Multimodal Transformer for Nursing Activity Recognition

Momal Ijaz[1], Renato Diaz[1], Chen Chen[1,2]
[1]Department of Computer Science, University of Central Florida, USA
[2]Center for Research in Computer Vision, University of Central Florida, USA
{im.momil, diazrenato2001}@knights.ucf.edu, chen.chen@crcv.ucf.edu

Abstract

In an aging population, elderly patient safety is a primary concern at hospitals and nursing homes, which demands for increased nurse care. By performing nurse activity recognition, we can not only make sure that all patients get an equal desired care, but it can also free nurses from manual documentation of activities they perform, leading to a fair and safe place of care for the elderly.

In this work, we present a multimodal transformer-based network, which extracts features from skeletal joints and acceleration data, and fuses them to perform nurse activity recognition. Our method achieves state-of-the-art performance of 81.8% accuracy on the benchmark dataset available for nurse activity recognition from the Nurse Care Activity Recognition Challenge. We perform ablation studies to show that our fusion model is better than single modality transformer variants (using only acceleration or skeleton joints data).

Our solution also outperforms state-of-the-art ST-GCN, GRU and other classical hand-crafted-feature-based classifier solutions by a margin of 1.6%, on the NCRC dataset. Code is available at `https://github.com/Momilijaz96/MMT_for_NCRC`.

1. Introduction

Elderly care and safety is a primary concern at health care centers, which highly demands for increased nurse care. Performing nurse activity recognition is an important task as it can aid in the process of monitoring the health care plans compliance for each patient and frees up nurses from the task of manual reporting and documentation. Activities performed by nurses tend to be more complex and longer than straightforward actions or gestures available in benchmark data sets like walking, running, eating, sleeping, and waving hello [24, 29].

Human activity recognition is a widely researched area in computer vision as it has applications in human computer interaction or video understanding, etc. [4, 5, 35]. Over the past few years, skeleton-based action recognition has gained popularity because of it's good estimate on human body's dynamic movements and is also more robust to illumination variations and background noises [11, 31].

Skeletal action recognition has been previously performed using hand crafted features [34] or manually structuring data as a pseudo image and passing it to a Convolutional Neural Network (CNN) [11], or as a sequence of coordinates vectors which are fed to a Recurrent Neural Network (RNN) [25, 37]. Other works have explored the benefits of using divided space-time feature extraction by using a Spatio-Temporal Graph Convolutional Network (ST-GCN), which models the spatial configuration and temporal dynamics of skeletons [39]. Recently, researchers have been trying to import the capabilities of transformers [33] from Natural Language Processing (NLP) to vision domain. Among all other variants, recent Vision Transformers [10] stood out, as they showed a convolution-free transformer network can show comparable performance to CNNs in vision tasks. Similar studies have been done by researchers for coming up with a pure transformer architecture for skeletal action recognition as well [27, 36].

In addition to skeletal poses, acceleration signal has been used in quite a few works [1, 2, 12] for performing action recognition and acceleration has proven to be quite effective for the task. The Nurse Care Activity Recognition Challenge (NCRC) dataset [18] comprises of data from acceleration sensors, location sensors and skeletal joints. Previous works on the dataset used different combinations of these modalities with hand crafted features, using simple classification algorithms like KNN or Random-Forests [20, 23]. Other advanced works on the dataset used ST-GCNs [3] and Gated Recurrent Units (GRUs) [15]. All these works used different combinations of data modalities (i.e. skeletal joints, location, and acceleration), but none of the works explored fusion of the two strongest signals for action recognition, i.e. acceleration & skeletal joints.

In this paper, we present a multimodal transformer network that fuses acceleration and spatio-temporal skeletal features to perform activity recognition on the NCRC

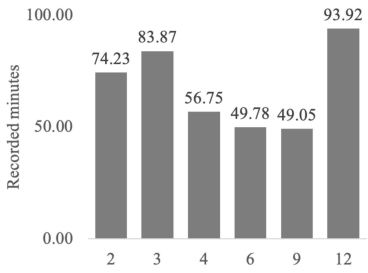

Figure 1. Individual recorded time for each activity. Activity2: Vital Signs Measurement, Activity3: Blood collection, Activity4: Blood glucose measurement, Activity6: Indwelling drip retention and connection, Activity 9: Oral care, Activity 12: Diaper exchange and cleaning of area.

dataset. The main contributions of our work are summarized as follows:

- We show that fusing acceleration signal and skeletal joints data leads to improved performance for action recognition, as compared to using single modality. Additionally, we present pure transformer-based single modality networks for skeletal joints and acceleration data, and an efficient dual modality network for both signals. Our dual modality transformer, using both acceleration and skeletal joints data, outperforms single modality networks by 5.2%.

- We present a novel attention-based fusion technique for fusing spatio-temporal skeletal features with acceleration features, for exploiting correlations between acceleration and skeletal joints; to develop better semantic understanding of actions being performed. Our fusion method outperforms simple fusion baseline, by 6.8%.

- Our proposed dual modality transformer outperforms state-of-the-art GRUs, ST-GCNs, and handcrafted feature-based classifiers, like KNNs, and achieves the highest performance on the NCRC dataset of 81.8%.

