scientific reports

OPEN

Temporal action localisation in video data containing rabbit behavioural patterns

Semyon Ilin[1], Julia Borodacheva[2,3], Ildar Shamsiev[2,3], Igor Bondar[2,3] & Yulia Shichkina[1,3]✉

In this paper we present the results of a research on artificial intelligence based approaches to temporal action localisation in video recordings of rabbit behavioural patterns. When using the artificial intelligence, special attention should be paid to quality and quantity of data collected for the research. Conducting the experiments in science may take long time and involve expensive preparatory work. Artificial intelligence based approaches can be applied to different kinds of actors in the video including animals, humans, intelligent agents, etc. The peculiarities of using these approaches in specific research conditions can be of particular importance for project cost reduction. In this paper we analyze the peculiarities of using the frame-by-frame classification based approach to temporal localisation of rabbit actions in video data and propose a metric for evaluating its consistency. The analysis of existing approaches described in the literature indicates that the aforementioned approach has high accuracy (up to 99%) and F1 score of temporal action localisation (up to 0.97) thus fulfilling conditions for substantial reduction or total exclusion of manual data labeling from the process of studying actor behaviour patterns in video data collected in experimental setting. We conducted further investigation in order to determine the optimal number of manually labeled frames required to achieve 99% accuracy of automatic labeling and studied the dependence of labeling accuracy on the number of actors presented in the training data.

Keywords Automatic detection, Video data, Rabbit, Behavior, Neural networks, Actor

Behaviour is the interface between the animal's brain and the environment, allowing for rapid adaptation to its conditions and, as a consequence, determining the organism's chances for survival and reproduction. One of the tasks of the nervous system is to choose behavioural programs most fitting to the present state and future goals of the organism. This choice is influenced by an enormous variety of factors unequally distributed in time and space. Therefore, accurate analysis of behaviour plays a key role in biological research at various levels of complexity.

The most widely used animal behaviour recording method in laboratory research is video recording. Frame-by-frame analysis of the animal behaviour is especially important when studying behavioural events lasting less than a second. In particular, determination of the precise temporal boundaries of behavioural patterns is of utmost importance when searching for correlations with neural activity. Some brain structures exhibit rapid changes in relation to behaviours executed by the animal itself, or a conspecific, or another animal. Comparison of neural activity with animal behaviour analysed frame-by-frame[1,2], in bouts of hundreds of milliseconds[3], and in periods of seconds and tens of seconds[4,5] can be found in the literature.

In order to more accurately study the neural activity underlying behavior or behavioral perception, there is a necessity to develop and implement methods for the most accurate frame-by-frame temporal localisation of the beginning and ending of behavioral events. Works on the topic of behaviour detection highlight a number of challenges[6–9]. First, there are segmentation and classification challenges. Behaviour is a continuum, so it is difficult to formulate clear rules for separating one behavioural pattern from another. The very term—"pattern"—is not clearly defined, although it is generally accepted that it should represent some kind of holistic movement and be composed of simpler and shorter stereotypical behavioural "syllables". Second, there are a number of technical difficulties, many of which stem from conceptual problems. Even among experienced observers, incomplete consistency has been found both within and across collectives[6,10]. Also, humans have difficulty recognizing fast patterns that may be meaningful to animals. At the same time, the workload required to qualitatively distinguish

[1]Saint Petersburg Electrotechnical University "LETI", Faculty of Computer Science and Technology, Saint Petersburg 197022, Russian Federation. [2]Institute of Higher Nervous Activity and Neurophysiology, RAS, Moscow 117485, Russian Federation. [3]Julia Borodacheva, Ildar Shamsiev, Igor Bondar and Yulia Shichkina contributed equally. ✉email: shichkina@co-evolution.ai

different behavioural patterns is enormous. Depending on the complexity of the task, labeling 1 h of video footage of a behavioural experiment can take hours, tens of hours[9], and, in some cases, even several months[2]. In our experience, extracting 1 behavioural pattern with 1-frame accuracy took about 1 h for a 15-min video.

Many of the aforementioned issues can be resolved by applying state-of-the-art video analysis methods. In this paper, we consider the application of artificial intelligence (AI) technologies for automatic frame-by-frame temporal action localisation in video data using rabbit rearing as an example, and propose a new metric for labeling consistency evaluation. We deliberately use the term "actor" and not "animal" to underline the fact that the proposed approach can be implemented for annotation not only of animal behaviours but also particular events in video recordings of humans, plants and other agents.

The paper is organised as follows: in the first section we review the literature on the subject highlighting challenges in the field of automatic frame-by-frame video labeling and approaches to their verification, in particular, the estimation of labeling consistency quality. In the second section we describe the experimental design during which the rabbit behaviour video data was collected. The third section contains the description of the AI technologies used for automated temporal action localisation in the collected video data. In the fourth section, we propose a novel approach to evaluate the quality of rabbit rear annotation based on the estimation of temporal mismatch between precise times of behavioural pattern onset and offset as determined by either a human or an automated algorithm. In the fifth section we sum up the results of the study and outline further directions in the research.

Related works

Historically, the first systematic attempts to study and measure complex animal behaviour were made by ethologists[11]. They relied mainly on observation, photography, and, later, video recording. This work grew largely out of amateur animal observation[12] and inherited many of its features. Since then video technologies have advanced immensely, thus expanding opportunities for behaviour recording. However, video data labeling remains a very labour-consuming and tiresome process. Therefore a search for new approaches to improve the accuracy and speed of automatic labeling of video data is of the utmost importance. Recent advances of AI technologies, particularly in machine learning (ML), deep learning (DL) and machine learning operations (MLOps), may prove instrumental in addressing this challenge and create great new opportunities for discoveries in the field of animal behavioral patterns analysis.

