Capturing Spatial and Temporal Context for Fine Grained Canine Behaviour Action Recognition

Anthony THEODORE
Samuel TONG

23rd April 2019

Contents

1.0 Abstract

Video-based action recognition tasks have been gaining popularity based on the significantly high potential for a wide ranging types of real-world applications such as autonomous driving vehicles and video surveillance. However, the attempts to build frameworks for visual recognition tasks have been focused mainly on the actions of human subjects. With growing demand in the deployment of canine as service animals in a wide variety of industries, such as the healthcare and security enforcement sector, the automatic translation of canine behaviour would be meaningful and important for the canine human counterpart. In this paper, we introduced a new challenging dataset for fine-grained action recognition of canine behaviour. We also investigated the task of performing fine-grained recognition through the incorporation of both spatial and temporal context in our proposed four-stages neural network architecture. The third stage model achieved the best result amongst all other models due to its capability in combining spatial and temporal contextual information extracted from global and local data.

2.0 Introduction

With growing developments in the medical and healthcare industry, the prescription of animal-assisted therapy has been growing steadily over the years [1]. These animals, most commonly canines, have been designated the roles of therapy animals, emotional support animals and assistance animals for the purpose of providing assistance related to a person's disability. In the security sector, canines are playing an important role in the provision of security enhancement through its deployment in the police and military force. While there are dedicated canine trainers to ensure the suitability and readiness of a canine prior to its deployment as a service animal to its human counterpart, the lack of domain knowledge in proper canine training and canine behaviour recognition by the human counterpart may result in issues with the handling of potential canine behavioural problems. As a result, the automatic translation of a canine behaviour would be meaningful and important for the handlers of service dogs.

2.1 Challenges

Existing solutions for action recognition have been implemented for a diverse selection of real-world applications such as autonomous driving vehicles and video surveillance. However, there are still many challenges that researchers face when optimizing action recognition tasks:

High Computation Cost

According to D. Tran *et al* [2], the training of a 3DConvNet on the UCF101 [3] dataset and approximately two months on the Sports-1M [4] dataset. Therefore, this would make experimentation on building extensive architecture on action recognition non-trivial without exhausting resources.

Capturing Spatial-Temporal Context

While action recognition may seem like the extension of image classification, it involves the classification of specific actions from a sequence of frames, or images, in a video clip, instead of a static image without the consideration of the temporal difference between the prior and post image. Furthermore, various challenges that are often faced in image classification are extended to action recognition as well.

In this paper, one such challenge that we have faced is the problem of properly discerning fine-grained details that separates different classes of canine behaviour. Even though there are two significantly large datasets dedicated for action recognition, UCF101 [3] and Sports-1M [4], these datasets do not contain videos that are specific to fine-grained action recognition. Therefore, there is a lack of a large-scale video database for researchers to design more sophisticated algorithms to optimize fine-grained action recognition. However, there are several datasets available to allow for domain-specific action recognition tasks such as the NTSEL Database, and the Near-Miss Driving Recorder Database (NDRDB) [5] for fine-grained pedestrian action recognition, and the MLB-YouTube [6] dataset for sports analysis. In our case, since we are interested in classifying the behavioural traits of canine through action recognition, we introduced a challenging new dataset, Canine-Behaviour, which is designed for fine-grained activity detection. The comparison of our dataset against other fine-grained datasets is shown in Table 1 below.

Table 1: Comparison of Fine-Grained Recognition Datasets

	# of Video Clips	# of Categories
NTSEL	100	4
NDRDB	82	4
MLB-YouTube	4,290	9
Canine-Behaviour (Our Dataset)	**916**	**4**

While it is possible to treat each individual video frame as an image to be classified based on the application of CNNs as a trivial method to perform action recognition tasks when aggregating the score over the entirety of each video, such method would not be able to take into account of the temporal features between adjacent frames. Additionally, the presence of background noises in the video may lead to the learning of undesirable features that would not benefit the training process. However, if we attempt to omit the background entirely, we may lose information that may be useful for the capturing of temporal features.

