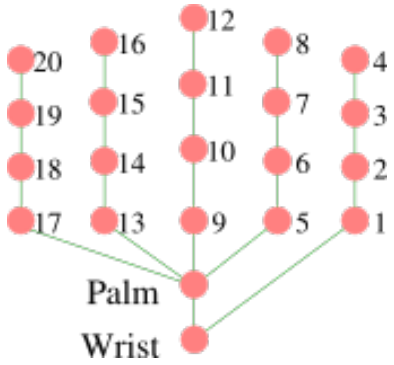
Learning Recurrent High-Order Statistics for Skeleton-Based Hand Gesture Recognition

Xuan Son Nguyen[†], Luc Brun[‡], Olivier Lezoray[‡], Sébastien Bougleux[‡]

† ETIS, Univ. Paris Seine, Univ. Cergy-Pontoise, ENSEA, CNRS, Cergy-Pontoise
‡ Normandie Univ, ENSICAEN, CNRS, UNICAEN, GREYC, Caen
France

Inputs



(a) Initial Graph

20	16	12	8	4
19	15	11	7	3
18	14	10	6	2
17	13	9	5	1

(b) Image encoding of joints. Each joint as a dim equal to 3.

19	15	11	7	3
20	16	12	8	4
19	15	11	7	3
18	14	10	6	2
17	13	9	5	1
18	14	10	6	2

(c) Duplication of lines so that each joint has a up and down neighbour.

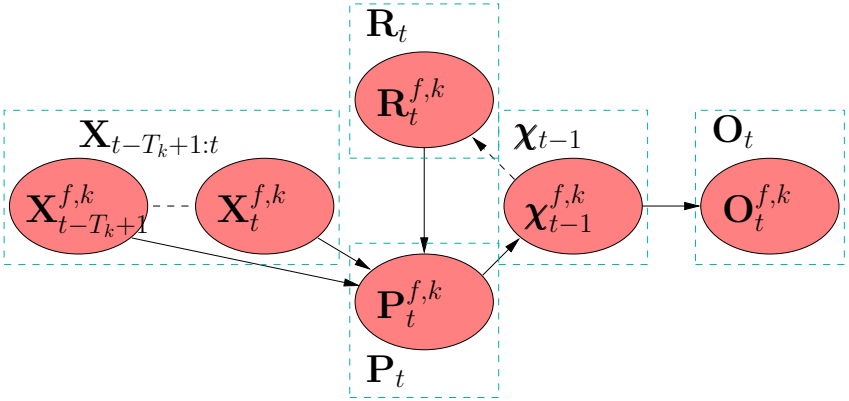
$X_t =$

20	16	12	8	4
19	15	11	7	3
18	14	10	6	2
17	13	9	5	1

(d) Final encoding by the concatenation of up and down neighbours' coordinates. Each join has a dim 9.

Our network

Statistical Recurrent Unit (SRU):



$$\mathrm{P}_t^{f,k} = \mathrm{ReEig}\left(\frac{(w_p^k)^2}{(w_p^k)^2 + (w_x^k)^2} \mathrm{R}_t^{f,k} + \frac{(w_x^k)^2}{(w_p^k)^2 + (w_x^k)^2} h^k(\mathrm{X}_t^f) \right) \quad (1)$$

