Species Classification in Thermal Imaging Videos

1st Hamish Duncanson
School of Computer Science
University of Auckland
Auckland, New Zealand
hdun603@aucklanduni.ac.nz

2nd Charles Tremlett
School of Computer Science
University of Auckland
Auckland, New Zealand
ctre019@aucklanduni.ac.nz

3rd Saahil Deshpande
School of Computer Science
University of Auckland
Auckland, New Zealand
sdes343@aucklanduni.ac.nz

4th Yahui Cai
School of Computer Science
University of Auckland
Auckland, New Zealand
ycai745@aucklanduni.ac.nz

5th Goutham Menon
School of Computer Science
University of Auckland
Auckland, New Zealand
gmen233@aucklanduni.ac.nz

***Abstract*—We examine a labelled dataset of thermal imaging recordings of animals in New Zealand bush, for the purpose of developing models which can automate the labelling of species. After approaching the problem with generic video classification models, we propose two novel contributions:**

- **This dataset differs from conventional video classification benchmarks, in that the subject of the recording (the animal) generally only occupies a small portion of the field of view. We address this by cropping each individual frame around the animal, and extracting the movement of the cropped region to retain information that would otherwise be lost in the cropping procedure. We show that this cropping can be implemented with a simple temperature threshold applied to the thermal images, requiring no additional expert information.**
- **We adapt well known image augmentation techniques so that they are suited for the domain of video classification, and investigate which combination of transformations is most performant.**

Experiments on the provided dataset support the efficacy of these contributions. We present a Inflated 3-D Convolutional Network (I3D), which combined with these techniques, can correctly label the species of over 90% of the thermal imaging recordings.

***Index Terms*—thermal imaging, video classification, species classification, computer vision**

I. Introduction

Many of New Zealand's native species are threatened by predation from introduced predators [15]. These non-indigenous species include, but are not limited to, rodents, stoats, possums, and feral cats. A challenge in predator control is the deployment of traps which have a high interaction rate, with many trapping methods being largely ignored by their targets [13]. The interaction rate can be increased by the use of a call or lure designed to attract a specific species. However, as there are generally multiple target species in a trapping area, it is not known which species will be present near a particular trap. To solve this problem, the Cacophony Project [14], a non-profit organisation, is attempting to design a trap which can identify target species in its vicinity, and then release the appropriate lure.

The Cacophony Project has identified that thermal imaging cameras are suitable for capturing recordings of pest species in the New Zealand bush, and that the performance of these cameras greatly exceeds the typical trail cameras used for capturing images of large game. The objective of this project is to design an algorithm which can identify the species present in a short video clip from a thermal imaging camera with a high degree of accuracy.

The Cacophony Project has kindly shared a large database of videos captured with thermal cameras. Sample frames from these videos are shown in Fig. 1. We assess whether existing approaches to video classification are appropriate for this data set, and identify possible improvements which can be made. This task falls under video classification, which has been widely researched in other computer vision contexts.

II. Hypotheses

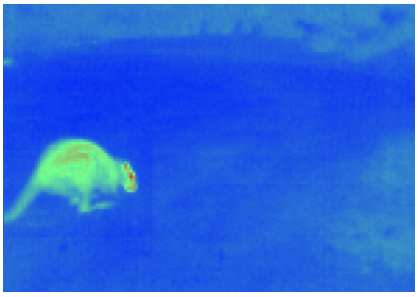
We propose that we can improve the performance of state-of-the-art algorithms on the species identification data set with the following contributions:

1) Cropping individual frames around the animal using a temperature threshold, and extracting the movement of the cropped region to retain information that would otherwise be lost during the cropping.
2) By adapting image augmentation techniques to be used in the domain of video classification.

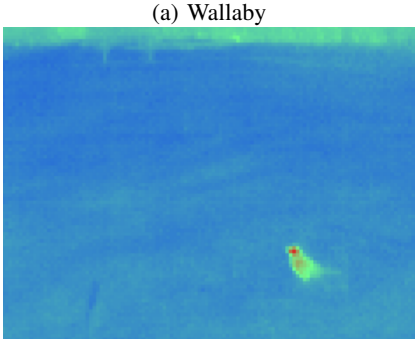
III. Related Research

The objective of the project falls under video classification, for which there is a large body of prior research. However, video classification is still very much an open problem. There is as of yet no clear consensus on the state-of-the-art framework, and accuracy on benchmark tasks continue to see frequent improvements [16]. Furthermore, a thorough search of the literature found few papers proposing methods specifically tailored for the classification of thermal imaging videos, and we did not find any previous work regarding classifying animal species from thermal images.

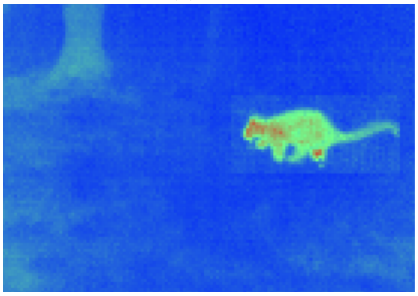
Shi et al. [3] propose using convolutional long short-term memory (convLSTM) layers to classify videos, which embed 2-dimensional convolutions into the transitions of an LSTM



(a) Wallaby



(b) Bird



(c) Possum

Fig. 1: Examples of single frames taken from thermal imaging videos.

module. This produces models that can interpret the image features within each frame, as well as the temporal correlations between frames, while being trained end-to-end.

