Deep Learning-Based Motion Estimation for Uninterrupted Tracking of Awake Rodents in PET

Shisheng Zhang, *Student Member, IEEE*, Mehala Balamurali and Andre Kyme, *Member, IEEE*

***Abstract*–The ability to image the brain of awake rodents using motion-compensated positron emission tomography (PET) presents many exciting possibilities for exploring the links between brain function and behavior. A key requirement of this approach is obtaining accurate estimates of animal pose throughout a scan. Our present motion tracking approach suffers crucial line-of-sight limitations which leads to tracking "drop-out" and subsequent loss of motion information that can be used for motion correction. The proportion of a scan affected can range anywhere from 5% for sedentary subjects up to >50% for highly active subjects. The aim of this work was to investigate the feasibility of augmenting optical motion tracking with a video-based deep learning motion estimation method to mitigate the impact of tracking drop-out.**

A deep convolutional neural network (CNN) based regression approach for estimating six rigid-body motion parameters is proposed. We tested our model using multi-view camera images of a rat phantom under robotic control. The commanded robot motion provided the labels for our data. We compared the performance of deep learning-based motion estimation for simulated gaps in the motion sequence against the robot ground truth. We also compared deep learning to naïve linear interpolation of motion across the gaps. Deep learning provided promising alignment with the ground truth motion, in many cases sub-degree/sub-mm. The root mean square error for the deep learning and interpolation methods versus ground-truth was 1.26° and 23.64° (y-axis rotation) and 0.77 mm and 6.57 mm (z-position), respectively.

Deep learning-based rigid-body motion estimation from multi-view video appears promising as a solution for augmenting optical tracking. Future work will focus on (i) the use of a Long Short-Term Memory (LSTM) unit to better model temporal information in the motion trace and (ii) incorporation of the known camera calibration to further constrain pose estimates.

I. Introduction

We have developed some key technologies that enable the brain of a rodent to be imaged while the animal moves freely inside a positron emission tomography (PET) scanner [1]. The technique has enormous potential to improve our understanding of how brain function and behaviour relate to each other in mammals. A vital component of the technique is the use of motion tracking to accurately estimate an animal's head motion during a scan. The optical tracking method we currently use, although accurate in principle, relies on attached optical markers which have a crucial line-of-sight limitation. This results in intermittent drop-out of motion tracking. The proportion of a scan affected by drop-out can range anywhere from 5% for sedentary subjects up to 50% or more for highly active subjects, such as those being scanned during an amphetamine drug challenge.

To address the problem of interrupted optical motion tracking, we investigated the feasibility of using a multi-view video-based deep learning motion estimation approach to augment the sporadic optically-derived measurements [2]. This is convenient because we typically have ready access to video data of the animal from the motion tracking system and/or video cameras used to record behavioural responses during a study. Our hypothesis is that a convolutional neural network can be trained on video images labelled with rigid-body head pose measurements from the optical tracking system in order to provide accurate estimates of head pose for frames acquired when the optical system drops out. We tested this hypothesis using multiple perspective views of a robotically controlled rat phantom and compared the resulting motion estimates with ground truth data.

Manuscript received December 1, 2018.

Shisheng Zhang is with the School of AMME, Faculty of Engineering & IT, University of Sydney, Sydney, Australia. Email: szha4847@uni.sydney.edu.au.

Mehala Balamurali is with the School of AMME, Faculty of Engineering & IT and Australian Centre for Field Robotics, University of Sydney, Sydney, Australia. Email: mehala.balamurali@sydney.edu.au.

Andre Kyme is with the School of AMME, Faculty of Engineering & IT and Brain & Mind Centre, University of Sydney, Sydney, Australia. Email: andre.kyme@sydney.edu.au

II. Methods

A. Image Data and Data Labelling

Synchronized frames were collected by four monochrome CCD cameras viewing a rat head phantom being manipulated by a six-axis robot (Epson C3, Seiko Corp, Japan) according to measured rat head motion. We used a sequence of 500 images from each camera. This corresponded to 15 s of rat motion since the camera frame rate was 30 Hz. A representative pose from each camera is shown in Fig. 1. Since the camera frames were synchronized with the robot, each frame was labelled with the known (commanded) robot rigid-body pose (six parameters: x, y and z displacement and x, y and z rotation).

B. Model Design

Deep convolutional neural networks are hierarchically organized models that map inputs to outputs in a highly non-linear fashion via successive functional transformations within layers of the network [3]. Layers consists of many computational units called neurons that receive input from previous layers and output a single value. The neurons in a layer constitute the internal feature representation of the layer.

