2024 SAIL Seminar

Conflux LSTMs Network: A Novel Approach for Multi-View Action Recognition

Amin Ullah[a], Khan Muhammad[b], Tanveer Hussain[a], Sung Wook Baik [a,*]

Neurocomputing 2021

SAIL
Senseable Artificial Intelligence Laboratory

순천향대학교 미래융합기술학과

Senseable AI Lab

석사과정 김병훈

0. Reasons for selection

Similarity of datasets

- This paper used a similar dataset to the CRC research

Extensibility

- Achieve extensibility with general action recognition

Previous research

- Under-researched areas(MVAR)

Index
Contents

1 Overview

2 Background

3 Proposed methodology

4 Experimental results and discussion

5 Conclusion and future work

1 Overview

1. Overview

Previous problems

- Most of the existing research on Human action recognition has used a single view approach (SVAR), but the problem is that action models trained on a single view do not perform well on other views
- The multi-view approach (MVAR) was difficult to perform well due to feature variations from different perspectives, the presence of unseen areas (occlusions) in each view, and the use of multiple videos, which made the computation of the model heavy

Contribution

- Select the middle layer (Conv5_4) of the pre-trained VGG19 to effectively extract features from the image
- Fusion structures using separate LSTM models for each view
- Apply inner operations to independent sequence patterns to find correlations

2 Background

2. Background

Define a Task

Human Action Recognition (HAR)

- Methodology for solving the problem of recognizing current actions
  - ✓ Classify whether a person is walking, running, etc.

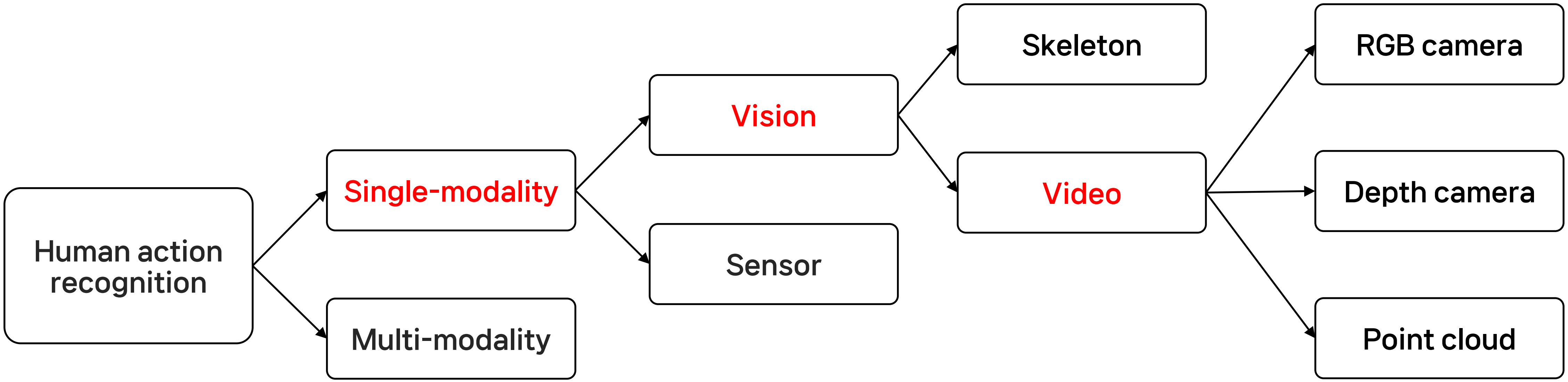
- Learning process
  - ✓ Input: Modality data extracted from actions
  - ✓ Output: Action Recognition result

※ Modality data: Data type (image, sensor, etc.)