Donahue et al. [1] propose a similar framework called Long-term Recurrent Convolutional Networks (LRCNs). These also utilize LSTMs, but first use a Convolutional Neural Network (CNN) to extract features from the individual frames in a video, which are then used as inputs to the LSTM, rather than the frames themselves.

Bourdev et al. [2] propose a simple approach to video classification which is not recurrent, instead using 3-dimensional convolutional layers. These can capture the temporal features of a video by convolving across the time dimension. They use these layers to construct a compact video classification model named C3D which can perform inference as much as an order of magnitude faster than competing models. Carreira et al. [16] extended C3D by using 3-dimensional convolutional and pooling layers to translate CNN architectures designed for image inputs into architectures suitable for video inputs. Their proposed I3D model shows very strong results on a number of video classification benchmarks.

Ng. et al. [8] propose a framework similar to LRCN, but instead of using an LSTM to interpret the frame-level features, they apply a max-pooling operation, followed by fully connected layers which make the classification. A surprising insight is that this method is temporally invariant, as shuffling the frames of a video generates exactly the same output. They also demonstrate that including optical flow (a hand-crafted input channel encoding pixel level movements between frames), can improve accuracy. Generating the optical flow involves a computationally expensive preprocessing step, although Tang et al. [9]. show that this can be mitigated with their "motion hallucination" networks. These approximate optical flow with a neural network, thereby accelerating inference.

Specific research into classifying thermal imaging videos is scarce. Janssens et al. [12] used CNNs to monitor rotating machinery, although do not appear to have tested modern video classification models on their dataset. Batchuluun et al. [17] use LRCNs to automate surveillance in dark environments, and demonstrate the efficacy of a handful of preprocessing techniques on their dataset, including cropping the videos, and using pose estimation techniques to generate additional input channels.

In summary, many of these video classification frameworks include integrated CNN architectures [1], [5], [8]. This allows them to leverage the extensive recent successes in image classification research [4], and also enables the models to be pre-trained on large image datasets (such as ImageNet). In some cases, this use of transfer learning was seen to improve accuracy, or to at least accelerate training [1], [8].

Overall it appears that the I3D model [16] produces the state-of-the-art results on many benchmark tasks, suggesting this is an appropriate starting point for future research.

Although some research suggests that using hand-crafted features such as optical flow or temporal differences as additional inputs improves performance [9], [10], other papers cast doubt on the necessity of these features [3], [7].

Augmenting videos with transformations is an often unused technique for video classification, with only a few papers utilizing horizontal flipping and cropping to augment their dataset [1], [8], which is still only a subset of the augmentation methods often seen in image classification algorithms [4].

There have been few papers specifically researching the classification of thermal imaging recordings [12], [17], and these have generally only involved generic video classification techniques, without exploiting the unique characteristics of thermal images.

A promising avenue for further research is in automating the cropping of videos around the region of interest, with the VideoLSTM model proposed by Li Et al. [6] demonstrating high performance on action recognition tasks. Batchuluun et al. reached similar findings with their thermal surveillance dataset [17], although their approach of manually cropping the videos is not a scalable technique.

IV. Methodology

A. Data Preparation

The full data set consists of 15,465 recordings, covering 18 classes. Each video has a resolution of 180x120, is captured at 9 frames per second, and contains a single channel consisting of the raw temperature readings.

In order to give our model a reasonable chance of identifying the target animal, we included only videos with at least 5 seconds of capture time. The rationale behind this restriction is related to the ultimate purpose for which this algorithm is designed, that is, to deploy an effective trap or lethal response to capture or kill a predator with a high degree of certainty. For videos longer than 5 seconds, we trimmed the videos to the 5 second period when the animal was closest to the camera. This was effected by selecting the sequence of 45 consecutive frames with the highest cumulative mass, where mass is the number of pixels occupied by the animal. As well as reflecting the objective of this project, standardising the length of the recordings in the dataset was convenient for training models which assume a fixed input length.

The raw temperature readings are poorly scaled and inappropriate for use with neural networks, which generally train faster with normalised inputs [4]. We preprocessed the temperature readings in three different ways, generating three input channels:

1) Min-max normalising the temperatures across the dataset
2) Min-max normalising the temperatures within each frame
3) Temperature readings after subtracting the background temperature, where the background is the frame recorded just before the animal entered the field of view.

These three channels provide complementary information, and were each found to contribute to improved performance. This justifies including all three channels as input. An example of these image normalisation techniques can be found in Fig. 2.

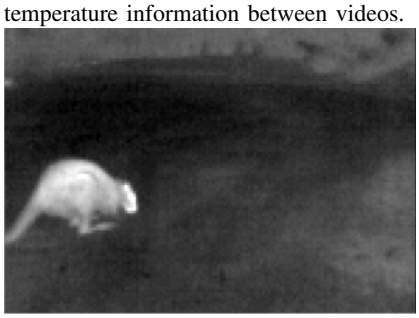
Another benefit to decomposing the raw temperature readings into three channels is this facilitates the re-use of generic model architectures which typically expect three channels as input (representing the primary colours "RGB"). This also enables transfer learning from models pre-trained on colour image datasets.

