Applying Deep Learning Models to Action Recognition of Swimming Mice with the Scarcity of Training Data

Ngoc Giang Nguyen[1], Mera Kartika Delimayanti[1,2], Bedy Purnama[1,3], Kunti Robiatul Mahmudah[1], Mamoru Kubo[4], Makiko Kakikawa[4], Yoichi Yamada[4] and Kenji Satou[4]

[1]*Department of Electrical Engineering and Computer Science, Kanazawa University, Kanazawa, Japan*
[2]*Department of Computer and Informatics Engineering, Politeknik Negeri Jakarta, Jakarta, Indonesia*
[3]*Telkom School of Computing, TELKOM University, Bandung, Indonesia*
[4]*Institute of Science and Engineering, Kanazawa University, Kanazawa, Japan*

Keywords: Swimming Mouse Behaviour Recognition, Deep Learning, Transfer Learning, Data Scarcity.

Abstract: Deep learning models have shown their ability to model complicated problems in more efficient ways than other machine learning techniques in many application fields. For human action recognition tasks, the current state-of-the-art models are deep learning models. But they are not well-studied in applying for animal behaviour recognition due to the lack of data required for training these models. Therefore, in this research, we proposed a method to apply deep learning models to recognize the behaviours of a swimming mouse in two mouse forced swim tests with a limited amount of training data. We used deep learning models which are used in human action recognition tasks and fine-tuned them on the largest publicly available mouse behaviour dataset to give the models the knowledge about mouse behaviour recognition tasks. Then we fine-tuned the models one more time using the small amount of data that we have annotated for our swimming mouse behaviour recognition tasks. The good performance of these models in the new tasks proved the efficiency of our approach.

1 INTRODUCTION

We have been using many animals from mice, fish to primates to study biology, psychology or develop new types of medicines. In these researches, to answer questions such as how animals behave in specific environments or how their behaviours change after using a new drug, we have to watch and annotate many hours of their recorded videos. These tasks are time-consuming and costly but essential for the researches. Another problem is that human assessments are not always consistent, so the experiments become harder to reproduce. Therefore, we need automated animal behaviour recognition systems to delegate these frustrating works to computers which are less erroneous and more consistent in their assessments.

The works presented in (Jhuang et al., 2010) and (Jiang et al., 2017) both proposed automated mouse behaviour recognition systems based on the same approach. First, relying on expert knowledge, they created feature extractors to extract important information from input data, such as the relative position of the mouse and some specific points in the scene (water faucet or food gate). Then they used the extracted features and a classifier to predict the behaviours of mice. One drawback of this approach is that the feature extractors are only designed for specific experiments. For this reason, we can not apply these systems for our experiments unless we exactly mimic the setups of their experiments or redesign the feature extractors and retrain the classifiers from scratch. We can avoid this drawback by using deep learning models to create our systems. Deep learning models are classifiers with built-in feature extractors. Thus, these models can learn to extract important features from input data without any requirement of expert knowledge. The current state-of-the-art models in human action recognition tasks are deep learning models. However, to achieve those high performances, these deep learning models have to learn the actions from thousands of minutes of annotated videos. Unfortunately, we do not usually have such extensive annotated datasets

Nguyen, N., Delimayanti, M., Purnama, B., Mahmudah, K., Kubo, M., Kakikawa, M., Yamada, Y. and Satou, K.
Applying Deep Learning Models to Action Recognition of Swimming Mice with the Scarcity of Training Data.
DOI: 10.5220/0007567602700275
In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019)*, pages 270-275
ISBN: 978-989-758-353-7

available to train these models for our specific experiments.

In order to palliate the issue related to training data scarcity, we proposed a method that does not require a large amount of training data to apply deep learning models to mouse behaviour recognition tasks. In our method, we utilize deep learning models pre-trained from human action recognition tasks. First, we retrain these models using the largest of the current publicly available mouse behaviour datasets (Jhuang et al., 2010). We use this step to give the models the knowledge about mouse behaviour recognition tasks. Then, we use data of our specific tasks to fine-tune these models. Because the retrained models after the first step have learned knowledge about mouse behaviours, we do not need a large amount of data to train them for our specific tasks in the second step.

In the next section, we describe the deep learning models and the mouse behaviour dataset we used in the first step of our proposed method. In Section 3, we present the swimming mouse behaviour recognition tasks we used to evaluate our method and the results of our experiments. Finally, in Section 4, we state our conclusions.