The basic layer types in our model are the convolution layer, dropout layer and dense layer. A convolution layer computes an output via a 2D convolution operation. Higher

978-1-5386-8495-5/18/$31.00 ©2018 IEEE

level convolution layers receive input from previous layers and are capable of representing more complex patterns than lower layers. A dropout layer is a regularization technique aimed at reducing the complexity of the model and preventing overfitting [3]. Use of a dropout layer randomly deactivates certain neurons in a layer according to a particular probability constraint. In our study we set this probability to 50%, forcing half the activations of a layer to zero in a given feed-forward pass through the neural network during training. Finally, a dense layer is a regular layer of neurons in a neural network, receiving input from all neurons in the previous layer.

Figure 2 shows the architecture of our model. It comprises eight convolution layers with 5×5 kernels, three dense layers to integrate information from our raw images, a concatenate layer to horizontally stack the output of the third dense layer, and a regression output layer which allows the prediction of rigid-body pose.

C. Image Preprocessing

The original images (1280x1024) were resized to 50x50 and zero-mean normalized to improve model convergence. Rotated versions of the input images, in 45° increments, were used to augment the raw data set and improve prediction performance [4]. Thus, we supplied eight images per camera per frame for the training dataset.

D. Model Training

To train the model we minimized an objective function defined as the mean square error (MSE) computed across each of the estimated parameter vectors and its corresponding ground-truth vector. We used the Adam Optimizer to reduce overfitting and provide faster convergence [5].

III. Results

Figure 3(a) and 4(a) demonstrate two 100-frame segments of the estimated y-rotation and z-position compared with the ground truth. Images in the 100-frame validation sequence were not included in the training set and were selected randomly from the 500-image sequence as an initial test. Figure 3(b) and 4(b) show the estimated y and z position for a contiguous (rather than random) sequence of 250 frames. This is the situation encountered in motion tracking drop-out and therefore is of chief interest in understanding how well a deep learning-based motion estimation approach can fill in the gaps caused by drop-out of an optical tracking system. Also shown in Fig. 3 and 4 is a comparison of the deep learning-based pose estimates with a naïve linear interpolation of pose within the gap caused by drop-out of the optical tracking system. The root mean square error relative to the ground truth, computed for the deep learning and interpolation methods, was 1.26° and 23.64° (for y rotation) and 0.77 mm and 6.57 mm (for z displacement), respectively.

IV. Discussion and Conclusion

We have described a deep learning method to estimate the missing rigid-body pose data due to line-of-sight limitations in optical motion tracking. It is clear from our preliminary investigation that this method is feasible and potentially offers an accurate and useful way to estimate motions. There are some limitations in estimating all degrees-of-freedom, in particular the y-position (data not shown) exhibits a systematic discrepancy with respect to the ground truth. We are currently investigating reasons for this.

Several important improvements can be made to what we have presented here and these are the focus of our future work: (i) the use of a Long Short-Term Memory (LSTM) unit to model the temporal information inherent in the motion of the animal, and (ii) incorporation of the known spatial calibration of the cameras to further constrain pose estimates.

Acknowledgement

The authors thank Fabio Ramos from the School of IT, University of Sydney, for useful discussions on implementing the deep learning approach.

References

[1] V. Zhou, J. Eisenhuth, A. Kyme, M. Akhtar, R. Fulton and S. Meikle. "A motion adaptive animal chamber for PET imaging of freely moving animals." *IEEE Transactions on Nuclear Science* 60, no. 5, pp. 3423-3431, 2013.

[2] H. Su, C. R. Qi, Y. Li and L. J. Guibas. "Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views." *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2686-2694. 2015.

[3] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." *The Journal of Machine Learning Research* 15, no. 1, pp. 1929-1958, 2014.

[4] T. Amaral, L. M. Silva, L. A. Alexandre, C. Kandaswamy, J. M. de Sá and J. M. Santos. "Transfer learning using rotated image data to improve deep neural network performance." *International Conference Image Analysis and Recognition*, pp. 290-300, 2014.

[5] D. Kinga and J. B. Adam. "A method for stochastic optimization." *International Conference on Learning Representations* (ICLR), vol. 5. 2015.

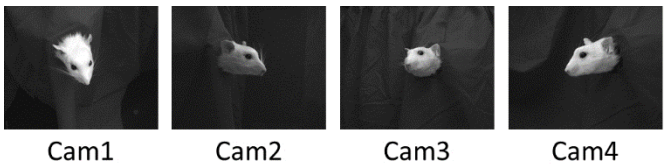

Fig. 1. Rat head pose from four cameras.

