Robustness analysis of 3D convolutional neural network for human hand gesture recognition

Dang-Manh Truong[1], **Huong-Giang Doan[2]**, Thanh-Hai Tran[1], Hai Vu[1] and Thi-Lan Le[1]

[1]Computer Vision Department, MICA, HUST

[2]Faculty of Control and Automation, Electrical Power University

International Research Institute MICA

Multimedia, Information, Communication & Applications

UMI 2954

Hanoi University of Science and Technology

1 Dai Co Viet - Hanoi - Vietnam

Outline

- Context
- Related Works
- Proposed framework
- Experimental results
- Conclusion and Discussion

Context

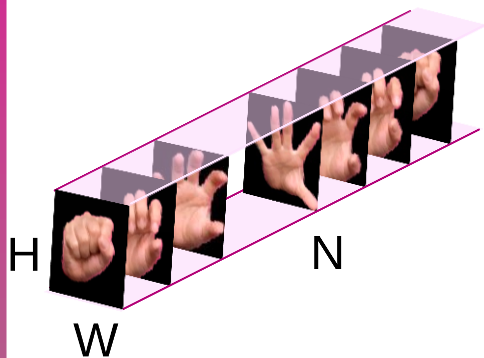
- **Controlling devices in different view points**

→Very challenge to obtain a invariant dynamic hand gesture recognition of hands on multi-views:

  - Complex background
  - Non-robust with variable view-points
  - Different, non-rigid and small hand shapes
  - Require some constraints of datasets

- **Current hand posture/gesture recognition approaches:**
  - Hand-crafted feature extraction: preferred on small datasets with some specific characteristics
  - Deep learning methods: robust with large dataset

Related works



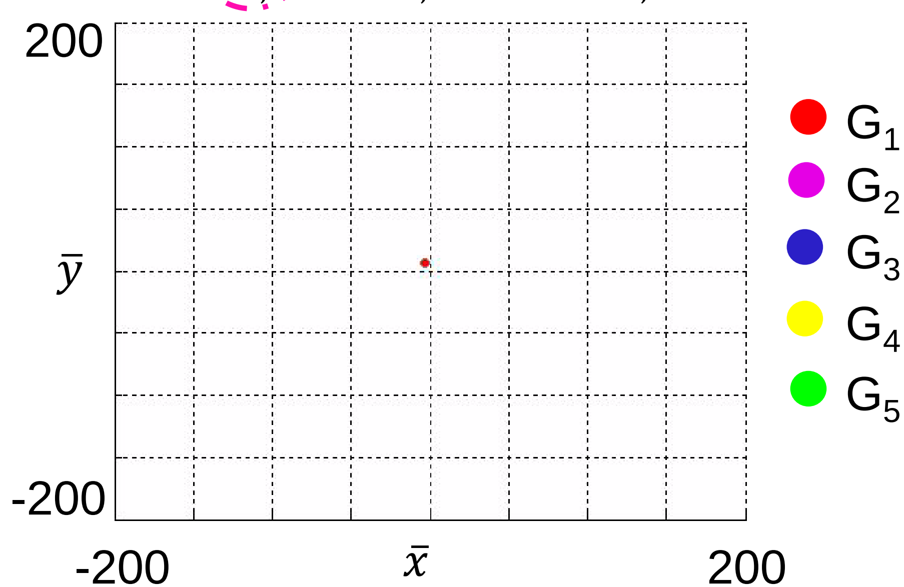
WxHxN
feature space
100x100xN =
10000xN

- **Hand-crafted features extraction [HuongGiangDoan2017]:**
  - **Preferred on small datasets with some specific characteristics**

Temporal T_1^{Tr}

$$G_N^{TS} = \left\{ \begin{pmatrix} \overline{x_1} \\ \overline{y_1} \\ Y_{1,1} \\ Y_{1,2} \\ Y_{1,3} \end{pmatrix}, \begin{pmatrix} \overline{x_2} \\ \overline{y_2} \\ Y_{2,1} \\ Y_{2,2} \\ Y_{2,3} \end{pmatrix}, \dots, \begin{pmatrix} \overline{x_N} \\ \overline{y_N} \\ Y_{N,1} \\ Y_{N,2} \\ Y_{N,3} \end{pmatrix} \right\}$$

Spatial Y_1^S



$\mathcal{R}^{10000xN} \Rightarrow \mathcal{R}^{5xN}$

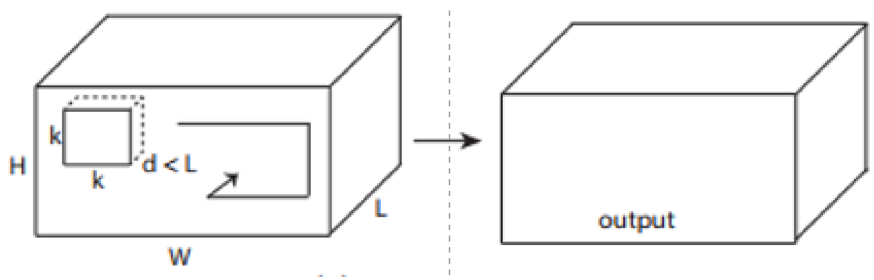
MICA 2018

Huong Giang Doan, Van Toi Nguyen, Hai Vu, and Thanh Hai Tran. *A combination of user-guide scheme and kernel descriptor on rgb-d data for robust and realtime hand posture recognition*, Journal of EAAI, vol. 49, no. C, pp. 103-113, 2016.