A lot of studies in the aforementioned direction are conducted in experimental settings. For example, in one study[13] a frame-by-frame analysis of video footage containing behavioural patterns of *Drosophila melanogaster* is evaluated. The authors used convolutional neural networks to classify Drosophila behaviour patterns on a frame-by-frame basis. To evaluate classification quality, the authors calculated model predictions accuracy and error rate. It is emphasised that with the application of artificial neural networks the detection task was solved with an accuracy of more than 99.9%. Special attention in the article is paid to the influence of human factor on the quality of manual annotation and to the ability of neural networks to detect potential errors in it. In another laboratory animal study[14] the authors used convolutional neural networks to detect grooming in video recordings of laboratory mouse behaviour. In order to further improve the quality of automatic annotation, the content of video frames was analysed using several networks, whose predictions were then summarised to obtain the final result. Authors compared their method with existing approaches and showed an increase in the quality of automatic annotation both in terms of prediction accuracy and correct detection of grooming.

In addition to experimental settings, studies are often conducted in domestic settings. For example, one study is addressing agonistic interactions of rabbits on a farm[15]. The authors developed a pipeline consisting of three key steps to detect relevant behavioural patterns. The first step involved segmenting the rabbits' actions and removing video fragments that lacked interaction data. The second step consisted in automatic detection of rabbit actions in the frame. Finally, the third step dealt with categorisation of the actions with purpose to detect the fact of agonistic interactions and determine the role of each rabbit in them. At the rabbit action detection step the algorithm reached precision level of 0.77 and recall level of 0.85 within 5-min intervals in comparison with manual annotation. Additional emphasis was placed on the methodological difficulties of manual annotation.

In addition to the abovementioned results, there is a substantial body of literature on localisation of temporal boundaries of animal behaviours in naturalistic habitats. In particular, one of the articles deals with data from photo and video traps designed to facilitate the human-free collection of data on animal behaviour in nature[16]. The authors developed a temporal action detection system consisting of two components. The first is a video segmentation system designed to detect and locate animals in the frame. The second component is a custom modification of the SlowFast neural network architecture for animal actions recognition. After testing the system on several datasets the authors concluded that in terms of the quality of automatic temporal action detection it surpasses analogs, including the widely used transformer model MViT.

An important distinction in the domain of temporal localisation of behavioural patterns in video data is the distinction between unsupervised and supervised approaches to data labeling. In both types of approaches, annotation can be performed by assigning certain behavioural patterns to the contents of individual video frames, e.g., on the basis of extracted pose or configuration of the animal's limbs (e.g.,DeepLabCut[17]). Unsupervised learning suggests using algorithms for autonomous clustering of behaviour, with subsequent analysis of the clusters by a human to determine their biological or behavioural significance[6,18,19]. The advantage of this approach consists in the ability to capture the whole spectrum of behavioural patterns and even "syllables" executed by the animal instead of attributing them to a predetermined set. The disadvantages are, firstly, reliance on human judgement though only at the final stages of analysis, secondly, a greater dependence on the properties and inclination angle of the recording camera, and, thirdly, difficulties related to interpretation of the results. There are examples of a very thoughtful analysis of data obtained using the latter approach with recruitment of experimental methods to even identify subtle changes in animal behaviour due to the introduction of new stimuli in the environment or as a result of a manipulation with its nervous system[18]. Another example of unsupervised learning approach can be found in the articles on the B-SOiD algorithm and the A-SOiD platform[20,21]. B-SOiD is positioned as an intelligent tool designed to use key points on the body of an animal to automatically segment and classify its behaviours[20]. In addition, A-SOiD is described as a platform with a user interface allowing to take advantage of active learning and significantly reduce the amount of manually labeled training data when solving a task of animal behaviour annotation[21].

An alternative view on organising intelligent tools for action localisation is reflected in the supervised learning approaches. Here, researchers rely entirely on using human manual labeling to train the algorithm[6,19]. The main advantage of this approach is the relative simplicity and unambiguous interpretation of the obtained results. This approach is the most appropriate for localisation of behaviour in tasks with low ambiguity of conditions in which key behavioural patterns can be easily determined. Such tasks include the majority of classical behavioural tests (open field, mazes), although it is sometimes pointed out that even in this case it is possible to miss important subtle features of behaviour[8]. Main disadvantages of this approach include the inevitable narrowness of the original behavioural pattern classification given by a human expert[18] and a possibility for the algorithm to adopt labeling features and traits unique to a certain person. The latter might negatively affect the performance of the algorithm when used on other datasets[9].

An example can be given by the DeepEthogram package[22]. While B-SOiD is based on unsupervised learning, DeepEthogram uses supervised learning and manually labeled data for training. Also, at the video analysis stage, DeepEthogram processes the whole frames and their contents instead of relying solely on actor body key points like B-SOiD. The authors compared DeepEthogram and several other methods with manual labeling on a number of different datasets and concluded that DeepEthogram's predictions are closer to the manual labeling than predictions from methods utilising unsupervised learning. The DeepEthogram's labels were found to be of a comparable quality to that of an expert, as the accuracy between them was higher than 85% and F1 score exceeded 0.7 (in some cases—0.9) for various types of behaviour.

The ideas behind DeepEthogram are further developed in DeepAction, another intelligent tool for detecting actions in video data relying on deep neural networks and supervised learning[23]. It involves a series of steps such as data extraction from video files, classifier training and its application to unlabeled data. When working with unlabeled data, the user can adjust the content of the predictions on the basis of their confidence score. The researchers highlight that DeepAction demonstrates better results than other methods reaching the level of manual labeling (average accuracy above 70%) with a small amount of training data and when working with unbalanced datasets.

Based on the literature review, we come to the following conclusions. The task of temporal action localisation is of high importance in the field of animal behaviour studies. The above mentioned studies are evaluated in different settings, the majority of which suggests that the researchers precede experiments with some preparatory work. Due to the nature of experiments in the field of animal behaviour studies, the preparatory work can imply financial costs and resource costs. Some of these costs can be reduced when organizing research with regard to the peculiarities of algorithms and approaches for behavior pattern analysis in use.

