Development of a Smartphone-Based Object Recognition Prototype Application "VocabEyes" for Trilingual Vocabulary Learning for the Visually Impaired

Warupin Butkan
Faculty of Information Technology
Thai-Nichi Institute of Technology
Bangkok, Thailand
2463110342@tni.ac.th

Saprangsit Mruetusatorn
Faculty of Information Technology
Thai-Nichi Institute of Technology
Bangkok, Thailand
*Corresponding author : saprangsit@tni.ac.th

***Abstract*—This research presents the development of an Android-based prototype application named "VocabEyes" to assist the visually impaired in learning trilingual vocabulary (Thai, Japanese, and English) from their surrounding environment. Addressing the limitations of existing assistive tools that often require internet connectivity and lack continuous multilingual support, this system operates entirely offline through on-device processing. The application utilizes computer vision technology for whole-frame image classification, employing a MobileNetV3 model pre-trained on the ImageNet-21K-P dataset (11,221 classes), which was selected for its low latency and high hardware efficiency on mobile devices. To convert visual data into language, the system integrates a custom dictionary based on WordNet semantic mapping, linking image labels to approximately 10,000 English, 8,500 Thai, and 6,000 Japanese words. The results are then continuously read aloud using multilingual Text-to-Speech (TTS) technology. The user interface follows Universal Design principles, incorporating full-screen gesture controls and high-contrast visuals to ensure accessibility. Technical evaluation was conducted utilizing 600 test images of 10 targeted objects across six distinct scenarios, including complex cases such as background clutter, partial occlusion, and low-light conditions. Based on these 10 objects, the results revealed a high average Top-1 Accuracy of 93.83% and a Top-5 Accuracy of 98.33%. As a notable limitation, the preliminary usability study was conducted with a sample group of only 5 blindfolded normal-vision volunteers to simulate visual impairment. Although the evaluation yielded an average System Usability Scale (SUS) score of 89.00 (Excellent level) and a user satisfaction rating of 4.80 out of 5.00, these results should be interpreted with caution. These findings highlight the potential of the offline architecture and primarily serve for a proof-of-concept as an introductory foreign language learning tool. Comprehensive clinical validation with actual visually impaired users remains essential in future work to definitively confirm its real-world usability.**

Keywords—image classification, mobilenetv3, text-to-speech, visually impaired, trilingual learning

I. INTRODUCTION

Vision is the most crucial sensory modality in all aspects of daily life. According to data from the Department of Empowerment of Persons with Disabilities, Thailand has over 171,953 visually impaired individuals who frequently encounter limitations in perceiving their surroundings. [1] [2]. In the current era of borderless communication technologies, learning foreign languages, particularly English, acts as a vital bridge for the visually impaired to access social information and educational opportunities [3]. English serves as a universal language for accessing global accessibility tools and extensive information resources. Meanwhile, Japanese is significant as Japan is a global leader in accessible technology. This aligns with the Japanese government's educational support policies, which provide funding through the Grant Assistance for Grassroots Human Security Projects (GGP), aiming to enhance the quality of education for all visually impaired individuals in Thailand [4].

Modern smartphone technology has advanced to include high-performance processors capable of supporting computer vision tasks, effectively acting as "eyes" for users [5] Although applications for the visually impaired currently exist, such as Seeing AI and Lookout, most face limitations regarding continuous multilingual reading support, internet dependency, and a lack of direct focus on vocabulary education [6].

Therefore, this research aims to develop a prototype Android application named "VocabEyes" designed to classify objects from photographs and continuously read the vocabulary aloud in three languages (Thai, Japanese, and English). The application emphasizes entirely on-device processing and User Experience/User Interface (UX/UI) design tailored specifically for accessibility by the visually impaired users. The target for this application encompasses individuals across varying degrees of visual impairment. This includes individuals with low vision, who can benefit from the application's high-contrast visuals, as well as totally blind individuals, who are fully supported through the implementation of full-screen touch gestures and startup auditory guidance.