4

Related works

- **C3D feature extraction [Tran2015]**: 3D convolution kernels can exploit temporal pattern besides spatial information, while eliminating the need for secondary temporal modeling techniques
  - **C3D net:** uses 3D convolution

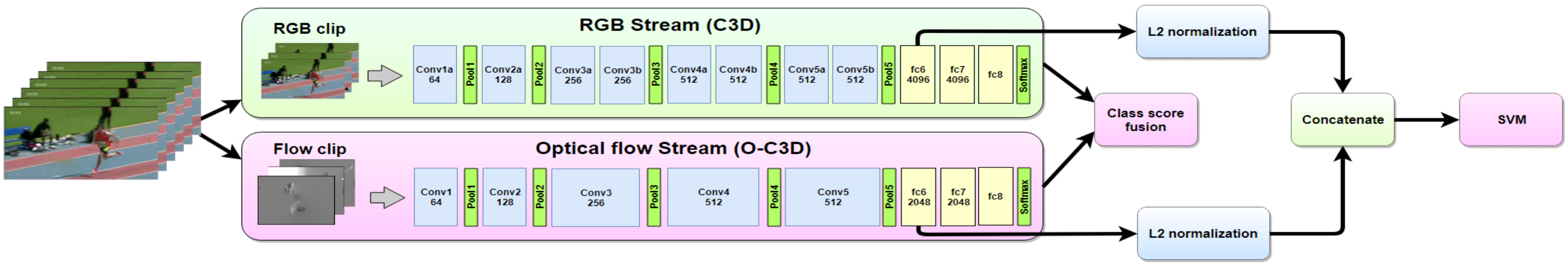


  - **Input**: a 16-frame clip, each frame of size 128x171

D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, ICCV, 2015, pp. 4489–4497.

Related works

- **Two stream C3D [Khong2018]:**
  - C3D uses only **RGB data**
  - **Optical Flow** or other stream could be additional information for recognition => two streams C3D (RGB+Optical Flow)
- **Two stream C3D architecture:**