2. Related Work

Skeletal Action Recognition: Skeletal action recognition has lately been a preferred method for performing action recognition since it is more robust to illumination variations and other background noises. Older methods such as manual extraction of hand crafted features [34], crafting pseudo images out of the skeletal poses and feeding them into CNNs [11], or other RNN based methods [25] have become outdated after the huge success of graph-based methods [39]. Yan et al. introduced ST-GCNs which are able to map spatial correlations and temporal changes of a human skeleton for performing action recognition. GCN-based approaches [9, 30] use topographical features of skeleton to extract and combine spatial skeletal features and temporal dynamics. In contrast to ST-GCN-based methods, transformers can directly learn correlations between joints in a frame and complete skeletal poses across frames. Using a divided space-time attention mechanism, researchers have shown different variants of transformer architecture for performing skeletal action recognition. In [27], authors use pure transformer architectures to map correlation between joints in one frame and across frames, using two different transformers for spatial and temporal feature extraction. In [36], authors group joints into parts. They use a single transformer encoder block for computing spatial and temporal features of the skeletal joints data. Their proposed method involves computing correlations between joints in one part, across parts in one frame and across frames for same part, using a modified intra-inter part attention mechanism.

Acceleration-based Action Recognition: Acceleration signal has proven to be useful for performing action recognition. Most earlier works used hand crafted statistical features from acceleration signal with simple classifiers like Support Vector Machines (SVMs) [1, 2, 12]. However, novel deep-learning-based techniques have outperformed the classical approaches significantly. In [14], the authors use a 3-layered CNN, followed by a Long Short Term Memory (LSTM) block for performing activity classification, and show that their method is better than SVM trained on similar features. In [8], the authors present a pure CNN architecture which exceeds classical feature extraction pipelines. The proposed CNN architecture has a modified convolutional kernel to adapt to the triaxial acceleration signal. However, as mentioned in [32], accelerometer-based activity recognition is considered a dead end as the sensor offers limited information. We extend this idea and explore the effectiveness of fusing skeletal joints data with acceleration signal for activity recognition.

Nurse Care Activity Recognition: Best performance on the NCRC dataset was obtained by using handcrafted features extracted from skeletal data and location sensor with an ensemble of K-Nearest Neighbors (KNNs) models [20]. Other works explored using RandomForests [23] with just acceleration signal. Among deep-learning approaches, baseline set up by competition organizers comprised of a CNN backbone which used all data modalities [23]. Other works used ST-GCNs [3] with skeletal joints data, and GRUs [15] with skeletal joints and location sensor data. We explore the fusion of acceleration signals with skeletal joints with a transformer-based network.

Transformers: Researchers have been trying to explore the capabilities of revolutionary transformers [33] in vision domain, to see if these are a strong competitor against the

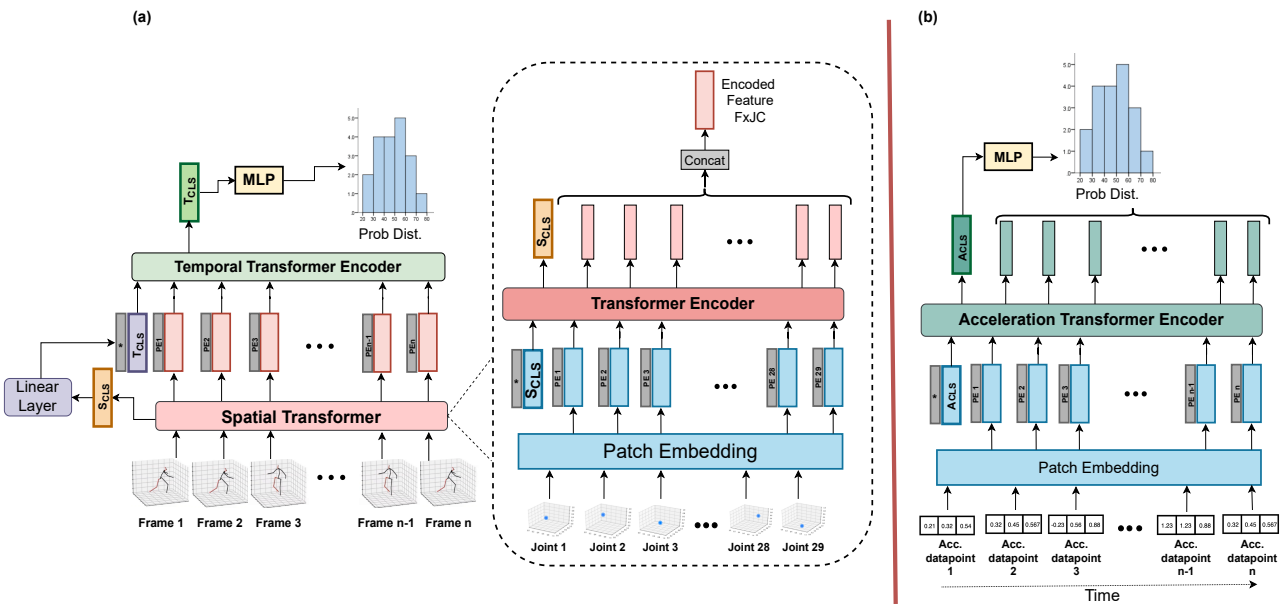

Figure 2. **Single Modality Models:** Single modality models use only acceleration or skeletal joints data. **(a) Spatio-Temporal Skeleton Model:** The skeletal model comprises of two transformer blocks: spatial and temporal encoders, for computing spatial and temporal features from skeletal joints of given action sample. **(b) Acceleration Model:** The acceleration model has one transformer block, which computes correlation across acceleration data-points for a given action sample.