2. Background

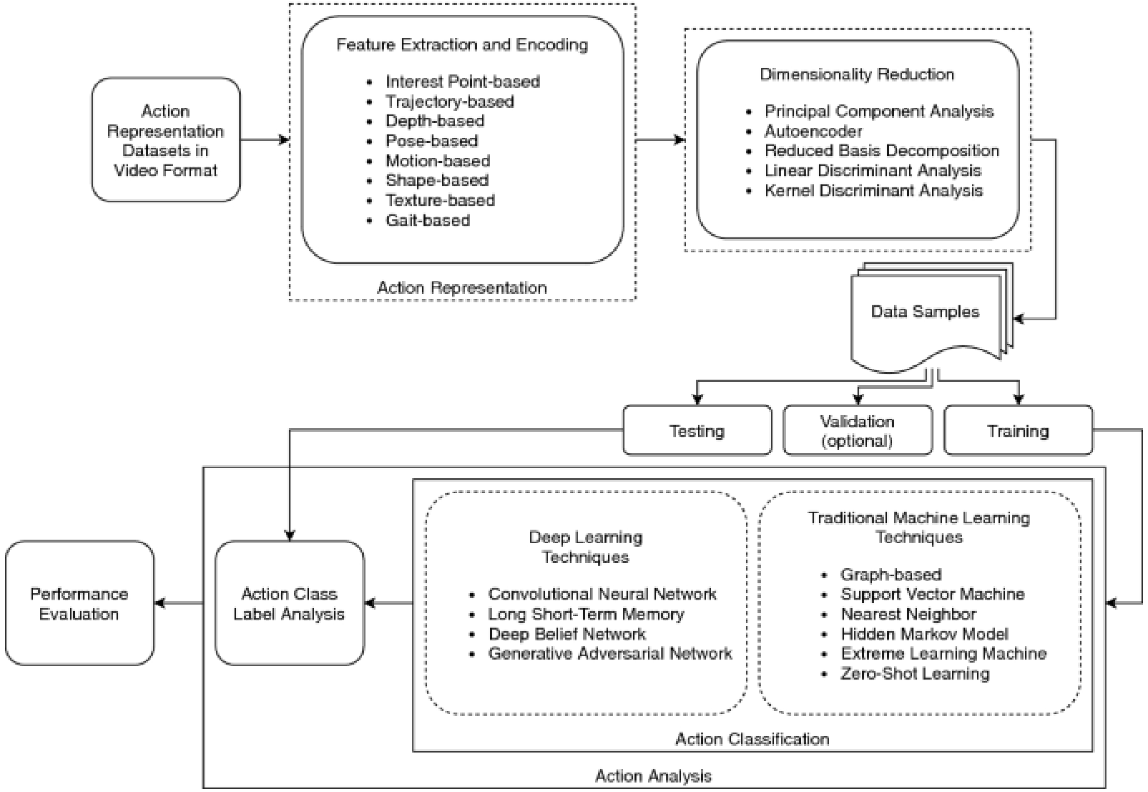
Approach

Human Action Recognition (HAR)



2. Background

Video-based action recognition



Author	Methods	Datasets	Performance (%)
Ijjina et al. [66]	MOCAP, CNN	Berkeley MHAD [67]	99.248
Wang et al. [68]	WHDMM, Deep CNN	MSR-Action3D [43], MSRDailyActivity3D [44], UTKinect-Action [69]	100.00, 85.00, 90.91
Du et al. [70]	RNN, LSTM	MSR-Action3D [43], Berkeley MHAD [67], Motion Capture Dataset HDM05 [71]	94.49, 100.00, 96.92
Zhang et al. [72]	MTRL	SARCO [73]	Mean = 0.5156
Yang et al. [74]	MTL	MSR-Action3D [43], UTKinect-Action [69], Florence3D-Action [75]	95.62, 98.80, 93.42