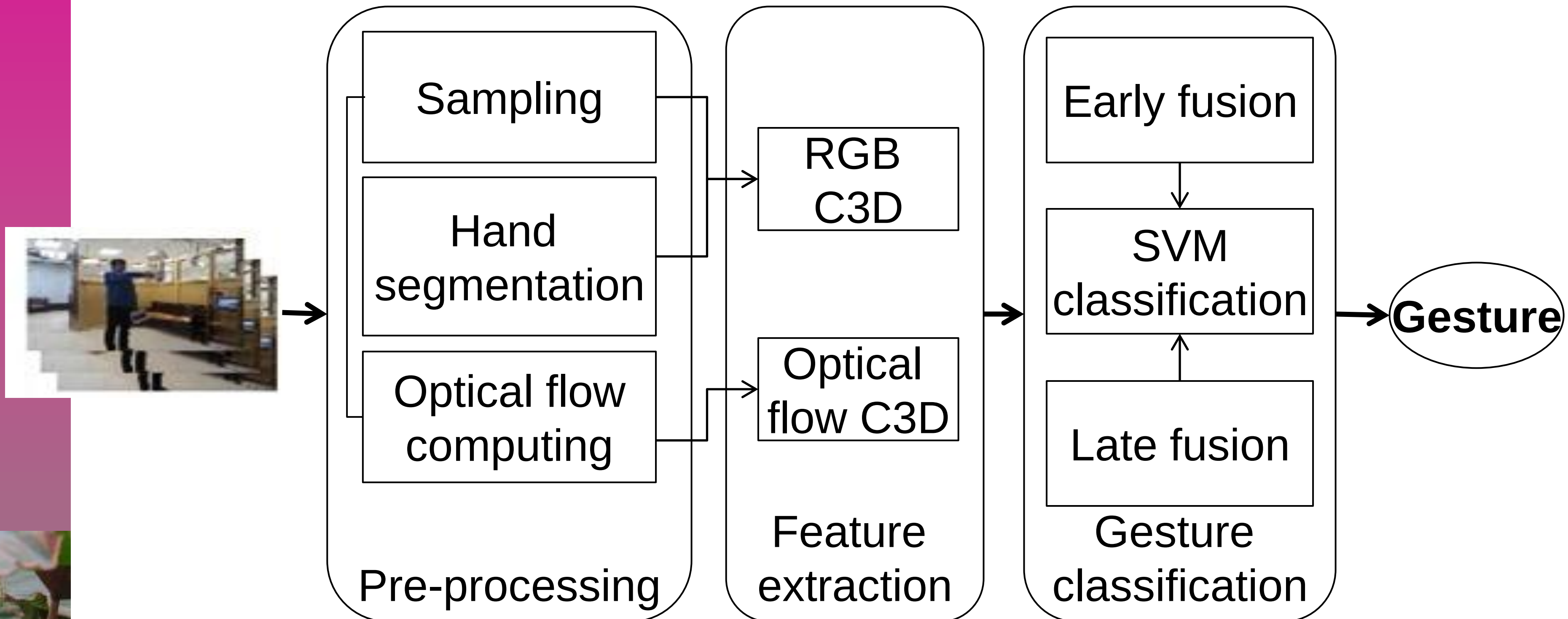
nated

Both [Tran2015] and [Khong2017] work only on human action **recognition** (UCF101, HMBD51)

... no evaluation of C3D to hand gestures under different viewpoints

V. Khong, T. Tran, Improving human action recognition with two-stream 3d convolutional neural network, MAPR, 2018, pp. 1–6.