The research contributions include

1. The development of a trilingual educational prototype application capable of 100% offline processing (On-device processing), thereby eliminating internet dependency constraints.
2. The creation of a trilingual semantic mapping dataset, constructed specifically by the researcher utilizing the WordNet structure, linking ImageNet image labels to approximately 6,000-10,000 WordNet vocabularies to translate image classification results into multiple languages for practical use.

3. An accessibility-focused application design employing Universal Design principles, utilizing full-screen gestures instead of button-groping, and incorporating high-contrast colors.
4. The design of testing in complex environments, evaluating the model's accuracy across 5 challenging real-world scenarios (e.g., low light, occlusion, background clutter) to determine the model's true potential and limitations.

II. LITERATURE REVIEW

This research integrates multidisciplinary technological knowledge to construct a comprehensive system:

A. MobileNetV3 Image Classification Model

MobileNetV3 is a Convolutional Neural Network (CNN) architecture specifically designed for mobile devices, offering high accuracy with low latency [7]. This study selected the MobileNetV3-Large model, pre-trained on the ImageNet-21K-P dataset, which encompasses 11,221 object classes [8]. It serves as a ready-to-use model equipped with an extensive visual knowledge base suitable for everyday life applications. This research opted for whole-frame image classification rather than object detection (e.g., YOLO) because visual impairment scenarios often present multiple objects in a single frame. Classification extracts only the most prominent object as the primary answer, reducing complexity and preventing auditory information overload for the user. Furthermore, MobileNetV3 outperforms other state-of-the-art models such as MnasNet, ProxylessNas, and MobileNetV2 [9].

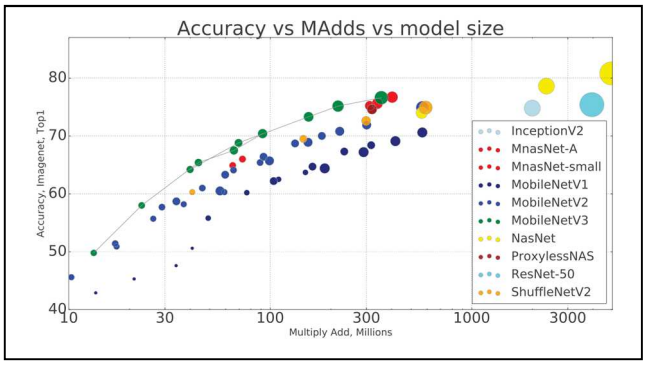

Fig. 1. Accuracy vs MAdds vs model size of each model [9]

Network	Top-1	MAdds	Params	P-1	P-2	P-3
V3-Large 1.0	**75.2**	**219**	5.4M	**51**	**61**	**44**
V3-Large 0.75	73.3	155	4.0M	39	46	40
MnasNet-A1	75.2	315	3.9M	71	86	61
Proxyless[5]	74.6	320	4.0M	72	84	60
V2 1.0	72.0	300	3.4M	64	76	56
V3-Small 1.0	67.4	66	2.9M	15.8	19.4	14.4
V3-Small 0.75	65.4	44	2.4M	12.8	15.6	11.7
Mnas-small [43]	64.9	65.1	1.9M	20.3	24.2	17.2
V2 0.35	60.8	59.2	1.6M	16.6	19.6	13.9

Fig. 2. Floating point performance comparison on mobile devices [9]

As detailed in the Fig. 2, MobileNetV3-Large demonstrates superior hardware efficiency. Specifically, while achieving an identical Top-1 accuracy of 75.2% to MnasNet-A1, it requires significantly fewer computational resources (219 MAdds vs. 315 MAdds) and delivers substantially lower inference latency across mobile devices (e.g., 51 ms vs. 71 ms on Pixel-1). This optimal balance between model parameters, computational cost, and real-world hardware latency strictly aligns with the efficient on-device processing requirements of this study [9].