2. Background

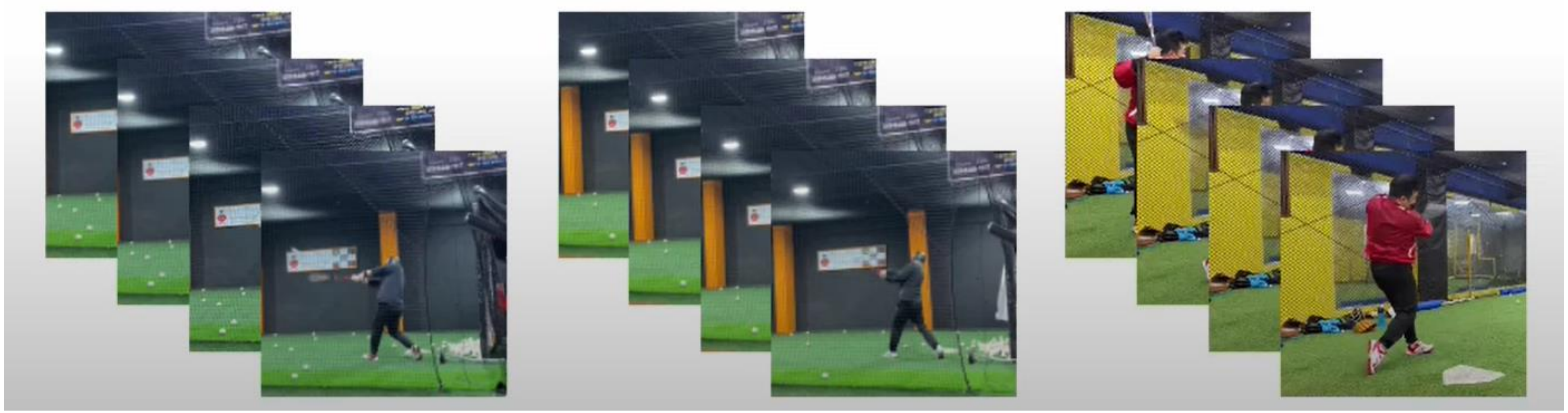
Limitations

Limitations of video-based action recognition

- Performance is highly dependent on camera angle, background and human body size variations

  For this, multiview-based action recognition methods have emerged

  However, when computation overload / parallel processing due to increase in data, convergence needs to be discussed

