SURGICAL ACTIVITIES RECOGNITION USING MULTI-SCALE RECURRENT NETWORKS

Ilker Gurcan, Hien Van Nguyen

Department of Electrical and Computer Engineering
University of Houston

ABSTRACT

Recently, surgical activity recognition has been receiving significant attention from the medical imaging community. Existing state-of-the-art approaches employ recurrent neural networks such as long-short term memory networks (LSTMs). However, our experiments show that these networks are not effective in capturing the relationship of features with different temporal scales. Such limitation will lead to sub-optimal recognition performance of surgical activities containing complex motions at multiple time scales. To overcome this shortcoming, our paper proposes a multi-scale recurrent neural network (MS-RNN) that combines the strength of both wavelet scattering operations and LSTM. We validate the effectiveness of the proposed network using both real and synthetic datasets. Our experimental results show that MS-RNN outperforms state-of-the-art methods in surgical activity recognition by a significant margin. On a synthetic dataset, the proposed network achieves more than 90% classification accuracy while LSTM's accuracy is around chance level. Experiments on real surgical activity dataset shows a significant improvement of recognition accuracy over the current state of the art (90.2% versus 83.3%).

Index Terms— Multi-Scale Recurrent Network, Surgical Activity, Scattering Convolution Network

1. INTRODUCTION

Automated surgical-activity recognition is valuable for various high-level objectives such as assessment of surgical skills. Earlier pioneering work [1] uses the Markov structure of a surgical task as an indicator of skill. Later work used hidden Markov models (HMMs) learned from kinematic data (hand-movement) [2, 3]. Recent work [4, 5, 6] introduced conditional random fields and other variants as an alternative discriminative to HMMs. These approaches all model each gesture using latent variables. They differ only in how observations are modeled within each gesture.

Recent research [7, 8, 9, 10] has showed that deep networks and reinforcement learning significantly outperformed traditional approaches, which use gestures analysis, on surgical-activity recognition. For example, [9] proposes a temporal convolution network to capture features across multiple time scales. This architecture, however, requires the input data to have a fixed length. [7] employs an LSTM and its bidirectional variant to achieve a high recognition accuracy and edit distance on JIGSAWS dataset [11]. However, these recurrent networks do not have an explicit mechanism to capture information across different temporal scales. While gated recurrent networks theoretically can capture multi-scale temporal information, our experiments on a synthetic data demonstrate that this capability is limited.

Our paper makes the following contributions: a) We design experiments with synthetic data to show the limitation of existing recurrent neural networks; b) To overcome the limitation, we propose a novel recurrent network with an explicit mechanism of encoding multi-scale temporal features; c) We validate the effectiveness of the proposed network on both real and synthetic datasets.

2. BACKGROUND ON INVARIANT SCATTERING CONVOLUTION NETWORKS

A major challenge in image classification comes from the high variability caused by rigid transformations such as translations, rotations, or scaling. The key idea of scattering convolution networks (ScatNet) is to create translation invariant representation of the input signals using a cascade of convolutions with wavelets. This effectively results in descriptors with multi-scale and multi-direction co-occurence information. Recent work demonstrated that these descriptors are highly effective for classification tasks [12, 13, 14, 15]. In what follows, we provide a brief review of ScatNet to facilitate the discussion.

Wavelets: A wavelet is a waveform localized both in time and frequency, as opposed to the Fourier sinusoidal waves which are only localized in frequency. A wavelet transform computes convolutions of input signals with wavelets such as Gabor filters [16]. We will use 2D wavelets as an example for our discussion as in [13]. Let $R_\gamma u$ be the rotation of $u \in R^2$ by an angle γ. Directional wavelets are obtained by rotating a single band-pass filter ψ by different angles γ, and scaling by 2^j, which we can write as follows:

$$\psi_{j,\gamma}(u) = 2^{-2j}\psi(2^{-j}R_\gamma u) \tag{1}$$

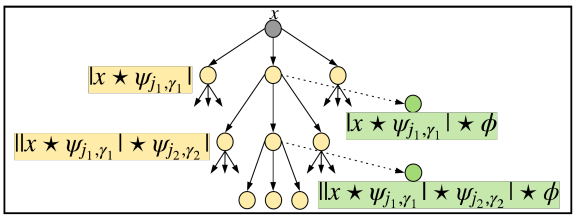

Fig. 1: Scattering convolution networks.