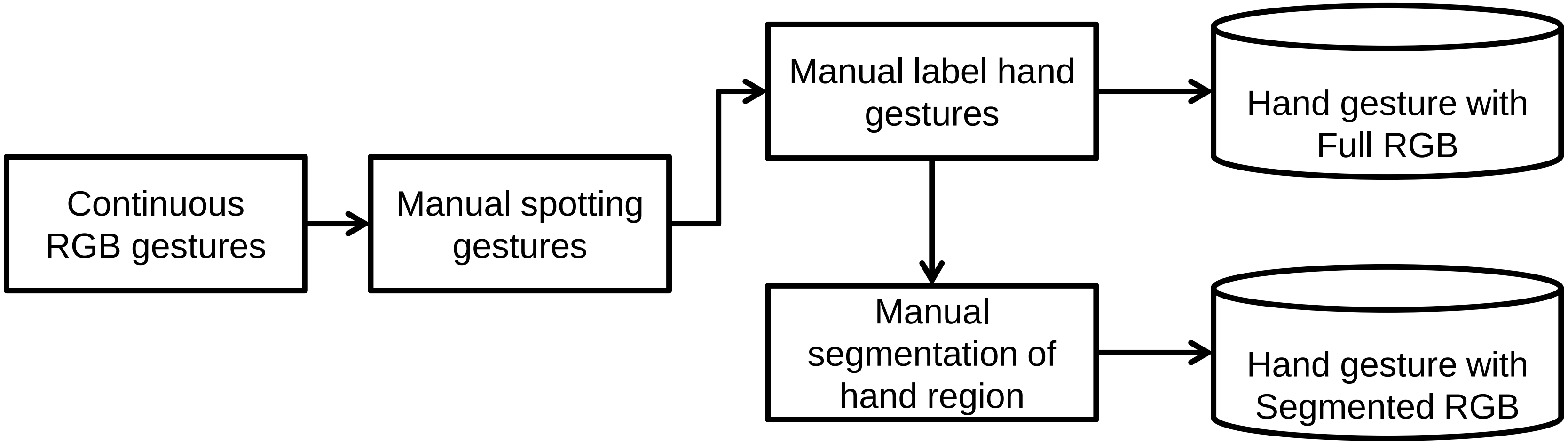
Proposed framework



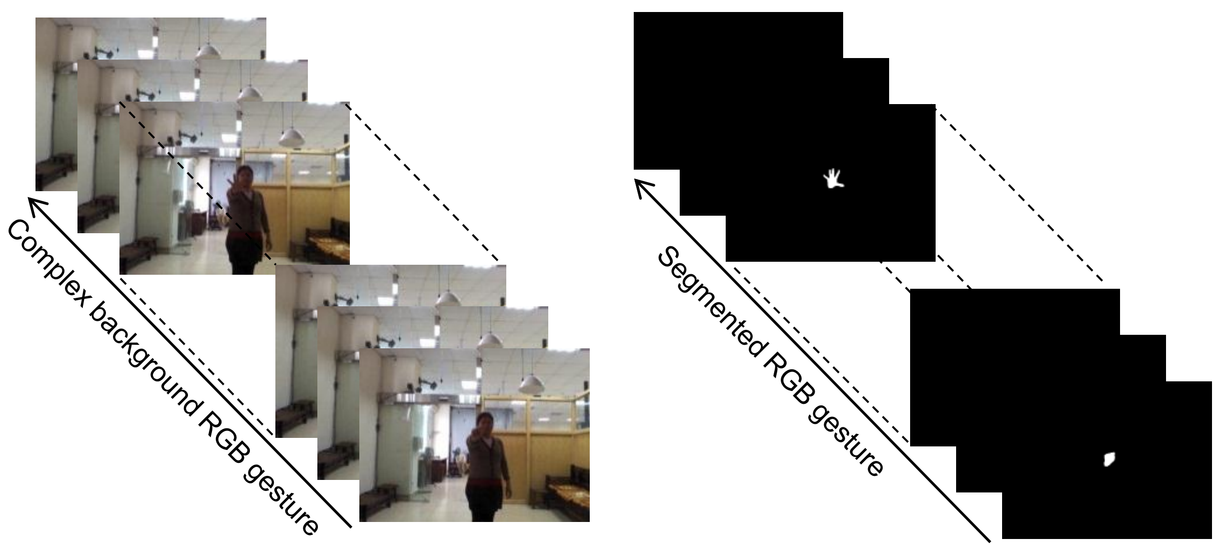
(1) Would C3D, a deep architecture tested on general action recognition datasets, still be suitable for hand gesture recognition where hand has a relatively low spatial resolution and it is the only moving object in the scene?

(2) The original C3D network has been trained and tested on human action datasets without considering the impact of viewpoints.

Hand segmentation



Pre-processing dynamic hand gesture



(a) (b)

Sampling method

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Duplicate the last frame								
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	23	23	23	23	23	23	23	23	23
The first 16 frame clip																The second 16 frame clip															

Sampling method used by [Khong2017]

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Take previous overlapped frames								
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
The first 16 frame clip																The second 16 frame clip															

The proposed sampling method

Computing and stacking optical flow

- **Manually spotting of gestures from stream**
- **Computing Optical Flow stream based on [1]**
  - **Optical flow:** characterize movement of pixels between consecutive images

$$I_\tau(\mathrm{u},\mathrm{v},2\mathrm{k}-1) = \mathrm{d}^x_{\tau+k-1}(u,v) \qquad I_\tau(\mathrm{u},\mathrm{v},2\mathrm{k}) = \mathrm{d}^y_{\tau+k-1}(u,v)$$

Two **consecutive** frames and two optical flows in **vertical** and **horizontal** dimensions

- **Stack Optical Flow as a 3D volume (d_x, d_y, 0)**

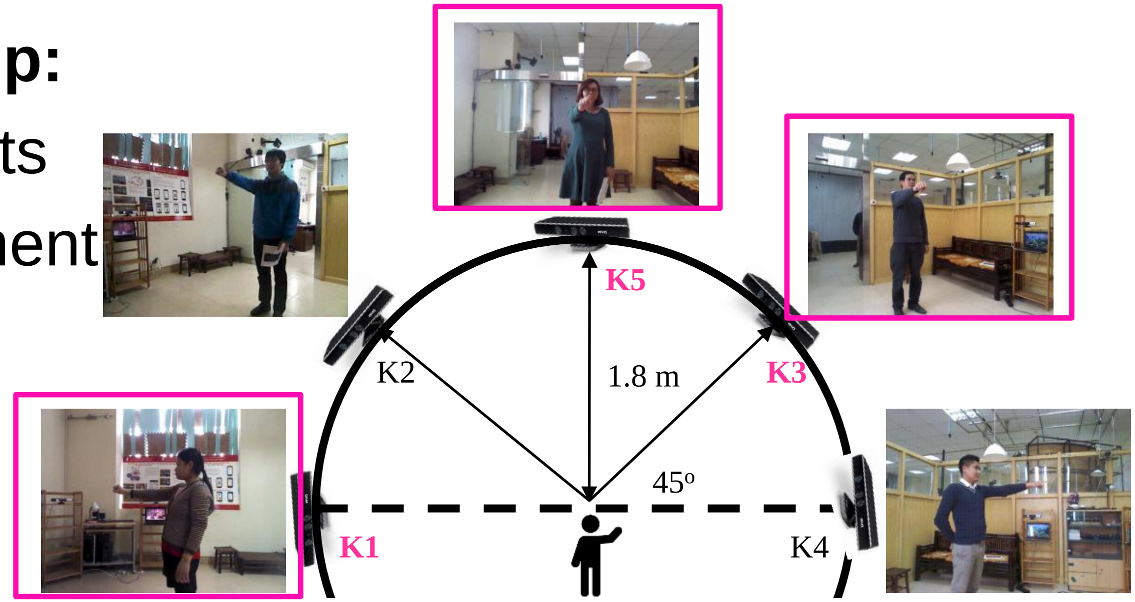
[1] http://www.robots.ox.ac.uk/~vgg/software/two_stream_action/

