotic technology is evolving towards more intelligent robots [1]–[5]. While robots were once primarily used in industrial settings, their applications are now extending into our everyday lives. This shift has been made possible due to the robust performance of deep neural network-based visual perception modules, which outperform traditional methods.

Previous researches have touched upon the importance of visual perception in robotics, but few have delved into optimizing it for real-time mobile operations. Our research fills this gap, presenting an approach that not only recognizes but anticipates human interactions using visual cues Robots like delivery bots and service robots no longer simply function within fixed standards and predefined environments as their industrial counterparts do. Instead, they are expected to flawlessly execute tasks in varied and unpredictable settings. However, one of the largest variables in these settings is humans themselves. Thus, it's imperative for modern intelligent robots to be developed with human-robot interaction in mind. In this paper, we present a range of neural network-based computer vision algorithms essential for facilitating human-robot interactions. We delve into how these algorithms can discern interactions rooted in specific relationships. Given the constraints of robotic systems with limited hardware capabilities, it is vital to employ an appropriate neural network configuration. To this end, neural network optimization, or '*lightening*', is imperative. We put forth a comprehensive pipeline that encapsulates all these considerations.

At its core, to assess human-robot interactions, we chose to gather information that can be visually ascertained, rather than relying on verbal cues from human speech. For instance, as a robot navigates a path and comes across a myriad of individuals, deciphering their spoken expressions for interaction could be computationally demanding due to the varied inputs from multiple sources. Moreover, it's uncommon for passersby to engage in profound conversations with a robot in such settings [6], [7]. Consequently, our proposed technique seeks to gather data on various individuals via a camera. By inferring cues from the interactions they exhibit, we aim to discern whether their interactive intent is directed towards the robot or towards another individual or entity.

To discern human interactions, we gather several key data points. First, we identify the human figure and the surrounding objects. This enables us to approximately gauge a person's position or context using information about both the individual and the nearby objects. For example, detecting a person beside a car can suggest various interactions such as entering the car, exiting, avoiding the vehicle, or opening the trunk. Accordingly, through object detection techniques, we identify both human and non-human entities within the robot's surroundings. We then employ an '*object detection perception module*' to either directly or indirectly deduce their interrelations. Secondly, our approach *identifies key points on a human's face*, serving as the sole method to discern facial expressions. For many, the face is the quickest canvas for emotions, especially during interactions with robots. It's crucial to understand that our proposed human-robot interaction system doesn't signify deep connections; rather, it pertains to the myriad interactions robots have with diverse individuals while performing tasks. As such, uncommon scenarios like maintaining a 'poker face'—where emotions are deliberately concealed—are not typical in daily interactions [8], [9]. Hence,

Dr. Taehyeon Kim is the corresponding author and also first author.
This work was supported by Korea Evaluation Institute of Industrial Technology(KEIT) grant funded by the Korea government(MOTIE). (No 20009760. Development of human friendly multipurpose service robot and new market creation by applying human robot interaction design)

merely detecting facial expressions can be significantly helpful in gauging a person's feelings. Finally, we *detect human posture*. Humans convey emotions through their arms and legs and their posture can also hint at their current actions. Furthermore, when deducing emotions, considering posture can help rectify potential misinterpretations. For example, if we detect an expression indicating discomfort but observe the person in a running posture, we can surmise that the discomfort isn't a result of the robot's presence but likely due to the physical exertion from running.

Our proposed system functions as a pipeline that integrates multiple computer vision algorithms. Facilitating its real-time operation on mobile devices presents a formidable challenge. To address this, we've employed lightweight algorithms tailored for each module, allowing the system to function seamlessly on mainstream smartphones. Given its strong compatibility with the Robot Operating System (ROS), our human-robot interaction pipeline is well-suited for integration with various robotic platforms [10].

II. Proposed Framework

A. Object Detection Module

In our proposed framework, the object detection module incorporates a model provided by MLKit, which utilizes a technique derived from RCNN (Region-based Convolutional Neural Networks). RCNN is a groundbreaking method in object detection that combines region proposals with convolutional neural networks. Instead of treating object detection as a single regression problem, RCNN first identifies potential object-bound regions in an image using a method called Region Proposal Network (RPN) [11], [12]. These proposed regions are then passed through a CNN to classify objects and refine bounding box coordinates. RPN is a neural network that scans an image and proposes candidate object-bound regions. It does so by sliding a small window across the image and, at each position, predicting multiple bounding boxes and their associated objectness scores. One of the key strengths of the RCNN and its variants is the flexibility in the classifier. By substituting the classifier, one can adapt the model to operate in diverse environments and detect various object categories. MLKit's variation on RCNN maintains this versatility, ensuring that the object detection model is adaptable and can function in a wide range of scenarios.