2 METHOD

As described in the previous section, our proposed method has two steps. In the first step, we fine-tune deep learning models which were used for human action recognition tasks by using the largest publicly available mouse behaviour dataset. Then in the second step, we train these models again using the data we prepared for our swimming mouse behaviour recognition tasks. In this section, we give the information about the deep learning models and the mouse behaviour dataset we used in the first step of our method.

2.1 The Two-Stream I3d Model

Carreira and Zisserman introduced the Two-Stream Inflated 3D ConvNets (Two-Stream I3D model) (Carreira and Zisserman, 2018), one of the current state-of-the-art deep learning models for human action recognition tasks. As reported, the Two-Stream I3D models achieve 98% of accuracy on UCF-101 human action recognition dataset (Soomro, Zamir and Shah, 2012) and 80.9% of accuracy on HMDB-51 human action recognition dataset (Kuehne et al., 2011). These models are derived from the Inception-V1 model which uses the Inception module architecture (Szegedy et al., 2015). Layers of the Inception modules combine filters of different sizes and pooling kernels to utilize all their good effects in feature extraction.

To create an I3D model, all 2D filters and pooling kernels of an Inception-V1 model are inflated to 3D by endowing them with an additional temporal dimension, i.e. $n \times n$ filters become $n \times n \times n$ filters, and the weights of the 3D filters are bootstrapped by repeating the weights of the respective 2D filters n times along the new temporal dimension. This bootstrap method let the I3D models benefit from the learned parameters of the pre-trained 2D models.

In this research, we used the same I3D models' architectures as described in the paper of Carreira and Zisserman (Carreira and Zisserman, 2018). The models were pre-trained on ImageNet data (Russakovsky et al., 2015) for the first step of our method. Also reported in the research of Carreira and Zisserman, using optical flow data computed from RGB data to train a complementary model for the model trained on RGB data can help to improve the prediction accuracy. Therefore, in this research, we also utilized optical flow data, and we experimented on various fusion ratios of RGB data trained models and optical flow trained models to find the best fusion ratio for the swimming mouse behaviour recognition tasks. To compute optical flow data from the RGB data we used the TV-L1 algorithm (Zach, Pock and Bischof, 2007).

2.2 The Mouse Behaviour Dataset

In the work of Jhuang H. et al. (Jhuang et al., 2010), they created a dataset to train their mouse behaviour recognition system. They have recorded and annotated more than 9000 video clips (~10 hours of video) of single housed mice in a home cage. There are 8 types of behaviour annotated in this dataset: “drink”, “eat”, “groom”, “hang”, “micro-movement”, “rear”, “rest” and “walk”.

From the recorded video clips, they selected 4,200 clips (~2.5 hours of video) that have the most unambiguous examples of each behaviour to create a subset called “clipped database”. In this research, we used this subset for the first step of our method.

3 EXPERIMENTS & RESULTS

The mouse forced swim tests are rodent behavioural tests used to study about antidepressant drugs, antidepressant efficacy of new compounds, and experimental manipulations aimed at preventing depressive-like states as described in the research of Can A. et al. (Can et al., 2012). In a forced swim test, a mouse is placed in a transparent cylinder which is filled with water for a certain period of time. Naturally, when being placed in such a dangerous environment, the mouse will become panic and try to escape. But if the mouse has taken some antidepressant drugs before, it will be less panic than a mouse which hasn't taken any drug. Therefore, by measuring the duration of each behaviour of the mouse, such as mobile behaviour and immobile behaviour, we can estimate the effect of the drug in the mouse.

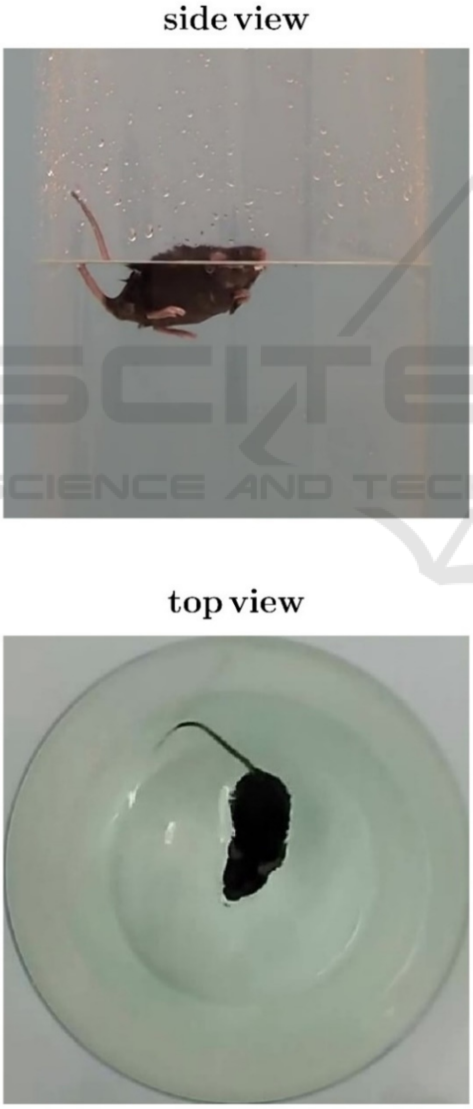

Figure 1: Example scenes of the side view data and the top view data.