We made a number of adjustments to the original classes, to remove those which did not classify the target, and either combine or remove classes with a very small number of instances. The original and final class distributions are detailed in Table I and Table II. After discarding the videos from classes we chose to remove, and videos shorter than 5 seconds, the total number of instances in the final data set is 10,664.

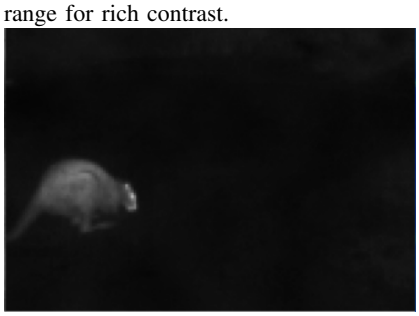
B. Metadata

The provided dataset consisted of the thermal recordings, as well as a handful of pieces of metadata associated with each video. This metadata included the longitude and latitude of the camera, the time of day, and the month of the year the recording took place.

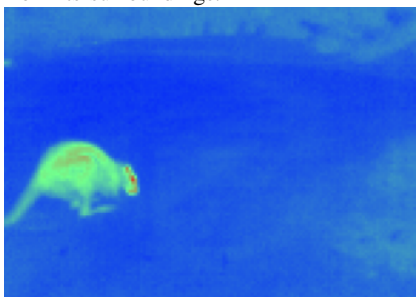
(a) The temperatures normalised across the dataset, which preserves the relative temperature information between videos.



(b) The temperatures normalised within the frame, which maximises dynamic range for rich contrast.



(c) The temperatures after subtracting the background, which isolates the animal from its surroundings.



(d) Collating the three channels into a "pseudo-RGB" image. This is the preprocessed input used throughout this project.

Fig. 2: An example of channels present in the raw thermal imaging videos, and their collation.

TABLE I: Class distribution in the raw data set.

Class	Count	Action
rodent	3990	Retained
false-positive	2928	Retained [a]
possum	1262	Retained
unknown	1238	Removed [b]
hedgehog	1175	Retained
bird	1059	Retained
wallaby	901	Retained
leporidae	778	Retained
cat	692	Retained
mustelid	616	Retained
insect	410	Retained
human	142	Retained
dog	102	Retained
sheep	80	Retained
part	61	Removed [b]
bird/kiwi	22	Merged with class "bird" [c]
poor tracking	8	Removed [b]
sealion	1	Removed [d]

[a] The false-positive class was retained in the data set, as it is important the model can identify a false positive recording triggered by wind or other sources of anomalous movement.
[b] These classes were removed, as they did not contain sufficient information to describe what was observed. Note that the "unknown" class, which was removed, is different to the "false-positive" class, which was retained.
[c] The "bird/kiwi" class was merged with the "bird" class. Kiwi are not the only nocturnal ground-dwelling bird, and the response triggered by the model (to deploy a trap or not) should be identical.
[d] As there is only one instance of the "sealion" class in the data, we have removed it from the data set.

TABLE II: Class distribution in the final data set.

Class	Count
rodent	3018
false-positive	1852
possum	1105
hedgehog	1047
bird	871
wallaby	779
cat	606
leporidae	599
mustelid	371
insect	215
human	71
dog	70
sheep	60

Although useful for classification, we decided to omit this information from our analysis. This is in line with the objectives of the project, since our eventual algorithm is intended to function effectively with a camera placed anywhere in New Zealand's bush. Therefore it would be undesirable to include the location metadata, since this would cast doubt on the generalisability of the models to new locations. This concern is exacerbated since many of the locations in the dataset were found to have severely imbalanced classes, which suggests a model being provided with this information might tend to overfit to the particular set of locations, rather than learning general patterns about the distribution of different species across New Zealand.

Including the time of day or time of year metadata would be associated with these same problems, since several cameras in the dataset were only operating for a few months, and so the time of year is correlated with the location information which we are meaning to exclude.

C. Training Procedure

1) Train/Validate/Test Split: We split the data into three sets, for training (7,664 instances), validation (1,500 instances), and testing (1,500 instances). The testing set is also referred to as the hold-out set, and is not used for training or validation. Stratified sampling was applied such that the three sets each exhibit a similar class distribution.

Models were trained on the training set, and evaluated on the validation set at the end of each epoch. The validation accuracy was used to assess training progress, to reduce the learning rate upon plateauing, and to terminate training in the absence of recent progress.

After the termination of training, the model state at the end of the epoch with the highest accuracy rate against the validation set was retained as the best model. The accuracy of the test (hold-out) set evaluated against this model was used to compare models against one other, and all accuracies presented in this report relate to this same hold-out set.

2) k-fold Cross-Validation: In order to better assess the performance of models against each other during training, we used k-fold cross-validation. With this technique, we combined the training and validation data sets, and then partitioned the data into k sets. Each model was run k times, with a different partition used as the validation set each run.

We set $k = 5$ for all cross-validation, which obtained stable results. A visualisation of our use of k-fold cross-validation can be found in Fig. 3.