The goal of our research was to develop a system for automatic temporal action localisation in video data tailored to analyse the animal behaviour in order to detect a certain behavioural pattern and its temporal boundaries with high F1 score.

The approach to temporal action localisation in use is based on the classification of single frames as either belonging to a certain predetermined behavioural pattern or not. We further investigate the peculiarities of solving the task with the use of different artificial neural network architectures. Then we compare the results obtained with different types of training data (balanced and imbalanced). Special attention is paid to the number of actors needed to improve the quality of automatic labeling results. Finally, we experiment with the size of training dataset to specify optimal number of manually labeled frames needed when solving the task with the use of the above mentioned approach. Also, in addition to the most widely used methods for evaluating the accuracy of automatic annotation based on frame-by-frame comparison, we propose a new method that allows us to assess the precision of behavioural pattern temporal boundaries localisation.

Methods

All required animal's manipulations were approved by the Ethics Committee of the Institute of Higher Nervous Activity and Neurophysiology of Russian Academy of Science at its meeting on 26th of October 2021 (protocol No.4) in line with the procedure established at the institute. All methods were carried out in accordance with relevant guidelines and regulations and in accordance with ARRIVE guidelines.

Video data collection

4 rabbits of the "Soviet chinchilla" breed (*Oryctolagus cuniculus* sexually mature males weighing 4–5 kg), further referred to as "actors", participated in the study. All rabbits were maintained in a state of moderate food deprivation and were trained to perform an instrumental conditioned reflex. Animals were trained to rear in response to an auditory stimulus and press the pedal with its front paws to obtain food reinforcement. The experiment was conducted for 3 months, with experimental sessions lasting up to 20 min each.

Behaviour of rabbits performing the instrumental conditioned reflex in the experimental chamber was recorded on a video camera (at 30 frames/s). We also would like to underline that video data used in this study had several constraints due to the peculiarities of the experimental procedure and according to the common animal welfare rules.

1. Rears constituted the majority of behavioural patterns in the video data as they were the conditioned responses which the demonstrator rabbit was trained to perform in order to get food reinforcement.

2. Single video recording was 15 min long due to the limited capacity of the automatic feeder which provided food reinforcement to the demonstrator rabbit after performing a correct conditioned response (a rear).Analysis of behaviour consisted in manual temporal action localisation aimed to detect a single behavioural pattern (rearing) using original software with 1 frame precision. For every rear, the frames containing its onset and offset were determined. Behaviour of rabbits performing the instrumental conditioned reflex in the experimental chamber was recorded on a video camera (at 30 frames/s). We selected rearing due to the unambiguous temporal localisation criteria as well as due to the fact that it is an important behavioural pattern oftenly considered while interpreting animal activity in various experimental paradigms[10,24,25].

From an ecological point of view, rears constitute a special type of exploratory activity aimed at obtaining fuller visual and olfactory information about the environment[26]. Behaviourally, a rear is defined as a movement of an animal in the vertical plane with both paws lifted from the ground[27,28] Therefore, we formulated the following criteria for temporal localisation of this pattern: the first frame in which the animal lifted both front paws from the floor was considered to contain the onset of a rear; similarly, the first frame in which the animal put at least one of its front paws back to the floor was considered to represent the offset of the rear. With rabbit pressing the pedal, the first frame after the pedal was fully pressed, was considered to represent the offset of the rear.

Video data processing by AI methods

For video data processing, we developed a system for automatic temporal action localisation based on two-class classification of single video frames with regard to the criterion of presence or absence of the desired action (rear) in the frame.

Input data. The input data were video files obtained during the abovementioned experiment with rabbits. A total of 6 video files were used. Each video lasted from 14 to 16 min. The total number of animals participating in the experiment was 4; however, in every video file only one actor performed. Thus, data on the behaviour of 4 rabbits was taken for further analysis. Total number of frames in all the videos was 152,513. Examples of video data are shown in Fig. 1.

In addition to video data the files containing the results of the manual labeling of the videos by human experts were used in the course of the study. The manual labeling files were .csv documents that included the following information for each annotated moment:

1. time in microseconds elapsed since the start of the video (calculated from the number of the frame);
2. numeric code of the pattern recorded at the given time;
3. numeric code of the pattern execution stage: onset or offset.The algorithm for processing the video files and manual labeling data included three stages: data preprocessing, training of artificial neural networks, and evaluation of the best model obtained on the training stage (Fig. 2).

Stage 1. Data preprocessing

For the main part of the study input data was splitted into training and evaluation datasets, the training dataset containing 3 video files and 3 corresponding files with manual labeling data, while the evaluation dataset containing 1 video file and 1 file with manual annotation data.

The evaluation dataset was preserved in original form for the purposes of final evaluation of the training results. On the other hand, files in the training dataset were further divided into segments. In the case of video files, segments were constituted by the sequences of consecutive frames. For each video file in the training dataset, all segments except the final segment contained 2000 frames. Final segments were shorter and contained the remaining frames of the corresponding video files. For manual labeling data, segments were .csv files containing annotation data for the video segments.

In the course of the training stage, segments from different video files were used in different combinations. When segments from more than one video file were in use, the data from the video files were extracted in equal proportions, thus the number of segments from each individual video file was reduced.

Segments 0–7 (frames 0–15,999) of each video file as well as the corresponding manual labeling data segments were used in the context of training, validating, and testing artificial neural networks. These segments were further divided into frames that were randomly split into train, validation and test subsets in the proportion of 80% : 10% : 10% of the total number of frames, respectively. All frames in the subsets were labeled with the use of manual labeling data included into the training dataset. Frame transformations applied at the training stage included changing the resolution of images, and, in some cases, normalising data in the frame. The frame resolution was set to 224 × 224 px.

Segments 8–12 (frames 16,000–23,999) of the video files in the training dataset were preserved in the video format and were used to compare different neural network architectures during model tests in the context of the training stage. The corresponding manual labeling data served as a reference for evaluating the quality of automatic annotation by different neural networks.