Therefore, we hypothesise that it is essential to extract features of frames at both a global and local level, while balancing between the loss of spatial information and temporal information, for accurate video classification. In addition, to compensate for the possible loss of temporal information due to the extraction of local images, we incorporate the use of optical flow images computed over adjacent frames to provide for the addition of temporal features, while also retaining spatial information. Our contributions can be summarised as follow:

1. We introduced a challenging fine-grained action recognition dataset that is designed specifically for the identification of canine behaviour.
2. We built an end-to-end pipeline to allow for the extraction of local images and optical flow images from a video, to be fed through each deep neural network architecture for training and testing.
3. We experimentally compared various approaches for capturing spatial and temporal context for fined-grained action recognition. These approaches include the study of how different combination of spatial and temporal contextual information extracted from the data can contribute to the accurate prediction of canine behaviour. These approaches will be discussed in the section 4 below.

3.0 Related Work

3.1 Fine-Grained Action Recognition Dataset

Due to its huge potential for real-world application, action recognition has been a popular research topic in various forms of application [5, 6, 7, 8]. According to Piergiovanni, AJ & S. Ryoo, Michael [6], action recognition has a huge potential in computer vision but much of fine-grained data that are domain-specific are not publicly available. Therefore, the paper also introduced a new dataset that focuses on fine-grained action recognition. In addition, the authors of the paper approached the issue of fine-grained action recognition by comparing various recognition approaches to capture temporal structure in videos, based on the classification of segmented videos, and extending those approaches to continuous videos. The rationale behind this is that the exploration of various methods of temporal feature aggregation for segmented videos allow for easier classification tasks since each frame would correspond to an action, and therefore there is no need for the classification model to determine when an action begins or ends.

3.2 Capturing Spatial Context: Global-Local Video Feature Representation

Other works have also experimented with extracting entire video frames and cropped images of target object in the foreground to exclude noise from the background [5]. Such methods thus allows the CNN to track and focus on important gestures that would provide better representation for better action recognition.

3.3 Capturing Temporal Context: Optical Flow and LSTM

According to the Ng, Y.H. *et al*, the incorporation of optical flow images that are computed over adjacent frames allows for the capturing of motion information and at the same time retain global spatial representation of the video frames. Additionally, it was noted by the authors of the paper that despite the fact that optical flow images may be noisy, they have proven to work well with Long Short Term Memory (LSTM) [9] models in conjunction with raw image features.

However, there were no previous attempts at capturing both spatial information and temporal information for a fine-grained action recognition task by incorporating raw image features at a global level (global image), local images that captures the object of interest and optical flow images through a CNN-LSTM based model. Additionally, instead of attempting to approach the fine-grained action recognition tasks by directly learning an LSTM model on top of deep features extracted by CNN, we experimented and evaluated several deep neural network architectures with a combination of different image information (global images, local images and optical flow images).

4.0 Approach

4.1 Dataset: Canine-Behaviour

The Canine-Behaviour dataset contains 916 video clips of four different canine behavioural traits, under two super-categories. One super-category contains two negative emotions – angry and sad. On the other hand, the other super-category contains two positive emotions – happy and submissive. Specifically, canine that expresses angry emotions tend to bare their teeth and are typically tensed. Canine that are sad tend to lay flat on the ground with ears and tails down. Canines that are happy tend to move at a rapid pace with tails lifted or wagging. Submissive canines tend to lay on their back with their bellies exposed.

However, the dataset poses multiple challenges to the optimization of canine behavioural recognition. For example, different species of canine may look vastly different from one another. The different physical attributions of different canine breeds may thus contribute to unwanted noise during deep feature learning that do not help in explaining the behaviour expressed. Additionally, due to the fact that physical movements, such as the movement of mouth, head, tail and legs, that correlate to different behavioural expressions may be subtle and is thus difficult for the neural networks to pick up.