B. WordNet Vocabulary Database and TTS Technology

The system correlates visual results to language utilizing the WordNet database structure [10], which facilitates semantic mapping from English to Thai and Japanese. Subsequently, the text is converted into speech using multilingual Text-to-Speech (TTS) technology.

C. Universal Design

User Interface (UI) development adhered strictly to accessibility and haptic perception principles. To accommodate users' tactile discrimination limitations, the design avoids reliance on precise button-groping. Instead, it employs full-screen gesture controls (e.g., tapping anywhere on the screen) as enlarged touch targets. Additionally, automated audio guidance upon application launch provides immediate acoustic feedback, ensuring comprehensibility and usability even for completely blind users [11].

III. METHODOLOGY

This experimental Research and Development (R&D) study was conducted under a conceptual framework divided into 5 primary phases:

A. Phase 1: Information Study and System Requirements Definition

The researcher established the following core system requirements:

- *Functional Requirements:* The system must classify objects via the smartphone camera or gallery, display results on-screen, and output speech via TTS in 3 languages (Thai, Japanese and English).
- *Non-Functional Requirements:* The system must possess 100% offline capability.
- *User Features:* Support for full-screen touch gestures, high-contrast UI for low vision users, and an automated tutorial guide upon opening the application.

B. Phase 2: System Design and Development

A cross-platform application structure was developed using Flutter, integrating the primary system pipeline:

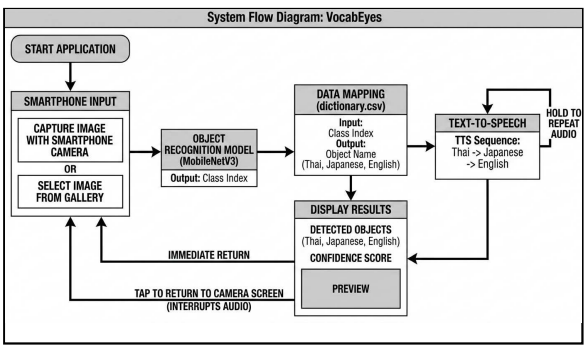

Fig. 3. System Flow Diagram

- *Image Classification Model:* Images are imported and processed using MobileNetV3-Large, converted into

a .tflite file for Flutter integration, outputting the Class index of objects.

- *Data Mapping:* The obtained label is matched against a custom dictionary (.csv) created by the researcher based on WordNet ID mapping, containing approximately 10,000 English words, 8,500 Thai words, and 6,000 Japanese words. Regarding cross-linguistic ambiguity (polysemy), the current prototype addresses this by automatically retrieving the primary (first) vocabulary entry linked to the specific WordNet ID in each language. The comprehensive refinement of these multi-meaning entries for optimal practical usage is further discussed in the future development suggestions.
- *Speech Synthesis (TTS):* Google TTS is commanded to sequentially read the results (Thai→ Japanese → English). Users can press and hold to repeat or tap to interrupt the audio and immediately return to camera mode.

C. Phase 3: Technical Evaluation Design

To measure accuracy, 600 images were captured from 10 designated everyday objects, with 10 images per object across specific scenarios. The 10 objects represent common everyday items to reflect real-world usage: 1. Curtain, 2. Cat, 3. Apple, 4. Pineapple, 5. Tree, 6. Bicycle, 7. Electric Fan, 8. Iron, 9. Egg, 10. Broom. These were selected based on OTPF-4 and Haptic Perception theories covering ADLs/IADLs, Tactile Discrimination, and Orientation & Mobility. The evaluation was divided into Simple and Complex Cases:

- *Simple Case (S1):*

  Single object with a plain background (100 images).
- *Complex Cases (C1-C5):*

  5 complex scenarios including Background Clutter (C1), Partial Occlusion (C2), Multiple Objects (C3), Small Object with large background (C4), and Low-light (C5), consisting of 100 images each. Standard metrics used were Top-1 and Top-5 Accuracy (totaling 500 images).