B. Facial Key-points Estimation Module

The Facial Key-points Estimation Module boasts a comprehensive suite of functionalities critical to human-robot interaction. By harnessing advanced facial feature recognition, this module can accurately determine the coordinates of significant features such as the eyes, ears, cheeks, nose, and mouth. This detailed recognition assists the robot in understanding not only the orientation of a face but also where a person might be focusing their gaze. In addition to feature recognition, the module meticulously maps the contours of the face, detailing the eyes, eyebrows, lips, and nose. This in-depth analysis provides the robot with a nuanced understanding of facial expressions, equipping it to detect and interpret subtle emotional shifts [13], [14]. Emotions are integral to interactions, and the module's proficiency in recognizing specific facial expressions like smiles or closed eyes offers invaluable insights into the individual's emotional state. Such insights enable the robot to tailor its responses to resonate with the user's emotions and expectations.

Consistency is paramount in dynamic scenarios, especially in video streams. The module's capability to track faces across video frames, attributing a unique identifier to each, ensures that interactions remain consistent and coherent. Such continuous tracking guarantees that the robot remains engaged and responsive, even if the individual changes their position or orientation. Furthermore, the emphasis on real-time video frame processing ensures seamless and rapid operations. This means the robot can promptly interpret facial cues and respond in real time, mirroring the pace of human interactions and ensuring fluid communication. In conclusion, the advanced features of the Facial Key-points Estimation Module allow the robot to achieve a deeper understanding of human users. By astutely interpreting facial cues and emotions, the robot can engage in interactions that are not only responsive but also empathetic, cultivating a more intuitive and natural human-robot interaction experience.

C. Posture Detection Module

Our posture detection module utilizes MLKit's pose estimation model to detect a total of 33 key points on the human body. These key points provide a comprehensive representation of human posture, capturing crucial joints and body parts including the head, shoulders, elbows, hips, knees, and feet, among others. Identifying these key points and understanding the spatial relationships between them allows us to infer the direction a person is facing or moving towards [15]. For instance, the relative positions of the shoulders and hips can give insights into the orientation of the torso, while the positioning of the feet can indicate the direction of movement. In the context of human-robot interaction, discerning the direction a person is facing or heading is of paramount importance. A robot's ability to recognize if a person is facing towards or away from it can drastically affect its decision-making process [16]. If a person is facing the robot, it might indicate potential engagement or interaction. On the other hand, if they are facing away, the robot might interpret it as disinterest or intent to move in another direction.

Furthermore, understanding a person's orientation can assist the robot in anticipating human actions and consequently, in making proactive decisions. For instance, if a person's posture indicates they are about to turn towards the robot, the robot can prepare for potential interaction, perhaps slowing down or stopping if it's in motion. In essence, capturing these 33 key points and interpreting them correctly forms the crux of effective human-robot interaction. It ensures that the robot operates in harmony with human intent, paving the way for smoother and more intuitive interactions.

D. Mobilizing Visual Perception

For our framework to operate efficiently on mobile devices, we have applied quantization techniques, including Quantization-Aware Training (QAT) [17]. Quantization is the process of constraining an input from a large set to output in a smaller set. In the context of neural networks, it usually means reducing the precision of the weights and activations to reduce memory and computational costs. Mobile devices, with their limited computational resources and memory, can struggle to run large deep learning models. Additionally, the power consumption associated with running these models can be a significant issue for battery-operated mobile devices. While reducing the size of the neural network is one solution, it often comes at the cost of model accuracy [18], [19]. Quantization, on the other hand, provides a balance. It reduces the model size and computational requirements without significantly compromising the model's performance. Quantization-Aware Training (QAT) is an advanced technique where the training process itself is aware of the quantization, ensuring that the quantized model maintains the accuracy of the original model.

One might wonder why not just reduce the network's size instead of quantizing? While reducing the network's size can lead to faster inference, it might not always result in reduced memory usage, especially when considering the storage of model parameters. Quantization not only speeds up inference but also drastically reduces the memory footprint of the model. This is particularly crucial for mobile devices where both speed and memory are at a premium. In conclusion, quantization, especially with techniques like QAT, is essential for deploying deep learning models on mobile devices. It ensures that the models are not only fast but also memory-efficient, allowing for real-time operations without draining the device's battery. Furthermore, we utilize TensorFlow Lite (TF Lite) for deploying our models on mobile platforms [20]. TF Lite is TensorFlow's lightweight solution for mobile and embedded devices. It allows deep learning models to run on-device, ensuring low latency and small binary sizes. By leveraging TF Lite, we ensure that our framework is not only accurate but also efficient and easily deployable across a variety of devices.

