A Transformer-Based Network for Dynamic Hand Gesture Recognition

Andrea D'Eusanio [1], Alessandro Simoni [1], Stefano Pini [1], Guido Borghi [2], Roberto Vezzani [1], Rita Cucchiara [1]

[1] Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia (Italy)

[2] Department of Computer Science and Engineering, University of Bologna (Italy)

INTRODUCTION

Transformer-based neural networks:

- Use a successful self-attention mechanism that achieves state-of-the-art results in language understanding and sequence modeling.
- Are rarely applied to visual data and, in particular, to the **dynamic hand gesture recognition** task.

In this paper:

- We propose a **transformer-based architecture** for the dynamic hand gesture recognition task.
- We show that our method achieves state-of-the-art results using **a single active depth sensor**, using the provided **depth maps** and the **surface normals** estimated from them.
- We test the method with other data types usually provided by RGB-D devices, such as infrared and color data.
- We test the framework on **two automotive datasets**, namely NVidia Dynamic Hand Gestures and Briareo.

GOAL

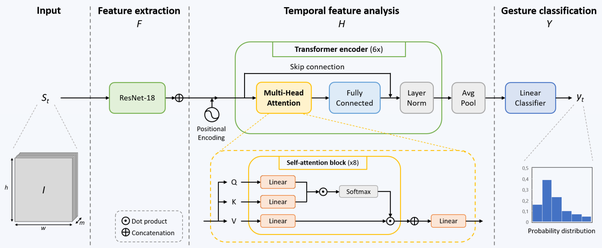
Our goal is the development of a *Natural User Interface* system for the **human-car interaction**:

- The system must work even in presence of **sudden light changes**, as often occurs during the car driving, due to low-light or bad weather conditions.
- **Inexpensive and compact cameras**, which can be easily integrated in the car cockpit, are an optimal choice in order to avoid obstructions to the driver's movements or gaze.
- Through NUIs, the interaction with the infotainment system of a car can significantly **reduce** the driver's manual and visual **distraction**.

DATA TYPES

RGB | Infrared | Depth Map | Optical Flow | **Surface Normals**

PROPOSED METHOD



- **Input**: a set of *m*=40 frames of a given modality (RGB, IR, depth, or surface normals)
- **Feature extraction**: a pretrained *ResNet-18* is used to extract features at frame level
- **Feature concatenation**: visual features extracted from the 40 frames are concatenated together
- **Positional Encoding**[1,2]: incorporates the temporal information regarding the sequential order of the frames
- **Transformer Encoder**: analyzes the 40-frame feature block and outputs a joint representation for each frame
- **Pooling**: combines the joint representations of the single frames into a single feature vector
- **Linear Classifier**: given the global representation of the gesture, a linear classifier maps the feature vector to the number of gesture classes
- **Multimodal fusion**: multiple modalities are combined with a late fusion approach

SURFACE NORMALS

The pixels of the surface normals encode the three components of the estimated normal vector in that point, computed from the depth map:

- Given a depth map D, we define $Z(x,y)$ as one of its pixel values. The direction $d = \langle d_x, d_y, d_z \rangle$ is calculated as:

$$d = \left(-\frac{\partial Z(x,y)}{\partial x}, -\frac{\partial Z(x,y)}{\partial y}, 1\right)$$

$$\frac{\partial Z(x,y)}{\partial x} \approx Z(x+1,y) - Z(x,y)$$

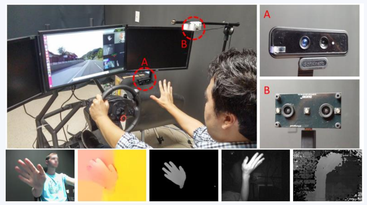
$$\frac{\partial Z(x,y)}{\partial y} \approx Z(x,y+1) - Z(x,y)$$

- Then, the normal vector is normalized by a factor $B = \sqrt{d_x^2 + d_y^2 + 1}$ obtaining: $\hat{v} = \frac{1}{B}(d_x, d_y, 1)$

DATASETS

NVIDIA Dynamic Hand Gesture Dataset (NVGestures)[4]

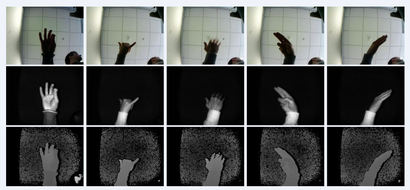
- Acquisition devices:
  - SoftKinetic DS325A, **active RGB-D** sensor, front-view
  - DUO 3DB, **infrared stereo** camera, top-view
- 5 streams: color, **depth**, color mapped on depth, **IR** left, IR right (and disparity map)
- **25 gestures** performed by **20 subjects**



Briareo Dataset[5]

- Acquisition devices:
  - *Pico Flexx,* a compact **active depth** sensor (ToF)
  - *Leap Motion,* an **infrared stereo** camera
- 5 streams: **depth** and **infrared** ampl., left and right **IR**, color
- Recording devices placed in a central tunnel console looking upwards
- **12 gestures** performed by **40 subjects**

Download here



PERFORMANCE ANALYSIS

The proposed method **runs in real time** with acceptable memory usage, even when applied on multiple modalities, running in parallel on the same hardware.