Experimental results

- **Sub-dataset:**
  - RGB images from Kinect sensor (640x480)
  - Data (5 subjects, 5 gestures on 5 Kinects) is separated following one-leave-out method:
    - 4 subjects for training, 1 subject for testing
- **Implementation on:**
  - Mask R-CNN: Open source Github
  - GPU GTX 1080Ti (Vram 12Gb)
- **Evaluations:**
  - Transfer learning on hand gesture dataset
  - Different view points evaluation
  - Late and Early fusion strategies

Dataset

- **Environment setup:**
  - Five fixed Kinects
  - Indoor environment

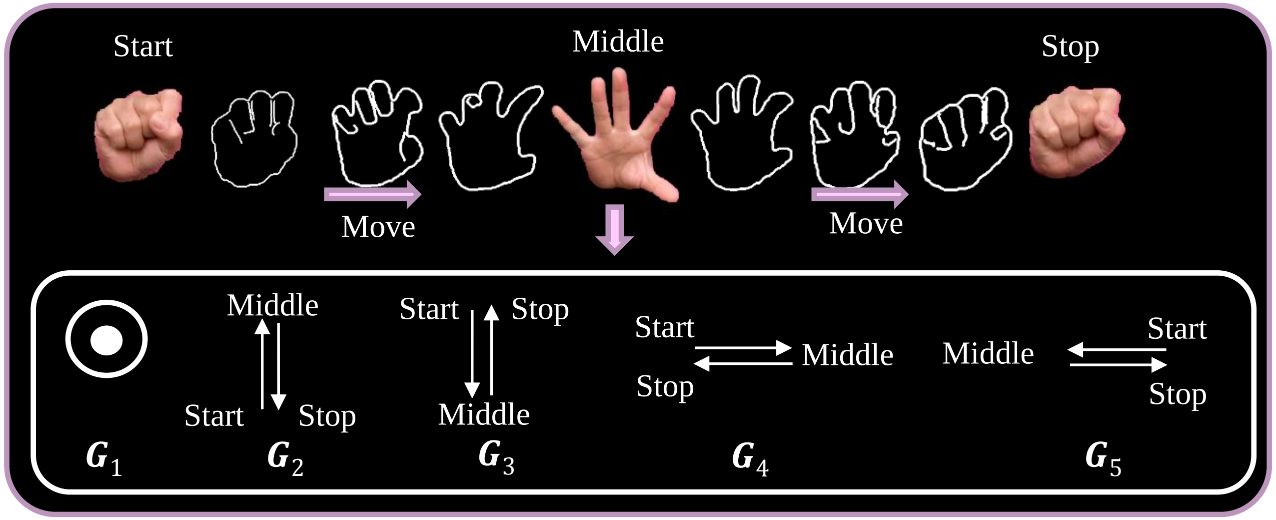


- **Captured database:**
  - Showroom – MICA Institute - HUST
  - The defined 12 hand gestures
  - Each gesture implements in 3 – 5 times
  - 20 people: 13 males, 8 females

05 gestures
5 subjects
5 views

Sub-Dataset

- **5 gestures**
- **5 subjects**
- **5 views**



View\Gesture	G_1	G_2	G_3	G_4	G_5
K1	26	22	33	26	23
K3	26	22	33	26	23
K5	26	22	33	26	23

Evaluation procedure

Leave-one-out cross validation and cross-view[Doan2017]

5 subjects (P1,..P5)
3 views
(K1, K3, K5)
5 gestures

Cross-view:

- Train: View K_i, subjects $\{P1,..P5\}/\{P_j\}$
- Test: View K_k, subjects P_j

Leave-one-out:

- Train: View K_i, subjects $\{P1,..P5\}/P_j$
- Test: View K_i, subjects P_j

Evaluation metric: $$Accuracy = \frac{\sum corrects}{\sum total}$$

Doan H.-G., Vu H., and Tran T.-H. (2017). Dynamic hand gesture recognition from cyclical hand pattern. 97–100

Transfer learning on hand gesture dataset

- **Fine tuning on RGB stream**
  - Initialize RGB-C3D with C3D model pre-trained on Sport1M
  - Fine tune
    - **Several layers** of RGB-C3D using gesture dataset
    - **All layers of RGB-C3D** using gesture dataset