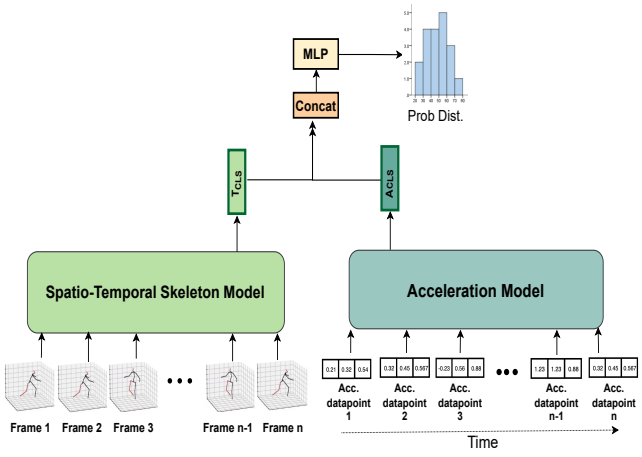

Figure 3. **Multi-Modal Transformer - Simple Fusion**. Multi-modal transformer uses both modalities, acceleration and skeletal joints. Skeletal features are extracted by single modality skeleton model, whereas the acceleration features are extracted by single modality acceleration model, and skeletal and acceleration features are added to perform classification.

widely used CNNs. Different variants of transformers have shown promising results for various vision tasks including but not limited to video and image classification, semantic segmentation, and object localization [6, 10, 26].

Transformer-based Fusion Strategies: Various mechanisms have been studied for the exchange or fusion of information between two transformer blocks. In [28], the authors fuse audio and visual signal using a simple early fusion technique to perform video classification. In CrossViT [6], authors utilize tokens to exchange information between two transformer blocks that process images of two different resolutions. The fusion technique used in CrossViT is based on cross-attention, in which tokens of one branch attend to encoded features of other branches for sharing information. As for video classification, in the state-of-the-art Multi-view Transformer [38], the authors compare three different fusion techniques to exchange information between different resolution of video tokens. They found cross view fusion to be the best fusion approach, in which tokens of larger resolution attend to tokens of smaller resolution, at selected layers in a transformer block. Our proposed fusion approach, for combining information from skeletal joints and acceleration signal, is majorly inspired from this technique.

3. Dataset

The dataset used for the study along with it's posed challenges is as follows.

3.1. Description

For this experimentation, we are using the dataset from the Nurse Care Activity Recognition Challenge [18]. In this particular dataset, 6 different activities have been recorded

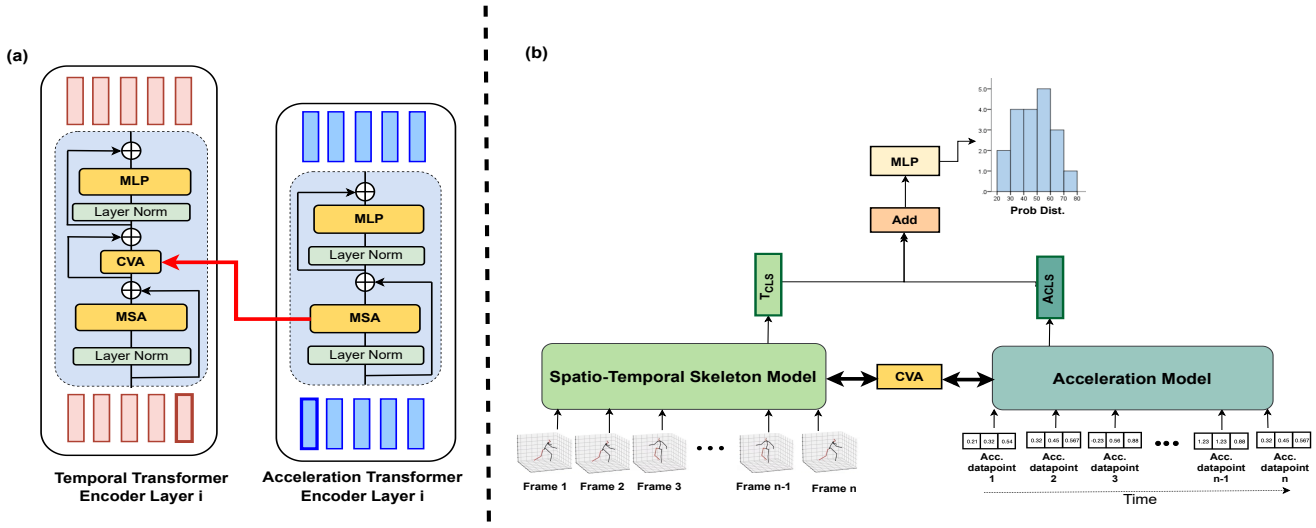

Figure 4. **(a) CrossView Fusion**. Cross View fusion is performed between corresponding encoder layers of acceleration block from acceleration model and temporal transformer block from spatio-temporal skeleton model. **(b) Multi-Modal Transformer - Cross View Fusion**. Cross view fusion-based multimodal transformer is exactly similar to simple fusion transformer (Figure-1), except for an added cross view fusion mechanism between acceleration and skeletal joints branch.