D. Phase 4: Preliminary Usability Study

An evaluation usability study was conducted with a sample group of 5 blindfolded volunteers to simulate visual impairment. According to the Nielsen Norman Group (NN/g) principles [13], a sample size of 5 users is sufficient for identifying qualitative usability issues during this preliminary phase. The assessment utilized the standard 10-item System Usability Scale (SUS) strictly independent from the custom 5-point Likert scale satisfaction questionnaire (ranging from 1 = strongly disagree to 5 = strongly agree) to preserve the integrity of its standardized positive-negative calculation formula and ensure reliable baseline metrics.

E. Phase 5: Data Collection, Analysis, and Conclusion

Quantitative and qualitative data were systematically analyzed to evaluate both the technical performance and the user experience of the prototype. Specifically, the quantitative analysis involved calculating the average percentages for Top-1 and Top-5 accuracy metrics across the 600 tested images to assess the model's reliability in diverse scenarios. Additionally, descriptive statistics, including means and standard deviations, were applied to evaluate the System Usability Scale (SUS) and overall user satisfaction scores. Parallelly, qualitative feedback gathered from the preliminary testers such as observations on TTS audio clarity, dictionary translation contexts, and UX/UI interactions was rigorously synthesized. This holistic analysis enabled the researchers to effectively identify current system limitations, including cross-linguistic polysemy and hardware resource consumption, and consequently formulate grounded recommendations for future clinical validations and on-device optimizations.

IV. RESULTS

Following the development and testing, data was collected and analyzed to evaluate the system's performance. The details are presented as follows:

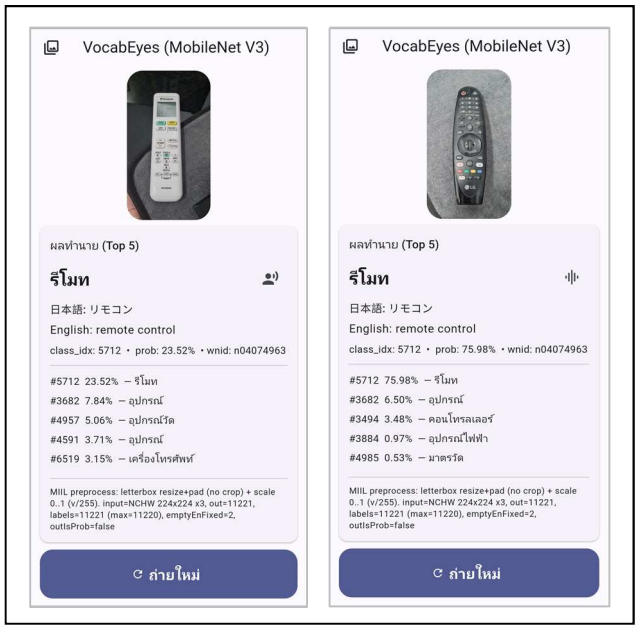

Fig. 4. Sample Pictures of Result Display

A. Accuracy Evaluation

From the classification test of 600 images, the system demonstrated high efficiency, achieving an average Top-1 Accuracy of 93.83% and a Top-5 Accuracy of 98.33%, as presented in Table I.

TABLE I. ACCURACY EVALUATION RESULTS ACROSS SCENARIOS

Scene	Images	Top-1 (%)	Top-5 (%)
S1: Simple Case	100	100	100
C1: Background Clutter	100	97	100
C2: Partial Occlusion	100	90	100
C3: Multiple Objects	100	89	90
C4: Small Object	100	90	100
C5: Low-light	100	97	100
Overall	600	93.83	98.33

