Self-Attention Enhanced Spatio-Temporal Convolution for ADL Recognition In Elderly Care

Ghazala Rafiq
Center of Self-Organizing Software
Kyungpook National University
Daegu, 41566, South Korea
ghazala@knu.ac.kr
0000-0002-1045-8715

Rafiq Muhammad
Department of Game Software
Keimyung University
Daegu, 42601, South Korea
rafiq@kmu.ac.kr
0000-0001-6713-8766

Soon Ju Kang*
Center of Self-Organizing Software
Kyungpook National University
Daegu, 41566, South Korea
sjkang@ee.knu.ac.kr
0000-0002-8171-195X

***Abstract*—The increasing aging population necessitates intelligent monitoring systems that can accurately detect and classify daily activities, enabling proactive healthcare interventions and promoting independent living for the elderly. To this end, this study presents ADL recognition system for elderly care, leveraging advanced 3D CNNs with skip connections and self-attention mechanism. Our proposed model processes RGB video data from the ETRI-Activity3D dataset. The architecture incorporates skip connections to mitigate the vanishing gradient problem and a self-attention mechanism to capture complex spatio-temporal features. Experimental results demonstrate the superiority of our approach, achieving 99.0% accuracy on the training set and 96.1% on the validation set. Notably, our model outperforms the baseline ETRI(FSA-CNN) implementation by 6%. The study provides both quantitative and qualitative evaluations, including accuracy metrics, loss curves, and visual representations of classification outcomes. These thorough results highlight the robustness and effectiveness of our proposed system in real-world scenarios. This research contributes to the evolving landscape of technology-assisted elderly care, offering a foundation for future developments in anomaly detection, risk management, and personalized alert systems. The proposed ADL recognition system has the potential to greatly improve the quality of life for elderly individuals and streamline healthcare resource allocation.**

***Index Terms*—Activity of Daily Living (ADL), action recognition, Convolutional neural network, healthcare, self-attention, spatio-temporal features.**

I. Introduction

The older adult population is continuously increasing around the world [1]. This phenomenon of global aging poses certain challenges including physical health deterioration, cognition decline, mental health issues and preference for aging in place with freedom, safety and independence [3], [4]. The activities of daily living (ADLs) [2] , is a term used in healthcare industry to refer to essential and routine activities that most people should be able to perform without any assistance [5] indicating their well-being [6]. ADL's monitoring allows caregivers for timely assistance provision at the time of need [7].

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2018R1A6A1A03025109).

* Corresponding Author: Soon Ju Kang (email: sjkang@ee.knu.ac.kr)

The development of smart homes [8], [9], [10] and smart cities offers a promising framework for implementing ADL recognition technologies. Smart homes, equipped with a range of sensors, IoT devices, and intelligent algorithms, can provide continuous, real-time monitoring of an individual's activities. In the context of healthcare, smart homes have the potential to transform how care is provided to the elderly by facilitating proactive and preventive care models. Through continuous monitoring, smart homes can detect deviations from normal activity patterns that may indicate the onset of health issues or the occurrence of accidents like falls. These early warnings can prompt timely interventions, potentially preventing hospitalizations and reducing the overall healthcare burden.

RGB video-based ADL recognition offers several advantages, including the ability to monitor activities [10] without requiring the individual to wear any devices or significantly alter their living environment. These systems utilize video cameras strategically placed within the home to capture daily activities [11], which are then analyzed using advanced computer vision techniques to recognize specific ADLs. The non-intrusive nature of RGB video monitoring [12] is particularly appealing in the context of elderly care, as it minimizes disruptions to the individual's routine and preserves their sense of privacy and dignity.

Fig. 1. Sample frames of various ADL classes from the ETRI-Activity3D dataset, showcasing a diverse range of activities captured under varying environmental conditions.