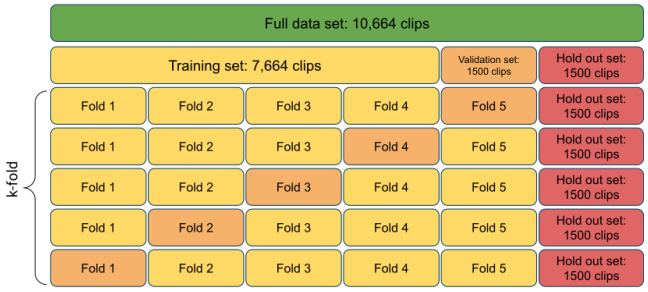


Fig. 3: Illustration of train/validate/test splits, and k-fold training procedure.

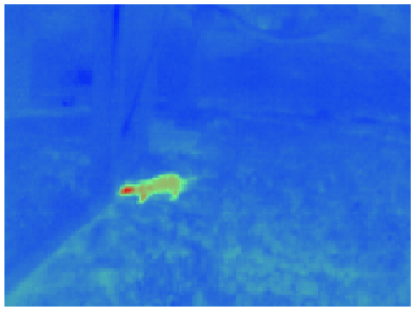
3) Hyperparameter Optimisation: For the best performing model types, we used a grid search to identify the best hyperparameters, including layer size, number of layers, and amount of regularisation applied through dropout.

D. Cropping

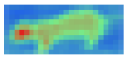
A key property of thermal images is that animals can be easily isolated from the surroundings, since both birds and mammals are warm-blooded and so tend to produce higher temperature readings than the background. We found that this property provides an effective mechanism for cropping the

recordings around the animal, by simply placing a rectangle around the pixels where the raw temperature readings exceeded a threshold. This threshold is chosen for each camera according to the typical temperature of its field of view. This cropping procedure was applied to each frame in a video individually, to produce a video tightly centred on the animal, regardless of its nearness to the camera. See Fig. 4 for an example of our cropping procedure.

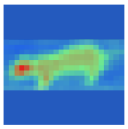
In order to use these cropped videos as inputs, it was necessary to standardise their resolution. This was implemented by first "letterboxing" the cropped regions to make them square (using the minimum temperature as the padding colour), and then using bilinear interpolation to scale each frame to a resolution of 24x24. Since the animals in the dataset were generally relatively distant from the camera, this usually meant that the frames were upscaled, which prevents losing information during this preprocessing.



(a) An uncropped frame.



(b) The frame cropped around the mustelid.



(c) The cropped region after letterboxing and scaling to 24 x 24 resolution.

Fig. 4: An example of the application of cropping to a frame in a video.

E. Movement Data

A disadvantage of the cropping is that much of the information describing the movement of the animal is lost, since it is challenging to ascertain movement patterns from a video that is tightly cropped around the subject. We hypothesize that this loss of information can be remedied by incorporating the movement of the cropped region as an input to the models.

Our approach is to extract and normalise three pieces of information at each time step:

1) The size of the cropped region
2) Horizontal velocity
3) Vertical velocity

To justify this, we first ran a series of experiments to establish whether this movement data was useful for classification. We trained LSTM models to classify the species, each given only one of these three pieces of information as input.

TABLE III: Performance of movement features on the hold-out data set.

Movement Feature	Hold-out Accuracy
Size of the cropped region	41.6%
Horizontal velocity	52.9%
Vertical velocity	48.1%
Combined movement features	62.7%

The size of the cropped region alone does not provide sufficient information to accurately identify the animals, only marginally outperforming a simple majority classifier. This is unsurprising, as the size of the animal itself cannot be ascertained since we do not know how close the animal is to the camera.

The horizontal velocity is more informative than the vertical velocity, which likely reflects the fact that most species are ground-dwelling and so exhibit little vertical movement.

An LSTM provided with all three of these types of movement information gives encouraging results, confirming that this data is indeed useful for classification, and can be effectively interpreted with an LSTM. Note that this model still falls short of all of our video classification models, although we posit that this information could be used to supplement the cropped videos as an additional input, enabling better informed classifiers.

F. Video Augmentation

We applied online data augmentation with four possible transformations for each video: horizontal flipping, acute random rotations, random cropping, and translocating. This augments the training set with additional examples, which is a technique widely used in the domain of image classification to improve generalisability [4]. We applied the same transformation to each frame in a given video, to ensure the movement remained coherent.

Appendix A details the improvement given by the application of different types and combinations of augmentation.

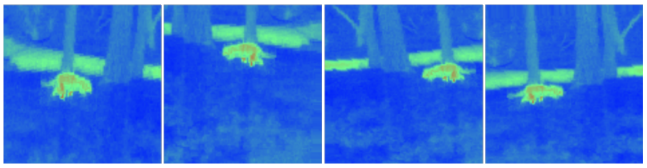


Fig. 5: Four different transformations applied to a frame extracted from a recording.

G. Model Design

We developed model architectures based on four state-of-the-art research papers. These are ConvLSTM (Convolutional

LSTM) [3], LRCN (Long-term Recurrent Convolutional Networks) [1], 3D-Conv (3D ConvNet) [2], and I3D (Inflated 3D ConvNet) [16]. We used these four models as baselines, which we sought to improve with our novel contributions.