We also decided to extend our study to evaluate the ability of the model to work with video data from different time points. For this additional experiment we added 2 new combinations of video files with corresponding manual labelling files. One pair was used for training of the model and another was implemented for evaluation of the model.

Stage 2. Training of artificial neural networks

At this stage, the training dataset was used in the context of transfer learning for the customization of pretrained artificial neural networks. Training task was defined as a two-class classification problem with networks learning

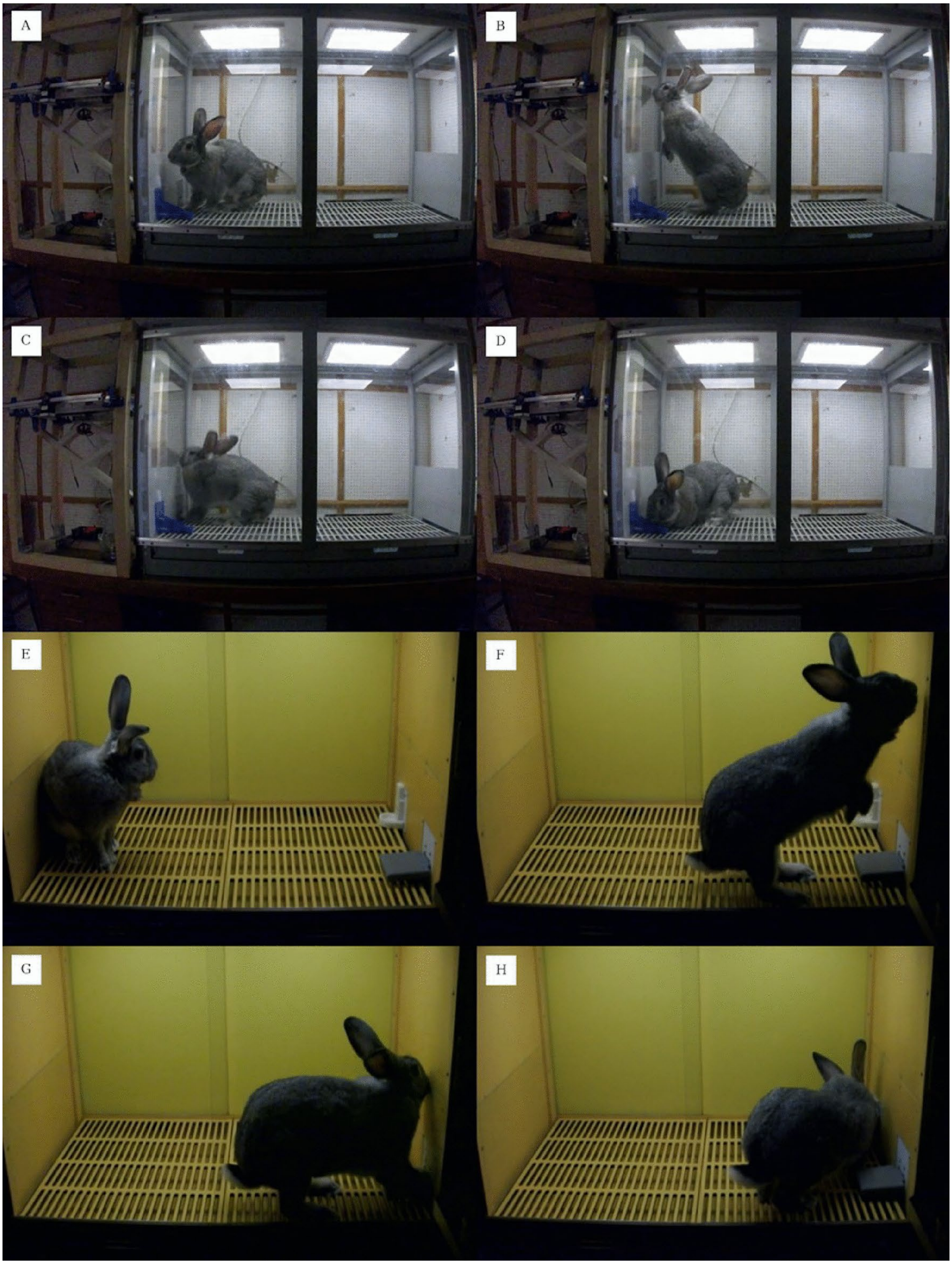

Fig. 1. Example of a video content. (**A**) The rabbit is at the onset of the rear; (**B**) the rabbit is at the highest point of the rear; (**C**) the rabbit finishes the rear; (**D**) the rabbit is receiving reinforcement from the feeder—a behavioural pattern distinct from rear. (**E**–**H**). The rabbit in a new experimental setting.

to predict whether the action of interest is presented in the video frame or not. All layers in the networks except the output classifier remained freezed during training.

Training was carried out using PyTorch 2.2 and the fastai library[29]. The computations were performed using Kaggle platform[30] on an Nvidia Tesla P100 graphics card. The default parameters of the training dataset were as follows. Total number of frames = 16,000, batch size = 16, number of epochs during training = 10 and number