3D CNNs [13] play key role in capturing spatio-temporal features from RGB videos for improved ADL recognition accuracy. Unlike traditional 2D CNNs [14] that focus solely on

spatial features within individual frames, 3D CNNs extend this capability by processing sequences of frames simultaneously, capturing both spatial and temporal information. This ability to learn spatio-temporal features [15] is particularly advantageous in ADL recognition, where understanding the sequence and timing of actions is crucial for accurately identifying specific activities. For example, distinguishing between sitting down and standing up requires not only spatial recognition but also an understanding of the movement over time, which 3D CNNs are well-equipped to handle. Further improvements in ADL recognition accuracy can be achieved through the integration of self-attention mechanisms into the 3D CNN framework. Self-attention modules improve the network's ability to concentrate on the most relevant aspects of the input data, efficiently filtering out noise and highlighting the key features most indicative of specific ADLs. By allowing the model to selectively attend to important temporal and spatial aspects of the video data, self-attention mechanisms can significantly boost the precision of ADL recognition, leading to more reliable and robust outcomes.

TABLE I
STATISTICS OF THE ETRI-ACTIVITY3D DATASET UTILIZED FOR THE EVALUATION OF PROPOSED ADL RECOGNITION SYSTEM

Dataset	ETRI-Activity3D
Videos	112, 620
Action Classes	55
Subjects	100
Mode	RGB Videos
Training Set	90,064 videos
Validation Set	11,280 videos
Test Set	11,276 videos

This research, situated at the intersection of computer vision, gerontechnology (elderly care), and healthcare informatics, aims to advance the state-of-the-art in ADL recognition technologies. It seeks to contribute to the broader fields of smart homes and smart cities, with a particular focus on non-intrusive RGB video-based methods and the application of advanced deep learning techniques, including skip connection-enhanced 3D CNNs [16] and self-attention modules.

Problem Statement: The objective of this research is to develop an advanced ADL recognition system utilizing function f such that

$f : V \rightarrow A$ that maps a video sequence V to an ADL class A, maximizing the classification accuracy:

$argmax(f) \sum(V, y) \in D|(f(V) = y)/|D|$

where I is the indicator function, i.e.,

$f(V) = y \rightarrow 1$ (predicted class matches true class)

$f(V) = y \rightarrow 0$ (predicted class does not match true class)

The key contributions of this research work to the field of ADL recognition are:

- We propose an architecture comprising a 3D CNN augmented with skip connections. This design facilitates efficient extraction of spatio-temporal features from RGB video inputs, subsequent to the application of diverse preprocessing methodologies.
- To enhance the model's capacity for comprehending complex scenarios, we incorporate a self-attention mechanism into the 3D CNN framework. This integration allows the network to selectively emphasize the most salient aspects of the input data, thereby improving its overall performance in ADL recognition task.
- We conduct a comprehensive evaluation of the proposed ADL recognition system utilizing the ETRI-Activity3D dataset. Our analysis encompasses both qualitative and quantitative assessments, providing a thorough examination of the model's efficacy and performance characteristics.

While our current research focuses on accurate ADL recognition, we envision extending this work to enhance its impact on elderly care. Future directions include developing anomaly detection algorithms to identify deviations in behavior, creating a risk management framework to predict and mitigate health and safety risks, and designing a personalized alert system to notify caregivers based on individual needs. Additionally, we plan to implement detailed ADL reporting mechanisms to provide comprehensive insights into daily activities, aiding healthcare providers in assessing health and adjusting care plans. These extensions aim to develop a more comprehensive and proactive elderly care solution, enhancing the quality of life for elderly individuals and alleviating the strain on traditional healthcare models.

In the following sections, we will detail our methodology, including the overall model architecture, the dataset, implementation detail used for training and evaluation, and quantitative and qualitative experimental results. We will also explore potential avenues for future inquiry within this research paradigm, with a particular emphasis on enhancing elderly care practices and improving health-related outcomes for the aging population.