Kinect 3

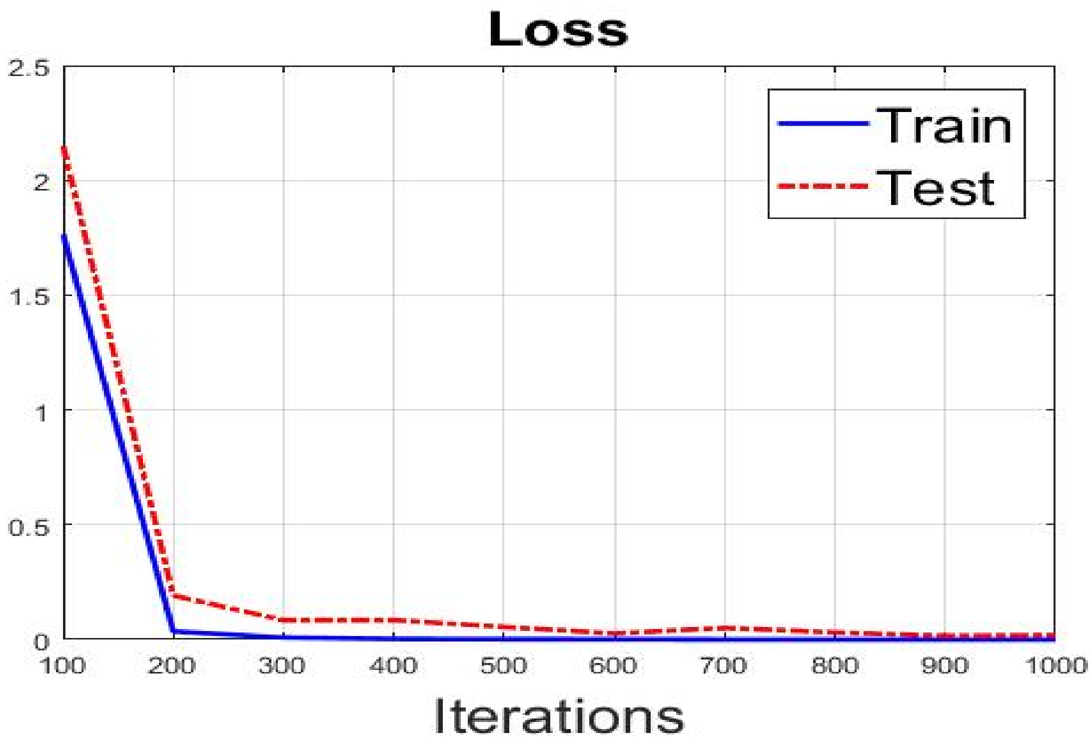
FCs only	All layers
64.10%	94.00%

Use fine-tuning of all layers of RGB-C3D for evaluation

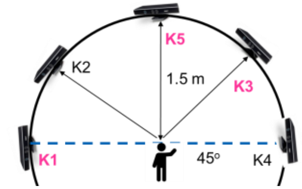
Transfer learning on hand gesture dataset

Fine tuning two streams C3D

- **Fine tuning on RGB stream (all layers)**



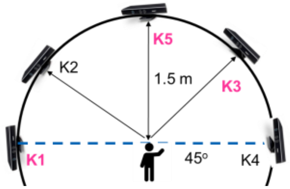
Different viewpoint &modalities



Train	RGB			Segmented RGB			OF		
Test	K1	K3	K5	K1	K3	K5	K1	K3	K5
K1	**76.27**	45.60	50.07	**89.97**	65.37	54.09	**70.98**	35.45	39.31
K3	47.76	**93.60**	76.04	50.67	**99.38**	89.84	47.63	**95.68**	71.51
K5	30.41	65.77	**96.67**	42.49	93.29	**99.05**	38.49	89.47	**93.28**
Avr	64.68			**76.01**			64.64		

- **Single view (K3, K5) is good results. K1 gives the worst result:**
  - Hands are occluded
  - Or out of camera field of view
  - Movement of the hand is not discriminative
- **Background has strong impact on classification result.**
- **Optical Flow gives competitive performance with RGB**

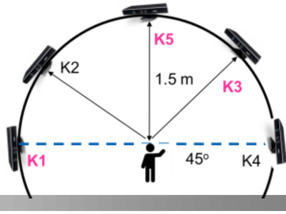
Late and Early fusion performance



Train	OF-RGB early			OF-RGB late		
Test	K1	K3	K5	K1	K3	K5
K1	**75.67**	55.30	47.53	**74.10**	52.07	45.61
K3	47.72	**94.43**	81.12	51.59	**94.36**	80.07
K5	36.70	68.17	**100.0**	41.55	70.83	**100.0**
Avr		**67.40**			**67.79**	

- **Combined RGB and OF can boost performance**
- **Could use early or late fusion strategy**