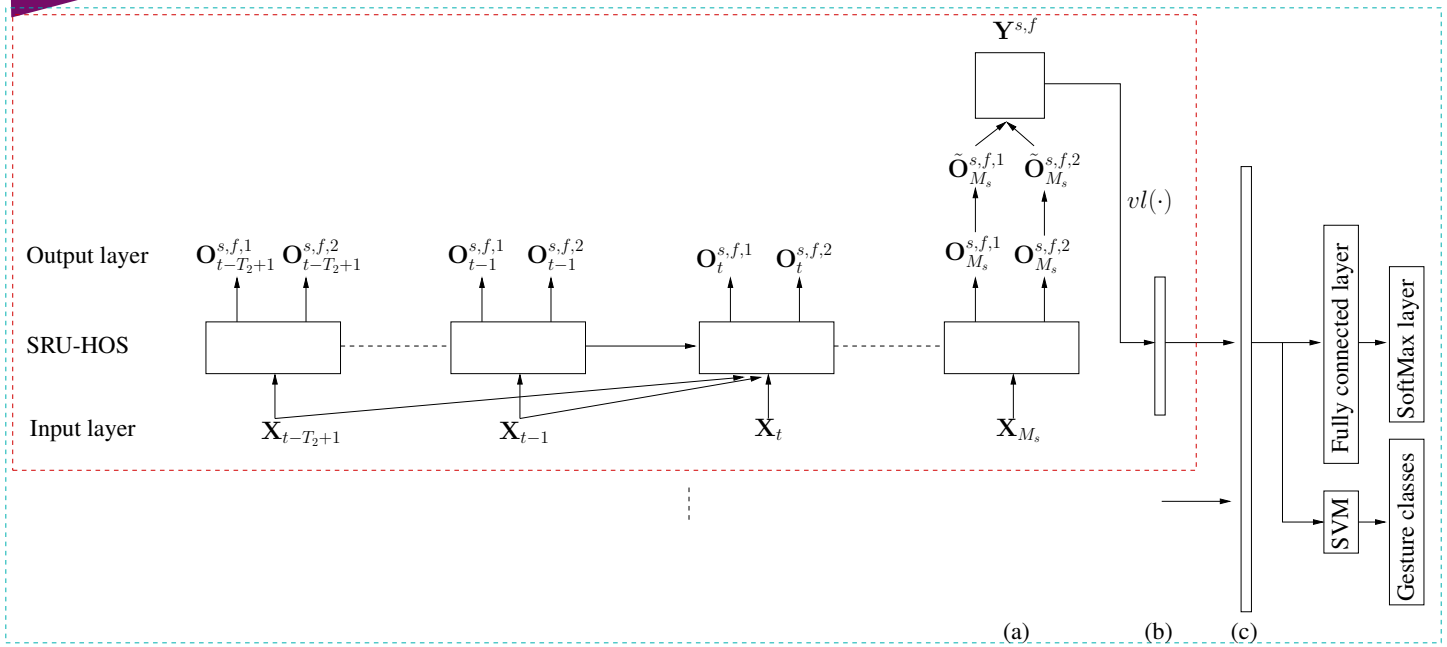
Finger f, time t, statistics of order k.

$$h^1(\mathrm{X}_t^f) = \begin{bmatrix} \boldsymbol{\Sigma}_t^{f,1} + \boldsymbol{\mu}_t^{f,1}(\boldsymbol{\mu}_t^{f,1})^T & \boldsymbol{\mu}_t^{f,1} \\ (\boldsymbol{\mu}_t^{f,1})^T & 1 \end{bmatrix} \quad \text{over the interval } [t - t_1, t].$$

$$h^2(X_t^f) = Cov(vl(h^1(\mathrm{X}_t^f)) \quad \text{over the interval } [t - t_2, t].$$

$$\forall (f,k) \in \{1,\ldots,5\} \times \{1,2\} \; O^{f,k} = SRU\_HOS_k(X^f)$$

Our network



For all $(s, f) \in \{1, \ldots, 6\} \times \{1, \ldots, 5\}$:

$$\mathrm{Y}^{s,f} = \begin{bmatrix} \tilde{\mathrm{O}}^{s,f,2}_{M_s} + \mathrm{vl}(\tilde{\mathrm{O}}^{s,f,1}_{M_s})\mathrm{vl}(\tilde{\mathrm{O}}^{s,f,1}_{M_s})^T & \mathrm{vl}(\tilde{\mathrm{O}}^{s,f,1}_{M_s}) \\ \mathrm{vl}(\tilde{\mathrm{O}}^{s,f,1}_{M_s})^T & 1 \end{bmatrix},$$

