Affective and Pleasurable Design, Vol. 123, 2024, 224–231
https://doi.org/10.54941/ahfe1004694

Developing AI Video Analysis Systems to Explore Human Behavior in Infant and Ethnographic Footage

Yuta Ogai[1], Yuto Ono[1], Yasushi Noguchi[2], Sayaka Tohyama[3], Hideaki Kondo[4], and Masayuki Yamada[5]

[1]Tokyo Polytechnic University, 5-45-1 Iiyamaminami, Atsugi, Kanagawa, Japan
[2]Tokyo Polytechnic University, 2-9-5 Honcho, Nakano, Tokyo, Japan
[3]Shizuoka University, 3-5-1 Johoku, Chuo-ku, Hamamatsu, Shizuoka, Japan
[4]Kanda University of International Studies, 1-4-1 Wakaba, Mihama-ku, Chiba, Chiba, Japan
[5]Kyushu Institute of Technology, 680–4 Kawazu, Iizuka-shi, Fukuoka, Japan

ABSTRACT

This research explores the application of AI in extracting meaningful insights from video data, focusing on developmental studies from daily life videos and artistic interpretations from ethnographic footage. Through case studies, including infant behavior analysis and the "Diverse and Universal Camera" project, we demonstrate the capability of AI to manage and interpret large datasets, overcoming challenges like video resolution and data size. Our findings suggest the significant potential of AI in enhancing human behavior research to indicate promising avenues for future exploration of various video analysis applications.

Keywords: AI video analysis, Infant behavior, Ethnographic footage

INTRODUCTION

The advancement of information and communication technology (ICT) has enabled the storage of large volumes of video data. In recent years, research has focused on technologies for extracting video data in formats suitable for specific purposes. For example, it is possible to derive insights regarding developmental processes from daily life video data or extract specific ethnographic footage segments for artistic expression.

Scholars stress the importance of analyzing daily videos of infants to understand their developmental processes (Roy, 2009), but watching all infants' daily videos would require an enormous amount of human effort. Vong et al. proposed a method for analyzing infant language acquisition using AI for video and audio from a camera attached to an infant's head (Vong et al., 2024). However, the recorded data in this study was limited to 61 hours. Therefore, in this paper, we consider how to use AI technology efficiently process long video data. This presentation provides two examples of our AI-based video analysis to demonstrate the possibilities and challenges involved.

INFANT BEHAVIOR ANALYSIS

In our first study, which focused on an infant's movements, we examined the possibility of using object-detection, action-recognition, and caption-generation AI to detect infant movements for developmental research and monitoring (Ogai et al., 2023; Yamada et al., 2023). We used videos of infants rolling over and standing up. The videos used are presented in Table 1. The videos were cut into still images every 0.5 seconds using FFmpeg, and YOLOv8 (YOLOv8, 2024) was applied to each still image. YOLOv8 used the pre-trained model yolov8n.pt. Since yolov8n.pt does not have a label for directly detecting infants, we used the label "Person". YOLOv8 identified areas where "Person" was detected in the video with a confidence of 0.25 or higher and cropped the area, which was enlarged three times in both the vertical and horizontal directions, as an image. The reason for using cropped images is that when the entire image is applied to the caption-generation AI, it often does not yield captions that focus on the infant. The number of still images and the number of images recognized as "Person" are also presented in Table 1.

Table 1. Videos used, number of still images, and number of "Person" images recognized.

Video	Time(sec)	Images	"Person" Images
RollingOver1.mp4	19.52	39	94
RollingOver2.mp4	10.04	20	2
RollingOver3.mp4	29.73	59	0
StandingUp1.mp4	30.49	60	22
StandingUp2.mp4	15.97	32	44
StandingUp3.mp4	18.56	37	116

As examples of the results of cropping images, we present the images cropped around the infant, the adult, and nothing in Figures 1, Figure 2, and Figure 3, respectively.