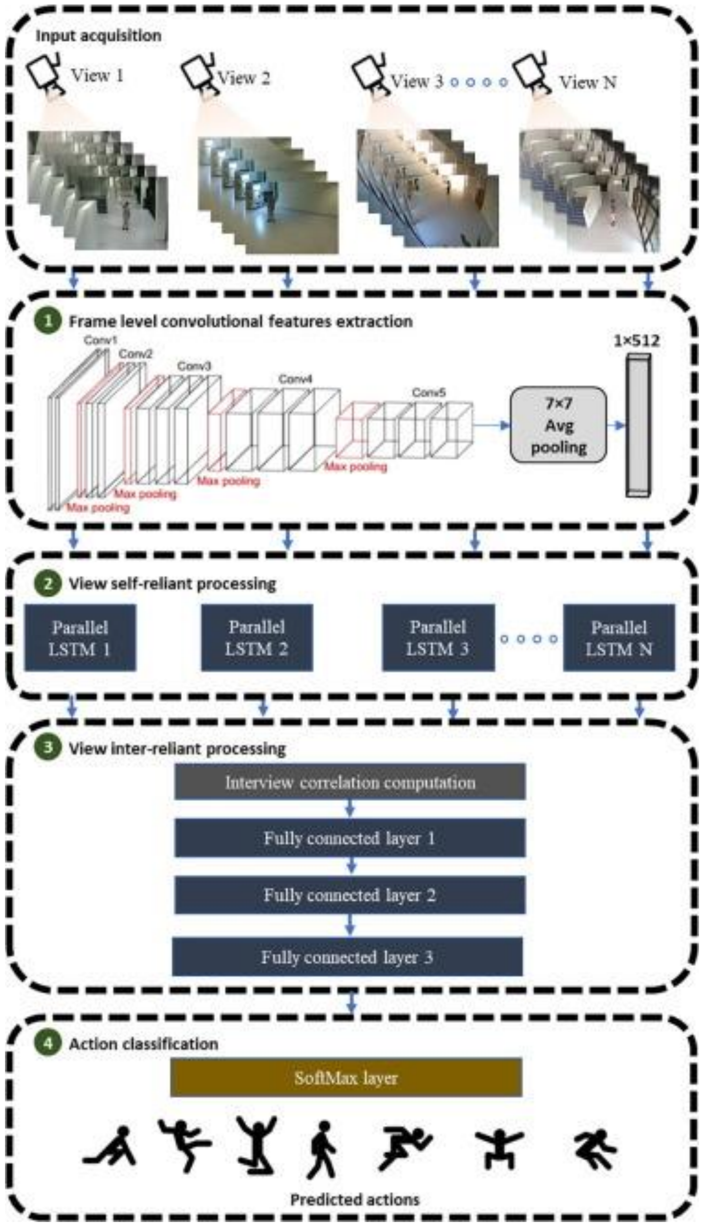


3 Proposed methodology

3. Proposed methodology

Overall architecture

1. Convolutional features extraction for sequence representation
2. View self-reliant network
3. View inter-reliant network
4. Action classification



3. Proposed methodology

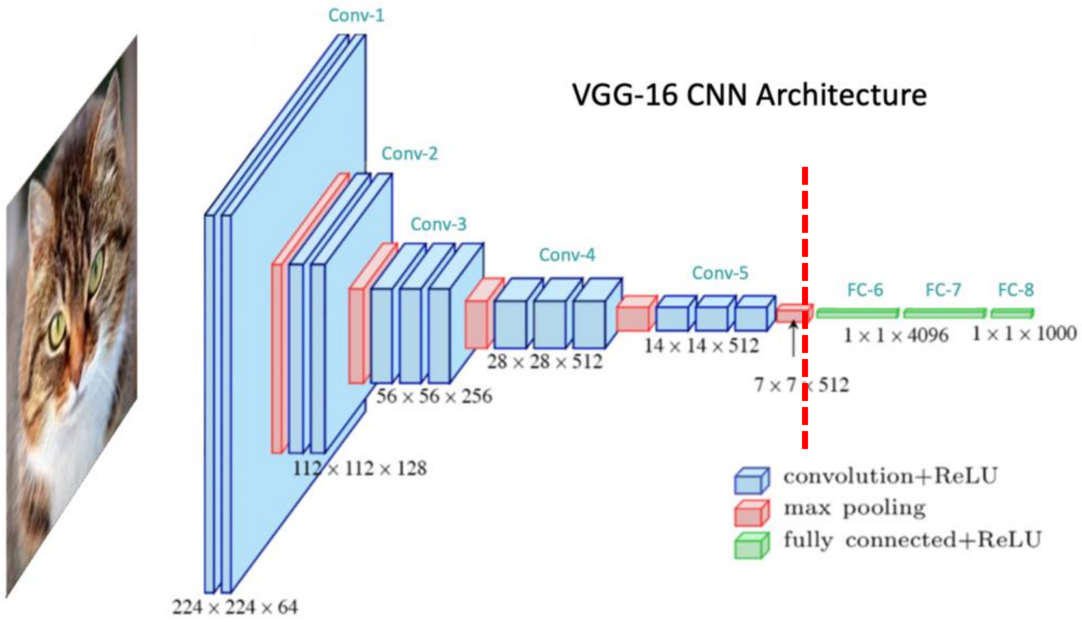
Convolutional features extraction for sequence representation

CNN based feature extraction

- Using the convolutional layer of the VGG-19 model to represent the level of the frame
  - ✓ Global information in a sequence of frames changes slowly, but there is a lot of local movement
  - ✓ FC layer learns more global features
  - ✓ Convolution layer is sensitive to local features