Scattering Convolution Networks: In classification, it is often beneficial to have translation-invariant representation of input signals. Unfortunately, wavelet transforms are not translation invariant. ScatNets overcome this problem by using a cascade of wavelet transforms, followed by non-linear modulus and averaging operators. Fig. 1 provides an illustration of ScatNets. Concretely, the scattering descriptor is computed by convolving the input signal x with a family of wavelets at different scales and orientations: $\{x \star \psi_{j,\gamma}(u)\}$. The output of this operations are 2D response maps containing both real and imaginary parts. ScatNets perform modulus operations on these maps to obtain the magnitudes. This helps reduce the variation due to their complex phases. It then performs an average operation over all pixel locations to obtain a translation-invariant representation. That are $\{|x \star \psi_{j,\gamma}(u)| \star \phi\}$, where ϕ denotes an averaging operator. Although the resulting coefficients are translation-invariant, it does not contain sufficient information for classification purposes as high-frequency features are lost after the averaging operations. To restore the high-frequency information, ScatNets repeatedly convolve the output of the wavelet transforms with other family of wavelets: $|x \star \psi_{j_1,\gamma_1}(u)| \star \psi_{j_2,\gamma_2}(u)$. After each convolution, ScatNets apply modulus and averaging operations to obtain an additional set of translation-invariant coefficients capturing higher-frequency features: $||x \star \psi_{j_1,\gamma_1}(u)| \star \psi_{j_2,\gamma_2}(u)| \star \phi$. These coefficients are called *scattering coefficients* because they result from interactions of x with two or more wavelets. ScatNets compute higher-order coefficients by further iterating on convolutions and modulus operations as shown in Fig. 1. A ScatNet is similar to a multi-layer feed-forward network, except that each layer uses a per-determined family of wavelets to transform the input data.

The concatenation of all scattering coefficients create a rich translation-invariant representation. Moreover, it is also stable to deformations, making it suitable for classification [13]. In other words, ScatNets' coefficients change smoothly for small deformations of the input x. Prior work showed that ScatNets achieve competitive results on hand-written digits classification using only a small number of training samples [13, 12].

3. MULTI-SCALE RECURRENT NETWORK

Motivation: Kinematic data recorded from surgical robots contain motions at different temporal scales. For example, surgical gripper's position might vary more often than its rotation matrix. Being able to extract co-occurence information across temporal scales, as in ScatNet, is important for understanding the on-going surgical activity embedded within kinematic data. Unfortunately, current state-of-the-art recurrent networks such as long-short term memory networks (LSTMs) and gated recurrent units (GRUs) [17] have limited capability in capturing information across multiple temporal scales. To corroborate our claim, we conduct a classification experiment on a set of 1D signals. Each signal contains two structures, one with a short duration and one with a longer duration. More detail on signal generation is provided in the experimental section. The networks have to classify whether the short-duration structure is inside or outside of the long-duration one. To achieve high classification accuracy, networks must capture the relationship of features across multiple temporal scales. Our results show that existing recurrent networks do not perform well on this simple task. For example, both LSTM's and GRU's classification accuracies are around chance level for input signals of 552 dimensions.