by 8 nurses working in a controlled, monitored environment. These activities are as under:

- Vital signs measurement
- Blood collection
- Blood glucose measurement
- Indwelling drip retention and connection
- Oral care
- Diaper exchange and cleaning of area

The dataset comes with a testing and training split. The training set contains all the aforementioned activities performed by 6 subjects, which sums up to a total of 282 action samples. Whereas the testing set contains actions performed by 2 different subjects, summing up to a total of 116 samples. Motion-capture cameras, accelerometer chips, and location sensors are used to record each action. The motion capture camera captures 29, 3D, joint locations at 100 Hz frequency. The accelerometer chip captures the acceleration of the subject along x, y and z axes at 4Hz frequency. The location sensor captures the x and y location coordinates of the subject and changes in air pressure at 20Hz frequency. Acceleration data is captured using a sensor, placed in upright position in the right pocket of the subject, whereas skeletal data is collected using IR-based motion capture cameras. Figure 1 shows the recorded minutes of each activity.

3.2. Challenges Posed by the Dataset

This dataset is the only one of it's kind that fulfills all of our requirements, i.e. being designed for nurse care activity recognition. However, there are a few challenges posed by this dataset, detailed below.

- The sampling rate of sensors are widely different, Figure 5. Skeletal data was recorded at 100Hz, giving 6000 skeletal poses per action, whereas acceleration data was recorded only at 4Hz, giving roughly 150 data points per action. That's 0.25 data points for acceleration and 100 for skeletal joints in a second.
- Acceleration data was very noisy. It had null values and was entirely missing for 2 action samples performed by subject 2.
- Overall size of the dataset was very small, there were just 282 training and 216 testing samples available, which proved to be a main road-block in training data hungry transformer-based architecture.
- Skeletal data also had entirely missing joints for some action samples and was noisy as well.

4. Method

In this study, we explore the fusion of acceleration features with spatio-temporal skeletal features for performing nurse activity recognition. The acceleration data and skeletal joints data are convenient and economical to collect. For acceleration, we have lightweight accelerometer chips or smartphones. For skeletal data, motion capture cameras like Kinetics [40] and RealSense [21] can do a pretty decent job even in a hustling nursing environment.

4.1. Single Modality Transformers

We present 2 single modality transformer models, which are trained only on acceleration or skeletal joints data. Each single modality transformer model comprises of a class token, like ViT [10], which is used for performing final classification.

4.1.1 Spatio-Temporal Skeleton Model

Figure 2(a) demonstrates the architecture of a spatio-temporal model, which is mainly inspired from the PoseFormer [41] network. The spatio-temporal skeletal model is a single modality transformer that performs action recognition using only skeletal joints data.

The model comprises of two transformer blocks, spatial and temporal. Spatial transformer takes each frame as an input and computes correlation between 29 individual 3D joint points in a frame. We pass each 3D joint coordinate through a linear patch embedding layer and add position encoding before passing it to a standard transformer encoder block. We also append a spatial CLS token S_{cls} to the inputs. The spatial transformer outputs feature vectors for each joint, which are concatenated together. This concatenated feature vector is a representation of each frame computed by the spatial transformer. This process is repeated for all frames in the video sample and finally we pass all encoded frames to the temporal transformer block for computing temporal correlation across frames. In the spatial transformer, each token is a joint, whereas for temporal transformer each token is a feature vector representing one frame.

The spatial CLS, S_{cls}, token is passed through a linear layer to project it up to the temporal embedding dimension. This token with temporal embedding dimension is called T_{cls} and is passed to the temporal transformer encoder along with other encoded frames. T_{cls} is used for final classification and hence is passed through a simple linear MLP classification head and gives probability distribution of labels.

4.1.2 Acceleration Model

This model, Figure 2(b), attempts to perform action recognition using just the acceleration of the performer. Each acceleration data point comprises of acceleration value, recorded every 4 seconds, along the x, y, and z dimensions. We interpolate the acceleration signal using simple linear interpolation. Next, we denoise it using a moving average window of size 40, and fill in the samples with missing acceleration data.

The acceleration-only model is similar to the spatial transformer model, although here each token is an acceleration data point (3D vector). We encode each data point using a linear embedding layer, append position encodings, and an acceleration CLS token, A_{cls}, with inputs, which is passed through acceleration transformer block. The output is the encoded feature vector A_{cls}, which is passed through a MLP head for the prediction of the target class.

Model	Learning Rate	Drop	Stoch. Drop	Attn. Drop
Skeleton only	0.02	0	0.2	0
Acceleration only	0.02	0.	0.2	0.
Simple Fusion	0.0025	0.05	0.2	0.05
CrossView Fusion	0.0025	0.	0.2	0.

Table 1. **Training Hyper-parameters**. Single Modality models performed well without strong regularization, whereas fusion models converged well with non-zero drop rates.