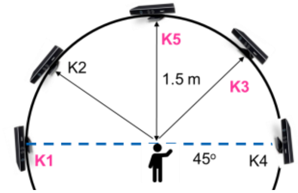
Sampling strategies



Train	16 frame clips [Tran2015]			16 randomly selected [4]			16 key frames [Khong2017]		
Test	K1	K3	K5	K1	K3	K5	K1	K3	K5
K1	**76.27**	45.60	50.07	**75.59**	56.60	41.97	**71.15**	51.79	33.61
K3	47.76	**93.60**	76.04	48.78	**95.34**	77.45	47.98	**95.34**	78.79
K5	30.41	65.77	**96.67**	36.15	56.25	**97.33**	38.31	61.51	**96.67**
Avr	64.68			**65.05**			63.90		

16 randomly selected gives the best results while taking the smallest computational time

Example results (1)



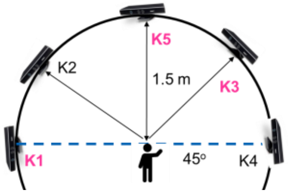
Investigation of the effects of complex background on recognition accuracy

- **Comparison with hand-segmented RGB frames**

If there is a drastic increase in accuracy, we can conclude that the environment does have an effect on recognition accuracy

Use a separate phase to extract regions of interest before recognition

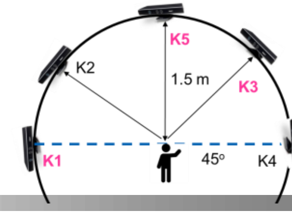
Example results(2)



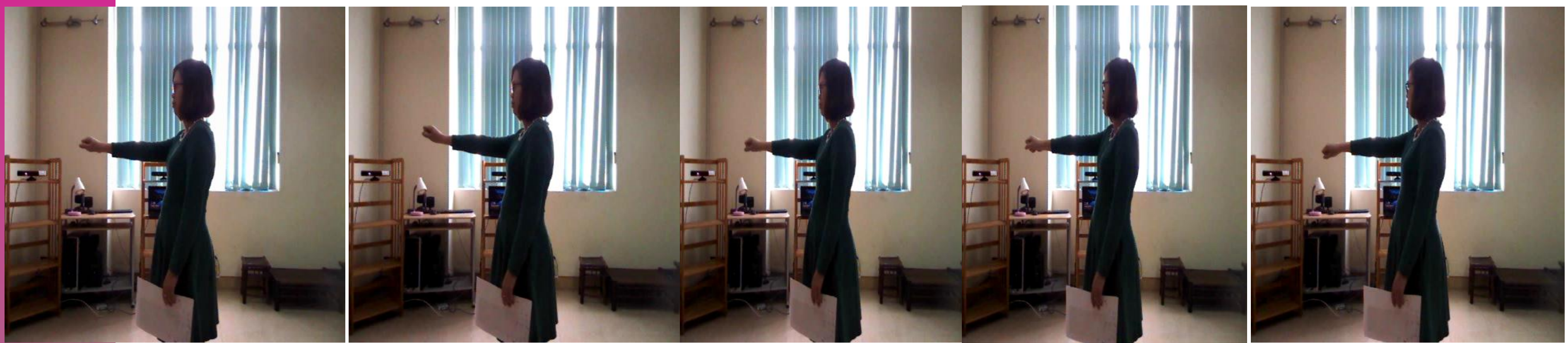
Ground truth: DOWN   Prediction: ON_OFF

This is because the hand is opened before it goes down. Therefore C3D mistakes this with ON_OFF

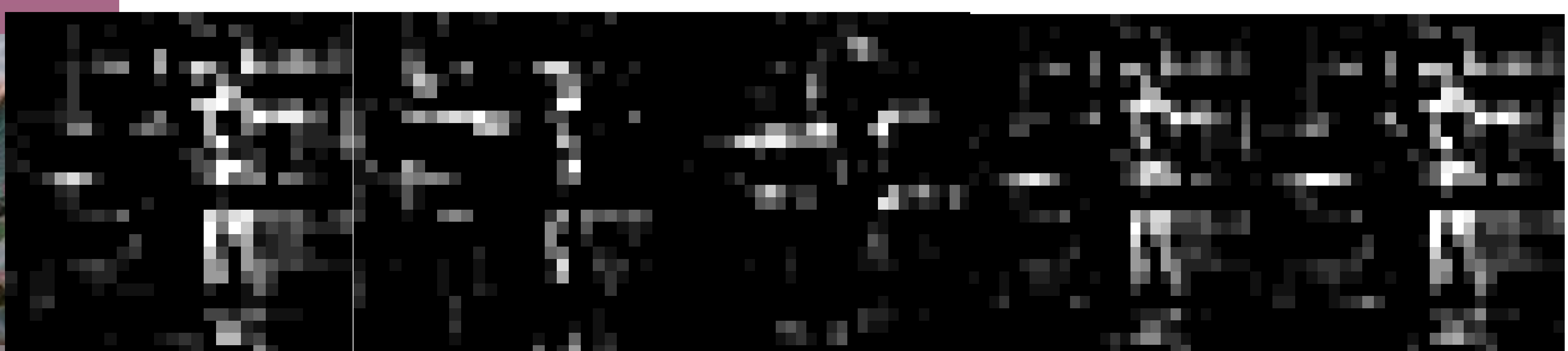
Experimental results (3)