II. RELATED WORK

Motivated by the huge success of deep learning algorithms in different computer vision fields, i.e., scene classification

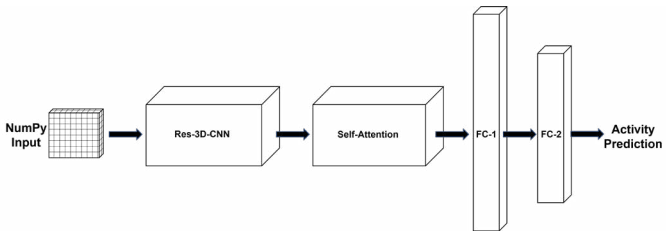

Fig. 2. Model overview for ADL recognition. The 3D NumPy array is input to a 3D-CNN with residual connections, which captures spatiotemporal features from the video. The extracted features are then refined using a self-attention block that enhances the model's focus on relevant aspects of the activity. Finally, the processed features are passed through fully connected layers, with the output probabilities generated by a softmax layer, providing the final predictions for different ADL classes.

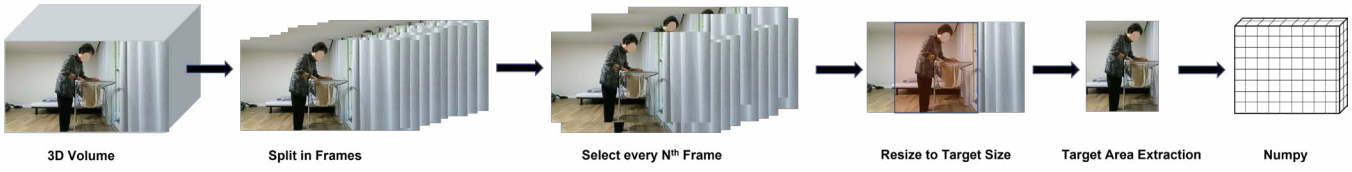

Fig. 3. Data preprocessing pipeline. The 3D video volume is first split into individual frames. From these frames, every nth frame is selected to reduce redundancy and focus on key moments of the activity. The selected frames are then resized to the target dimensions. After resizing, the relevant portion of each frame is extracted. Finally, the processed frames are saved as NumPy arrays, preparing them for input into the activity recognition model.

[17], action recognition , language and vision related tasks [18], [19], researchers adopted these mechanisms for ADL recognition as well. particularly, CNN based architectures demonstrated great results. Where 2D CNNs outperformed other algorithms for images, 3D-CNN are considered best for video handling because of their ability to learn spatio-temporal representations.

Numerous studies have creatively applied residual connections within CNNs in diverse ways, improving information flow, processing, and optimization in both images [20], [21], [22] and videos [23], [24], [25]. RGB cameras, when integrated with advanced and sophisticated computer vision and deep learning techniques demonstrate superior capabilities in recognizing a wide spectrum of complex ADLs. This comprehensive recognition is crucial for providing a holistic understanding of an individual's daily routines and potential health indicators [26].

TABLE II
QUANTITATIVE RESULTS

	Accuracy	**Loss**
Training Set	99.0	0.028
Validation Set	96.1	0.110
ETRI(FSA-CNN) [31]	90.1	-

III. METHODOLOGY

This section presents the proposed architecture for ADL recognition, which leverages 3D Convolutional Neural Networks augmented with skip connections (Res-3D-CNN) [27] as the primary feature extraction mechanism. The utilization of 3D CNNs with skip connections is particularly well-suited for this task, as it facilitates the model capacity to effectively learn both spatial and temporal information [28] from video data which is a critical requirement for accurate ADL recognition. Subsequent to feature extraction, a self-attention mechanism [29] is employed to further refine the extracted features. This integration of Res-3D-CNNs and an attention network further enhances the model's ability to capture underlying patterns in the videos, thereby improving the action classification accuracy. Figure-2 provides a schematic representation of the overall architecture of the proposed system.

This research aims to develop a robust system for recognizing and monitoring ADLs of elderly individuals using only camera-captured video data. The primary objective is centered on healthcare-related activities, with a secondary focus on the surveillance of general daily routines.