To evaluate the efficiency of our method, we annotated two videos from mouse forced swim tests, each video has a length of about 5 minutes. The first video was recorded from a side view. For this side view video we annotated three behaviours: "swim", "struggle" and "float". The difference between "struggle" behaviour and "swim" behaviour is that when a mouse is struggling, it just slightly moves one of its feet. When a mouse is floating, it is immobile. The second video was recorded from a top view, and for this video, we annotated two behaviours: "mobile" and "immobile". We showed an example scene of each video in Figure 1.

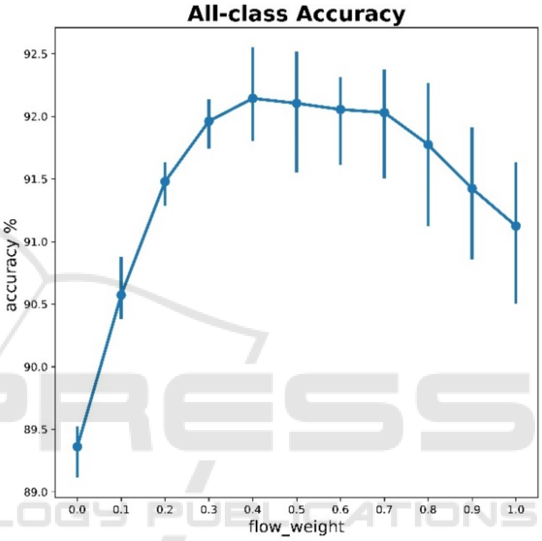

Figure 2: Prediction accuracy in the side view data.

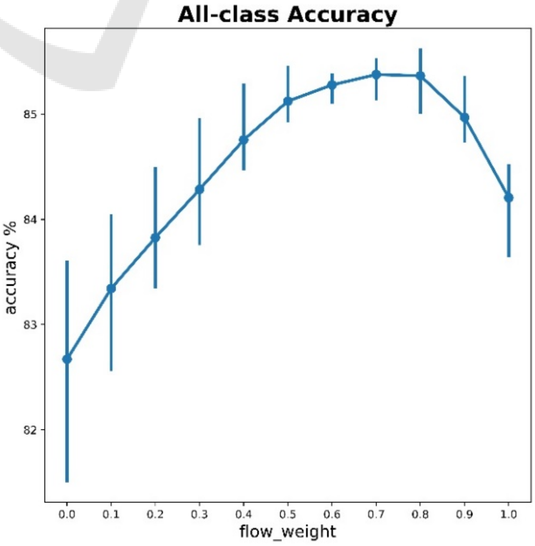

Figure 3: Prediction accuracy in the top view data.