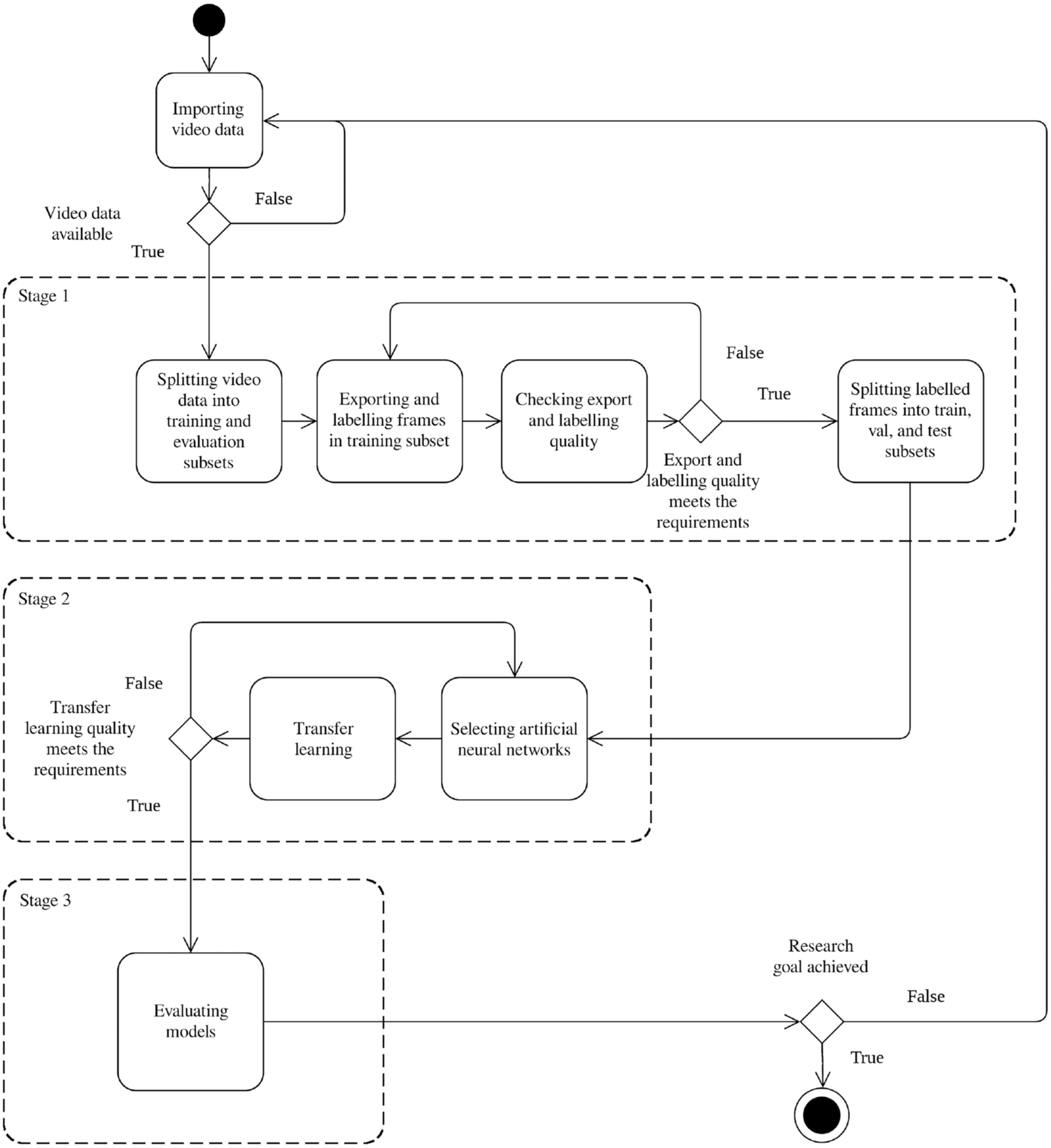
Importing video data
Video data available
False
True
Stage 1
Splitting video data into training and evaluation subsets
Exporting and labelling frames in training subset
Checking export and labelling quality
False
True
Export and labelling quality meets the requirements
Splitting labelled frames into train, val, and test subsets
Stage 2
False
Transfer learning quality meets the requirements
True
Transfer learning
Selecting artificial neural networks
Stage 3
Evaluating models
Research goal achieved
False
True

Fig. 2. Stages of data processing and model selection in the course of the study.

of epochs without progress for early stopping of training = 5. These parameters were varied in the course of the neural networks training.

Stage 3. Evaluation of the chosen artificial neural network model

At this stage, the video in the evaluation dataset was analysed with the use of the model that obtained the best test results during the stage 2. Labels obtained in the course of model evaluation were then compared with the manual labels in the evaluation dataset. As an evaluation metric we used F1 score calculated according to the formula (1):

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{1}$$

where F_1 is the F1 score measure, and *TP*, *FP* and *FN* stand for true-positive, false-positive and false-negative predictions of the model, respectively.

Results

Main results obtained in the course of the training of artificial neural networks with regard to the aforementioned research strategy can be summarised in 5 steps:

1. Selection of the best model from 8 variants differing in architecture and number of parameters;
2. Selection of the approach to training the model according to the criterion of balance of the training sample;
3. Selection of a model training approach based on the criterion of using data on different number of agents;
4. Selection of training approach by the criterion of training sample size.
5. Model evaluation under the data drift conditions.The results of each step are summarized below.

Step 1. Selection of the best model

In order to analyse the rabbit behavioural patterns, we used pre-trained models for image classification presented in the PyTorch Image Models (TIMM) library[31]. We considered model variants with high top-1 accuracy ($acc@1 \geq 80\%$) from the following catalogues:

- Fastest timm models > 80% Top-1 ImageNet-1k;
- Fastest timm models > 83% Top-1 ImageNet-1k;
- Fastest timm models > 86% Top-1 ImageNet-1k;
- Fastest timm models > 88% Top-1 ImageNet-1k.In addition, we took into account peculiarities of the models' architecture, training and evaluation speed, and system requirements. However, the task of comprehensive analysis of the listed catalogues was not among the goals of the study.

At the first step of the study 8 models with default weights were selected for further comparison. Namely, BEiT-large v.2[32], ConvNeXt-B[33], DeiT-III-medium[34], EfficientFormer-L1[35], RegNetY-16GF[36], TresNet-L v. 2[37], ViT-base, and ViT-large[38].

When training the selected models, the following results were obtained. It was confirmed that in the case of rabbits two-class classification of video frames can be used to perform temporal action localisation. Video segments preserved for model testing were processed successfully in the case of all models, with predictions obtained for each frame. The details on the models' performance are shown in Table 1.