1) Convolutional LSTM: Shi et al. [3] propose a convolutional LSTM layer, which is built on the LSTM framework, with convolutional structures embedded into the transitions. This produces models which have the 2-dimensional convolutions typically used in image classification, as well as a recurrent component that can capture patterns between frames in a video. We found that stacking a second layer improved results, although this improvement quickly saturated with additional layers, so our chosen implementation utilized two convolutional LSTM layers.

2) Long-term Recurrent Convolutional Networks: Donahue et al. [1] propose Long-term Recurrent Convolutional Networks (LRCNs), which function by first using a Convolutional Neural Network (CNN) to extract features from each frame in a video individually. An LSTM then takes the features as inputs, and outputs a classification for a given video. We opted to use ResNet-18 as the CNN, which is a widely-used model for extracting features from images [4].

3) 3D Convolutional Networks: Bourdev et al. [2] propose using 3-dimensional convolutional layers for video classification. These layers convolve across a third dimension - in the context of video classification, this is the time dimension - and so can capture both the visual features and the temporal features of a video. We reimplemented the "C3D" model described in the paper.

4) Inflated 3D Convolutional Networks: Carreira et al. [17] propose utilizing 3-dimensional convolutional and pooling layers to convert CNN's designed for image classification into models suitable for video inputs, and suggest using 'GoogLeNet' as a base CNN. This can be seen as an advancement of the original Conv3D model.

On performing a detailed review of the I3D model, we determined that the architecture was optimised for videos of a much higher resolution than those from the thermal imaging cameras. However, we found that prepending the model with an up-sampling layer averted this issue, and so our following I3D results all use this small addition. There were a number of hyperparameters needing optimisation for this model, which we solved with a grid search and k-fold cross validation, documented in Appendix B.

H. Integrating Movement of the Cropped Region

We experimented with a number of different approaches to integrating the movement inputs into these models:

1) Encoding the movement information for each frame as pixels, and pasting these onto the videos.
2) Separately training an LSTM with the movement information, and then using this with the video model as an ensemble.
3) Modifying the video models to take an extra vector of inputs with each frame, so that the movement information could be fed into the models directly.

We found that the third approach was the most effective. Our specific implementation was to remove the last layer of the given video model so that it extracts a set of features from the video rather without classifying it, and then using an LSTM to extract a set of features with a matching shape from the movement inputs. These two sets of features are then concatenated, and the last layer of the model is finally applied to make the classification.

For example, from a single video, the I3D model extracts a set of features with shape (512, 9). We then use an LSTM over the movement inputs to extract 512 features from each of the 45 frames, and apply average pooling to aggregate the movement information into shape (512, 9). Finally, the two sets of features are concatenated into shape (1024, 9), before a fully connected layer outputs a classification.

The success of this approach is presented in the Results section.

V. Results

A. Baseline Models

We first present the results obtained on the dataset using our implementations of the four state-of-the-art models without any of our contributions. These numbers comprise the baselines which we sought to improve upon.

TABLE IV: Baseline model results.

Model	Hold-out Accuracy
ConvLSTM	69.3%
LRCN	79.3%
3D-ConvNet	72.4%
I3D	85.9%

B. Cropping

We achieved substantial improvements in accuracy with these models by cropping around the animal in each frame of the videos. This confirms our hypothesis that models benefit from receiving inputs that more prominently feature the animal, rather than the background context. As these videos are captured at night, the background of the thermal images provides little signal, and is unlikely to add any contextual information to improve the classification.

TABLE V: Improvement over baseline with cropping.

Model	Hold-out Accuracy
ConvLSTM	72.3% (+3.0)
LRCN	80.5% (+1.2)
3D-ConvNet	78.6% (+6.2)
I3D	86.6% (+0.7)

C. Movement Information

We found that integrating the movement of the cropped region as a complementary input into the models was an effective approach for retaining the movement information that is lost in the cropping procedure. Including these inputs

produced a substantial improvement in accuracy above simply feeding the models the cropped videos.

TABLE VI: Improvement over just cropping with addition of movement features.

Model	Hold-out Accuracy
ConvLSTM	74.9% (+2.6)
LRCN	85.1% (+4.6)
3D-ConvNet	80.4% (+1.8)
I3D	90.2% (+3.6)

D. Video Augmentation

Testing different combinations of online augmentations revealed that applying only horizontal flipping and random rotations produced the best results. Some details of this analysis can be found in Appendix A. We found that this data augmentation technique achieved modest improvements in the accuracy of our baseline models, however improvements because less significant as the performance of the baseline model improved.

TABLE VII: Improvement over just cropping with the use of video augmentation.

Model	Hold-out Accuracy
ConvLSTM	72.8% (+0.5)
LRCN	81.4% (+0.9)
3D-ConvNet	79.4% (+0.8)
I3D	86.4% (-0.2)

E. Combining Contributions

Finally, we combined all three of these techniques (cropping, movement inputs, and video augmentation) to obtain our final results. The contributions appear to have a complementary effect, since using them in tandem gives the best results for all of the models. In the case of 3D-ConvNet, the combined improvement over the baseline was actually marginally greater than the sum of the improvements separately, highlighting the synergies of our contributions.

TABLE VIII: Improvement over baseline with the application of all contributions.

