CNN-LSTM WITH TEMPORAL ATTENTION AND DCT-BASED FREQUENCY ENHANCEMENT FOR ADL RECOGNITION IN TRAILER LIVING LAB

HAN SHI[1], YUNING SHI[1], QINGWEI SONG[1], NAOYUKI KUBOTA[1]

[1]Department of Mechanical Systems Engineering, Graduate School of Systems Design, Tokyo Metropolitan University,
6-6 Asahigaoka, Hino-shi, Tokyo 191-0065, Japan
E-MAIL: {shi-han, shi-yuning, qingwei-song}@ed.tmu.ac.jp, kubota@tmu.ac.jp

Abstract:

Human action recognition is important for health monitoring, home-based rehabilitation, and daily behavior analysis. In particular, robust recognition of activities of daily living (ADL) in realistic home-like environments is a fundamental step toward practical support for elderly care, daily life monitoring, anomaly detection, and lifelog analysis. However, recognizing ADL in such environments remains challenging because of constrained space, cluttered backgrounds, furniture occlusion, limited viewpoints, and confusion among visually similar actions. To address this issue, this study constructs a self-collected Trailer Living Lab (TLL) ADL Dataset in a compact home-like environment. Based on a CNN-LSTM baseline with temporal attention, a discrete cosine transform (DCT) branch is introduced to enhance temporal representation from a frequency-domain perspective. Experiments under a cross-person setting show that the proposed method improves ADL recognition performance on the TLL ADL Dataset, providing a useful recognition component for future elderly behavior monitoring and long-term lifelog-based analysis.

Keywords:

Activities of daily living; Elderly care; Daily life monitoring; Home-based rehabilitation; Lifelog analysis

1. Introduction

Human action recognition is important for health monitoring, home-based rehabilitation, and daily behavior analysis [1]. In particular, recognizing activities of daily living (ADL) in home environments is essential for evaluating functional status and supporting independent living and rehabilitation [2, 3]. ADL understanding is also crucial for elderly care and long-term monitoring, as continuous observation of daily activities can support health management and daily rhythm regulation [4].

Compared with laboratory settings or public benchmark datasets, ADL recognition in realistic home environments is more challenging due to constrained space, cluttered backgrounds, furniture occlusion, limited viewpoints, and confusion among visually similar actions [3, 5]. Many public datasets focus on sports or generic behaviors, such as UCF101, HMDB51, and Kinetics [6, 7, 8], and cannot fully represent real household scenarios. From the viewpoint of elderly care, robust ADL recognition is not only a pattern recognition problem but also a key component for life log analysis and anomaly detection [9].

To address these issues, this study constructs the Trailer Living Lab (TLL) ADL Dataset in a compact home-like environment. Based on a CNN-LSTM baseline with temporal attention [2, 10], a DCT branch is introduced to complement time-domain modeling with frequency-domain motion information [11, 12]. Experiments under a cross-person setting demonstrate the effectiveness of the proposed method on the TLL ADL Dataset.

2. Related Work

Video-based human action recognition has been widely studied for surveillance, intelligent interaction, rehabilitation, and behavior analysis [1]. Existing approaches include CNNs for spatial feature extraction, LSTMs for temporal modeling, and 3D CNN or attention-based methods for spatiotemporal learning [10, 14, 15]. Although these methods achieve strong performance on benchmark datasets such as UCF101, HMDB51, and Kinetics [6, 7, 8], ADL recognition in realistic home environments remains challenging.

Datasets such as Charades and Toyota Smarthome emphasize daily action recognition in household scenes [5, 3]. Previous studies in our laboratory also explored daily life monitor-

ing and life log analysis for elderly care based on informationally structured space (ISS) [4, 9], motivating this work from the perspective of future support systems. In addition, frequency-domain features can complement temporal modeling by capturing global motion trends and reducing sensitivity to local disturbance [11, 12].

Motivated by these studies, this work introduces a DCT-based frequency branch into a CNN-LSTM framework for ADL recognition in a compact home-like environment, while comparisons with more advanced models such as 3D CNNs and Transformers are left for future work.

3. Proposed Method

3.1. Method Overview

Fig. 1 shows the overall architecture of the proposed framework. Given an input video clip, a fixed number of frames are uniformly sampled and sequentially fed into a CNN-LSTM backbone. The CNN extracts frame-wise visual features, and the LSTM models temporal dependencies to produce shared sequence features.