Let $V = \{v_1, v_2, ..., v_T\}$ represents a sequence of video frames over time T, where v_t represents the t^{th} frame. After pre-processing, these frames are processed to extract spatial-temporal features using 3D convolutional neural networks. The extracted features $fi, j, k(V)$, representing the activation of the $k-th$ filter in the $j-th$ layer for the $i-th$ input feature map , form the basis for recognizing ADLs.

The ADL recognition model F can be described by the function:

$\hat{y} = F(V; \theta)$

where $\hat{y}$ is the predicted activity label, and θ denotes the model parameters. The learning objective is to minimize the cross-entropy loss function:

$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \iota(y_i, F(V_i; \theta))$

where N is the number of training samples, y_i is the true label for the $i-th$ sample, and ι represents the cross-entropy loss. 3D-CNN represents an advanced deep learning architecture specifically designed to process the spatio-temporal data inherent in video sequences. Its application in the recognition of Activities of Daily Living (ADL) for elderly individuals within smart home environment is of particular significance, as it facilitates the simultaneous extraction of both spatial and temporal features from video inputs. The architecture of 3D CNNs mirrors that of traditional 2D CNNs, with the key distinction being the incorporation of an additional temporal dimension. This allows 3D CNNs to capture the progression of visual information over time by performing convolution operations across the three dimensions of height, width, and time. As a result, the model effectively captures the spatio-temporal information embedded within video sequences, providing a robust feature set for accurately recognizing ADLs.

The integration of skip or residual connections into 3D CNNs confers several distinct advantages. Residual connections, originally popularized in ResNet [16] architectures, enable the network to bypass one or more layers, allowing it to learn identity mappings more effectively. This design mitigates the problem of vanishing gradients [30], a phenomenon of particular concern in deep networks, thus enabling the training of more profound model architectures. In the context of ADL recognition, where accurately identifying subtle and varied actions over time is crucial, residual connections enhance the model's ability to capture complex spatio-temporal features, improving both convergence speed and overall performance.

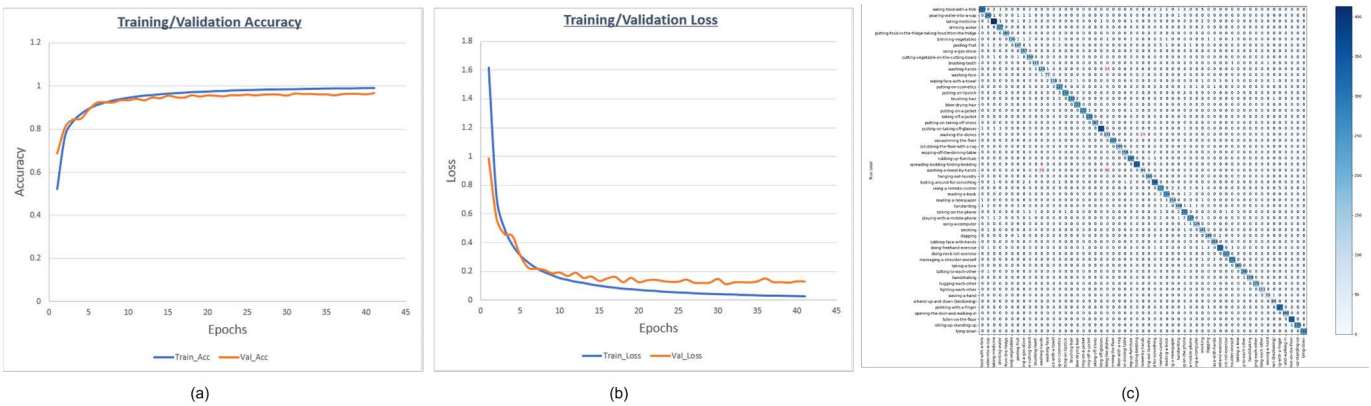

Fig. 4. (a) Training and validation accuracy, and (b) training and validation loss graphs. The graphs depict the model's performance over the course of training and validation, with accuracy increasing and loss decreasing as the model learns. (c) Confusion matrix for the test set, highlighting the classification performance across various activities. Activities such as 'washing hands,' 'washing towel by hands,' and 'washing the dishes' are among the most misclassified (presented in red), indicating challenges in distinguishing between these similar actions. The matrix provides insights into the model's strengths and weaknesses in recognizing different Activities of Daily Living.