4.2. Multi-Modal Transformers

We use the single modality transformer models to create 2 different dual modality transformer models, which utilize both acceleration and skeletal joints data. The first dual modality transformer is a simple feature baseline, which concatenates the respective class tokens from acceleration and skeletal joints branch to perform classification. The second dual modality transformer is similar to the first one with just the addition of cross view fusion mechanism. This mechanism allows for the exchange information between skeletal joints and the acceleration branch.

4.2.1 Simple Fusion

Figure 3 illustrates the simple fusion model, inspired from the early fusion technique presented in [28], for fusing visual and audio signal. In the simple fusion model, we take the single modality spatio-temporal skeleton model and acceleration model. Skeleton model takes skeletal joints as input and computes spatio-temporal skeletal features and gives a temporal CLS token, T_{cls}, as output. The acceleration model takes acceleration of the same action as input and outputs an acceleration CLS token A_{cls}. We simply concatenate these two CLS tokens and pass them to the MLP classification head, which gives us the resultant class of action sample.

4.2.2 CrossView Fusion

In CrossView fusion model, along with simple aggregation of the CLS tokens from both branches, we fuse information between acceleration and skeletal encoders. Particularly, the tokens of the temporal transformer block of the spatial-temporal skeleton model act as queries, and the tokens of acceleration encoder block act as key and value pairs. This fusion technique is majorly inspired from the CrossView attention presented in Multi-view transformer [38] paper,

for fusing information from multi resolution input patches. CrossView fusion model allows the temporal skeletal joint features to attend to acceleration features for developing a better understanding of the action being performed. For CrossView fusion, we keep the embedding dimension and depth of both acceleration and temporal encoders similar, which eliminates the need to project up or down the tokens before passing to other branch. CrossView fusion introduces an additional cross attention operation after multi-headed self attention (MSA) mechanism in each layer of the temporal skeletal encoder. This block attends to MSA encoded acceleration tokens from the corresponding layer. Mathematically, each i^{th} layer in temporal skeletal encoder attends to the i^{th} layer of acceleration encoder, as shown in the equation below.

$$z^{temporal_i} = CVA(z^{temporal_i}, z^{acc_i})$$

$$CVA(x, y) = Softmax(\frac{W^Q x W^K y^T}{\sqrt{d_k}})W^V y$$

Here, CVA stands for CrossView Attention, $z^{temporal}$ is temporal skeletal encoder tokens, and z^{acc} are acceleration encoder tokens, with W^Q, W^K and W^V as the weights of CVA block for computing query, key and value representations.

5. Experiments and Results

5.1. Implementation Details

The NCRC dataset has a limited number of samples, and training a transformer based network requires strong data augmentation or pre-training. Although, we were able to converge our transformer models without using any of the two, by using adaptive sharpness aware minimization [22] (ASAM). This technique has been tested out on ViT [7], and authors in this work were able to make ViTs outperform ResNet without augmentation or pre-training using sharpness aware minimization. We used ASAM with neighborhood size of 0.5, to smooth out the loss function and avoid over-fitting. ASAM focuses on finding optimal neighborhoods for network parameters instead of optimal values, which ultimately leads to a much smoother loss function and better generalization.

Along with various other regularization techniques, we also used stochastic depth [16], which is known to facilitate the convergence of deep transformers. [13] We set the stochastic depth rate in the range of 0.1 - 0.2. Additionally, we also used drop outs and attention drop rates, and found them key factor for allowing our fusion models to converge and generalize well. However, single modality models performed well with drop rates set to zero. We used a batch size of 16, with SGD optimizer, a weight decay rate of 5e-4, and Cosine Annealing learning rate scheduler. The rest of the hyper-parameters for converging every model are given in Table 1.

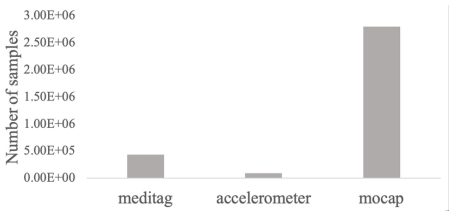

Figure 5. Number of Observations by each sensor

5.2. Comparison with State-of-the-art

Earlier works that have also performed on this dataset are summarized in Table 2. The NCRC dataset comprises of three different modalities: acceleration, skeletal joints (Motion Capture), and location sensors. Among different combinations of these modalities, the fusion of acceleration with the skeletal joints performs the best and gives the highest validation accuracy. Reliance on just one signal like skeletal joints or acceleration does not provide adequate results. The second best performance is obtained by an ensemble of a KNN-based method [20], which uses location and skeletal joints data with hand crafted features. We did not use the location signal in our method as the collection of location data in a nursing environment is relatively harder in the real world as compared to skeletal joints or acceleration data. Deep learning-based solutions like CNN [23], ST-GCN [3] and GRU [15] perform poorly compared to our transformer-based solution due to the smaller size of the dataset. The usage of ASAM [22] clearly helps our model to avoid over-fitting compared to these solutions.

In terms of class wise performance, summarized in Table 3, our method performs best in terms of accuracy and F1-score on all classes except for class 3, which is blood collection. As shown in the confusion matrix in Figure 6, this class is mostly confused with blood glucose measurement. This is due to the fact that blood collection and blood glucose measurement are quite similar actions. CrossView fusion model performs best on class 12, which is diaper exchange and cleaning of area, reflecting this activity has highest variation from all other activities in the dataset. Overall, mean accuracy and F1 scores of class-wise performance for all classes, of our approach is 12.5% better than the state-of-the-art KNN-based solution and the second-best ST-GCN-based solution.