III. Experimental Setup

In order to validate the effectiveness and efficiency of our proposed human-robot interaction framework, we conducted experiments on a mainstream smartphone, the Galaxy A24. The choice of this device is strategic. It represents a typical mid-range smartphone, ensuring that our system is not tailored only for high-end devices but can be applied broadly across a spectrum of devices that people use daily.

Device Specifications Relevant to our Experiments:

- *Processor:* MediaTek Helio G99 MT6789 SoC. This mid-tier processor is representative of the general processing capabilities of most smartphones in the market.
- *Memory:* 4 GB LPDDR4X SDRAM and 128 GB UFS 2.2 internal storage. The memory ensures adequate speed for real-time operations.

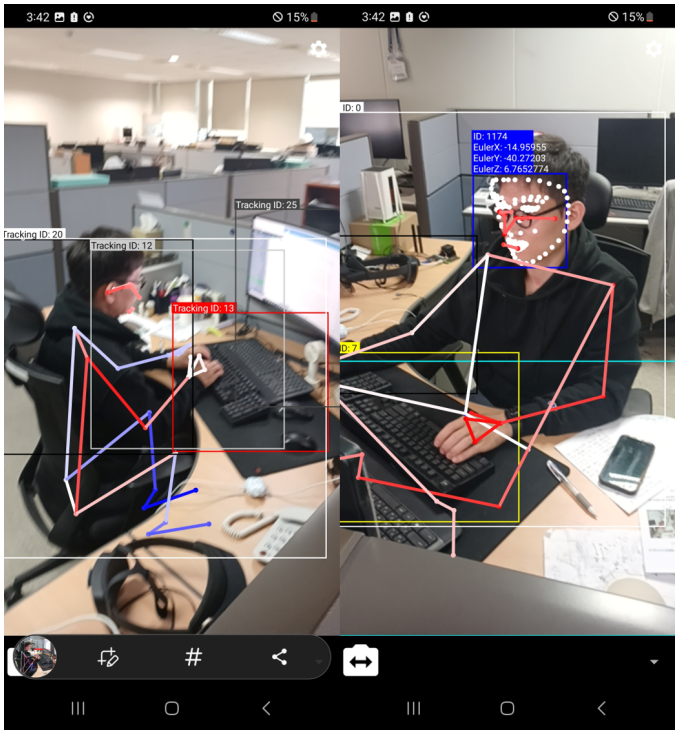

Fig. 1. Screenshot of our system in action on the Galaxy A24.

- *Display:* 6.5-inch 19.5:9 aspect ratio 2340 x 1080 Infinity-U Display (396 ppi). The display size and resolution are ideal for visual interactions.
- *Camera:* Front camera with 13 million pixels (F/2.2). Essential for our framework's visual perception module.
- *Battery:* Integrated Li-Ion 5,000 mAh. This ensures that our system can run for prolonged periods without draining the device's battery quickly.
- *Operating System:* Android 13 with Samsung One UI 5.1. This modern OS ensures compatibility with the latest algorithms and tools.

The Galaxy A24, while boasting decent specifications, does not have the computational prowess of high-tier smartphones. Yet, it is crucial for human-robot interaction systems to be deployable on such devices, given their widespread use. By choosing the Galaxy A24, we deliberately placed our system in a challenging environment. Our decision to refactor all our programs into Java further ensured the framework's compatibility with the Android ecosystem. This optimization is paramount as Java is the primary development language for Android, offering direct integration with the device's hardware and software functionalities. In essence, if our proposed system can efficiently function in real-time on the Galaxy A24, it signifies that our framework is not just robust but also highly optimized. Demonstrating effectiveness on such a device underscores the practicality and universality of our approach, making it a viable solution for a myriad of real-world applications.

In Figure 1, we present a snapshot taken directly from the

Galaxy A24, showcasing our system in action. The system impressively achieved a frame rate of 40 FPS, which is near real-time for most human-robot interaction scenarios. The image vividly depicts a person seated and engaging with a computer. Our system, true to its capabilities, successfully detects various objects in the scene, such as the desk, computer, person, and keyboard. Beyond mere object detection, the granularity of information acquired paints a clear picture of the individual's activity: working at a desk using a computer. This demonstration underscores the prowess of our approach. Not only does our system accurately detect and discern objects, but it also effectively contextualizes human activities. Such context is pivotal for robots to interpret human intentions and subsequently make informed decisions. The ability to run this sophisticated system on the Galaxy A24, a mainstream smartphone with the given hardware constraints, serves as a testament to the efficiency of our system. It also highlights the significance of our work as we endeavor to make human-robot interaction more intuitive and widespread. Given these promising results, we envision our framework being employed in various real-world scenarios, fostering seamless interactions between humans and robot