The formula for convolutional feature extraction

$$C_F(K) = \frac{1}{(w,h)} \sum_{i=1}^{w} \sum_{j=1}^{h} FM^K(i,j)$$



3. Proposed methodology

View self-reliant network

Self-reliant network

- Parallel processing of LSTMs to learn behavior patterns

Layer	Dimensions	No. of parameters
Input	$\lVert 15 \quad 512 \rVert \times 3$	–
LSTM (View1)	$\left\lVert \begin{matrix} 512 & 256 \\ 256 & 128 \end{matrix} \right\rVert \times 3$	2,286,276
LSTM (View2)	$\left\lVert \begin{matrix} 512 & 256 \\ 256 & 128 \end{matrix} \right\rVert \times 3$	2,286,276
LSTM (View3)	$\left\lVert \begin{matrix} 512 & 256 \\ 256 & 128 \end{matrix} \right\rVert \times 3$	2,286,276
Correlations	$\lVert 128 \quad 128 \quad 128 \rVert$	–
FC 1	$\lVert 1 \times 128 \rVert$	16,384
FC 2	$\lVert 1 \times 64 \rVert$	4,096
FC 3	$\lVert 1 \times 18 \rVert$	324
SoftMax	$\lVert \textbf{\textit{no of classes}} \rVert$	**6,817,220**

3. Proposed methodology

View inter-reliant network

Inter-reliant network

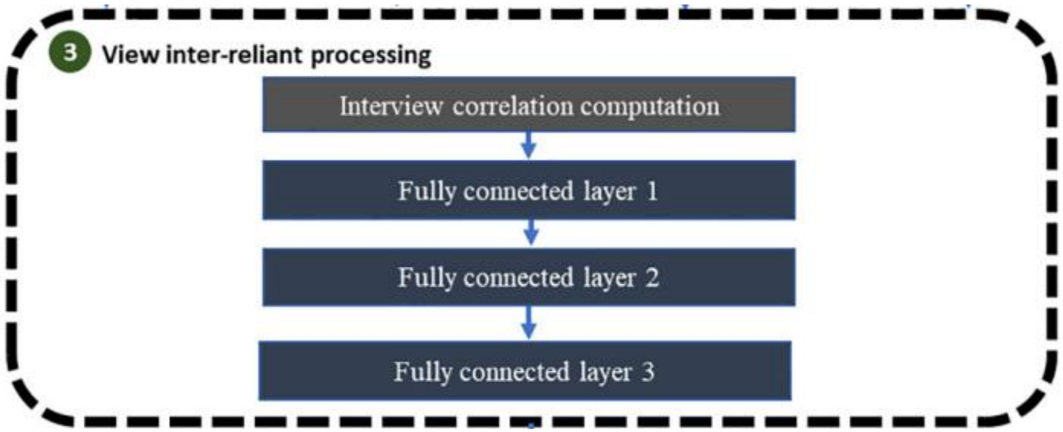
- Additional Inter-reliant networks to extract higher-level features
  - ✓ Perform pairwise operations between sequences
  - ✓ FlowNet: performs correlation analysis between two sequential feature maps when generating optical flows

Algorithm 1: Conflux LSTMs Network

Input: Multi-view video streams $\{V_1, V_2, V_3, V_4, \ldots, V_n\}$
Output: Predicted action class along with the probability score
Preparation:
1. Acquire synchronized multi-view frames
2. Load pretrained VGG19 CNN model M_1
3. Initialize trained Conflux LSTMs network M_2

Steps:
while (video frames $(\{V_1, V_2, V_3, V_4, \ldots, V_n\})$)
1. Read frames $\leftarrow \{f_i \in V_j\}s$
2. Forward $f_i \in V_n$ frames to M_1
3. $FM \leftarrow M_1$
4. Apply 7×7 average pooling to FM^K for frame level visual features extraction using Eq. (1).
5. Repeat step 2, 3, and 4 for sequence of frames of V_n.***note:** sequence length in our experiments is 15
6. Combine frame level sequential features F_v from $\{V_1, V_2, V_3, V_4, \ldots, V_n\}$
7. Labeled action class $\leftarrow$ Forward propagate F_v to M_2
8. Show predicted action along with probability score

end while



4 Experimental results and discussion

4. Experimental results and discussion

Datasets and Experimental environments

Datasets

- MCAD : 5 cameras / 18 action class
- Northwestern-UCLA multi-view action 3D dataset: 3 Kinect cameras / 10 action class