Multi-Scale Recurrent Network: Motivated by the observation of LSTM and GRU's shortcomings, we propose a new recurrent architecture, called multi-scale recurrent network (MS-RNN). The central idea is to incorporate scattering coefficients into the network's computation to facilitate the learning of multi-scale temporal features. We use scattering coefficients because they are co-occurence of multiple wavelets, therefore naturally encode interaction of features across multiple scales. In addition, scattering coefficients are translation invariant, and have rigorous theoretical justification in terms of stable deformation. Learning model trained on these coefficients have been showed to generalize well with much smaller number of training examples [12, 13]. This is a desirable property as biomedical datasets, including surgical-

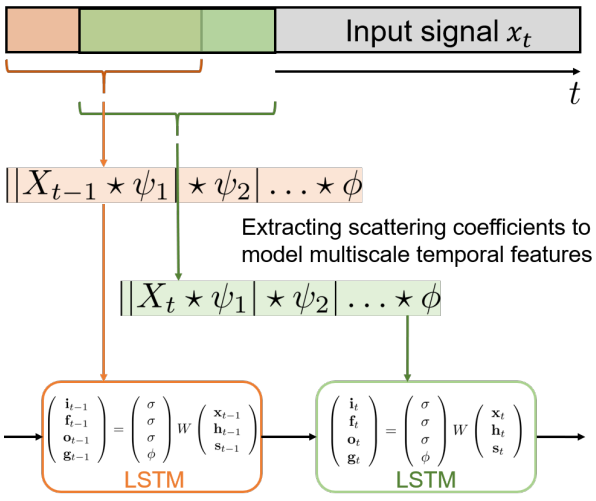

Fig. 2: Illustration of multi-scale recurrent network.

activity recognition datasets, often do not have a large number of samples.

Fig. 2 provides an illustration of the proposed network on 1D signals. Our network uses LSTM as the basic building block. The main difference between MS-RNN and LSTM is in how hidden units and gates are updated at each time step. Let $X_t = [x_{t-p}, \ldots, x_t]$ denotes the buffer data by concatenating the current input and p past inputs. Scattering coefficients, denoted by s_t, are computed by applying a ScatNet on X_t. Specifically, we convolve X_t with multiple sets of filters $\{\psi_i^{(\ell)}\}_{i=1}^{K_\ell}$ along the temporal dimension, where K_ℓ is the number of filters in layer ℓ of the ScatNet. We apply modulus and averaging operations on the convolution output each layer ℓ to obtain a set of scattering coefficients. These coefficients then serve as additional input variables to our network. The hidden units and gates in MS-RNN is updated as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{si}s_t + b_i), \quad (2)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{sf}s_t + b_f), \quad (3)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{so}s_t + b_o), \quad (4)$$
$$g_t = \phi(W_{xc}x_t + W_{hc}h_{t-1} + W_{sc}s_t + b_c), \quad (5)$$
$$\text{where } s_t = \text{ScatNet}(X_t) \quad (6)$$

ScatNets utilize pre-determined wavelets to compute co-occurences of all scales and orientations, which can be highly redundant. To increase the compactness of the extracted features, MS-RNN jointly optimizes LSTM's parameters and the ScatNet's filters $\{\psi_i^{(\ell)}\}_{i=1}^{K_\ell}$. Learned filters are more compact since they are tuned to data, therefore reduce the number of additional input variables in our networks and make its training less prone to over-fitting. MS-RNN is much simpler than recently proposed methods [18, 19, 20]. Intuitively, one can think of MS-RNN as an augmentation of the existing LSTMs. This is in the same vein to Neural Turing Machines [21] which achieve the capability of learning long-term dependency by augmenting LSTMs with an external memory.

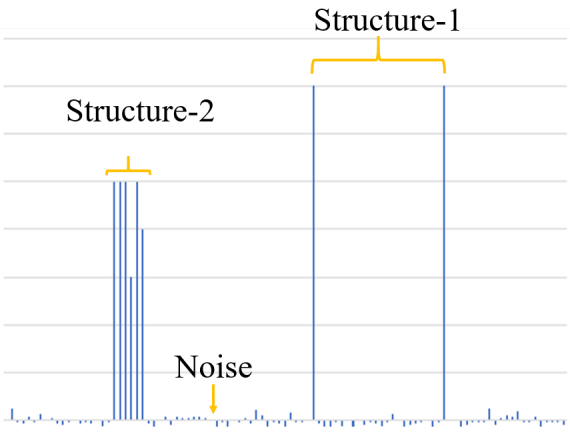

Fig. 3: An example of synthetic signals. The label is 0 as structure-2 falls outside of structure-1.