Nevertheless, it must be cautiously noted that this high accuracy is specifically bounded within the limited scope of the 10 targeted everyday objects. The performance may vary when generalized to a broader and more diverse range of real-world items. The results indicate that even in environments with multiple interfering objects (C3), the model maintained a Top-1 accuracy of 89% and a Top-5 of 90%. This extremely narrow 1% margin in the C3 scenario explicitly highlights the model's exceptional precision in correctly prioritizing the primary object as its top prediction. Furthermore, considering the overall performance, the narrow 4.5% margin between Top-1 and Top-5 accuracy implies that the system consistently provides the correct answer within its top 5 predictions. However, experiments revealed that accuracy degradation primarily stemmed from model confusion, where the model occasionally prioritized secondary objects that appeared more visually prominent than the primary target. While both metrics are presented, Top-1 Accuracy is fundamentally the most critical indicator in the context of visually impaired users. Providing a single and highly accurate auditory result is essential to prevent user confusion during real-world vocabulary learning.

Fig. 5. Sample Picture of Complex case 3

B. Preliminary Usability Evaluation Result

The average score from the SUS assessment was 89 out of 100 (S.D. = 13.06). The application's usability is rated as "Excellent," suggesting that the interface is simple and enables rapid user learning without expert assistance. However, given the small sample size (n=5) of blindfolded simulated users, these findings primarily indicate promising preliminary potential rather than definitive proof of usability for actual visually impaired individuals. Although the overall SUS score was excellent, the relatively high standard deviation (13.06) resulted from score disparities among test groups. The 3 Japanese testers provided exceptionally high scores (95-100), whereas the 2 Thai testers gave lower scores (72.5-77.5). This variance may be attributed to external factors such as pre-test fatigue, which was observed during preliminary discussions and potentially impacted concentration and the interpretation of complex or negatively phrased SUS questions. Furthermore, differences in cultural and linguistic contexts may have influenced UI interpretation, alongside participants' lack of familiarity with the blindfolding simulation.

Regarding the potential impact of TTS language differences or dictionary translation quality on user satisfaction, direct inquiries with the participants yielded insightful feedback. For audio quality, testers confirmed that the TTS pronunciation was clear and comprehensible. Regarding dictionary translation quality, Thai testers did not raise any concerns. Conversely, Japanese testers observed that while some translated words belonged to the correct semantic categories, they were not the most natural choices for everyday contexts. This limitation occurs because the current system automatically retrieves the primary (first) vocabulary entry linked to a specific WordNet ID, which might be a less common synonym. For instance, the English word "Refrigerator" was mapped to "アイスボックス" (aisu bokkusu, or Icebox in English) via WordNet. While this is structurally and semantically correct within its synonym set: アイスボックス (Icebox) | 冷蔵庫 (Refrigerator), it is unnatural in daily conversation; a more natural choice would be "冷蔵庫" (Reizouko or Refrigerator in English). Similarly, the word "Fan" was mapped to "ブロワ" (Blower) because it appears first in the synonym set: ブロワ (Blower) | ブロワー (Blower) | 扇風機 (Fan) | 電気扇 (Electric Fan), whereas "扇風機" (Senpuuki or Fan in English) is the appropriate everyday term. However, despite noting this linguistic nuance, the Japanese group still awarded exceptionally high SUS scores (95–100), whereas the Thai group, who faced no translation issues, provided lower scores (72.5–77.5). This inverse relationship suggests that, within this small preliminary group (n=5), translation quality might not be the primary factor negatively affect overall satisfaction or cause the score disparity. The only shared concern raised by participants pertained to occasional object misclassifications when encountering complex or out-of-domain items. However, after the researcher explained that the model's classification accuracy is evaluated as a distinct performance metric from the application's ease of use, the testers understood the scope and did not consider these misclassifications a usability barrier. Consequently, it is hypothesized that the disparity in SUS scores may be influenced by the aforementioned external factors rather than the system's linguistic or audio components.