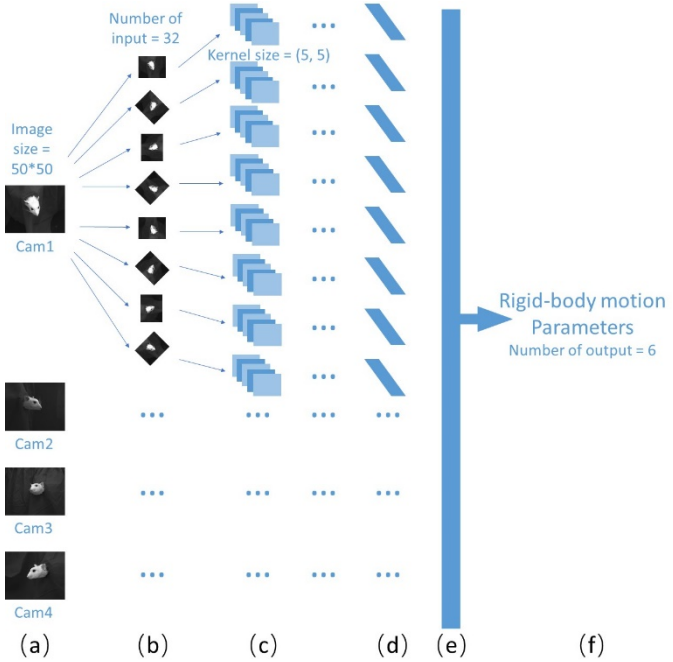

Fig. 2. The architecture of our network. (a) Input images are rotated in increments of 45°; (b) 32-image input data including normal images and rotated images. (c). Convolution layers and dropout layers that extract features and avoid overfit. (d). Each feature matrix is flattened to one array in a flatten layer. (e). Arrays are concatenated into one large feature matrix. (f). Prediction includes six rigid-body motion parameters.

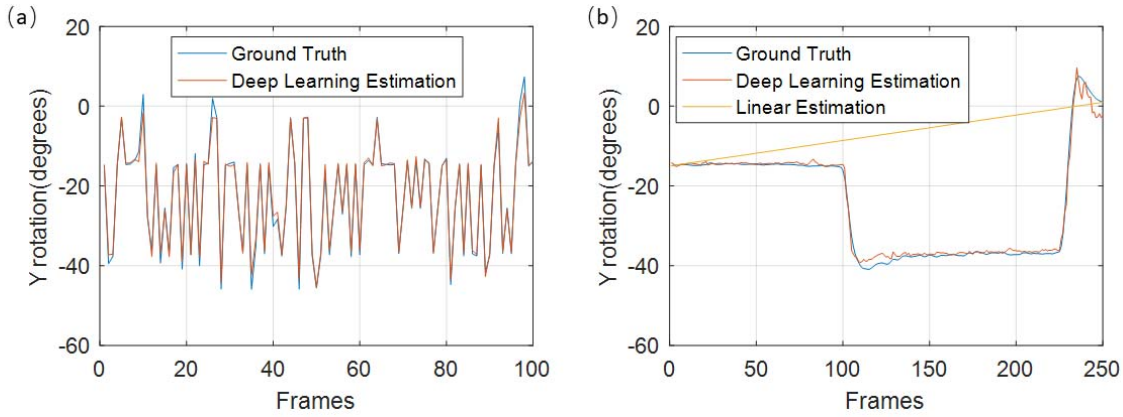

Fig. 3. Deep learning-based estimate of y-axis rotation for (a) 100 randomly chosen frames and (b) 250 sequential frames of the rat motion trace. Also shown in (b) is a comparison with simple linear interpolation of pose across the gap.

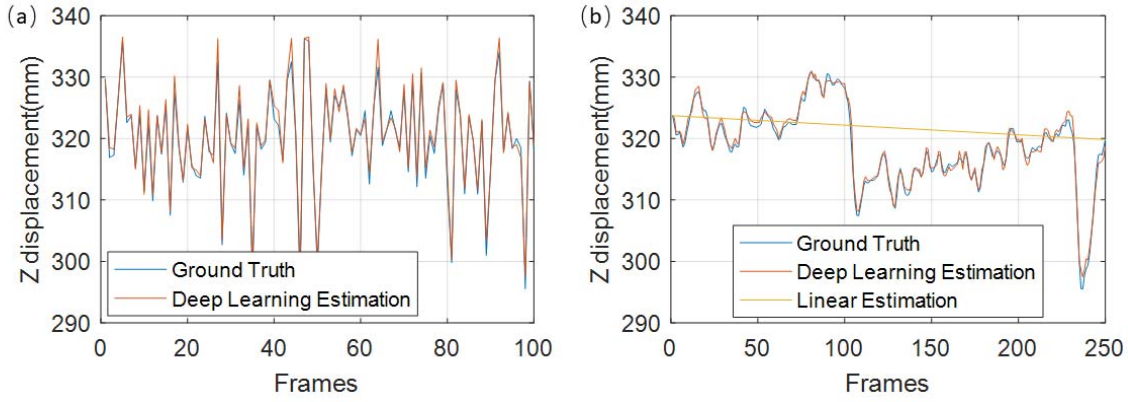

Fig. 4. Deep learning-based estimate of z-position for (a) 100 randomly chosen frames and (b) 250 sequential frames of the rat motion trace. Also shown in (b) is a comparison with simple linear interpolation of pose across the gap.