Table 1 demonstrates that the training and evaluation speed of the models generally decreased with increase in the number of the models' parameters. However, this rule is not strict, thus other peculiarities of models' architectures should also be taken into consideration. E.g., the maximum training speed among selected models was shown by the EfficientFormer-L1 architecture, which has slightly more parameters than RegNetY-16GF. At the same time, the latter demonstrated the maximum evaluation speed. The ViT-large turned out to be the slowest model in terms of both training and evaluation speed.

At the step of selecting between different models, the artificial neural networks were trained on data from one randomly selected actor (rabbit 1, Actor 1) and tested on data from two actors. First, the same actor whose action data networks were trained on. Second, another actor (rabbit 2, Actor 2) whose data was not included in the initial training dataset. Test results for the models after the aforementioned training are shown in Fig. 3.

Figure 3 demonstrates that all chosen models successfully categorised the majority of frames of the test video sequence into two classes. An increase in the number of model parameters was associated with an increase in classification quality. E.g. in the case of Actor 1, models with 10-20 million parameters obtained F1 Score=0.93, while models with number of parameters more than 300 million obtained F1 Score=0.97. As can be seen from the comparison of ViT-base and ViT-large, this association is manifested even in the case of the models from the same family. Also, transformer architectures showed slightly higher data generalisation ability. For the Actor 2 video sequence, the F1 Score for EfficientFormer-L1 is greater than that of RegNetY-16GF; the DeiT-III-medium F1 Score exceeds that of TresNet-L v.2; the ViT-base F1 Score turns out to be superior to the ConvNeXt-B. At the same time, most of the analysed models with <50 million parameters are inferior to larger models in their ability to handle new actor data. As the number of parameters increases, the difference gradually decreases. The best F1 value for an unfamiliar actor was obtained using the BEiT-large v.2 model.

Model	Number of parameters, millions	Average duration of 1 epoch at the training stage, s	Average duration of analysis of 1 batch of video frames at the stage of results evaluation, s
EfficientFormer-L1	12.3	61.1	1.103
RegNetY-16GF	11.2	124.3	1.097
DeiT-III-medium	38.8	137.9	1.169
TresNet-L v.2	46.2	170.2	1.158
ConvNeXt-B	88.6	717.5	1.266
ViT-base	86.6	262	1.269
BEiT-large v.2	304.4	840.1	1.445
ViT-large	304.2	1118.8	1.512

Table 1. Performance of artificial neural networks from the timm library in the context of classification based temporal action localisation.

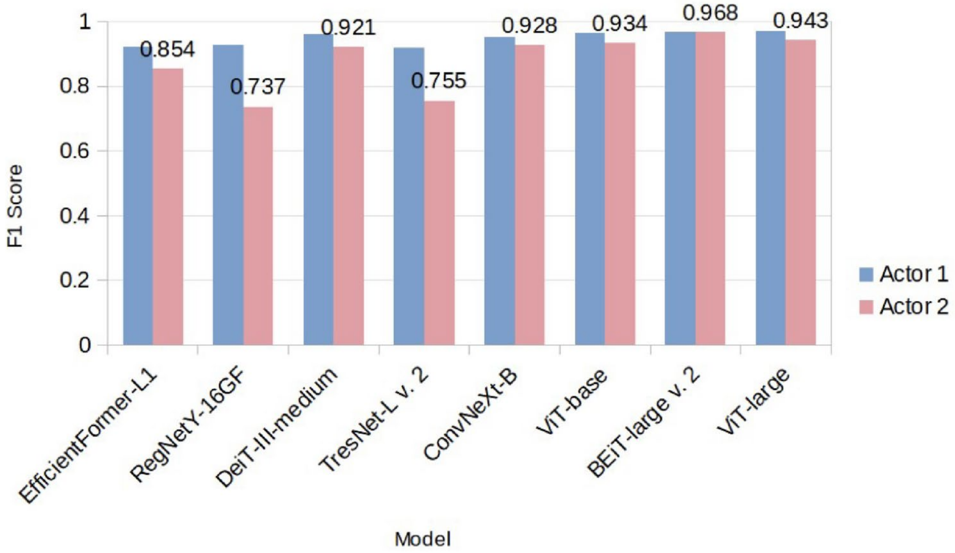

Fig. 3. Comparison of primary test results for timm models in the context of temporal action localisation.

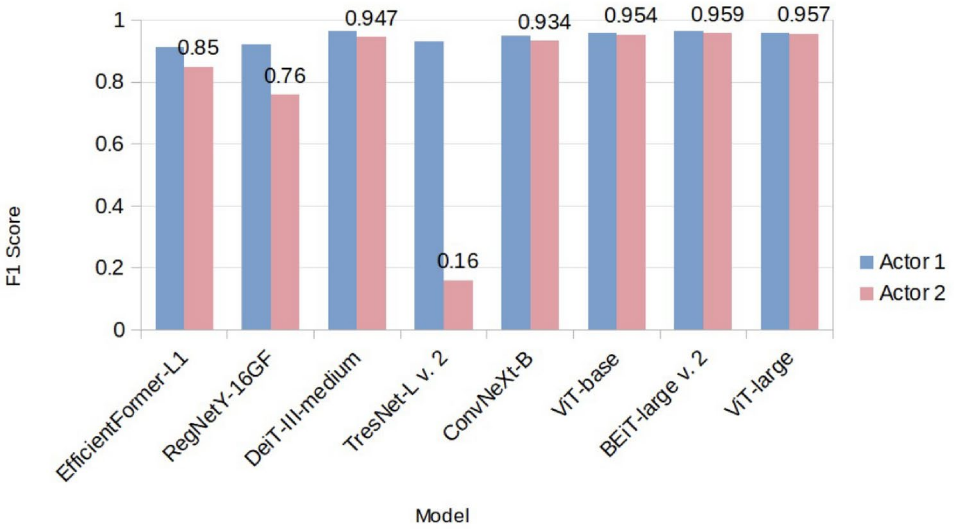

Fig. 4. Comparison of test results for timm models in the context of data normalisation.

Another option checked at the first step of training the models was adding the normalisation to the list of data transformations at the data preprocessing stage. Test results for the case of normalised data are presented in Fig. 4.