C. Satisfaction Evaluation Result

Overall user satisfaction averaged 4.80 out of 5.00. Users highlighted the full-screen gesture feature, which facilitates convenient picture-taking, and the smooth sequential audio reading as key strengths. Qualitative feedback from the preliminary testers explicitly highlighted that the application's operation method is remarkably simple and intuitive, making it highly suitable for the visually impaired. Furthermore, the preliminary testers suggested that due to its simplified interface and the ability to seamlessly connect vocabulary across three languages (Thai, Japanese, and English), the application possesses potential to be recommended as a play-

based language learning tool for highly curious young children who are still acquiring their mother tongue.

V. CONCLUSION AND SUGGESTIONS

This research achieved preliminary success in developing the "VocabEyes" smartphone prototype application to assist visually impaired individuals in learning trilingual vocabulary (Thai, Japanese, and English). Operating offline with the MobileNetV3-Large model, it delivered high accuracy (Top-1 accuracy 93.83%) and successfully met Universal Design goals with an excellent SUS score of 89 out of 100.

Limitations:

- *Testing and Target Group Constraints:* Accuracy evaluation was limited to 10 specific target objects (600 images) due to time and resource constraints, potentially not reflecting true comprehensive capabilities in diverse everyday scenarios. Since this research represents a prototype phase primarily focused on technical feasibility and baseline UX/UI preparation, the preliminary usability study utilized blindfolded volunteers with normal vision. The researcher profoundly acknowledges that the spatial imagination and interaction behaviors of blindfolded individuals differ completely from actual visually impaired users, meaning these behavioral insights from this preliminary study must be interpreted with caution. Additionally, the limited scale of both object diversity and participant sample size fundamentally constrains the generalizability of the current findings. Therefore, conducting comprehensive clinical validations with genuine visually impaired individuals in the future work is the most crucial next step to verify actual real-world usability.
- *Absence of Confidence Thresholds:* Since the primary objective of VocabEyes emphasizes continuous vocabulary learning rather than strict real-world object identification, the system intentionally omits a minimum confidence score threshold that would suppress results (e.g., stating 'Object not found' or 'Cannot identify' when the confidence is low). Prioritizing the continuous output of vocabulary words maintains the learning flow.
- *Battery Consumption:* Running deep learning models and TTS entirely on-device necessitates continuous CPU/GPU resources, which may lead to heat accumulation and high battery drainage during prolonged use. To quantify this, a preliminary test conducted at normal room temperature without air conditioning in Thailand during April revealed a 7% battery drop (from 86% to 79%) within a 10-minute period of continuous usage—capturing and processing approximately 10 images per minute. Additionally, the smartphone device became noticeably warm to the touch.
- *Object Class Limitations:* The model is constrained to 11,221 classes from the ImageNet-21K-P dataset. Updating the model for specialized or novel objects requires a complete application file update, as cloud-based updates are unsupported.

Suggestions for Future Development:

- *In-depth Testing with Visually Impaired Users:* Conduct clinical trials/real user studies with both totally blind and low-vision individuals to gather accurate data from the genuine target demographic, building upon this successful proof-of-concept.
- *Enhanced Result Communication:* To prevent potential misinformation, future development should implement a feature that audibly reads the AI confidence percentage alongside the result. This approach, which was directly recommended by a preliminary Japanese tester, ensures that users can continuously learn new words while remaining explicitly aware of the system's prediction reliability.
- *Language Database Refinement:* During the preliminary usability study, feedback from Japanese testers indicated that some vocabularies were not terms commonly used in everyday practical contexts. Therefore, relying solely on WordNet mapping may not fully satisfy practical usage requirements across different languages. This limitation arises from polysemy, where a single concept or ID links to multiple vocabulary entries or synonyms. Consequently, it is strictly necessary to collaborate with language experts or native speakers to manually clean and refine the vocabulary database. This process will involve filtering out unnatural synonyms and selecting the most contextually appropriate term for each WordNet ID, ensuring it is contextually accurate for actual real-world usage.
- *Device Optimization:* To directly mitigate the rapid battery drainage and heat accumulation issues identified during continuous operation, the implementation of ‘Model Quantization’ techniques is highly recommended. This approach will effectively compress the model size, optimize computational efficiency, and reduce power consumption without severely compromising accuracy, thereby ensuring more stable and prolonged real-world usability.
- *Customization Features:* Integrate functions such as allowing users to adjust voice speed and toggle specific languages on or off according to individual learning preferences.
- *Extended Language Support:* Scaling the framework to incorporate additional languages (e.g., Chinese) to enhance the breadth of the learning experience.