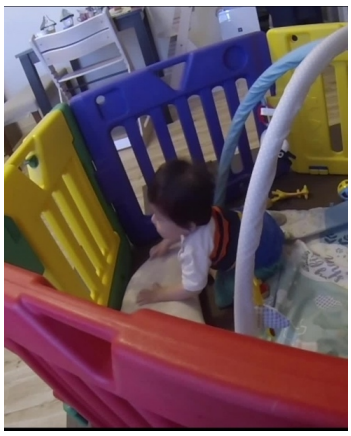

Figure 1: An image cropped around the infant.

Figure 2: An image cropped around the adult.

Figure 3: An image cropped around nothing.

The caption-generation AIs CATR (CATR, 2024) and BLIP (Li et al., 2022) were then used on each image to evaluate whether they could detect infants and provide information regarding their behavior. CATR and BLIP are not the latest caption-generation AIs, but we used them because they are relatively lightweight. Using Google Colab's TPU, CATR generated captions in approximately 10–15 seconds per image and BLIP in approximately 20–25 seconds. CATR used the pre-trained model v3. The following are the output results of CATR for each image.

Figure 1: A toddler sitting in a yellow play chair.

Figure 2: A person standing next to a red plastic chair.

Figure 3: A blurry photo of a blue chair in a room.

The output result of Figure 1 includes the words "toddler" and "sit." The output result of Figure 2 includes the word "person". The output result of Figure 3 does not include any words similar to "person". In this manner, it is possible to detect infants and estimate their behavior using the words included in the output results of CATR. However, in Figure 2, the adult is sitting, but the word "stand" is output. We believe that this is because the camera is set to mainly capture the infant and the adult is only partially visible, thereby resulting in poor recognition results.

The results of Infant Sensitivity, Infant Specificity, Infant+Sit Sensitivity, and Infant+Sit Specificity by CATR for the videos mentioned in Table 1 are presented in Figure 4. Here, captions containing the words "boy", "child", "baby", "toddler", and "girl" are detected as an infant. "Infant+Sit" refers to the case where the word "sit" co-occurs when an infant is detected. As depicted in Figure 4, Infant Sensitivity, Infant Specificity, and Infant+Sit Sensitivity have a high accuracy of over 96%. On the other hand, Infant+Sit Specificity has a low accuracy of 58.91%.

When BLIP was used, Infant Specificity had a low accuracy of 23.86%. In addition, the word "play," which can include various actions, was often used in infant behavior.

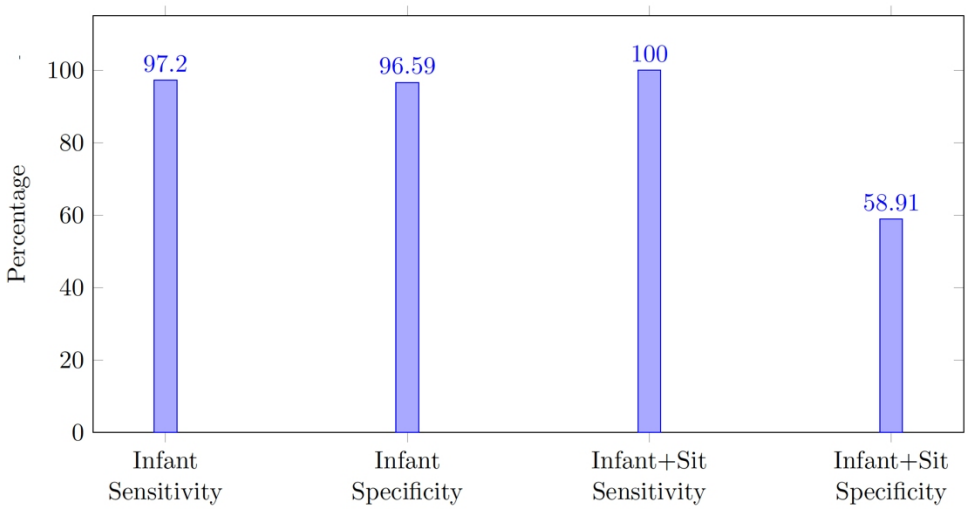

Figure 4: Results of infant sensitivity, infant specificity, infant+sit sensitivity, and infant+sit specificity using CATR.