Figure 4 reflects that working with normalised frames in most cases led to a decrease in the quality of automatic data labeling for the actor known to the models from training. In the case of EfficientFormer-L1 F1 Score decreased from 0.922 to 0.913; for RegNetY-16GF, from 0.929 to 0.922; and so on. On the contrary, the proportion of successfully detected frames with data on the unknown actor in some cases increased. In particular, for RegNetY-16GF—from 0.737 to 0.76; for DeiT-III-medium—from 0.921 to 0.947; for ViT-base—from 0.934 to 0.954; etc. Different dynamics was demonstrated by the model TresNet-L v. 2, for which F1 Score in the first case increased from 0.918 to 0.931, while in the second case it fell to 0.16. The maximum number of correctly recognized frames when working with unfamiliar agent data was again demonstrated by BEiT-large v.2. However, in its case, introduction of frame normalisation led to a decrease in the labeling quality.

Generalisation of behaviour patterns toward actors not included in the training datasets represents an integral and critical aspect of temporal action localisation. Thus, based on the model comparison conducted at the first step of the study it was decided to choose the BEiT-large v.2 model without video frame normalisation for further analysis.

Step 2. Selection between balanced and imbalanced datasets

At this step, we compared different options for balancing the number of event and non-event frames in the training dataset. In the initial data, the ratio of frames from different classes was 1:7, i.e. 13154 frames containing the event and 96090 frames without the event. Therefore, we investigated if it is preferable to use an equal number of frames from different classes at the risk of reducing the total size of the training dataset or to use the maximum number of available frames but maintaining the imbalance in the number of frames from different classes. At this step, 3 scenarios of training the selected artificial neural network model were implemented:

1. scenario 1: training on a balanced dataset while reducing the total number of frames;
2. scenario 2: training on an imbalanced dataset of size equal to that from scenario 1;
3. scenario 3: training on an imbalanced dataset of maximum available size.Training was again performed on behavioural data of actor 1, and testing was performed on data of actor 1 and actor 2. Scenarios 1 and 2 used datasets of size = 4624 frames; scenario 3 used 16000 frames. The results for all 3 scenarios are presented in Fig. 5.

Figure 5 shows that the use of balanced training dataset had a negative impact on the ability of the artificial neural network to analyse the actions of an unfamiliar actor (F1 Score fell below 0.9). With an imbalanced dataset of equal size, the quality of automatic labeling proved to be higher, but was inferior to the results of the model trained on the entire available amount of data. Thus, we conclude that the best scenario in this case is to use an imbalanced dataset of maximum size.

Step 3. Selection of number of actors to be represented in the training dataset

In the study special attention was paid to the usage of data of different actors when training the artificial neural networks. A comparison was made between 3 training scenarios for the case of BEiT-large v.2:

1. use training dataset that includes data only from 1 actor, namely, Actor 1;
2. use training dataset that includes data from 2 actors, namely, Actor 1 and 3;
3. use training dataset that includes data from 3 actors, namely, Actor 1, 3 and 4.Other training parameters in all 3 scenarios were kept unchanged, including the use of 16000 frames during training. In the second and third scenarios, data from different actors were used in equal proportion. The results of evaluation of the respective models are summarized in Table 2.

Table 2 demonstrates that the different model training options have their strengths and weaknesses. Training on the data of one agent allows the network to process its actions more effectively (F1 Score for the control video of Actor 1 is maximum in the case of the network that was trained on the data of this actor). At the same time, this training scenario creates a risk of performance degradation when required to process the actions of unknown actors (in the case of Actor 3 the F1 value dropped to 0.629). As the number of actors whose data is included in

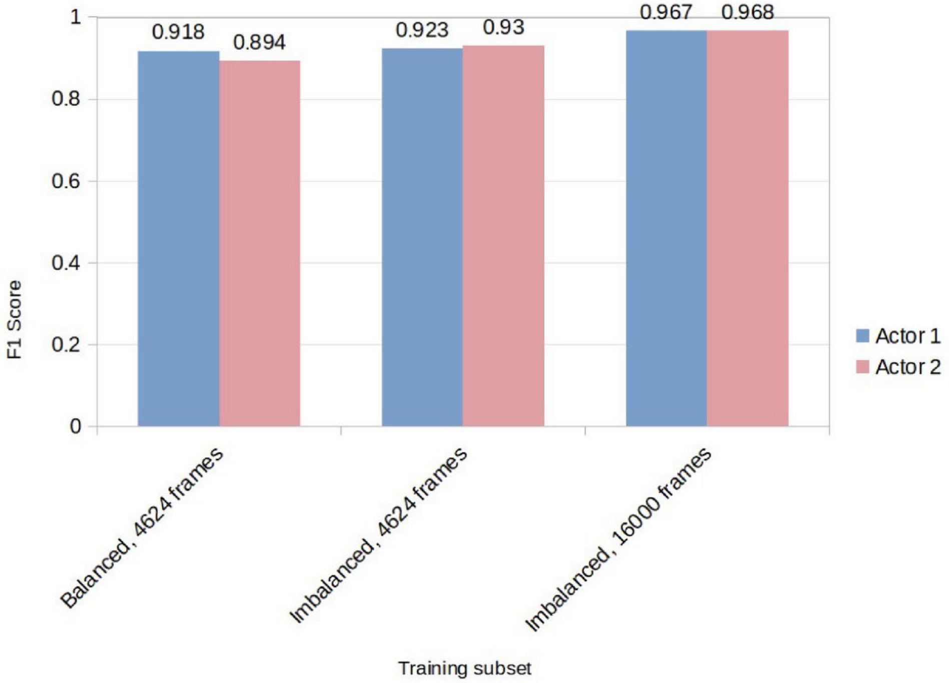

Fig. 5. Comparison of test results for the task of temporal action localisation in the context of using balanced and imbalanced training datasets.