4.2 Problem Setting

In this paper, we attempt to learn a function that transforms both spatial and temporal input into a sequence of feature vectors that is amenable to classification. This allows us to be able to understand the behavioural traits expressed by a canine at a particular point in time, or how it changes over time.

Formally, given the visual input of spatial and temporal information (S, T), where:

- $S = (s_1, s_2, \ldots, s_n)$
- $T = (t_1, t_2, \ldots, t_n)$

We want to minimize $E_c(S, T)$, where:

- E_c = classification error

To overcome the challenges posed by the dataset and to conduct fine-grained action recognition to classify canine behaviour, we attempted to capture both spatial and temporal information through the incorporation of three different variants of images:

1. Global Images – Represents the raw image of the entire video frame
2. Local Images – Represents the cropped image of the canine from the original raw image
3. Optical Flow Images – Represents the motion information between adjacent frames

The three variants of images will then be fed to different variants of neural network architectures, which will be discussed in section 4.6. Additionally, other than attempting to correctly predict each of the four canine behaviour, we also evaluate the performance of the model based on the ability to correctly separate between positive and negative emotion.

4.3 Gesture Detection for Extracting Local Images

Motivated by the success of deep learning object detectors [10, 11], we pre-trained a Single Shot MultiBox Detector (SSD) on the VOC2007[1] for the tracking and detection of canine gestures, which inherently depicts the behavioural expressions of canines. Based on each input of global images (of each frame in a video), the SSD outputs two resulting images – the original global image and the cropped image of the canine that was detected in each frame. The reason for giving the additional global output is to ensure consistency in the number of global and local images. For example, given a video, there might be some frames where the canine might not be detected simply because it was not within the frame.

The SSD based detection could also fail in certain cases where the canine is occluded by other objects in the foreground. However, our proposed method is largely unaffected by this problem as the canine subject is mostly contained within the centre of each frame.

Figure 1: Global image of an angry dog (left). Local image of the same frame (right)

[1] http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html

4.4 Retrieving Optical Flow Images

Based on the Lucas-Kanade method, we utilised the algorithm provided by OpenCV, a library of programming functions mainly for the support of computer vision, to compute the optical flow for all points in each adjacent frame.

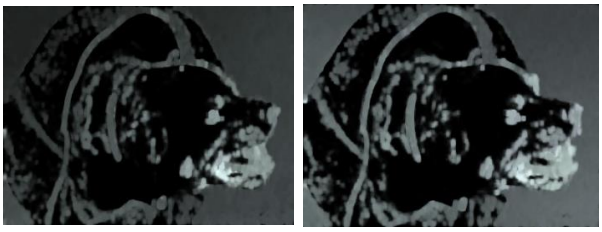

Figure 2: Examples of optical flow images

4.5 Base Network Architecture

As the first Runner Up for image classification in the ImageNet Large Scale Visual Recognition Competition (ILSVRC) in 2015, InceptionV3 [12] managed to achieve a lower error rate based on its capability to achieve high computational efficiency with relatively fewer parameters. Action recognition tasks, which requires high computational costs, would benefit greatly by adopting the InceptionV3 architecture. The factorization convolutions allows for the reduction of parameters without compromising on its network efficiency. Therefore, we used InceptionV3 as our base model, to which other proposed techniques will be added on top for ablation study. Additionally, since we are addressing the problem of action recognition, it would make sense to utilise any motion or temporal information that may allow us to discover and integrate useful information over time for better classification performance. Therefore, we investigated the usage of a recurrent neural network architecture, the LSTM, which allows for the learning of long-term dependencies, by retaining information for an extended period of time.

In the next section, we explore multiple variations of incorporating InceptionV3 in the conduct of our experiments in conjunction with LSTM networks to learn temporally ordered sequences.

4.6 Proposed Training Architectures

We divide our training architectures into four stages.

Stage One (Spatial)

In stage one, to ensure consistency in our experiments, we first introduce an InceptionV3 model, where we attempt to perform a naïve classification of each frame in a video individually. In the subsequent stages, additional complexity will be built on top of it sequentially. Based on the architecture of a single CNN, we are essentially training our model based on only spatial information. Therefore, the input of our model will be each individual global or local RGB video frame. For each frame the softmax layer outputs the predicted label. To obtain the final predicted label of the entire video clip sequence we take the maximum of all the predicted classes and select the class with the highest number of predictions.