IV. Conclusion

Human-robot interaction is at a crossroads, where the need for intuitive and seamless communication between humans and machines has never been more paramount. This paper sought to address this pressing need by introducing a comprehensive visual perception framework, specifically optimized for real-time operations on mainstream mobile devices. Throughout our research, we underscored the significance of detecting and interpreting human activities through visual cues, emphasizing that such insights are fundamental for robots to respond appropriately in real-world interactions. A primary distinction of our work lies in its deployment on the Samsung Galaxy A24, a mid-tier smartphone. By choosing a device that represents an average computational capability available to many, we showcased that our system is not just for the elite few with access to high-end devices but is a solution designed for the masses. Our experiments yielded impressive results, with the system achieving near real-time performance, even in the constrained environment of the Galaxy A24. Such results not only highlight the efficiency and optimization of our framework but also underscore its robustness. The ability to detect and contextually interpret complex human activities, like a person working at a desk, indicates the depth and granularity of our approach. Furthermore, the successful deployment on a popular smartphone operating system, through our decision to refactor our programs into Java, signifies a broader applicability. Android, being the dominant OS in the global market, ensures that our approach has the potential to be rolled out on a large scale, making human-robot interaction more accessible and widespread. However, the journey of refining human-robot interactions does not end here. While our framework offers a promising step forward, continuous iterations and improvements are essential to cater to the ever-evolving dynamics of human behaviors and technological advancements. It is also crucial to test our framework across a wider variety of devices and real-world scenarios to further ascertain its robustness. In conclusion, our research stands as a testament to the potential of integrating advanced computer vision techniques with mobile computing for revolutionizing human-robot interactions. By bridging the gap between humans and robots in real-world scenarios, we pave the way for a future where robots are not mere machines but collaborative partners that understand and resonate with human intentions and emotions.

References

[1] Sheridan, Thomas B. "Human–robot interaction: status and challenges." Human factors 58.4 (2016): 525-532.

[2] Hancock, Peter A., et al. "A meta-analysis of factors affecting trust in human-robot interaction." Human factors 53.5 (2011): 517-527.

[3] Dautenhahn, Kerstin. "Socially intelligent robots: dimensions of human–robot interaction." Philosophical transactions of the royal society B: Biological sciences 362.1480 (2007): 679-704.

[4] Steinfeld, Aaron, et al. "Common metrics for human-robot interaction." Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction. 2006.

[5] De Santis, Agostino, et al. "An atlas of physical human–robot interaction." Mechanism and Machine Theory 43.3 (2008): 253-270.

[6] Savery, Richard, and Gil Weinberg. "A survey of robotics and emotion: Classifications and models of emotional interaction." 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). IEEE, 2020.

[7] Kwon, Dong-Soo, et al. "Emotion interaction system for a service robot." RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication. IEEE, 2007.

[8] Lottridge, Danielle, Mark Chignell, and Aleksandra Jovicic. "Affective interaction: Understanding, evaluating, and designing for human emotion." Reviews of Human Factors and Ergonomics 7.1 (2011): 197-217.

[9] Brave, Scott, and Clifford Nass. "Emotion in human-computer interaction." Human-computer interaction fundamentals 20094635 (2009): 53-68.

[10] Quigley, Morgan, et al. "ROS: an open-source Robot Operating System." ICRA workshop on open source software. Vol. 3. No. 3.2. 2009.

[11] Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.

[12] Bharati, Puja, and Ankita Pramanik. "Deep learning techniques—R-CNN to mask R-CNN: a survey." Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019 (2020): 657-668.

[13] Sun, Yi, Xiaogang Wang, and Xiaoou Tang. "Deep convolutional network cascade for facial point detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2013.

[14] Jin, Sheng, et al. "Whole-body human pose estimation in the wild." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. Springer International Publishing, 2020.

[15] Murphy-Chutorian, Erik, and Mohan Manubhai Trivedi. "Head pose estimation in computer vision: A survey." IEEE transactions on pattern analysis and machine intelligence 31.4 (2008): 607-626.

[16] Haralick, Robert M., et al. "Pose estimation from corresponding point data." IEEE Transactions on Systems, Man, and Cybernetics 19.6 (1989): 1426-1446.

[17] Hubara, Itay, et al. "Quantized neural networks: Training neural networks with low precision weights and activations." The Journal of Machine Learning Research 18.1 (2017): 6869-6898.

[18] Kim, Taehyeon, Heungjun Choi, and Yoonsik Choe. "Automated Filter Pruning Based on High-Dimensional Bayesian Optimization." IEEE Access 10 (2022): 22547-22555.

[19] Kim, Taehyeon, and Yoonsik Choe. "Fast circulant tensor power method for high-order principal component analysis." IEEE Access 9 (2021): 62478-62492.

[20] Girija, Sanjay Surendranath. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." Software available from tensorflow. org 39.9 (2016).