Additionally, these connections contribute to better generalization, reducing the risk of overfitting by encouraging the reuse of features learned in earlier layers. This makes Res-3D-CNN a powerful and efficient approach for ADL recognition, providing a reliable foundation for monitoring and supporting the well-being of elderly individuals within smart home environments.

In the proposed methodology for recognizing Activities of Daily Living (ADL) in elderly individuals within smart home environments, 3D Convolutional Neural Networks with skip connections serve as the cornerstone for feature extraction, generating robust spatio-temporal features from the input video data. However, the gradual expansion of receptive fields in deeper layers, combined with their fixed nature, often results in inefficiencies in capturing global context, which can lead to inaccurate classification in CNNs. The integration of self-attention mechanisms with skip connections allows the model to capture long-range dependencies in a single operation, thereby improving both the accuracy and computational efficiency of the model. Moreover, skip connections facilitate effective training by mitigating the vanishing gradient problem, stabilizing the learning process by preserving the original input features, and maintaining a balance between learning new features and retaining essential original information.

A. Dataset

The ETRI-Activity3D dataset [31], collected through Kinect v2 sensor, was used to evaluate the proposed method. The dataset comprised of a total of 112,620 videos obtained from 100 individuals consist of 55 action classes performed in the living room, kitchen, and bedroom in a residential apartment environment. Sample frames from example videos of ETRI-Activity 3D dataset are shown in Figure-1. The dataset consists of three modalities of RGB videos, depth maps, and skeleton sequences. We selected the RGB video mode as input to our proposed model. In our experimental design, we implemented

TABLE III
IMPLEMENTATION DETAIL AND TRAINING PARAMETERS

Spatial dimension	224 × 224
Epochs	100
Batch size	10
Learning Rate	0.0001
Weight decay	1 x 10^{-4}
Warm up	05 epochs
Early stopping	10 epochs with Val Loss
Optimizer	Adam
Loss Function	Cross Entropy Loss
Evaluation Metric	Accuracy
Deep Learning Framework	Pytorch
Data preprocessing libraries	NumPy, OpenCV

a stratified data partitioning strategy for the dataset. Specifically, we allocated 80% of the data for model training, while reserving 10% for validation purposes and the remaining 10% for final testing. Dataset statistics along with train-val-test split information is desplayed in Table-I

B. Data Pre-processing

The 3D video volume undergoes an initial segmentation process, resulting in the extraction of individual frames. Subsequently, a systematic sampling procedure is applied, wherein every 5^{th} frame is selected. This strategic subsampling serves dual purposes: it mitigates data redundancy and focuses the model's attention on salient temporal moments of the activity under observation. Further, a spatial transformation is applied to each selected frame, re-scaling it to the target dimensions of 224×224, ensuring uniformity in input size. In the final preparatory stage, these processed frames are converted and stored as NumPy arrays facilitating efficient computational