Input

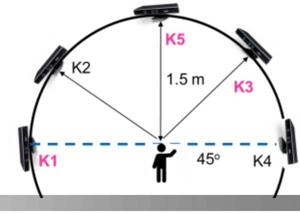


Conv3 layer



It is hard to discriminate between ON_OFF, LEFT, and RIGHT in the Conv3 layer. This only happens in K1 view

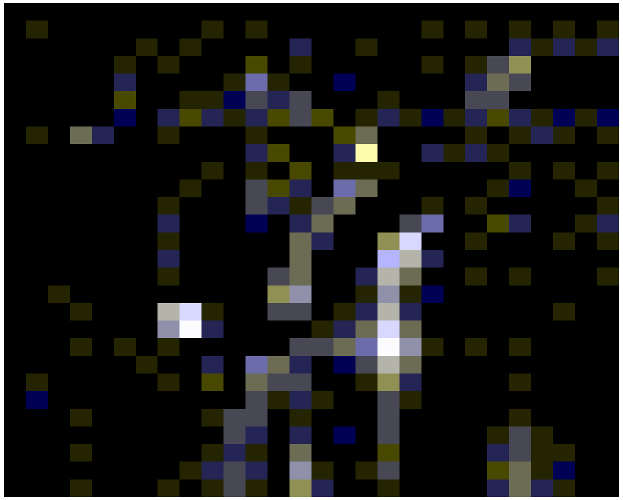
Experimental results (4)



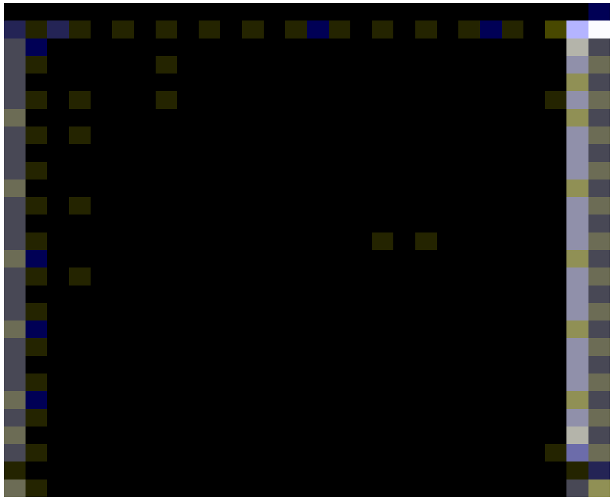
RGB

Optical flow

Input



Conv3 layer



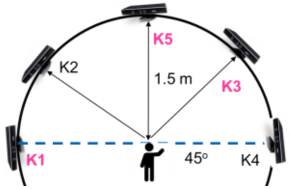
Here in this example incorporating optical flow Information helps us better recognize the action in the conv3 layer

Conclusions and Discussions

- **Conclusions**
  - The performance of C3D remains stable under a small change of viewpoint (**<= 45 degrees**)
  - Background has strong impact on classification performance (**increased 11.32%)**
  - Incorporating Optical Flow (OF) channel in a two streams C3D gives improved results (**increased 3.11%**)
- **Future works**
  - Evaluate on remaining views (K2, K4)
  - Comparison with existing method [Doan2017] (manifold based learning)
  - Testing with automatic segmented hand regions.
  - Adapting the C3D to be more robust to viewpoint change

THANKS FOR YOUR ATTENTION

Experimental results



Train	RGB			Segmented RGB			OF			OF-RGB early			OF-RGB Late		
Test	K1	K3	K5	K1	K3	K5	K1	K3	K5	K1	K3	K5	K1	K3	K5
K1	**76.27**	45.60	50.07	**89.97**	65.37	54.09	**70.98**	35.45	39.31	**75.67**	55.30	47.53	**74.10**	52.07	45.61
K3	47.76	**93.60**	76.04	50.67	**99.38**	89.84	47.63	**95.68**	71.51	47.72	**94.43**	81.12	51.59	**94.36**	80.07
K5	30.41	65.77	**96.67**	42.49	93.29	**99.05**	38.49	89.47	**93.28**	36.70	68.17	**100.0**	41.55	70.83	**100.0**
Avr	64.68			**76.01**			64.64			**67.40**			**67.79**		

- **Single view (K3, K5) is good results. K1 gives the worst result:**
  - Hands are occluded
  - Or out of camera field of view
  - Movement of the hand is not discriminative
- **Background has strong impact on classification result.**
- **Optical Flow gives competitive performance with RGB**
- **Combined RGB and OF can boost performance**