4. EXPERIMENTS

4.1. Synthetic Data

Data Generation: We first use synthetic data containing interaction of multiple temporal scales to compare the performance of the proposed network to LSTM and GRU. Our synthetic data are N-length one dimensional signals. Each signal consists of the embedding of two structures. Structure–1 has a longer duration and is composed of two pulses with equal magnitudes. The interval between two pulses, denoted by d_1, has variable length with minimum length of 45 and maximum length of 75. Structure-2 has a shorter duration, and consists of multiple pulses with irregular magnitudes, which are smaller than the magnitude of pulses in structure-1. The rest of the signal is filled with Gaussian noise with zero mean and unit standard deviation. We assign the label $y = 1$ to a signal if the structure-2 falls between two pulses of the structure-1, and $y = 0$ otherwise. Fig. 3 shows an example of generated signals. The task is to predict the label y given a signal. Intuitively, to achieve high classification accuracy, a network must understand the relationship between the two structures. This in turns requires the extraction of multi-scale features.

Training: We generate a dataset of 20,000 samples for training and the same number of samples for testing. We use Adam optimizer with learning rate of 0.001, and batch size of 16, to train all the networks. The best number of hidden units for all networks are selected by a 10-fold cross-validation. For MS-RNN, we set the number of hidden units to be $m = 30$, beyond which we did not observe any significant improvement in the classification accuracy. All weights are initiated using Xavier initialization. For computing scattering coefficients, we set the buffer signal length to $p = 80$. The numbers of layers and filters in ScatNet are respectively $L = 2$ and $K_\ell = 20$.

Results: In addition to LSTM, we also compare with a gated recurrent unit (GRU) [17]. GRUs have a simpler gating mech-

Network	LSTM		GRU		MS-RNN	
Sequence length	100	552	100	552	100	552
Validation accuracy (every 800 training iterations)	0.7773	0.5052	0.7564	0.4836	0.8164	0.5540
	0.7846	0.5235	0.7754	0.5195	0.8401	0.6101
	0.7315	0.5095	0.7685	0.5297	0.9597	0.7187
	0.7773	0.5058	0.7894	0.5184	0.9750	0.7974
	0.7993	0.5168	0.7873	0.5189	0.9677	0.8908
	0.7810	0.5143	0.7921	0.5197	0.9841	0.9201
Testing accuracy	**0.7799**	**0.5078**	**0.7925**	**0.5064**	**0.9848**	**0.9161**

Fig. 4: Comparison of LSTM, GRU, and MS-RNN classification accuracy on the synthetic signals of length 100 and 552.

anism than LSTM and has been showed to produce competitive results in various classification tasks. Fig. 4 summarizes the classification accuracies of different recurrent networks. We can notice that MS-RNN outperforms both LSTM and GRU by a significant margin. For example, MS-RNN achieves 98.48% accuracy compared to 77.9% of LSTM for 100-dimensional input signals. When the signal length increases to 552, LSTM's and GRU's performances are reduced to chance level, while the proposed network maintains more than 90% classification accuracy. We note that increasing the number of hidden units within LSTM and GRU does not improving their accuracies. The experiment demonstrates that MS-RNN is highly effective in dealing with complex interaction of features from different temporal scales.

4.2. Surgical-Activity Recognition

Description of Dataset: The JIGSAWS dataset consists of videos data, robot kimenatics, and manual annotations sampled at 30 Hz. These data were recorded when surgeons perform elementary tasks on a bench-top model in a laboratory using the *da Vinci* Surgical System [22]. There are 15 surgical tasks related to suturing, knot-tying, and needle-passing. These tasks are part of the surgical skills training curricula. Kinematic data include positions, rotation matrices, angular velocities, gripper's angles master tool manipulators (i.e. operated by surgeons) and patient-side manipulators.