Environments

- OS : Ubuntu-16.04
- CPU : Intel Core i5-6600
- GPU : GeForce TITAN X
- Other : Python 3.5 / Tensorflow-1.12

Experimental design

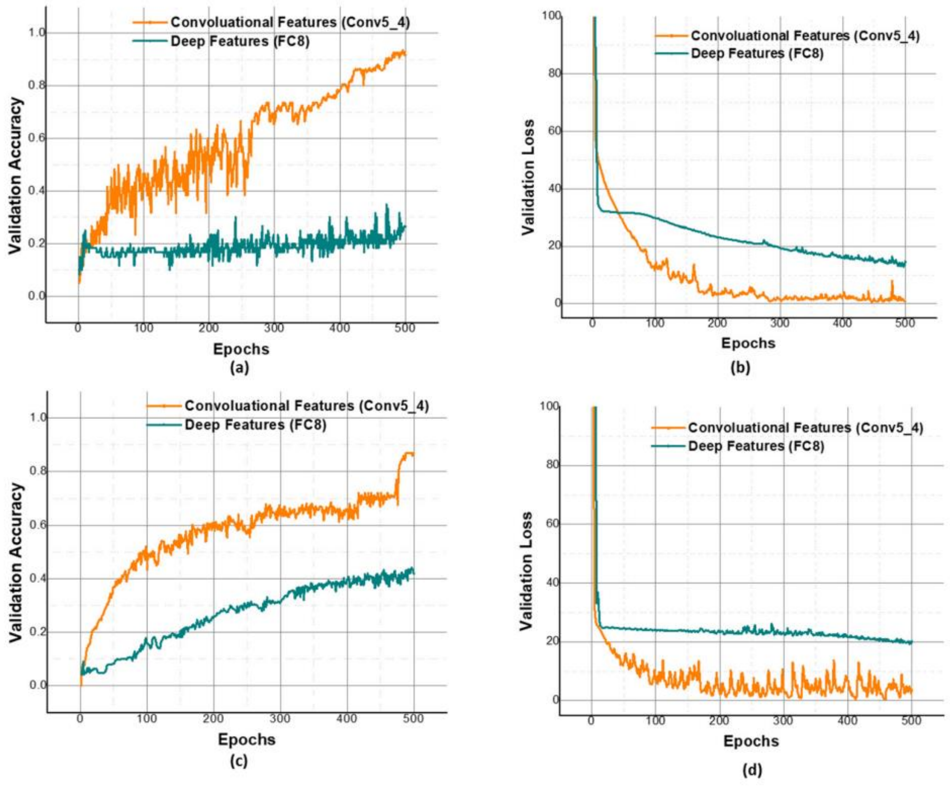
1. Comparing feature extraction methods
2. Results from open and closed test sets
3. Comparison with SOTA research