Figure 5: Infant behavior recognized by SlowFast. "Touch (an object)" with a confidence of 0.96 and "stand" with a confidence of 0.35 are outputs.

In this manner, it was suggested that infants could be detected in videos by using the caption-generation AI CATR for infants that YOLOv8, an object-recognition AI, did not directly learn. CATR and BLIP did not yield high accuracy in recognizing infant behavior. However, since infant behavior often changes slowly, we believe that accuracy can be improved by diffusing it forward and backward as time-series information. Kondo et al. (2024) conducted research on infant behavior detection using OpenPose (Cao et al., 2019), and it is possible to improve accuracy by combining this technology.

We are also considering using BLIP-2's Visual question answering AI, which answers questions regarding images. As depicted in Figure 5, we used SlowFast (Fan et al., 2020), an action-recognition AI, to detect infant behavior in the video.

Based on the results of these studies of individual AIs, we discuss the potential of combining them. Ethnographic video data presents similar challenges.

Ethnographic Footage Analysis

Another example of AI utilization is the experiential video installation entitled "Diverse and Universal Camera" by Noguchi and Ogai (TOP, 2024; Yoshino, 2023) (Figure 6). This project employed SlowFast and YOLOv8 to develop a system that automatically labels the actions of people and objects in videos and enables efficient video retrieval for exhibitions of ethnographic footage archives. A few labels were manually attached by humans, thereby making it a hybrid system. We used the footages of the Encyclopaedia Cinematographica (ECFilm, 2024) as an ethnographic video archive. Viewers can search and view videos by using cards linked to labels as input devices. The videos displayed when searched are only the parts of the video with labels, and short videos are displayed in a loop. The exhibition was held at the Tokyo Photographic Art Museum from November 11, 2023, to December 10, 2023.

The videos are shot utilizing different themes, each emphasizing varying subjects or objects. There are 1207 video files, with a total duration of over 225 hours. The longest video is approximately 1.7 hours, and the shortest is approximately 33 seconds.

Labelling with YOLOv8 was done after converting once to still images. We cut out still images every 0.5 seconds using FFmpeg and obtained approximately 1.62 million still images. YOLOv8 used the pre-trained model yolov8n.pt, and only labels with a confidence of 0.8 or higher were used. From the output labels, only labels that might be related to ethnic culture, such as "chair" and "dog", were retained.

SlowFast utilized the pre-trained model SLOWFAST_32x2_R101_50_50.pkl and detected only labels with a confidence of 0.7 or higher. The SlowFast program was modified to output labels and save them to a text file. Labels are output for each two-to-three-second video segment in SlowFast. From the output labels, only labels that appeared to be related to ethnic culture, such as "dance" and "paint", were retained.

The videos are often old, in black and white, and low-resolution. Therefore, YOLOv8 and SlowFast often missed detection. It is difficult to

appreciate the video when it is displayed in small pieces. Therefore, we processed the video segments and concatenated them when the parts of the video with labels were separated by a short duration. We also deleted video segments with durations that were too short. In this manner, by adding post-processing in the time-series direction to the video AI processing in accordance with the viewing method, we were able to build a system that is easy to appreciate.

Figure 6: Experiential video installation entitled "Diverse and Universal Camera."