Based on this shared representation, two branches are constructed. A temporal attention branch emphasizes informative time steps and produces a temporal feature, while a DCT branch transforms the sequence into the frequency domain and extracts low-frequency motion information. The two features are fused and fed into a fully connected layer for final action classification.

3.2. CNN-LSTM Baseline with Temporal Attention

As the backbone architecture, this study adopts a CNN-LSTM framework for joint spatial-temporal modeling. A ResNet18 backbone is used to extract high-level spatial features from individual frames, while the LSTM captures temporal dependencies across the frame sequence. As shown in Fig. 1, the sequence feature produced by the CNN-LSTM backbone is further processed by a temporal attention branch, which adaptively assigns different weights to different time steps and allows the model to focus on more informative parts of the action sequence.

Let the input clip be represented as

$$X = \{x_1, x_2, \ldots, x_T\} \tag{1}$$

where T denotes the number of sampled frames and x_t is the t-th frame.

Each frame is first mapped into a spatial feature vector by the CNN:

$$v_t = f_{\mathrm{CNN}}(x_t), \quad t = 1, 2, \ldots, T \tag{2}$$

where $v_t \in R^D$ denotes the feature representation of frame x_t, and D is the feature dimension.

These frame-wise features are then arranged in temporal order and fed into the LSTM:

$$h_t = f_{\mathrm{LSTM}}(v_t, h_{t-1}), \quad t = 1, 2, \ldots, T \tag{3}$$

where h_t is the hidden state at time step t, summarizing the temporal information up to the current frame.

To aggregate temporal information more effectively, an attention mechanism is applied over all hidden states. First, an attention score is computed for each time step:

$$e_t = f_{\mathrm{att}}(h_t) \tag{4}$$

The scores are normalized by the softmax function to obtain temporal attention weights:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{i=1}^{T} \exp(e_i)} \tag{5}$$

Then, the clip-level temporal representation is obtained as a weighted sum of all hidden states:

$$f_{\mathrm{temp}} = \sum_{t=1}^{T} \alpha_t h_t \tag{6}$$

For the baseline model, the aggregated temporal feature is directly fed into a fully connected layer followed by softmax for action classification:

$$\hat{y}_{\mathrm{base}} = \mathrm{Softmax}(W_{\mathrm{base}} f_{\mathrm{temp}} + b_{\mathrm{base}}) \tag{7}$$

where W_{base} and b_{base} are trainable parameters, and $\hat{y}_{\mathrm{base}}$ denotes the predicted class probability distribution of the baseline model.

Compared with using only the final hidden state, temporal attention enables the model to emphasize more discriminative frames and is therefore adopted in the baseline model.

3.3. DCT-Based Frequency Branch

To further enhance the baseline model, this study introduces a DCT-based frequency branch. As shown in Fig. 1, the DCT branch takes the shared sequence feature generated by

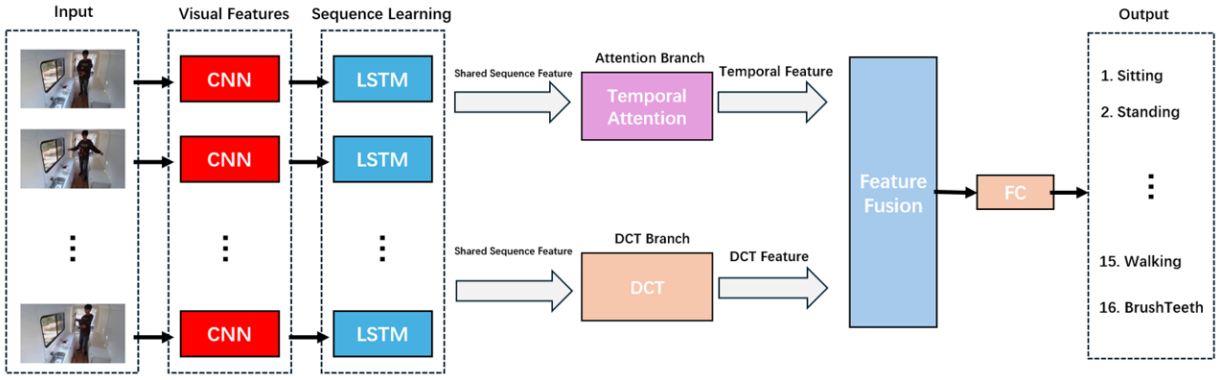

FIGURE 1. Overall architecture of the proposed action recognition framework. Input video frames are first processed by the CNN and LSTM to obtain shared sequence features. Then, a temporal attention branch and a DCT branch extract complementary temporal and frequency-domain representations, respectively. The fused feature is finally used for multi-class action classification.