Model	Hold-out Accuracy
ConvLSTM	75.3% (+6.0)
LRCN	85.2% (+5.9)
3D-ConvNet	82.1% (+9.7)
I3D	91.6% (+5.7)

F. Analysis of Final Results

The classification results of our best model (I3D with our contributions) provide insights into which classes pose the greatest challenges. Refer to Table IX and Table X.

The most commonly confused classes are "cat" with "possum", and "hedgehog" with "rodent". This is unsurprising given the similarity in appearance and movement patterns of these species. Fortunately, these are all pest species, so from the perspective of conservation, there is probably little harm in these misclassifications.

Overall the poorest performing class is "sheep", where the model achieved precision and recall scores of just 0.67 and 0.44 respectively. This reflects the limited data for this class, which has only nine instances in the hold-out set, and 60 in the complete data set. However, this is of limited concern since feral sheep are relatively rare and are not a pest species targeted by traps.

TABLE IX: Classification report on hold-out set from the best I3D model.

Class	Precision	Recall	F1-score	Support
Bird	0.91	0.91	0.91	123
Cat	0.89	0.87	0.88	85
Dog	1.00	0.70	0.82	10
False-positive	0.95	0.98	0.97	260
Hedgehog	0.91	0.92	0.91	147
Human	0.78	0.70	0.74	10
Insect	0.79	0.77	0.78	30
Leporidae	0.85	0.90	0.88	84
Mustelid	0.85	0.88	0.87	52
Possum	0.88	0.87	0.87	156
Rodent	0.94	0.94	0.94	424
Sheep	0.67	0.44	0.53	9
Wallaby	0.94	0.92	0.93	110

To investigate how often this model makes "costly" misclassifications (labelling a pest as a non-pest, or a non-pest as a pest), we also evaluated the model on a simpler binary classification problem, to classify a recording as a pest species or a non-pest species (see table XI). With our contributions, the I3D model was able to attain 97.6% accuracy on the hold-out set for this problem.

VI. Future Work

A. Augmenting the cropped regions' movement information

Although we thoroughly experimented with different combinations of transformations to apply as video augmentation, we did not transform the movement of the cropped region. This leaves open the possibility to increase the generalisability of the movement information by augmenting the training set with transformed instances (for example, negating the horizontal velocity of the cropped region is analogous to horizontally reflecting the original video). This could have the potential to derive further improvements from our movement inputs contribution.

B. Investigating optical flow

Optical flow is an additional input channel generated from the pixel level movements between frames. Although it does not introduce information beyond what is already obtainable from the video inputs, some recent research has suggested that it can improve performance of video classification models [9], [16]. If these results transfer effectively to our thermal

TABLE X: Confusion matrix on hold-out set for best I3D model. The rows indicate the true class of the sample, while the columns show the predicted class.

	Bird	Cat	Dog	False-positive	Hedgehog	Human	Insect	Leporidae	Mustelid	Possum	Rodent	Sheep	Wallaby
Bird	112	0	0	0	0	0	0	2	0	3	5	0	1
Cat	0	74	0	0	0	0	0	0	1	7	2	0	1
Dog	0	1	7	0	0	0	0	0	0	1	0	1	0
False-positive	0	0	0	254	0	0	6	0	0	0	0	0	0
Hedgehog	0	2	0	1	135	0	0	1	0	1	6	1	0
Human	0	1	0	0	0	7	0	0	1	1	0	0	0
Insect	0	0	0	7	0	0	23	0	0	0	0	0	0
Leporidae	1	0	0	1	1	0	0	76	2	0	2	0	1
Mustelid	0	0	0	0	0	1	0	0	46	0	5	0	0
Possum	3	1	0	0	7	1	0	0	1	135	4	0	4
Rodent	7	2	0	0	6	0	0	6	3	2	398	0	0
Sheep	0	1	0	1	0	0	0	0	0	2	1	4	0
Wallaby	0	1	0	2	0	0	0	4	0	2	0	0	101

TABLE XI: Assignment of classes to ”pest” and ”non-pest” meta-classes.

Pest	**Non-pest**
Possums	Birds
Rodents	Humans
Mustelids	Insects
Rabbits	False-positive
Hedgehogs	Sheep
Wallabies	
Cats	
Dogs	

recordings dataset, this could be a promising direction for future research.

One complication is that conventional approaches to computing the optical flow are not efficient enough for the real-time inference that would be needed for a trapping mechanism. One potential solution suggested by Tang et al. [9] is to train a neural network to approximate the optical flow channel for faster inference.

C. Generalisation to new data

Ideally, our models will remain just as effective at classifying species when tested on similar data sets of thermal imaging videos from the New Zealand bush. This ability to generalise is important for the practical application of this technology for conservation purposes.

If additional data sets are captured by the Cacophony project, for example from different locations or over a different time period, we would like to assess whether our models retain their level of accuracy.

D. Searching for a more efficient model architecture

Although I3D was found to be a highly accurate classifier, it is also computationally expensive, consisting of over 13 million parameters. This poses a practical problem for implementing the algorithm in a trap, since real-time classification is needed for the trap to quickly respond to a nearby pest. Real-time classification with this model would require a GPU to accelerate the neural network's inference, which unfortunately means that wide-scale adoption of this trapping technology might be prohibitively expensive. This suggests that a more efficient variant of this model, without significantly compromised accuracy, would be desirable. We posit that such an architecture could be discovered using neural architecture search methods, similar to the approach employed by Piergiovanni et al. [7] to evolve their “Tiny Video Networks”.