DISCUSSION

In these examples, AI tools process amounts of video data that are too large to be managed by humans, extracting parts of a video that merit human attention to provide a better under-standing of human behavior. One challenge is that installing multiple cameras in a household to capture everyday situations often necessitates reducing the video resolution due to storage and network bandwidth constraints. Moreover, because of the need to cover wide areas, an infant frequently appears small in the video and other objects commonly appear in the video, such as family members and the infant's bedding and toys. Ethnographic footage is also difficult to analyse, as it is typically old, in black-and-white, and in low resolution. Furthermore, each segment of footage, shot with different themes, emphasizes varying subjects or objects. To address these problems, we believe that a combination of various AI tools—such as cropping individual image segments, classification of video

by generated captions, diffusion in the time direction, and the accumulation of various technologies—will be necessary in the future. Therefore, there will be a need for reports on system development using AI and opportunities for sharing technology. AI that has undergone end-to-end learning may render these complex systems unnecessary. However, this is possible only after the usefulness of what has been achieved by these complex systems has been confirmed.

CONCLUSION

In this study, we employed AI technologies such as object detection, action recognition, and caption generation to analyse video data for insights into infant development and to artistically interpret ethnographic footage. Our findings demonstrated AI's effectiveness in processing large video datasets, addressing challenges like video resolution and size. We presented two case studies: analysing infant behavior for developmental research and using AI for the "Diverse and Universal Camera" project, showcasing AI's application in art. The results revealed AI's potential to enhance research on human behavior and suggest further exploration in video analysis applications.

ACKNOWLEDGMENT

This research was partially supported by the 2023 research grant from the KDDI Foundation, a public interest incorporated foundation, and the Science and Art Integration Research Grant from Tokyo Polytechnic University for the fiscal year 2023.

REFERENCES

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S. & Sheikh, Y. A. (2019), Openpose: Realtime multi-person 2d pose estimation using part affinity fields, IEEE Transactions on Pattern Analysis and Machine Intelligence.

CATR (2024), CATR: Image Captioning with Transformers, URL: https://github.com/saahiluppal/catr (visited on 02/12/2024).

ECFilm (2024), Encyclopaedia Cinematographica, URL: http://ecfilm.net/, in Japanese (visited on 02/12/2024).

Fan, H., Li, Y., Xiong, B., Lo, W. Y. & Feichtenhofer, C. (2020), 'Pyslowfast'. URL: https://github.com/facebookresearch/slowfast (visited on 02/12/2024).

Kondo, H., Ogai, Y., Yamada, M. & Tohyama, S. (2024), Examination of automated clustering methods for infant posture and movement using monocular camera images, Proceedings of the Japan Society for Educational Technology (JSET) Spring Conference 2024 (44th), in Japanese.

Li, J., Li, D., Xiong, C. & Hoi, S. (2022), Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in 'ICML'.

Ogai, Y., Ono, Y., Tohyama, S., Kondo, H. & Yamada, M. (2023), Investigating the use of image recognition and caption generation ai for infant motion detection, The 41st Workshop of Special Interest Group of Skill Science (SIG-SKL) of the Japanese Society of Artificial Intelligence (JSAI), pp. 5–8, in Japanese.

Roy, D. (2009), New horizons in the study of child language acquisition, Proceedings of Interspeech 2009, pp. 13–20.

TOP (2024), Tokyo Photographic Art Museum, Integrating Technology & Art through Photography : Tokyo Polytechnic University 100th Anniversary Exhibition, URL: https://topmuseum.jp/e/contents/exhibition/index-4590.html (visited on 02/12/2024).

Vong, W. K., Wang, W., Orhan, A. E. & Lake, B. M. (2024), Grounded language acquisition through the eyes and ears of a single child, Science 383(6682), 504–511. URL: https://www.science.org/doi/abs/10.1126/science.adi1374 (visited on 02/12/2024).

Yamada, M., Kondo, H., Ogai, Y. & Tohyama, S. (2023), A case study toward the development of an infant posture and movement discrimination system using automatic caption generation model, Proceedings of the Japan Society for Educational Technology (JSET) Autumn Conference 2023 (43rd), pp. 363–364, in Japanese.

YOLOv8 (2024), Ultralytics YOLOv8, URL: https://github.com/ultralytics/ultralytics (visited on 02/12/2024).

Yoshino, H. (2023), Integrating Technology & Art through Photography: Tokyo Polytechnic University 100th Anniversary, Crevis, in Japanese.