the CNN-LSTM backbone and transforms it into a frequency-domain representation. Instead of applying frequency analysis directly to raw image pixels, the proposed method performs DCT on the temporal feature sequence at a more semantic level while keeping the branch lightweight.

Let the CNN feature sequence be represented as

$$V = \{v_1, v_2, \ldots, v_T\}, \quad v_t \in R^D \tag{8}$$

where T is the number of sampled frames and D is the feature dimension.

For each feature dimension, DCT is applied along the temporal axis to transform the sequence from the time domain into the frequency domain. The transformed coefficient at frequency index k is defined as

$$c_k = \sum_{t=1}^{T} v_t \cos\left[\frac{\pi}{T}\left(t - \frac{1}{2}\right)k\right], \quad k = 0, 1, \ldots, T-1 \tag{9}$$

where c_k denotes the DCT coefficient corresponding to the k-th frequency component.

After transformation, only the first K low-frequency components are retained:

$$C = \{c_0, c_1, \ldots, c_{K-1}\}, \quad K < T \tag{10}$$

because low-frequency components mainly represent the global motion trend and rhythmic characteristics of an action sequence, whereas high-frequency components are more likely to reflect local variations or noise.

The retained frequency-domain representation is then projected into a compact feature space:

$$f_{\text{dct}} = \phi(C) \tag{11}$$

where $\phi(\cdot)$ denotes a learnable projection layer.

This branch complements the attention-based temporal representation by providing frequency-domain information about global motion trends and rhythmic patterns.

3.4. Feature Fusion and Classification

After the temporal and frequency-domain features are obtained from the two branches shown in Fig. 1, they are fused to form a more complete representation of the input action sequence. The temporal feature generated by the temporal attention branch captures sequential dependency and informative time-step weighting, while the DCT-based feature provides information about global motion trends and rhythmic patterns. In this study, the two features are combined at the feature level by concatenation:

$$f_{\text{fusion}} = [f_{\text{temp}}; f_{\text{dct}}] \tag{12}$$

where f_{temp} denotes the temporal attention feature and f_{dct} denotes the frequency-domain feature.

The fused representation is then fed into a fully connected classification head to predict the action category:

$$\hat{y} = \text{Softmax}(W f_{\text{fusion}} + b) \tag{13}$$

where W and b are trainable parameters, and $\hat{y}$ denotes the predicted class probability distribution of the proposed model.

Through this design, the model jointly exploits complementary information from the time and frequency domains, which helps improve recognition robustness for actions with similar visual appearance but different temporal patterns.

4. Dataset Construction

4.1. Dataset Overview

FIGURE 2. Overview of the TLL environment and representative ADL classes.

The Trailer Living Lab (TLL) ADL Dataset is a self-built dataset for action recognition in a realistic compact home-like environment. Compared with ordinary laboratory settings, TLL involves stronger spatial constraints, furniture occlusion, and viewpoint limitations, making recognition more challenging and closer to real home scenarios. The dataset contains 16 action classes and was collected from 8 participants, providing a benchmark for cross-person ADL recognition in compact home-like environments.

4.2. Data Acquisition and Preprocessing

The raw videos were recorded using an iPhone 15 from multiple fixed viewpoints in TLL. Each participant continuously performed all 16 actions in a single session, and the long videos were manually segmented into action-specific clips.

Each clip was further divided into 5-second segments with 50% overlap. For model input, a fixed number of frames were uniformly sampled from each clip, resized to 224×224, normalized using the ImageNet mean and standard deviation, and augmented by random horizontal flipping. The dataset was organized in a cross-person manner, where the training, validation, and test sets were split by participant identity.

5. Experiments

5.1. Experimental Settings

Experiments were conducted on the self-built TLL ADL Dataset under a cross-person setting, where the data were split into training, validation, and test sets according to subject identity to evaluate generalization to unseen subjects. The final splits contained 5150, 934, and 1952 clips, respectively.