For simplicity, we named our stage one models using global images and local images as *1SCNN(Global)* and *1SCNN(Local)* respectively, where 1S stands for one stream.

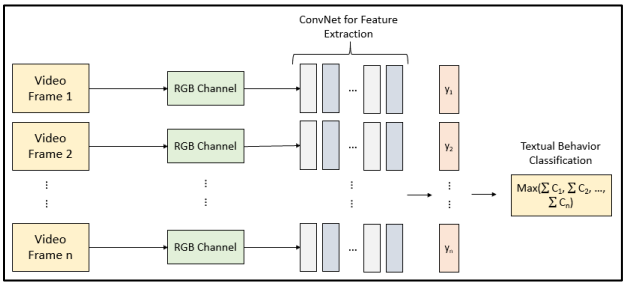

Figure 3: Overview of Stage One Architecture

Formally, the predicted label of each video clip is defined as follow:

$$Max(\sum c_1, \sum c_2, \ldots, \sum c_n)$$

Stage Two (Temporal)

In stage two, rather than obtaining the prediction at the end of the CNN softmax layer, we will feed the deep features of each global or local frame extracted from the last layer of the InceptionV3 into an LSTM. On a high level definition, the LSTM decoder sequentially updates LSTM state and internal memory, given previous state and current input data to predict a label y_t at time t. In this case, we are experimenting with the capabilities of the LSTM to pick up motion / temporal information from frame to frame.

For simplicity, we named our stage two models as *1SCNN+LSTM(Global)* and *1SCNN+LSTM(Local)* for the models using global images and local images respectively.

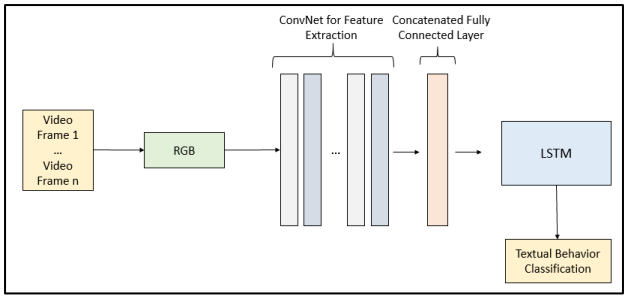

Figure 4: Overview of Stage Two Architecture

Stage Three (Aggregation of Global and Local Images)

In stage one, the action recognition tasks is trained based only on spatial information, while in stage two, the inclusion of the LSTM model allows for the learning of temporal features between frames. Since the input to the model in each of the two stages is based on either a global or local image, we are interested in investigating whether the aggregation of deep features of both global and local images of identical frame would lead to a better result.

Since the cropped local image has reduced background noise, we hypothesised that the temporal information on the movement of the canine with reference to the background would be lost. On the other hand, the global image could potentially benefit from the LSTM given that the temporal information of the canine subject relative to its background is retained. In this case, we will be implementing two different models, where the first model, *2SCNN*, is a merger of two CNNs (each with global and local images) to a fully connected layer for prediction. The second model, *2SCNN+LSTM*, is a merger of two deep features layer from two CNNs (each with global and local images) into an LSTM.