5.3. Ablation Studies

Single Modality vs. Dual Modality: The impact of fusing acceleration and skeletal joints signal can be observed by comparing it with the single modality transformer models, trained on just skeletal joints or acceleration signal. We

Sensors Used	Method	Validation Accuracy (%)
Motion Capture and Location	KNN	80.2
Motion Capture	ST-GCN	64.6
All modalities	CNN	46.5
Acceleration	Random Forest	43.1
Motion Capture and Location	GRU	29.3
Acceleration and Motion Capture (Our Approach)	**Transformers**	**81.8**

Table 2. **Comparison with state-of-the-art:** Comparing our approach with other modalities and methods on the NCRC dataset. Our proposed method, fusion of acceleration and skeletal joints using transformer-based method outperforms all other modalities and methods.

Activity id	Accuracy(%)					F1-Score				
	Our Approach	ST-GCN	RF	DTT	KNN	Our Approach	ST-GCN	RF	DTT	KNN
2	**80.00**	65.10	4.17	52.08	47.92	**80.00**	63.30	5.97	54.95	61.33
3	75.0	54.5	68.25	73.02	**95.24**	65.5	48.9	67.19	69.70	**76.43**
4	**71.0**	60.0	13.89	38.89	22.22	**72.8**	55.0	21.74	43.08	34.78
6	**85.8**	62.2	84.38	34.38	21.88	**82.8**	65.3	67.50	37.93	32.56
9	**71.5**	50.8	12.12	0.00	57.58	**77.0**	44.2	17.39	0.00	67.86
12	**87.1**	49.0	90.77	47.69	100	**93.2**	40.5	63.10	40.79	73.45
Class Wise Mean	**78.2**	57.0	50.54	45.85	65.70	**78.6**	52.9	43.82	44.92	61.61

Table 3. **Activity wise performance comparison with state-of-the-art**. Our method outperforms all existing solutions in terms of Accuracy and F1 Score except for Class 3 which is blood collection. Overall, Class Wise mean of accuracy and F1-scores of our approach is 12.5% better than state-of-the-art KNN based solution.

Model	Accuracy	F1-Score	Precision	Recall
Skeleton Model	76.7	67.0	69.1	70.5
Acceleration Model	45.6	10.9	9.3	14.9
Simple Fusion	75.0	71.6	75.6	72.3
Cross-View Fusion	**81.8**	**78.4**	**79.4**	**78.3**

Table 4. **Single Modality vs. Dual Modality Performance Comparison** Dual modality CrossView fusion model outperforms single modality and simple fusion models.

Nurse ID	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
2	59.2	65.8	59.9	67.4
3	68.3	67.9	66.9	72.7
4	68.7	68.4	66.9	80.0
5	72.6	73.1	70.3	74.6
6	**95.4**	**91.5**	**92.5**	**93.4**
7	*44.8*	*60.9*	*50.6*	*61.1*
8	79.2	77.2	78.0	79.3
9	82.7	79.4	76.9	82.3

Table 5. **LOSOCV Performance of CrossView Fusion Model**. The best performing subject ID is 6 (in bold), whereas the worst performing subject ID is 7(in italic).

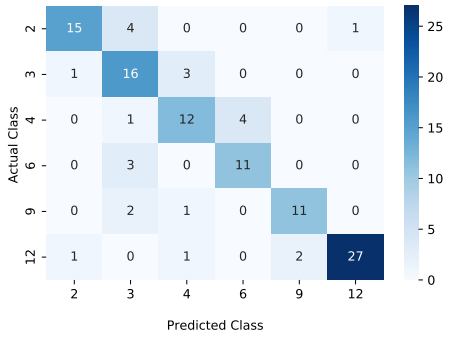

Figure 6. **Confusion Matrix** of CrossView Fusion Model on Validation set. Activity2: Vital Signs Measurement, Activity3: Blood collection, Activity4: Blood glucose measurement, Activity6: Indwelling drip retention and connection, Activity 9: Oral care, Activity 12: Diaper exchange and cleaning of area.

can see in Table 4, both spatio-temporal skeleton model and acceleration model give lower validation accuracy than CrossView fusion model. CrossView Fusion model out performs spatio-temporal skeleton model by 5%. However, the simple fusion method does not perform as good as single modality spatio-temporal skeleton model, which reflects that simple concatenation of the acceleration and skeletal joints feature vector, hurts the performance of model and makes it perform 1.7% lower than skeleton only model. Among single modality models, we can see that spatio-temporal skeleton model performs 31% better than acceleration model, and that makes sense because of the wide difference of sensor sampling rates (skeletal data was recorded at 100Hz and acceleration at just 4Hz, Figure 5), and more noise in acceleration data than skeletal joints data.