Recognition Task: The goal of this experiment is to recognize surgical activities given a sequence of robot kinematics in JIGSAWS dataset [11]. This is the same as classifying every frame of the sequence into one of the activities. We downsample the data from 30 Hz to 5 Hz as done in [7]. We use the standardized leave-one-user-out evaluation setup: for the i-th run, use i-th user data for testing, and the rest for training. We average the results over 8 runs, one for each user. We use Levenshtein distance to measure the performance on this dataset. Intuitively, this distance corresponds to the number of actions required for transforming a sequence of predicted labels into ground-truth sequence of labels. This is different from accuracy, which is the percentage of correctly-classified frames, without considering temporal consistency.

Hyperparameter Setting: Here we include the most relevant details regarding hyperparameter selection and training. For each run we train for a total of approximately 200 epochs. We use a batch size of 5 sequences for all experiments. We performed 10-fold cross-validation to determine the best parameter settings corresponding to the lowest edit distance. In particular, we set the number of hidden units $m = 1024$, ScatNet's number of layers $\ell = 2$, convolution filters $K_\ell = 30$, buffer length $p = 60$ for all experiments. Using a modern GPU (GTX 1080 Ti), training takes about 2 hours. At test time, the network takes approximately 0.5 seconds to compute the output for each minute of kinematic sequence (300 time steps).

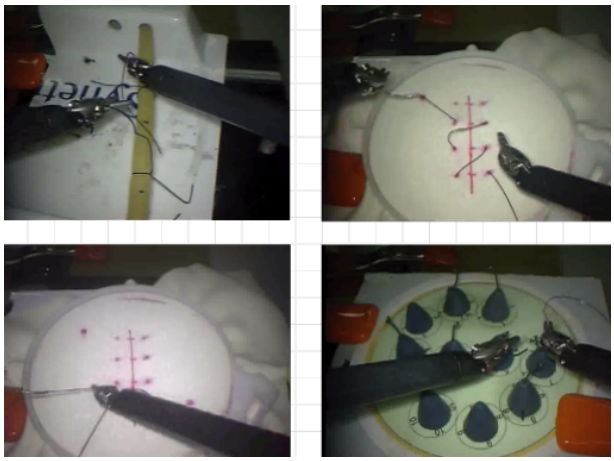

Fig. 5: Images from JIGSAWS dataset.

	Accuracy (%)	Edit Dist. (%)
MsM-CRF	72.6	-
SDSDL	78.7	-
SC-CRF	80.3	-
LC-SC-CRF	82.5 ± 5.4	14.8 ± 9.4
Forward LSTM	80.5 ± 6.2	19.8 ± 8.7
Bidir. LSTM	83.3 ± 5.7	14.6 ± 9.6
MS-RNN	$\mathbf{90.2 \pm 7.5}$	$\mathbf{10.5 \pm 9.8}$

Table 1: Quantitative comparisons of MS-RNN performances to prior work on JIGSAWS dataset. Our performances were averaged over 10 runs and all activities.

Results: Table 1 shows gesture recognition accuracies and the corresponding Levenshtein distance on JIGSAWS dataset. We compare with a forward LSTM, a bidirectional LSTM, and other state-of-the-art approaches. We include standard deviations where possible. The dataset has a large inter-user variation, where some users are highly challenging, regardless of the recognition method. Our method outperforms other approaches by a significant margin, both in terms of accuracy and edit distance.

5. CONCLUSIONS

This paper investigates the limitation of existing recurrent networks such as LSTMs and GRUs on capturing multi-scale interaction within data. We propose a simple yet effective multi-scale recurrent network. We demonstrate the effectiveness of the proposed approach using both real and synthetic data. For synthetic data, MS-RNN achieve higher than 90% accuracy while the performance of other methods are reduced to chance level. For surgical-activity recognition task, the MS-RNN outperforms state-of-the-art method using by a significant margin. In the future, we will combine kinematic and video data to further improve the recognition performance.