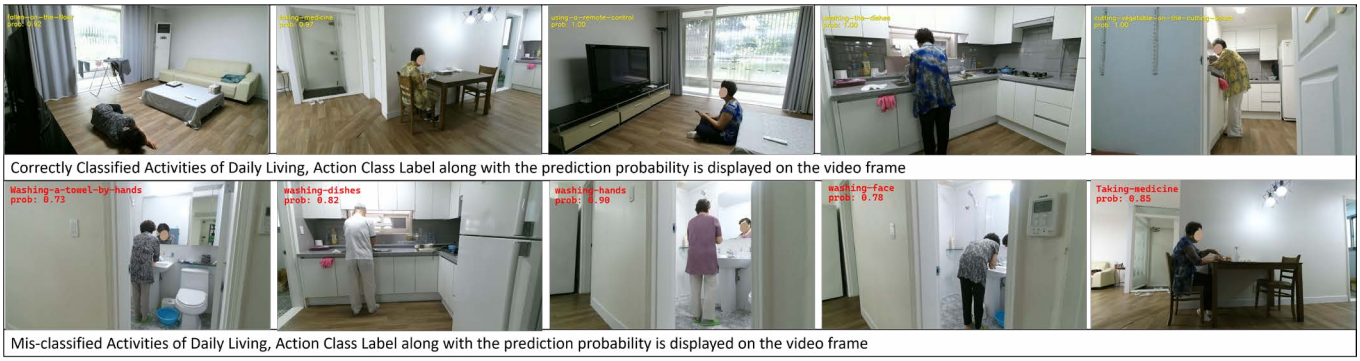

Fig. 5. Qualitative evaluation - examples of activity classification outcomes. The first row showcases successful classifications, whereas the second row illustrates instances of misclassified activities, highlighting the discrepancies between the actual and predicted action labels. In both rows, the action class label and predicted probability are overlaid on each frame, providing a visual representation of the model's performance in real-time activity recognition..

handling. This pre-processing workflow is shown in Figure-3

C. Implementation Detail

The implementation of the proposed ADL recognition system leveraged PyTorch as the primary deep learning framework. Table-III showcases the key training parameters and implementation details of our system. As detailed in this table, the model processes input data with a spatial dimension of 224 x 224 pixels over 100 epochs, using a batch size of 10. An early stopping mechanism that halts training after 10 epochs without improvement in validation loss. While we initially set up the training for 100 epochs, due to the early stopping mechanism, the actual training concluded after 41 epochs, indicating efficient convergence of the model. For data preprocessing, a combination of NumPy and OpenCV was employed. The model training was conducted on a high-performance NVIDIA GeForce RTX 4090 GPU, ensuring efficient computation for the complex 3D CNN architecture. Hyperparameter management and experiment tracking were facilitated through Weights & Biases, enabling systematic optimization of the model.

D. Evaluation Metrics

For the quantitative assessment of our model's performance, we primarily employed accuracy as the evaluation metric. Accuracy, defined as the proportion of correct predictions out of the total number of predictions made by the model.

$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100$

where TP are True Positives, TN are True Negatives, FP are False Positives, and FN are False Negatives.

IV. Experimental Results

Our proposed model demonstrates exceptional performance, as evidenced by both quantitative metrics and visual representations of the training process. Figure-4 (a) & (b) illustrates the training and validation accuracy curves, along with their corresponding loss graphs, depicting the model's performance over the course of training. These graphs show a consistent increase in accuracy and a concurrent decrease in loss as the model learns, indicating effective training dynamics. The training curves demonstrate the model's ability to fit the training data well, while the validation curves provide insight into its generalization capabilities on unseen data.

Table-II presents the quantitative results of our model's performance. Our approach achieved a remarkable 99.0% accuracy on the training set and 96.1% on the validation set, with corresponding low loss values of 0.028 and 0.110, respectively. These results highlight the model's strong predictive capabilities and its ability to generalize well to new data.

Notably, our model outperforms the existing ETRI(FSA-CNN) [31] implementation, evaluated on the same ETRI-Activity3D dataset. This significant improvement of 6% enhancement underscores the effectiveness of our proposed approach in recognizing ADLs from the ETRI-Activity3D dataset, marking a substantial advancement in the field of activity recognition for elderly care applications.

For more in-depth understanding of our model's performance across different activities, Figure-4 (c) presents the confusion matrix for the test set as true labels on the y-axis and predicted labels along the x-axis. This matrix offers valuable insights into the classification performance for various ADLs. While the model demonstrates high accuracy overall, it reveals specific challenges in distinguishing between certain similar actions. Activities such as 'washing hands,' 'washing towel by hands,' and 'washing the dishes' are among the most frequently misclassified, as indicated by the red cells in the matrix. These misclassifications highlight the inherent difficulty in differentiating between activities that share similar motion patterns and environmental contexts. The confusion matrix thus provides a detailed view of the model's strengths and areas for potential improvement in recognizing different ADLs.