Global representation:

$$[\mathrm{vl}(\mathrm{Y}^{1,1})^T, \ldots, \mathrm{vl}(\mathrm{Y}^{6,5})^T]^T$$

Experiments : Ablation study

- concatenation of joint's coordinates

Dataset	Feature concatenation	
	No	Yes
DHG (14 gestures)	85.36	**94.40**
DHG (28 gestures)	78.09	**89.52**
FPHA	83.48	**94.61**

- Relevance of $h^1()$ and $h^2()$ statistics

Statistics	DHG (14 gestures)	DHG (28 gestures)	FPHA
only $h^1(.)$	85.00	76.43	77.04
only $h^2(.)$	89.29	86.07	93.57
Full	**94.4**	**89.52**	**94.61**

- # of parameters

Model	Number of parameters
ST-TS-HGR-NET	672,243
SRU-HOS-NET	**18,894**

Experiments: Comparison with state of the art

- Performance of our method and state-of-the-art methods on DHG dataset.

Method	Year	Color	Depth	Pose	RNN/LSTM	Accuracy (%)	
						14 gestures	28 gestures
HON4D [Oreifej and Liu, 2013]	2013	✗	✓	✗	✗	78.53	74.03
Devanne et al. [Devanne et al., 2015]	2015	✗	✗	✓	✗	79.61	62.00
Huang et al. [Huang and Gool, 2017]	2017	✗	✗	✓	✗	75.24	69.64
De Smedt et al. [Smedt et al., 2016]	2016	✗	✗	✓	✗	88.24	81.90
Devineau et al. [Devineau et al., 2018]	2018	✗	✗	✓	✗	91.28	84.35
SRU [Oliva et al., 2017]	2018	✗	✗	✓	✓	82.02	76.31
SRU-SPD [Chakraborty et al., 2018]	2018	✗	✗	✓	✓	86.31	80.83
ST-TS-HGR-NET [Nguyen et al., 2019]	2019	✗	✗	✓	✗	94.29	89.40
SRU-HOS-NET		✗	✗	✓	✓	**94.40**	**89.52**

Experiments: Comparison with state of the art

FPHA dataset.

Method	Year	Color	Depth	Pose	RNN/LSTM	Accuracy (%)
HON4D [Oreifej and Liu, 2013]	2013	✗	✓	✗	✗	70.61
Novel View [Rahmani and Mian, 2016]	2016	✗	✓	✗	✗	69.21
1-layer LSTM [Zhu et al., 2016]	2016	✗	✗	✓	✓	78.73
2-layer LSTM [Zhu et al., 2016]	2016	✗	✗	✓	✓	80.14
Moving Pose [Zanfir et al., 2013]	2013	✗	✗	✓	✗	56.34
Lie Group [Vemulapalli et al., 2014]	2014	✗	✗	✓	✗	82.69
HBRNN [Du et al., 2015]	2015	✗	✗	✓	✓	77.40
Gram Matrix [Zhang et al., 2016]	2016	✗	✗	✓	✗	85.39
TF [Garcia-Hernando and Kim, 2017]	2017	✗	✗	✓	✗	80.69
JOULE-color [Hu et al., 2015]	2015	✓	✗	✗	✗	66.78
JOULE-depth [Hu et al., 2015]	2015	✗	✓	✗	✗	60.17
JOULE-pose [Hu et al., 2015]	2015	✗	✗	✓	✗	74.60
JOULE-all [Hu et al., 2015]	2015	✓	✓	✓	✗	78.78
Huang et al. [Huang and Gool, 2017]	2017	✗	✗	✓	✗	84.35
Huang et al. [Huang et al., 2018]	2018	✗	✗	✓	✗	77.57
SRU [Oliva et al., 2017]	2018	✗	✗	✓	✓	72.17
SRU-SPD [Chakraborty et al., 2018]	2018	✗	✗	✓	✓	78.96
ST-TS-HGR-NET [Nguyen et al., 2019]	2019	✗	✗	✓	✗	93.22
SRU-HOS-NET		✗	✗	✓	✓	**94.61**