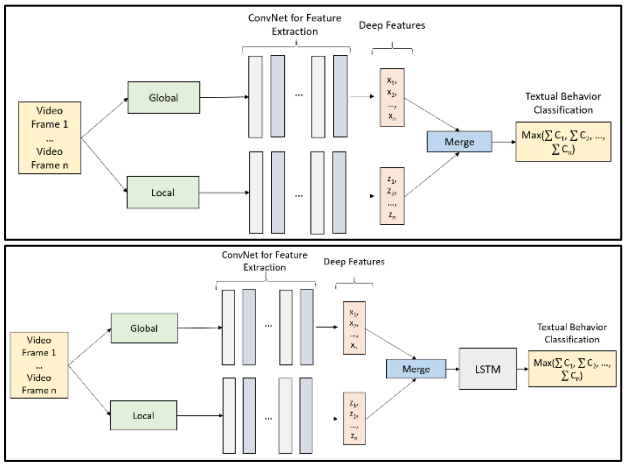

Figure 5: Overview of Stage Three Architecture. Aggregation of global and local images to a fully connected layer prior to final prediction based solely on spatial information (top). Aggregation of global and local images to LSTM prior to final prediction based on temporal information (bottom).

Stage Four (Three-Stream CNN-LSTM with Optical Flow)

In stage four, we explore the potential benefit that the deep features of the optical flow images can bring to the CNN-LSTM architecture.

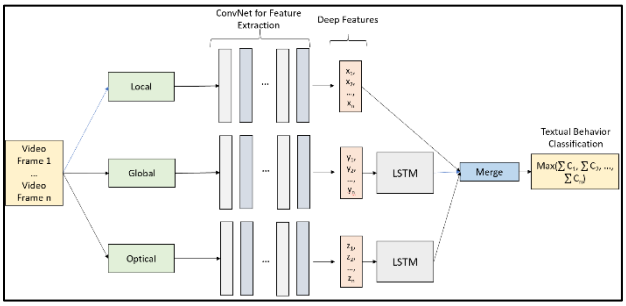

Figure 6: Overview of Stage Four Architecture. Aggregation of global, local and optical flow in a three-stream CNN-LSTM architecture.

Since optical flow is able to capture motion information and at the same time retain global spatial representation of the video frames, we hypothesise that its aggregation with both global and local deep features will allow for the encapsulation of both spatial and temporal features for better action recognition. Therefore, we implement a three-stream neural network, *3SCNN+LSTM*, where both global and optical flow images will be fed to a CNN and to an LSTM. The local image will only be fed into a CNN. Subsequently, the deep feature layers of all three images will be merged into a fully connected layer for prediction.

5.0 Experiments

5.1 Experimental Setting

In the Canine-Behaviour dataset, the videos are of 25 frames per second with resolution of 720 X 1280. Subsequently, each video was divided into 500-frames clips, and resized to 224 X 224. The parameters for each model in each of the four stages are optimized individually to achieve the highest accuracy.

5.2 Evaluation Metrics

Formally, the predicted label of each video clip is defined as follow:

$$Max(\sum c_1, \sum c_2, \ldots, \sum c_n)$$

Where c refers to the class predicted.

5.3 Results and Analyses

The evaluation results for each of the model are included in Table 2 below:

Table 2: Model Evaluation Results.

Methods	Validation Accuracy	Test Accuracy
Stage One		
1SCNN(Global)	0.323	0.281
1SCNN(Local)	0.386	0.418
Stage Two		
1SCNN+LSTM (Global)	0.353	0.274
1SCNN+LSTM (Local)	0.330	0.271
Stage Three		
2SCNN	0.345	0.291
2SCNN+LSTM	0.411	0.455
Stage Four		
3SCNN+LSTM	0.384	0.275

Discussion of Results

In stage one, the action classification is essentially a naïve image classification tasks where the temporal information between frames are not taken into consideration. Based on the architecture of a single CNN, we are essentially training our model based on only spatial information. Since the cropped local image has reduced background noise, it allows for the neural network to obtain features that are more representative of the canine subject itself. Therefore, the performance of a single stream CNN on the local image is higher than that of the single stream CNN on the global image.

In stage two, with the addition of the LSTM on top of the CNN, we hypothesised that the model will be able to learn temporal features that are pertinent in the recognition of sequential actions in this dataset. However, the results of both one stream CNN + LSTM models performed worse than the one stream CNN models.

On the other hand, the performance of the 1SCNN+LSTM on global image is better than that of the 1SCNN+LSTM on the local image. Since the cropped local image has reduced background noise, the temporal information on the movement of the canine with reference to the background would be lost and therefore result in a poorer performance than that of the global image.