6. REFERENCES

[1] Jacob Rosen, Blake Hannaford, Christina G Richards, and Mika N Sinanan, "Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills," *IEEE transactions on Biomedical Engineering*, vol. 48, no. 5, pp. 579–591, 2001.

[2] Julian JH Leong, Marios Nicolaou, Louis Atallah, George P Mylonas, Ara W Darzi, and Guang-Zhong Yang, "HMM assessment of quality of movement trajectory in laparoscopic surgery," in *MICCAI*. Springer, 2006, pp. 752–759.

[3] Carol E Reiley and Gregory D Hager, "Task versus subtask surgical skill evaluation of robotic minimally invasive surgery," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2009, pp. 435–442.

[4] Aristotelis Dosis, Fernando Bello, Duncan Gillies, Shabnam Undre, Rajesh Aggarwal, and Ara Darzi, "Laparoscopic task recognition using hidden markov models," *Studies in Health Technology and Informatics*, vol. 111, pp. 115–122, 2005.

[5] Balakrishnan Varadarajan, *Learning and inference algorithms for dynamical system models of dextrous motion*, The Johns Hopkins University, 2011.

[6] Lingling Tao, Ehsan Elhamifar, Sanjeev Khudanpur, Gregory D Hager, and René Vidal, "Sparse hidden markov models for surgical gesture classification and skill evaluation," in *International conference on information processing in computer-assisted interventions*. Springer, 2012, pp. 167–177.

[7] Robert DiPietro, Colin Lea, Anand Malpani, Narges Ahmidi, S Swaroop Vedula, Gyusung I Lee, Mija R Lee, and Gregory D Hager, "Recognizing surgical activities with recurrent neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 551–558.

[8] Colin Lea, Austin Reiter, René Vidal, and Gregory D Hager, "Segmental spatiotemporal CNNs for fine-grained action segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 36–52.

[9] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 47–54.

[10] Daochang Liu and Tingting Jiang, "Deep reinforcement learning for surgical gesture segmentation and classification," *MICCAI*, 2018.

[11] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al., "JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modeling," in *MICCAI Workshop: M2CAI*, 2014, vol. 3, p. 3.

[12] Joan Bruna and Stéphane Mallat, "Classification with scattering operators," in *CVPR, 2011 IEEE Conference on*. IEEE, 2011, pp. 1561–1566.

[13] Joan Bruna and Stéphane Mallat, "Invariant scattering convolution networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[14] Joakim Andén and Stéphane Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.

[15] Joakim Andén and Stéphane Mallat, "Multiscale scattering for audio classification.," in *ISMIR*. Miami, FL, 2011, pp. 657–662.

[16] Tai Sing Lee, "Image representation using 2D gabor wavelets," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 18, no. 10, pp. 959–971, 1996.

[17] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[18] Jan Koutnik, Klaus Greff, Faustino Gomez, and Juergen Schmidhuber, "A clockwork RNN," *arXiv preprint arXiv:1402.3511*, 2014.

[19] Junyoung Chung, Sungjin Ahn, and Yoshua Bengio, "Hierarchical multiscale recurrent neural networks," *arXiv preprint arXiv:1609.01704*, 2016.

[20] Asier Mujika, Florian Meier, and Angelika Steger, "Fast-slow recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 5915–5924.

[21] Alex Graves, Greg Wayne, and Ivo Danihelka, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, 2014.

[22] Gary S Guthart and J Kenneth Salisbury, "The intuitive/sup TM/telesurgery system: overview and application," in *Robotics and Automation, 2000. Proceedings. ICRA'00. IEEE Intl. Conference on*. IEEE, 2000, vol. 1, pp. 618–621.