Impact of CrossView Fusion: We analyze the impact of CrossView Fusion mechanism in our multi-modal transformer by comparing it with the Simple Fusion variant. In Simple fusion model, we simply concatenate the CLS tokens coming from the single modality transformers, whereas in CrossView fusion model, we add cross attention

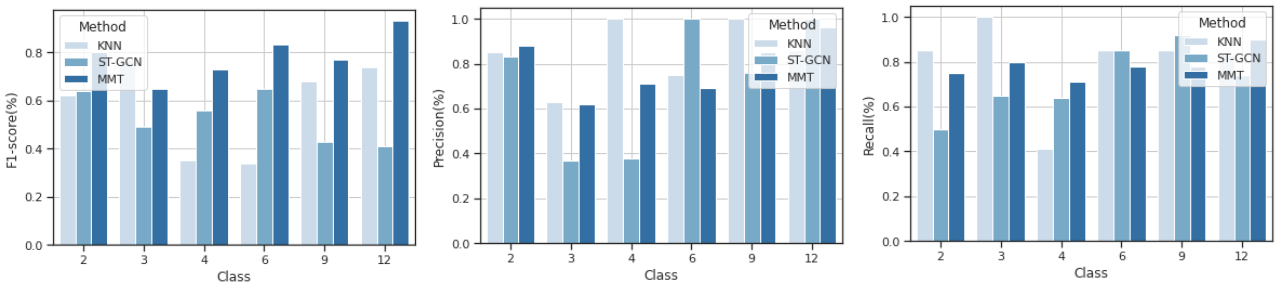

Figure 7. Class Wise F1-score, Precision and Recall comparison of **our Multi-Modal Transformer(MMT)** with top two solutions, **ST-GCN** and **KNN**. Our model MMT, outperforms ST-GCN and KNN in terms of F1-score for all classes and gives comparable performance in terms of precision and recall scores.

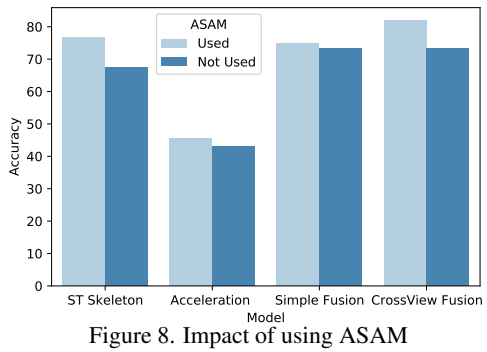

Figure 8. Impact of using ASAM

based fusion between acceleration encoder and temporal encoder of spatio-temporal skeleton model. The impact of this added fusion can be seen in Table 4, where the CrossView fusion-based model outperforms the simple fusion model by 6.8%. We can also see the features extracted by simple fusion method are different and not as diverse as CrossView fusion features, in Figure 4.

Leave-One-Subject-Out Cross Validation: For testing the generalizability of our proposed approach we perform leave-one-subject-out cross validation on the dataset. We combine test and train subjects and treat each subject as a test subject, while others as training subjects. Resultant performance of CrossView Fusion model for all subjects is reported in Table 5. Our proposed solution performed best for cross validation on subject 6, reflecting all other subjects made up a good diverse data for our solution to converge well and achieve 93.4% accuracy. The worst validation scores were obtained for subject 7, for which the model only gave 61.1% accuracy. Overall, we can see that the model is generalizing well and giving adequate performance on cross subject validation.

Impact of using ASAM: We tried training all dual and single modality transformer models with and without ASAM [22], as shown in Figure 8. Using ASAM allowed models to avoid over-fitting and generalize well. Single Modality models, Spatio-Temporal Skeleton model, and Acceleration model had less parameters so they have benefitted the least as compared to the CrossView fusion models which benefitted more from ASAM. The Acceleration model's accuracy improved by 2.6% and skeleton model's accuracy improved by 9% utilizing ASAM. We saw a boost of 2.7% for simple fusion and a boost of 8.53% for CrossView Fusion model by using ASAM. The CrossView fusion model had the largest gain from usage of ASAM, mainly because it has the highest number of parameters and a more bumpy loss function than all other models.

6. Conclusion and Discussion

In this work, we demonstrate the effectiveness of fusing acceleration and skeletal joints signals for performing skeletal action recognition. We present a novel multimodal transformer architecture with cross-attention-based fusion between skeletal joints and acceleration data. Our proposed multimodal fusion transformer model outperforms single modality and simple fusion baselines by a margin of 5-6%. We achieve state-of-the-art results on the Nurse Care Activity Recognition dataset and illustrate generalizing ability of our method in ablation studies.

Limitations and Future Work:

- A limitation of the dataset includes highly imbalanced sampling rates of skeletal joints and acceleration signals, Figure 5, which proved to be an obstacle in unlocking the full potential of our proposed method. Exploring the impact of multimodal fusion transformers on a dataset with a uniform number of observations from acceleration and skeletal joints sensors might result in improved performance.
- The small size of the dataset is also a potential issue to resolve for future works.
- One can also explore pre-training the skeletal branch on skeletal joints data like NTU-RGB+D60/120 [24, 29] and acceleration branch on NCRC-2 [17] or NCRC-3 [19] dataset to further improve the model's convergence.