4. Experimental results and discussion

Parameters selection for conflux LSTMs network

Comparing feature extraction methods

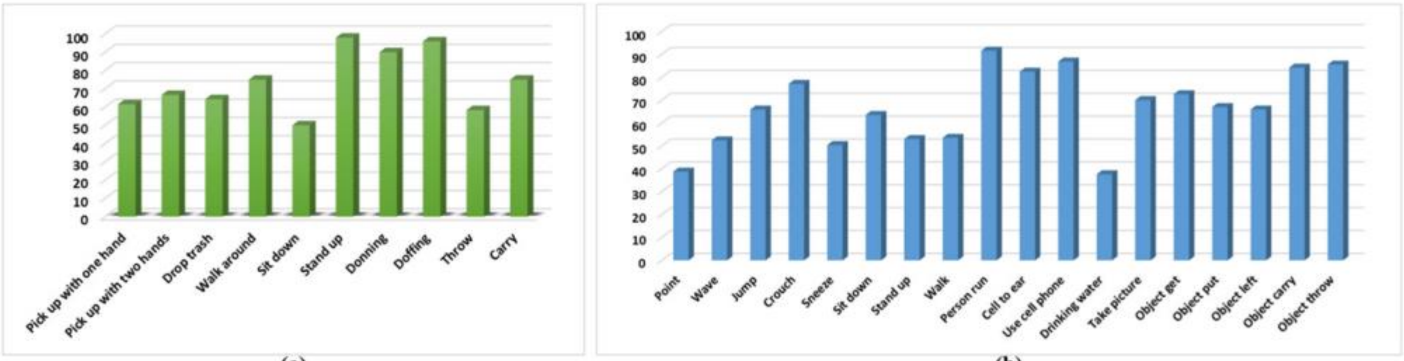
- Feature extraction with convolutional layer vs. FC layer



Feature extraction with a convolutional layer performs significantly better

4. Experimental results and discussion

Closed set, open set, and class-wise evaluation



(a) (b)

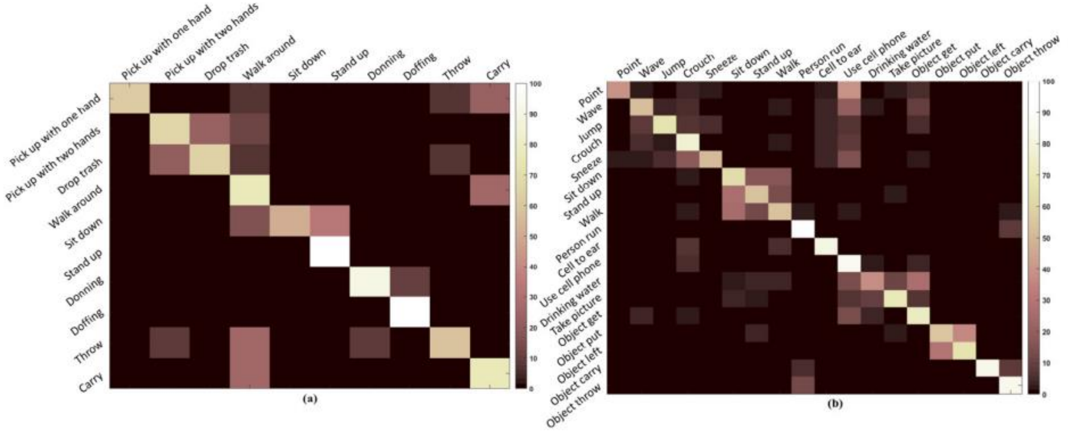


Fig. 4. The confusion matrix for the closed test set of (a) the northwestern-UCLA dataset and (b) MCAD dataset. The bar line displays the accuracy range from 0 to 100 where the classes that achieved a brighter color on its diagonal has better results, and the ones that are closer to a dark color are confused with other classes.