E. Further investigation into transfer learning

Throughout this project, we attempted to use models pre-trained on conventional image and video databases, which is known to produce models that train faster and with better generalisability [1], [8]. However, we consistently found that although pre-trained models were generally better after a single epoch, their training did not reach a higher asymptote, and so we did not reap any benefits from our attempts at transfer learning.

This finding can be attributed to the fact that images/videos from conventional datasets contain three channels corresponding to “RGB”. Although our dataset also contains three channels, these represent different methods of parsing the temperature readings (as discussed in the Data Preparation section), rather than the three primary colours. We hypothesize that this explains why transfer learning was ineffective.

To rectify this, we could try to find a large thermal imaging dataset so that we can transfer knowledge from a more closely related domain. Alternatively, it might be the case that our preprocessing technique of decomposing the raw temperature readings into three channels can be modified so that the resulting images are more similar to the images used for the pre-training.

F. Semi-supervised learning

Although this dataset was entirely labelled, if thermal cameras were ever to be deployed into New Zealand bush at a wide scale, it would quickly become impractical to have experts labelling every animal recording. This means that a future avenue of improvement would be to obtain a larger dataset (albeit not requiring labels), and investigate a semi-supervised learning approach to improve classification

accuracy. This would provide a scalable strategy for updating the model, since a very large number of videos could be processed without needing human input. Jing et al. [11] have proposed a promising algorithm for semi-supervised learning of video classification problems with as little as 10% of the labels, which could likely be adapted for this dataset.

VII. Conclusions

We have extended generic video classification techniques so that they are better suited for automating the labelling of animal species in thermal imaging recordings in New Zealand bush. We found that cropping each frame around the animal with a temperature threshold produced video inputs that more prominently featured the animal rather than the background context, improving the models' performance.

We then found that by extracting the movement of the cropped region at each frame, and passing this into the models as an additional input, we could retain movement information that would otherwise be lost in the cropping procedure, further improving accuracy.

We also showed that applying video augmentation, inspired by the transformations which are common-place in the domain of image classification, provided modest improvements in performance (although this benefit was fairly limited when used with optimised models).

We finally show that there is synergy between these contributions, and present a Inflated 3-D Convolutional Network (I3D), which combined with these techniques, can correctly label the species of 91.6% of the thermal imaging recordings across 13 classes.

Although this model has some difficulties distinguishing between the various pest classes, these misclassifications are largely inconsequential from a trapping perspective, and so we are confident our final model could be considered accurate enough for real-world use.

Appendices

A. Video augmentation benchmarking with 5-fold cross-validation.

We tested a number of combinations of video augmentation techniques, including horizontal flips, random rotations, random cropping and translocations. Even with 5-fold cross-validation (see Fig. 6), it is difficult to identify which combination was the most effective. We opted to use only horizontal flips and random rotations, since this combination is simple, and is seen to be competitive or better than the more complicated augmentation schemes.

B. Hyperparameter Optimisation Results (I3D)

We tested a small grid search over reasonable hyperparameters to optimise the I3D model (see Fig. 7). This included the width of the layers, the number of LSTM units, and amount of regularisation applied through dropout.

Acknowledgement

We sincerely thank the Cacophony Project for the use of their high quality data set.

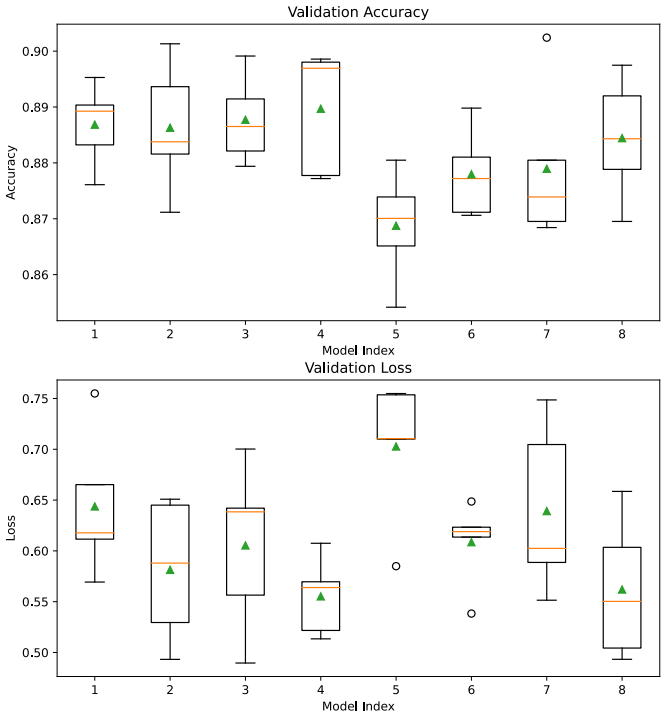


Fig. 6: 5-fold cross-validation over different combinations of video augmentation techniques.