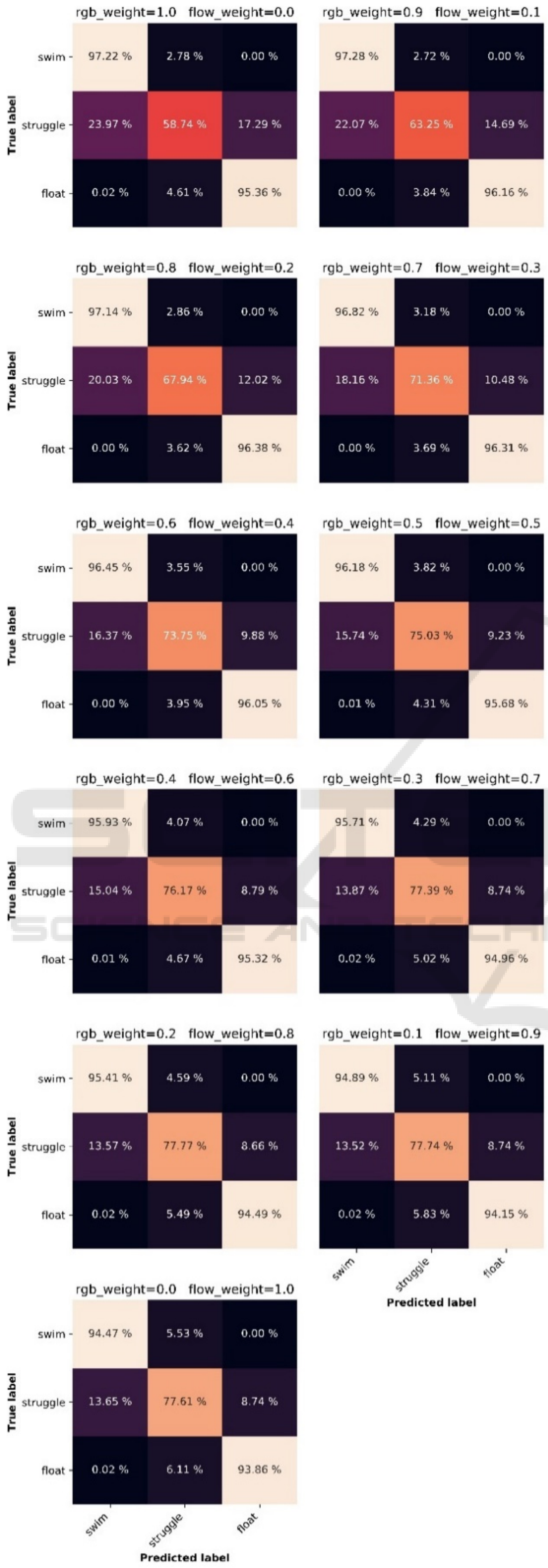

Figure 4: Confusion matrix of prediction in the side view data.

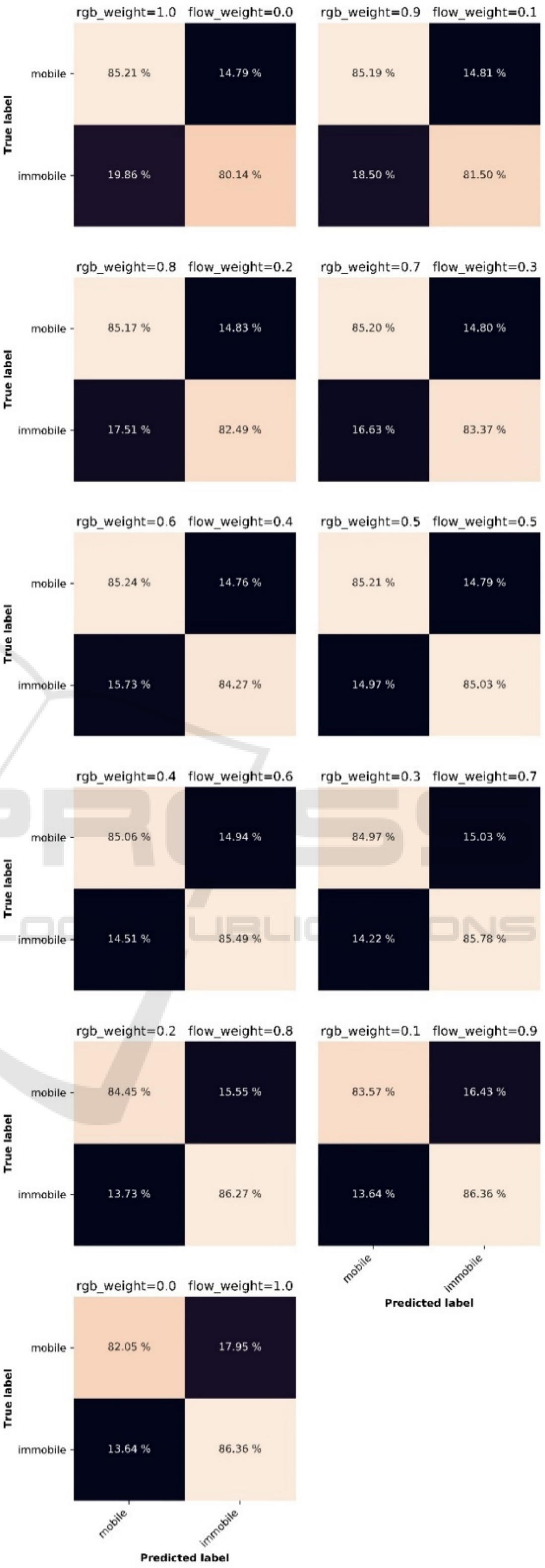

Figure 5: Confusion matrix of prediction in the top view data.