For each clip, 16 frames were uniformly sampled and resized to 224×224. Standard preprocessing and augmentation, including normalization and random horizontal flipping, were applied during training.

The models were implemented in PyTorch and trained using the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 32. A learning-rate scheduler and early stopping (patience = 8) were used, and the model with the best validation accuracy was selected.

Two models were compared: a CNN-LSTM baseline with temporal attention and the proposed DCT-enhanced variant. ResNet18 was used for spatial feature extraction, and LSTM modeled temporal dependencies. The proposed model additionally introduced a DCT branch, with $k = 8$, feature dimension $d = 256$, fusion weight $\alpha = 1.0$, and dropout rate 0.2. The temporal and DCT features were fused before classification into 16 ADL classes.

Performance was evaluated using Accuracy, Precision, Recall, and F1-score, with confusion matrices for detailed analysis.

5.2. Recognition Results

Table 1 summarizes the overall recognition performance of the baseline model and the tested DCT-enhanced variants on the TLL ADL Dataset. Compared with the baseline model, both DCT-based variants achieved better overall performance, indicating that frequency-domain enhancement provides useful complementary information beyond temporal attention. Among the tested settings, the model with $k = 8$ achieved the best performance and was selected as the final proposed model.

TABLE 1. Overall recognition performance on the TLL ADL Dataset.

Model	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
CNN-LSTM (Baseline)	68.39	77.41	67.77	68.07
CNN-LSTM + DCT ($k = 4$)	74.03	80.23	74.62	73.83
CNN-LSTM + DCT ($k = 8$, Proposed)	75.77	80.67	75.59	76.06

Specifically, the final proposed model improved the accuracy from 68.39% to 75.77%, while the F1-score increased

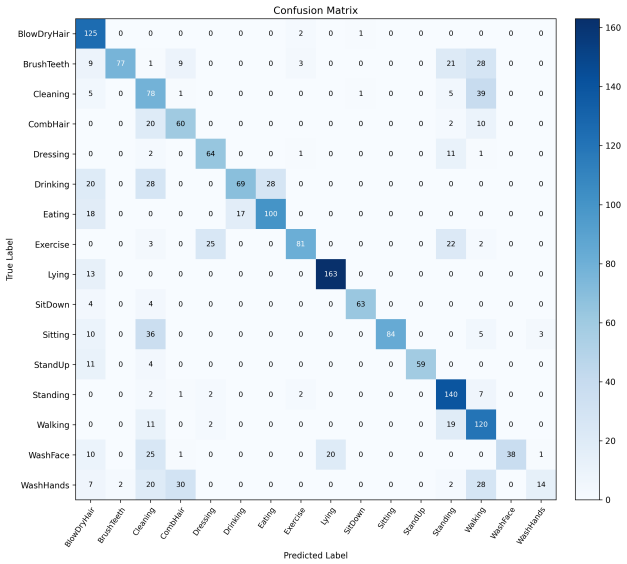

FIGURE 3. Confusion matrix of the baseline CNN-LSTM model under cross-person evaluation.

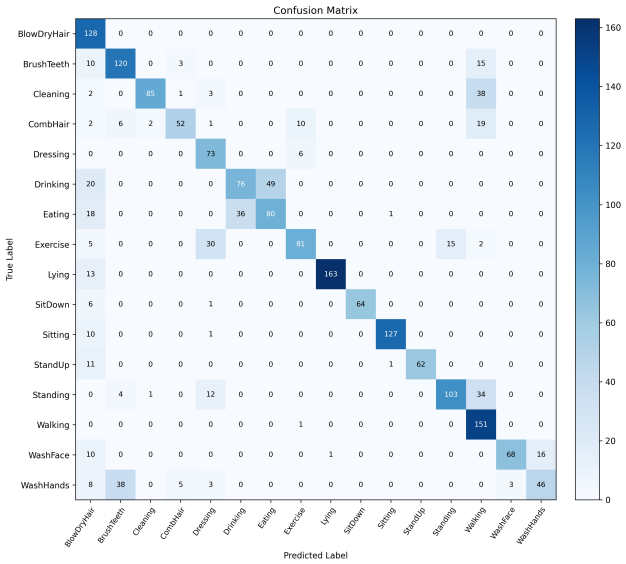

FIGURE 4. Confusion matrix of the proposed CNN-LSTM + DCT model under cross-person evaluation.

from 68.07% to 76.06%. Precision and recall also improved from 77.41% and 67.77% to 80.67% and 75.59%, respectively. These results indicate that the additional DCT branch provides useful complementary information beyond temporal attention by capturing motion tendency and temporal regularity.