Figure 6: Example of reduced background noise on two adjacent images that results in the reduction motion information

In stage three, we observed that when compared with the models in stage two and stage one, the performance of 2SCNN is higher. Additionally, the performance of 2SCNN+LSTM is significantly higher than all other previous models. Surprisingly, given the poor result of the ISCNN+LSTM in stage two, the addition of a second steam of CNN+LSTM actually contributed to a higher result. This could be due to the spatial-temporal nature of the data, that demands network structures to be highly complex [14].

In stage four, we observed that even though the validation accuracy of the 3SCNN+LSTM was relatively high as compared to the other models, its test accuracy was significantly lower than expected. One possible reason for its inability to generalize well to the test data, is due to the high model complexity which involves the merging of deep features from each of the three streams CNN into an LSTM. However, during training of the model, we observed that by fine-tuning the model, it has potential of achieving a better accuracy than the stage three models. Given enough resources, we believe that the stage four 3SCNN+LSTM will be able to perform well.

6.0 Conclusion

We introduced a new challenging fine-grained action recognition dataset that is designed specifically for the identification of canine behaviour. Additionally, we created a pipeline for the extraction of local and optical flow image for end-to-end training. Based on our experiments in building network architecture in fine-grained action recognition, we established that the combination of global, local and optical flow images through a two-stream combination of CNN and LSTM results in a higher performance of recognising fine-grained canine behaviour. We also show that using LSTM is not always helpful, unless a more sophisticated model is used to capture the complex spatial-temporal nature of the data to produce better results.

In the current models, only visual images are incorporated into the network to perform action recognition. In the future, more work can be done to explore the application of cross-modal learning in fine-grained action recognition. In particular, it would be interesting to explore the combination of audio, visual and pose estimation inputs to perform fine-grained action recognition tasks.

7.0 References

[1] Arkrow P. Animal-Assisted Therapy and Activities: A Study, Resource Guide and Bibliography for the Use of Companion Animals in Selected Therapies. 10th ed. P. Arkrow; Stratford, NJ, USA: 2011.

[2] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In ICCV, 2015

[3] Soomro, Khurram & Roshan Zamir, Amir & Shah, Mubarak. (2012). UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. CoRR.

[4] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 1725-1732.

[5] Kataoka, H., Satoh, Y., Aoki, Y., Oikawa, S., & Matsui, Y. (2018). Temporal and Fine-Grained Pedestrian Action Recognition on Driving Recorder Database. Sensors (Basel, Switzerland), 18(2), 627.

[6] Piergiovanni, AJ & S. Ryoo, Michael. (2018). Fine-grained Activity Recognition in Baseball Videos.

[7] Wang, Limin & Xiong, Yuanjun & Wang, Zhe & Qiao, Yu & Lin, Dahua & Tang, Xiaoou & Van Gool, Luc. (2016). Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. 9912. 10.1007/978-3-319-46484-8_2.

[8] Huang, J., Zhou, W., Zhang, Q., Li, H., & Li, W. (2018). Video-Based Sign Language Recognition Without Temporal Segmentation. *AAAI*.

[9] Hochreiter, Sepp & Schmidhuber, JÃ¼rgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

[10] Liu W. et al. (2016) SSD: Single Shot MultiBox Detector. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905. Springer, Cham

[11] Girshick, R.B., Donahue, J., Darrell, T., & Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 580-587.

[12] Yue-Hei Ng, Joe & Hausknecht, Matthew & Vijayanarasimhan, Sudheendra & Vinyals, Oriol & Monga, Rajat & Toderici, George. (2015). Beyond Short Snippets: Deep Networks for Video Classification. Cornell Univ. Lab..

[13] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818-2826.

[14] Ye, Hao & Wu, Zuxuan & Zhao, Rui-Wei & Wang, Xi & Jiang, Yu-Gang & Xue, Xiangyang. (2015). Evaluating Two-Stream CNN for Video Classification. 10.1145/2671188.2749406.