ACKNOWLEDGMENT

This research is submitted in partial fulfillment of the graduation requirements at the Faculty of Information Technology, Thai-Nichi Institute of Technology. I express my deepest gratitude to Dr. Saprangsit Mruetusatorn for his invaluable guidance and constant support. I also extend my sincere appreciation to the Thesis Examination Committee and all supporters for their insightful comments and advice, which significantly contributed to the completion of this work.

REFERENCES

[1] Department of Empowerment of Persons with Disabilities, "Status of Persons with Disabilities in Thailand," 2025. [Online]. Available: https://app.powerbi.com/view?r=eyJrIjoiNTJhMWE5YjAtZDBjMi00OTY1LTlkOGItN2U1MzllNzg4MDVjIiwidCI6IjE2ZTI4YTZhLWUxYzYtNDNhMi1hNjUwLWI0OTY3OWY2OGEyMCIsImMiOjEwfQ%3D%3D. [Accessed: Sep. 2, 2025]. (in Thai).

[2] World Health Organization (WHO), "Blindness and visual impairment," 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment. [Accessed: Sep. 2, 2025].

[3] A. Quatraro and M. Paiano, "New ways of language learning for blind or visually impaired children and teenagers," in *Proc. 3rd Int. Conf. ICT for Language Learning (ICT4LL2010)*, Italy, Nov. 2010.

[4] Embassy of Japan in Thailand, "Japan supports the project for constructing a media production and educational technology center for the blind in Khonkaen province," Nov. 30, 2020. [Online]. Available: https://www.th.emb-japan.go.jp/itpr_th/2020_27.html. [Accessed: Jan. 5, 2026]. (in Thai).

[5] H. M. Fadhil, A. Muayad, S. Majid, M. M. Jalal, and H. Husain, "Smartphone-based object recognition application for educational engagement," in *Proc. 4th Int. Conf. Distributed Sensing and Intelligent Systems (ICDSIS 2023)*, Dubai, UAE, Dec. 2023.

[6] S. Amartmontree and A. Suwatten, "Smartphone: Assistive technology for learners with visual impairment," *ECT Journal*, vol. 14, no. 17, pp. 9–19, Jul.–Dec. 2019. [Online]. Available: https://so01.tci-thaijo.org/index.php/ectstou/article/view/224648. [Accessed: Sep. 24, 2025]. (in Thai).

[7] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[8] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K pretraining for the masses," *arXiv preprint arXiv:2104.10972*, 2021.

[9] A. Howard *et al.*, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 1314–1324.

[10] T. Leelanoi, "The construction of Thai wordnet of 1st order entity common base concepts using a bi-directional translation method and with dictionaries of different compilational approaches," M.S. thesis, Chulalongkorn Univ., Bangkok, Thailand, 2008. (in Thai).

[11] W. Mamee, "Public toilet design criterior for users with walking disability in conjunction with universal design paradigm," M.Ind.Ed. thesis, Dept. Industrial Design, Faculty of Industrial Education, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, 2011, pp 13-15. (in Thai).

[12] J. Brooke, "SUS: A quick and dirty usability scale," in *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland, Eds. London, U.K.: Taylor & Francis, 1996, pp. 189–194.

[13] R. Budiu, "Why 5 Participants Are Okay in a Qualitative Study, but Not in a Quantitative One," *Nielsen Norman Group*, Jul. 11, 2021. [Online]. Available: https://www.nngroup.com/articles/5-test-users-qual-quant/. [Accessed: Mar. 14, 2026].