In the first step of our experiments, we used weights from I3D models' checkpoints that were pre-trained on ImageNet data to initialize parameters of the model. We retrained them using the "clipped database" dataset. For each sample data, we used 16 successive frames as an input to the I3D model (current frame, its 8 previous frames and its 7 next frames). To train the models, we used a momentum optimizer with a momentum value equals to 0.9 and a learning rate starts from 1e-3 and decays to 5e-5 after several thousands of iterations. We used the dropout technique in fully connected layers with a keep-probability of 36% to reduce the effect of overfitting when we train the models.

After the first step, we have two I3D models, one was trained on RGB data from "clipped database" and the other was trained on optical-flow data from "clipped database". In the second step, we fine-tuned the two models using RGB data and optical-flow data from the swimming mouse behaviour data that we have annotated. To fine-tune the models, we also used a momentum optimizer with a momentum value of 0.9 and a learning rate starts from 1e-3 and decays to 5e-3. We also used the dropout technique with a keep-probability of 36%.

To test the effect of different fusion ratios of RGB data trained models and optical flow data trained models, we used two parameters *flow_weight* and *rgb_weight*, i.e. *flow_weight = 0.3* then *rgb_weight = 0.7* and it means the fusion ratio is 30% of optical flow data trained models and 70% of RGB data trained models. In our experiments we examined flow_weight values from 0 to 1 with a step of 0.1 and respectively with *rgb_weight* values equal to *1 - flow_weight* values to find out the best fusion ratio.

To calculate the prediction accuracies of the models, we separated the data into 10 folds and conducted 10-fold cross-validation. The prediction accuracies of our method for each fusion ratio in the side view data and the top view data are shown in Figure 2 and Figure 3.

The confusion matrices of the models in the side view data for different fusion ratios are shown in Figure 4. For the top view data, the confusion matrices are shown in Figure 5.

For the side view data, the models perform well on predicting "swim" and "float" behaviours but have some difficulty in predicting "struggle" behaviours. For the top view data, the models have some problems in distinguishing "mobile" and "immobile" behaviours because in some "immobile" samples the water surface is still shaking as a result of the previous "mobile" behaviour. So we may need more data to help the models understand these cases.

In the side view swimming mouse behaviour recognition task, our method achieved the best performance at the fusion ratio of 40% of optical flow data trained models and 60% of RGB data trained models with the prediction accuracy of 92.14%. In the top view swimming mouse behaviour recognition task, the best performance of our method is 85.38% of prediction accuracy with the fusion ratio of 70% of optical flow data trained models and 30% of RGB data trained models.

4 CONCLUSIONS

In this research, we have proposed a method to apply deep learning models to mouse behaviour recognition tasks to achieve high prediction accuracies without the requirement of a large amount of training data. The results of the experiments proved the efficiency of our approach.

With this approach, we will attempt to create a framework to apply deep learning models to general mouse behaviour recognition tasks and also for other specific mouse related experiments.

In further researches, we will develop our method to apply deep learning models to other animal behaviour recognition tasks.

ACKNOWLEDGEMENTS

In this research, the super-computing resource was provided by Human Genome Center, the Institute of Medical Science, the University of Tokyo. Additional computation time was provided by the super computer system in Research Organization of Information and Systems (ROIS), National Institute of Genetics (NIG). This work was supported by JSPS KAKENHI Grant Number JP18K11525.

REFERENCES

Jhuang, H., Garrote, E., Yu, X., Khilnani, V., Poggio, T., Steele, A. D., Sere, T., 2010. Automated home-cage behavioural phenotyping of mice. In *Nature communications*.

Jiang, Z., Crokes, D., Green, B.D., Zhang, S., Zhou, H., 2017. Behaviour recognition in mouse videos using contextual features encoded by spatial-temporal stacked Fisher vectors. In *Proceeding of International Conference on Pattern Recognition Applications and Methods*.

Carreira, J., Zisserman, A., 2018. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going Deeper with Convolutions. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei, L. F., 2015. ImageNet Large Scale Visual Recognition Challenge. In *International Journal of Computer Vision*.

Can, A., Dao, T. D., Arad, M., Terrillion, E. C., Piantadosi, C. S., Gould, D. T., 2012. The Mouse Forced Swim Test. In *Journal of Visualized Experiments: JoVe*.

Zach, C., Pock, T., Bischof, H., 2007. A duality based approach for realtime TV-L1 optical flow. In *Proceeding of 29th DAGM Symposium of Pattern Recognition*.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T., 2011. HMDB: a large video database for human motion recognition. In *Proceeding of the International Conference on Computer Vision*.

Soomro, K., Zamir, A. R., Shah, M., 2012. UCF101: a dataset of 101 human actions classes from videos in the wild. In *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*.