Qualitative results in Figure-5 showcases examples of activity classification outcomes, offering a visual representation of the model's real-time activity recognition capabilities. The first row displays instances of successful classifications, demonstrating the model's accuracy in correctly identifying various ADLs. In contrast, the second row illustrates cases where the

model misclassified activities, highlighting the discrepancies between the actual and predicted action labels. For each frame, we overlay the action class label and the predicted probability, allowing for a detailed examination of the model's decision-making process and confidence levels across different scenarios. This qualitative evaluation complements our quantitative results, providing insights into both the strengths and limitations of our ADL recognition system in practical applications.

V. Conclusions & Future Works

In this research work, we proposed ADL recognition model for elderly care applications, leveraging advanced 3D CNN with skip connections and self-attention mechanisms. Our model demonstrated superior performance, achieving higher accuracy on the validation set and outperforming the baseline model. The comprehensive evaluation, including quantitative metrics, confusion matrices, and qualitative examples, underscores the robustness and efficacy of our proposed system in accurately recognizing a wide range of ADLs in real-world scenarios.

Future directions for this research are promising and multifaceted. We aim to develop anomaly detection algorithms to identify behavioral deviations, potentially indicating health issues or cognitive decline. Creating a risk management framework and designing a personalized alert system will enhance preventive capabilities and ensure timely interventions. Implementing detailed ADL reporting mechanisms will provide comprehensive insights for healthcare providers. These future endeavors aspire to transform our ADL recognition system into a more comprehensive and proactive elderly care solution, potentially improving the quality of life for elderly individuals and reducing the burden on traditional healthcare models.

References

[1] "Ageing and health," World Health Organization (WHO) . Accessed: Aug. 10, 2024. Online. Available: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

[2] Michelle E. Mlinac and Michelle C. Feng. 2016. Assessment of activities of daily living, self-care, and independence. Archives of Clinical Neuropsychology 31, 6 (08 2016), 506–516. https://doi.org/10.1093/arclin/acw049

[3] J. L. Wiles, A. Leibing, N. Guberman, J. Reeve and R. E. S. Allen, "The meaning of 'aging in place' to older people", Gerontologist, vol. 52, no. 3, pp. 357-366, Jun. 2012.

[4] M. J. Bárrios, R. Marques and A. A. Fernandes, "Aging with health: Aging in place strategies of a Portuguese population aged 65 years or older", Revista de Saúde Pública, vol. 54, pp. 1-11, Dec. 2020.

[5] Y. Cardinale, J. Manuel Negrete Ramírez, I. Garamendi Bragado and I. de Fez, "Streaming Processing for ADL Monitoring in Smart Home Environments," in IEEE Access, vol. 12, pp. 100700-100724, 2024, doi: 10.1109/ACCESS.2024.3430395.

[6] M. Hopman-Rock, H. van Hirtum, P. de Vreede and E. Freiberger, "Activities of daily living in older community-dwelling persons: A systematic review of psychometric properties of instruments", Aging Clin. Experim. Res., vol. 31, no. 7, pp. 917-925, Jul. 2019.

[7] N. Camp et al., "Technology used to recognize activities of daily living in community-dwelling older adults", Int. J. Environ. Res. Public Health, vol. 18, no. 1, pp. 163, 2021.

[8] A. Zielonka, M. Wozniak, S. Garg, G. Kaddoum, Md. J. Piran and G. Muhammad, "Smart homes: How much will they support us? A research on recent trends and advances", IEEE Access, vol. 9, pp. 26388-26419, 2021

[9] O. Djumanazarov, A. Väänänen, K. Haataja and P. Toivanen, "An overview of iot-based architecture model for smart home systems", Proc. Int. Conf. Intell. Syst. Design Appl., pp. 696-706, 2022

[10] G. Cicirelli, R. Marani, A. Petitti, A. Milella and T. D'Orazio, "Ambient assisted living: A review of technologies methodologies and future perspectives for healthy aging of population", Sensors, vol. 21, no. 10, pp. 3549, May 2021.