References

[1] Ling Bao and Stephen S. Intille. Activity recognition from user-annotated acceleration data. In Alois Ferscha and Friedemann Mattern, editors, *Pervasive Computing*, pages 1–17, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[2] Akram Bayat, Marc Pomplun, and Duc A. Tran. A study on human activity recognition using accelerometer data from smartphones. *Procedia Computer Science*, 34:450–457, 2014. The 9th International Conference on Future Networks and Communications (FNC'14)/The 11th International Conference on Mobile Systems and Pervasive Computing (MobiSPC'14)/Affiliated Workshops.

[3] Xin Cao, Wataru Kudo, Chihiro Ito, Masaki Shuzo, and Eisaku Maeda. Activity recognition using st-gcn with 3d motion data. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '19 Adjunct, page 689–692, New York, NY, USA, 2019. Association for Computing Machinery.

[4] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[5] Chen Chen, Kui Liu, Roozbeh Jafari, and Nasser Kehtarnavaz. Home-based senior fitness test measurement system using collaborative inertial and depth sensors. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4135–4138. IEEE, 2014.

[6] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification, 2021.

[7] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pretraining or strong data augmentations. *ArXiv*, abs/2106.01548, 2021.

[8] Yuqing Chen and Yang Xue. A deep learning approach to human activity recognition based on single accelerometer. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1488–1492, 2015.

[9] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 180–189, 2020.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021.

[11] Yong Du, Yun Fu, and Liang Wang. Skeleton based action recognition with convolutional neural network. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 579–583, 2015.

[12] Anitha Edison and Jiji C. V. Hsga: A novel acceleration descriptor for human action recognition. In *2015 Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, pages 1–4, 2015.

[13] Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. *ArXiv*, abs/1909.11556, 2020.

[14] Esther Fridriksdottir and Alberto G. Bonomi. Accelerometer-based human activity recognition for patient monitoring using a deep neural network. *Sensors*, 20(22), 2020.

[15] Md. Nazmul Haque, Mahir Mahbub, Md. Hasan Tarek, Lutfun Nahar Lota, and Amin Ahsan Ali. Nurse care activity recognition: A gru-based approach with attention mechanism. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '19 Adjunct, page 719–723, New York, NY, USA, 2019. Association for Computing Machinery.

[16] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016.

[17] Sozo Inoue, Sayeda Shamma Alia, Paula Lago, Hiroki Goto, and Shingo Takeda. Nurse care activities datasets: In laboratory and in real field, 2020.

[18] Sozo Inoue, Paula Lago, Shingo Takeda, Alia Shamma, Farina Faiz, Nattaya Mairittha, and Tittaya Mairittha. Nurse care activity recognition challenge. *IEEE Dataport*, 2019.

[19] Sayeda Shamma Alia; Kohei Adachi; Nhat Tan Le; Haru Kaneko; Paula Lago; Sozo Inoue. Third nurse care activity recognition challenge, 2021.

[20] Md. Eusha Kadir, Pritom Saha Akash, Sadia Sharmin, Amin Ahsan Ali, and Mohammad Shoyaib. Can a simple approach identify complex nurse care activity? In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '19 Adjunct, page 736–740, New York, NY, USA, 2019. Association for Computing Machinery.

[21] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel realsense stereoscopic depth cameras. *ArXiv*, abs/1705.05548, 2017.

[22] Jungmin Kwon, Jeongseop Kim, Hyunseong Park, and Inae Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *ICML*, 2021.

[23] Paula Lago, Sayeda Shamma Alia, Shingo Takeda, Tittaya Mairittha, Nattaya Mairittha, Farina Faiz, Yusuke Nishimura, Kohei Adachi, Tsuyoshi Okita, François Charpillet, et al. Nurse care activity recognition challenge: summary and results. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pages 746–751, 2019.

[24] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2020.

[25] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 816–833, Cham, 2016. Springer International Publishing.

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[27] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Spatial temporal transformer network for skeleton-based action recognition. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani, editors, *Pattern Recognition. ICPR International Workshops and Challenges*, pages 694–701, Cham, 2021. Springer International Publishing.

[28] Merey Ramazanova, Victor Escorcia, Fabian Caba Heilbron, Chen Zhao, and Bernard Ghanem. Owl (observe, watch, listen): Localizing actions in egocentric video via audiovisual temporal context. *arXiv preprint arXiv:2202.04947*, 2022.

[29] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[30] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Non-local graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:1805.07694*, 1(2):3, 2018.

[31] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, 2017.

[32] C. Tong, Shyam A. Tailor, and Nicholas D. Lane. Are accelerometers for activity recognition a dead-end? *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications*, 2020.

[33] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

[34] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2014.

[35] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. *ArXiv*, abs/1608.00859, 2016.

[36] Qingtian Wang, Jianlin Peng, Shuze Shi, Tingxi Liu, Jiabin He, and Renliang Weng. Iip-transformer: Intra-inter-part transformer for skeleton-based action recognition. *ArXiv*, abs/2110.13385, 2021.

[37] Chunyu Xie, Ce Li, Baochang Zhang, Chen Chen, Jungong Han, and Jianzhuang Liu. Memory attention networks for skeleton-based action recognition. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 1639–1645, 2018.

[38] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. *ArXiv*, abs/2201.04288, 2022.

[39] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *ArXiv*, abs/1801.07455, 2018.

[40] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE Multimedia - IEEEMM*, 19:4–10, 02 2012.

[41] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.