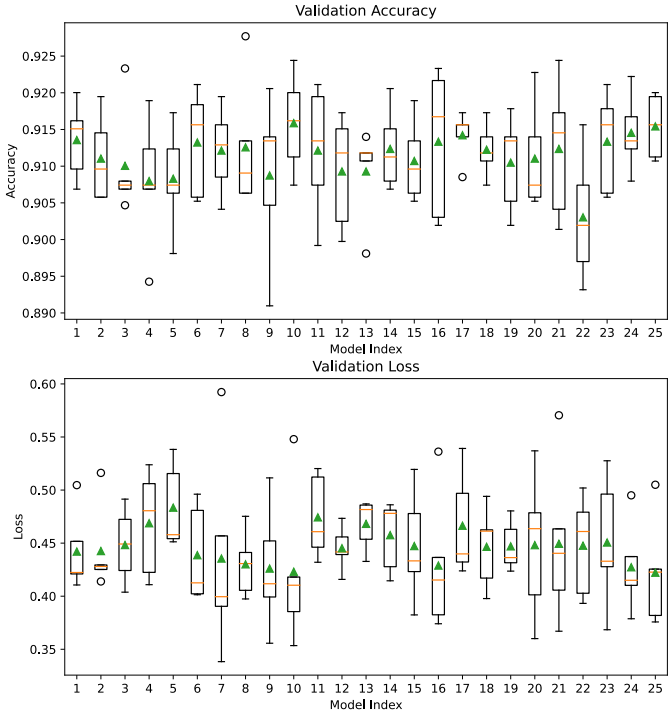


Fig. 7: 5-fold cross-validation over a sample of different I3D model architectures.

REFERENCES

[1] Donahue, Jeff & Hendricks, Lisa & Guadarrama, Sergio & Rohrbach, Marcus & Venugopalan, Subhashini & Darrell, Trevor & Saenko, Kate. (2015). Long-term recurrent convolutional networks for visual recognition and description. 2625-2634. 10.1109/CVPR.2015.7298878.

[2] Bourdev, Lubomir & Fergus, Rob & Torresani, Lorenzo & Paluri, Manohar. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. 4489-4497. 10.1109/ICCV.2015.510.

[3] Shi, Xingjian & Chen, Zhourong & Wang, Hao & Yeung, Dit-Yan & Wong, Wai Kin & WOO, Wang-chun. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting.

[4] He, Kaiming & Zhang, Xiangyu & Ren, Shaoqing & Sun, Jian. (2015). Deep Residual Learning for Image Recognition. 7.

[5] Wang, Heng & Torresani, Lorenzo & Ray, Jamie & LeCun, Yann & Paluri, Manohar. (2018). A Closer Look at Spatiotemporal Convolutions for Action Recognition. 6450-6459. 10.1109/CVPR.2018.00675.

[6] Li, Zhenyang & Gavves, Efstratios & Jain, Mihir & Snoek, Cees. (2016). VideoLSTM Convolves, Attends and Flows for Action Recognition. Computer Vision and Image Understanding. 10.1016/j.cviu.2017.10.011.

[7] Piergiovanni, AJ & Angelova, Anelia & Ryoo, Michael. (2019). Tiny Video Networks.

[8] Ng, Joe & Hausknecht, Matthew & Vijayanarasimhan, Sudheendra & Vinyals, Oriol & Monga, Rajat & Toderici, George. (2015). Beyond short snippets: Deep networks for video classification. 4694-4702. 10.1109/CVPR.2015.7299101.

[9] Tang, Yongyi & Ma, Lin & Zhou, Lianqiang. (2019). Hallucinating Optical Flow Features for Video Classification. 926-932. 10.24963/ijcai.2019/130.

[10] Ng, Joe & Davis, Larry. (2018). Temporal Difference Networks for Video Action Recognition. 1587-1596. 10.1109/WACV.2018.00176.

[11] Jing, Longlong & Parag, Toufiq & Wu, Zhe & Tian, Yingli & Wang, Hongcheng. (2020). VideoSSL: Semi-Supervised Learning for Video Classification.

[12] Janssens, Olivier & Van de Walle, Rik & Loccufier, Mia & Hoecke, Sofie. (2017). Deep Learning for Infrared Thermal Image Based Machine Health Monitoring. IEEE/ASME Transactions on Mechatronics. 23. 151 - 159. 10.1109/TMECH.2017.2722479.

[13] Brown, Samantha & Warburton, Bruce & Ekanayake, Jagath & Hough, Steve. (2013). Using radio frequency identification technology to measure possum interaction rates with traps. Kararehe Kino - Vertebrate Pest Research. Issue 22. 20 - 21.

[14] Moore's Law for New Zealand's birds — The Cacophony Project (2020). Retrieved from https://cacophony.org.nz/

[15] Animal Pests and Threats A-Z: Threats and Impacts (2020). Retrieved from https://www.doc.govt.nz/nature/pests-and-threats/animal-pests/

[16] Carreira, Joao & Zisserman, Andrew. (2018). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset.

[17] Batchuluun, Ganbayar & Nguyen, Dat Tien & Pham, Tuyen Danh & Park, Chanhum & Park, Kang Ryoung. (2019). Action Recognition from Thermal Videos. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2931804.