5.3. Ablation and Analysis

To further analyze the proposed method, this subsection examines the contribution of the DCT branch. The comparison between the baseline model and the proposed model confirms the effectiveness of frequency-domain enhancement. While the baseline model focuses on time-domain temporal modeling, the proposed model additionally captures low-frequency motion trends and rhythmic characteristics through the DCT branch. The variant with $k = 4$ achieved 74.03% accuracy and 73.83% F1-score, while the $k = 8$ variant achieved the best overall performance with 75.77% accuracy and 76.06% F1-score.

Fig. 3 shows the confusion matrix of the baseline model, where confusion remains for visually similar or transition-related actions such as Sitting/SitDown and Standing/StandUp. Fig. 4 shows the confusion matrix of the proposed model. Compared with the baseline, the proposed model achieves a more balanced pattern and reduces confusion among these categories, indicating that the DCT branch helps capture temporal transitions. However, confusion still exists for actions with similar contexts, such as Drinking/Eating and WashFace/WashHands, suggesting that further spatial-temporal modeling is needed.

6. Discussion and Conclusion

The experimental results indicate that the proposed CNN-LSTM with temporal attention and DCT framework is effective for ADL recognition in a compact home-like environment, consistently outperforming the baseline. This demonstrates that frequency-domain enhancement provides useful complementary information to time-domain temporal modeling.

The proposed method is particularly effective for actions that are visually similar but differ in motion trend or temporal continuity. Class-wise analysis suggests that the DCT branch improves recognition of transition-related actions, although the improvement is not uniform across all categories.

Several limitations remain. The current dataset is relatively limited in scale, and the generalization ability of the framework should be validated on larger datasets. Although the DCT branch improves overall performance, confusion still exists for actions with similar contexts. In addition, this study focuses on comparing the CNN-LSTM baseline with the DCT-enhanced variants, while comparisons with more advanced architectures such as 3D CNNs and Transformers are left for future work.

In conclusion, this study proposes a CNN-LSTM with temporal attention and DCT framework for ADL recognition in TLL, demonstrating the effectiveness of combining temporal attention with frequency-domain motion representation under realistic conditions.

Acknowledgements

This work was supported in part by the JST Moonshot Research and Development Program (Grant Number JPMJMS2034).

References

[1] P. Gong and X. Luo, "A Survey of Video Action Recognition Based on Deep Learning," *Knowledge-Based Systems*, vol. 320, p. 113594, 2025.

[2] G. Ercolano and S. Rossi, "Combining CNN and LSTM for Activity of Daily Living Recognition with a 3D Matrix Skeleton Representation," *Intelligent Service Robotics*, vol. 14, pp. 175–185, 2021.

[3] S. Das, S. Sharma, M. Kaul, D. K. Sethi, and F. Bremond, "Toyota Smarthome: Real-World Activities of Daily Living," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 833–842.

[4] D. Tang, Y. Yoshihara, T. Obo, T. Takeda, J. Botzheim, and N. Kubota, "Social Rhythm Management Support System Based on Informationally Structured Space," in *Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC)*, 2017.

[5] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 510–526.

[6] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in the Wild," *arXiv preprint arXiv:1212.0402*, 2012.

[7] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A Large Video Database for Human Motion Recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2556–2563.

[8] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.

[9] D. Tang, Y. Yoshihara, T. Takeda, J. Botzheim, and N. Kubota, "Informationally Structured Space for Life Log Monitoring in Elderly Care," in *Proceedings of the 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2015, pp. 1421–1426.

[10] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2625–2634.

[11] A. Tran *et al.*, "Action Recognition in the Frequency Domain," *arXiv preprint arXiv:1409.0908*, 2014.

[12] Q. Zhao *et al.*, "PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 8877–8886.

[13] C.-Y. Wu, M. Zaheer, H. Hu, R. Manmatha, A. J. Smola, and P. Krähenbühl, "Compressed Video Action Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6026–6035.

[14] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 20–36.

[15] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-Local Neural Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.