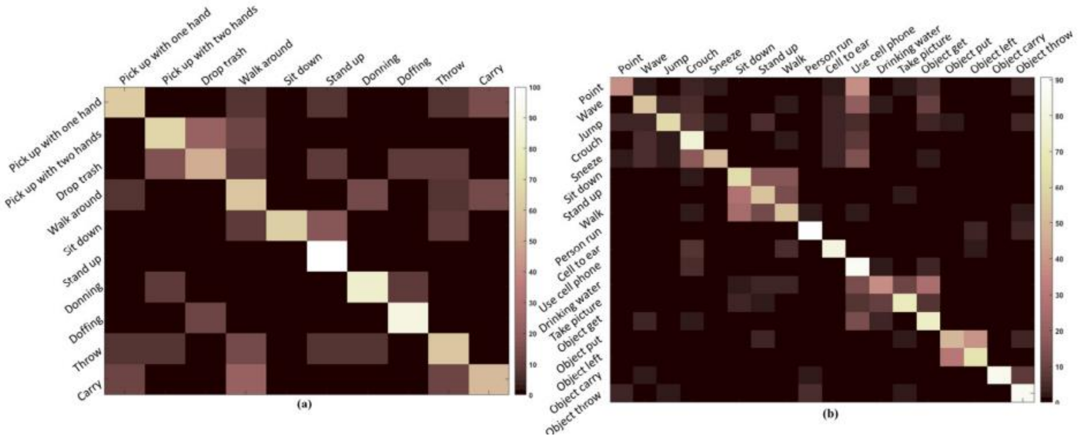


Fig. 5. The confusion matrix for open test set of (a) the northwestern-UCLA dataset and (b) the MCAD dataset

4. Experimental results and discussion

Comparison with the state-of-the-art

Table 2
Comparison of the proposed conflux LSTMs on the northwestern-UCLA multi-view action dataset via different view (V) settings with depth, pose, and RGB based methods, respectively.

Data	Methods	Train V_1 & V_2	Test V_3	Train V_1 & V_3	Test V_2	Train V_2 & V_3	Test V_1	Average
Depth	Virtual views [29]	58.5		55.2		39.3		51.0
	Virtual path [30]	60.6		55.8		39.5		52.0
	3D viewpoints [31]	91.9		75.2		71.9		79.7
Pose	Hierarchical RNN [32]	78.5		–		–		–
	View invariant HAR [33]	86.1		–s		–		–
	Temporal sliding LSTM [34]	89.2		–		–		–
RGB	3D pose motion [35]	68.6		68.3		52.1		63.0
	Knowledge transfer model [36]	75.8		73.3		59.1		69.4
	Glimpse global model [10]	85.6		84.7		79.2		83.2
	Glimpse clouds [10]	90.1		89.5		83.4		87.6
	Conflux LSTMs network	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	
		V_1 & V_2	**V_2 & V_3**	**V_1 & V_3**	**V_2 & V_3**	**V_2 & V_3**	**V_1 &V_3**	
		85.7		**92.5**		**88.6**		**88.9**

Table 3
Comparison with state-of-the-art methods using the overall recognition accuracy of the northwestern-UCLA multiview-3D dataset.

Method	Accuracy (%)
MST-AOG w/o Low-S [28]	65.3
MST-AOG w Low-S [28]	73.3
HOPC [37]	80.0
Multi-view dynamic images + CNN [13]	84.2
Conflux LSTMs network	**88.9**

Table 4
Comparison with state-of-the-art methods using the overall recognition accuracy of the MCAD dataset.

Method	Accuracy (%)
IDT [38]	84.2
Covariance matrices [39]	64.3
STIP [27]	81.7
Cuboids [27]	56.8
Conflux LSTMs network	**86.9**

5 Conclusion and future work

5. Conclusion and future work

Conclusion

1. In this paper, a Conflux LSTM network is proposed to solve the MVAR problem

2. Compared with recent SOTA techniques, it has better performance

3. However, multi-view data has high dimensionality and requires a lot of computation

4. In future work, the authors plan to lightweight the feature extraction model for embedded programming

5. Furthermore, they want to combine vision and sensor data for multimodal learning.

5. Conclusion and future work

How to apply?

1. 저자가 후속 연구로 제안한 부분을 적용하면 좋을 거 같음

2. Skeleton 그래프를 데이터로 활용하면 더 좋은 결과가 나오지 않을까?

   SGC 모델로 똑같이 컨볼루션 레이어에서 특징 추출하고, 똑같은 구조로 실행..

3. 논리구조를 잘 가져가고 이해하기 쉽도록 간결하게 적혀 있음 -> 논문 포맷을 차용 하면 좋을 거 같음

4. 이 주제도 좋으나 다른 접근법에 대해 계속 찾아봐야겠음