[11] V. Vijayan, J. P. Connolly, J. Condell, N. McKelvey and P. Gardiner, "Review of wearable devices and data collection considerations for connected health", Sensors, vol. 21, no. 16, pp. 5589, Aug. 2021.

[12] Vrskova R, Hudec R, Kamencay P, Sykora P. Human activity classification using the 3DCNN architecture. Appl Sci. 2022;12(2):931.

[13] Ji, Shuiwang, et al. "3D convolutional neural networks for human action recognition." IEEE transactions on pattern analysis and machine intelligence 35.1 (2012): 221-231.

[14] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.

[15] Qiu, Zhaofan, Ting Yao, and Tao Mei. "Learning spatio-temporal representation with pseudo-3d residual networks." proceedings of the IEEE International Conference on Computer Vision. 2017.

[16] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[17] Rafiq, Muhammad, et al. "Scene classification for sports video summarization using transfer learning." Sensors 20.6 (2020): 1702.

[18] Rafiq, Ghazala, et al. "DeepRide: Dashcam video description dataset for autonomous vehicle location-aware trip description." IEEE Access 10 (2022): 107361-107375.

[19] Rafiq, Ghazala, Muhammad Rafiq, and Gyu Sang Choi. "Spectral representation learning and fusion for autonomous vehicles trip description exploiting recurrent transformer." IEEE Access 11 (2023): 61437-61452.

[20] G. Wang, Y. Hu, X. Wu, and H. Wang, "Residual 3-D Scene Flow Learning With Context-Aware Feature Extraction," IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1–9, 2022.

[21] K. M. Hosny and M. A. Kassem, "Refined residual deep convolutional network for skin lesion classification," Journal of Digital Imaging, vol. 35, no. 2, pp. 258–280, 2022.

[22] Z. Chang, X. Zhang, S. Wang, S. Ma, and W. Gao, "Strpm: A spatiotemporal residual predictive model for high-resolution video prediction," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13946–13955, 2022

[23] Uchiyama, Tomoki, et al. "Visually explaining 3D-CNN predictions for video classification with an adaptive occlusion sensitivity analysis." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

[24] Tredowicz, Magdalena, et al. "PrAViC: Probabilistic Adaptation Framework for Real-Time Video Classification." arXiv preprint arXiv:2406.11443 (2024).

[25] Yosry, Shaimaa, et al. "Various frameworks for integrating image and video streams for spatiotemporal information learning employing 2D–3D residual networks for human action recognition." Discover Applied Sciences 6.4 (2024): 141.

[26] G. Cicirelli, R. Marani, A. Petitti, A. Milella and T. D'Orazio, "Ambient assisted living: A review of technologies methodologies and future perspectives for healthy aging of population", Sensors, vol. 21, no. 10, pp. 3549, May 2021.

[27] K. Hara, H. Kataoka and Y. Satoh, "Learning spatio-temporal features with 3D residual networks for action recognition", Proc. IEEE Int. Conf. Comput. Vis. Workshops, pp. 3154-3160, 2017.

[28] D. Tran, J. Ray, Z. Shou, S. F. Chang and M. Paluri, "ConvNet architecture search for spatio-temporal feature learning", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1-6, 2017.

[29] Vaswani, A. "Attention is all you need." Advances in Neural Information Processing Systems (2017).

[30] D. Balduzzi, M. Frean, L. Leary, J. P. Lewis, K. W. Ma, and B. McWilliams, "The shattered gradients problem: If resnets are the answer, then what is the question?," in International Conference on Machine Learning, pp. 342–350, 2017.

[31] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, and J. Kim, "ETRI-Activity3D: A large-scale RGB-D dataset for robots to recognize daily activities of the elderly," in IEEE International Conference on Intelligent Robots and Systems, Institute of Electrical and Electronics Engineers Inc., Oct. 2020, pp. 10990–10997. doi: 10.1109/IROS45743.2020.9341160.