Conclusion

- A new RNN model for skeleton-based hand gesture recognition
- integrate high-order statistics in the SRU for learning discriminative hand gesture representations
- competitive to the state of the art on DHG dataset
- we outperform the state of the art by 1.39 percent on FPHA.

Bibliography I

Chakraborty, R., Yang, C.-H., Zhen, X., Banerjee, M., Archer, D., Vaillancourt, D. E., Singh, V., and Vemuri, B. C. (2018).
A Statistical Recurrent Model on the Manifold of Symmetric Positive Definite Matrices.
In *NeurIPS*, pages 8897–8908.

Devanne, M., Wannous, H., Berretti, S., Pala, P., Daoudi, M., and Bimbo, A. D. (2015).
3-D Human Action Recognition by Shape Analysis of Motion Trajectories on Riemannian Manifold.
IEEE Transactions on Cybernetics, 45(7):1340–1352.

Devineau, G., Moutarde, F., Xi, W., and Yang, J. (2018).
Deep Learning for Hand Gesture Recognition on Skeletal Data.
In *IEEE International Conference on Automatic Face Gesture Recognition*, pages 106–113.

Du, Y., Wang, W., and Wang, L. (2015).
Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition.
In *CVPR*, pages 1110–1118.

Garcia-Hernando, G. and Kim, T.-K. (2017).
Transition Forests: Learning Discriminative Temporal Transitions for Action Recognition.
In *CVPR*, pages 407–415.

Bibliography II

Hu, J., Zheng, W., Lai, J., and Zhang, J. (2015).
Jointly Learning Heterogeneous Features for RGB-D Activity Recognition.
In *CVPR*, pages 5344–5352.

Huang, Z. and Gool, L. V. (2017).
A Riemannian Network for SPD Matrix Learning.
In *AAAI*, pages 2036–2042.

Huang, Z., Wu, J., and Gool, L. V. (2018).
Building Deep Networks on Grassmann Manifolds.
In *AAAI*, pages 3279–3286.

Nguyen, X., Brun, L., Lézoray, O., and Bougleux, S. (2019).
A Neural Network Based on SPD Manifold Learning for Skeleton-based Hand Gesture Recognition.
In *CVPR*.

Oliva, J. B., Póczos, B., and Schneider, J. (2017).
The Statistical Recurrent Unit.
In *ICML*, pages 2671–2680.

Bibliography III

Oreifej, O. and Liu, Z. (2013).
HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences.
In *CVPR*, pages 716–723.

Rahmani, H. and Mian, A. (2016).
3D Action Recognition from Novel Viewpoints.
In *CVPR*, pages 1506–1515.

Smedt, Q. D., Wannous, H., and Vandeborre, J. (2016).
Skeleton-Based Dynamic Hand Gesture Recognition.
In *CVPRW*, pages 1206–1214.

Vemulapalli, R., Arrate, F., and Chellappa, R. (2014).
Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group.
In *CVPR*, pages 588–595.

Zanfir, M., Leordeanu, M., and Sminchisescu, C. (2013).
The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Recognition and Detection.
In *ICCV*, pages 2752–2759.

Bibliography IV

Zhang, X., Wang, Y., Gou, M., Sznaier, M., and Camps, O. (2016).
Efficient Temporal Sequence Comparison and Classification Using Gram Matrix Embeddings on a Riemannian Manifold.
In *CVPR*, pages 4498–4507.

Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., and Xie, X. (2016).
Co-occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep LSTM Networks.
In *AAAI*, pages 3697–3703.