Model	Parameters (M)	Inference (ms)	VRAM (GB)
R3D-CNN [27]	38.0	30	1.3
C3D-HG [25]	26.7	55	1.0
Ours (1 modality)	24.3	26.7	1.8
Ours (2 modalities)	48.6	61.7	3.0
Ours (4 modalities)	97.2	108.3	5.3

RESULTS on NVGestures

- We compare our method with the literature in the **unimodal** and **multimodal** settings
- The combination of **depth maps** and **surface normals** obtains **state-of-the-art results**
- The additional use of color and infrared data slightly improves the overall accuracy

Unimodal setting

Modality	Method	Accuracy
color	Spat. st. CNN [40]	54.6%
	iDT-HOG [45]	59.1%
	Res3ATN [12]	62.7%
	C3D [42]	69.3%
	R3D-CNN [33]	74.1%
	Ours	**76.5%**
	I3D [8]†	78.4%
depth	SNV [47]	70.7%
	C3D [42]	78.8%
	R3D-CNN [33]	80.3%
	I3D [8]†	82.3%
	Ours	**83.0%**
infrared	R3D-CNN [33]	63.5%
	Ours	**64.7%**
flow	iDT-HOF [45]	61.8%
	Temp. st. CNN [40]	68.0%
	Ours	72.0%
	iDT-MBH [45]	76.8%
	R3D-CNN [33]	**77.8%**
	I3D [8]†	83.4%
normals	**Ours**	**82.4%**
color	*Human [33]*	*88.4%*

Multimodal setting

Method	Modality	Accuracy
Two-st. CNNs [40]	color + flow	65.6%
iDT [45]	color + flow	73.4%
R3D-CNN [33]	color + flow	79.3%
R3D-CNN [33]	color + depth + flow	81.5%
R3D-CNN [33]	color + depth + ir	82.0%
R3D-CNN [33]	depth + flow	82.4%
R3D-CNN [33]	all	83.8%
8-MFFs-3f1c [26]*	color + flow	84.7%
I3D [8]†	color + depth	83.8%
I3D [8]†	color + flow	84.4%
I3D [8]†	color + depth + flow	85.7%
$\text{MTUT}_{\text{RGB-D}}$ [1]†	color + depth	85.5%
$\text{MTUT}_{\text{RGB-D+flow}}$ [1]†	color + depth	86.1%
$\text{MTUT}_{\text{RGB-D+flow}}$ [1]†	color + depth + flow	86.9%
Ours	**depth + normals**	**87.3%**
Ours	**color+depth+normals+ir**	**87.6%**
Human [33]	*color*	*88.4%*

RESULTS on Briareo

- We compare our method with the literature in the **unimodal** and **multimodal** settings
- The combination of **infrared amplitude** and **surface normals** obtains **state-of-the-art results**
- The additional use of color data does not improve the overall accuracy

#	Modality	Accuracy
1	color	90.6%
	depth	92.4%
	ir	95.1%
	normals	**95.8%**
2	color + depth	94.1%
	depth + ir	95.1%
	color + ir	95.5%
	depth + normals	96.2%
	color + normals	96.5%
	ir + normals	**97.2%**
3	color + depth + ir	95.1%
	color + depth + normals	95.8%
	color + ir + normals	96.9%
	depth + ir + normals	**97.2%**
4	**color + depth + ir + normals**	**96.2%**

Method	Modality	Accuracy
C3D-HG [31]	color	72.2%
C3D-HG [31]	depth	76.0%
C3D-HG [31]	ir	87.5%
LSTM-HG [31]	3D joint features	94.4%
Ours	normals	95.8%
Ours	depth + normals	96.2%
Ours	**ir + normals**	**97.2%**

REFERENCES

1. Gehring et al, *Convolutional sequence to sequence learning*. ICML 2017.
2. Vaswani et al., *Attention is all you need*. NIPS 2017.
3. P. Molchanov et al., *Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network*. CVPR 2016.
4. F. Manganaro et al., *Hand gestures for the human-car interaction: the Briareo dataset*. ICIAP 2019.