	F1 Score when analyzing the control video sequence		
Control video	For training on the data of 1 actor	For training on the data of 2 actors	For training on the data of 3 actors
Actor 1 data	0.967	0.958	0.956
Actor 2 data	0.968	0.933	0.973
Actor 3 data	0.629	0.899	0.953
Actor 4 data	0.862	0.816	0.952
General	0.857	0.902	0.959

Table 2. Comparison of neural networks trained on data from a different number of actors.

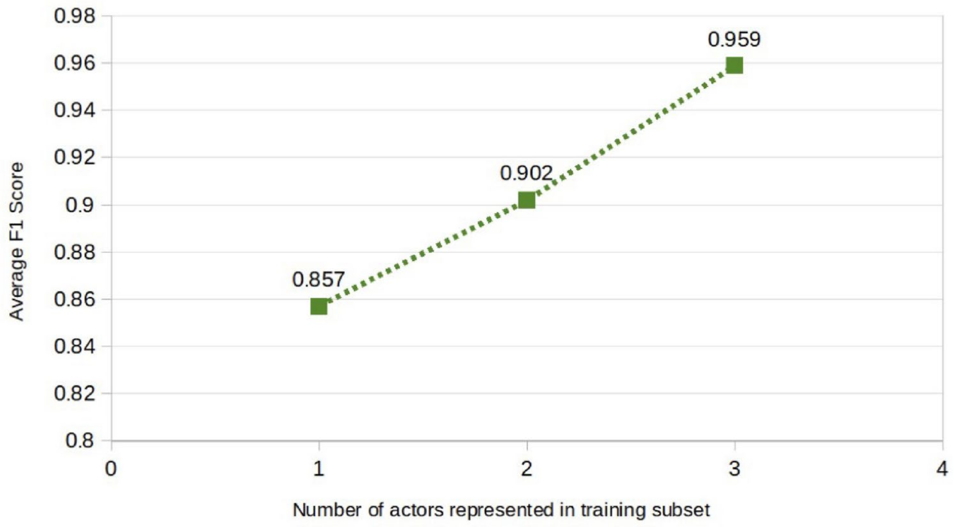

Fig. 6. Average F1 scores for models trained on data from a different number of actors.

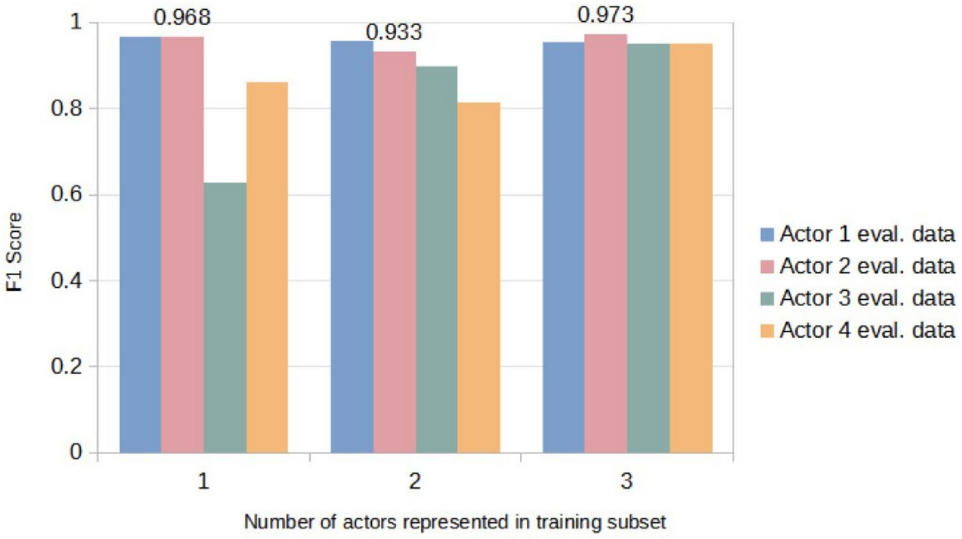

Fig. 7. Evaluation results for models trained on data from a different number of actors.

the training dataset increases, its performance improves, although in some cases it may be inferior to the similar performance of models with a focus on one or several actors (Fig. 6).

Processing data from an actor unknown to the network at the training stage is of particular interest. We evaluated the network with data from Actor 2 (it was not present in any of the training datasets) and provided the results in Fig. 7.

The best result was demonstrated by the model trained on the data of three actors. The second-best result was achieved by the model trained on the data from one actor. The model trained on the data from two randomly selected actors demonstrated a weaker result. We can assume that such an outcome is explained by the difference in the individual style of performing actions by different actors.

In view of the above, the scenario of model training using data from the maximum available number of actors was found to be the most promising.

Step 4. Selection of number of frames to be added to the training dataset

At this step, the question of the optimal amount of data required for successful model training was addressed. We implemented a scenario with model training on data from three actors but manipulated the size of the training dataset. Possible sample sizes included 1 thousand, 2 thousand, 4 thousand, 8 thousand, 16 thousand or 32 thousand frames equally distributed among the data of different actors. The changes in the average F1 Score as a function of the training dataset size are summarised in Fig. 8.

Figure 8 illustrates the fact that increasing the number of data in the training dataset does lead to better quality of frame classification, but as the sample grows, the effect from the sample size increase weakens. When increasing the sample from 1 thousand to 2 thousand frames, the increase in the average F1 Score is 0.16. When changing from 4 thousand to 8 thousand frames, the increase is 0.088, and so on (Fig. 9).

These results were additionally analysed using the cubic spline interpolation, that allowed to estimate the average F1 Score values depending on the number of frames in the training dataset in the range 1...32 thousands frames.

Special attention was paid to the identification of the optimal ratio between the amount of resources invested in model training and the quality of the resulting automatic labeling. To identify the aforementioned ratio, the threshold value ϵ can be defined equal to such an increment in the expected average value of F1 Score after artificial neural network model evaluation, that model can be considered justifiable to train in terms of time and/or other costs for the training. Based on the current study, 0.001 proved to be a promising value for ϵ. Reaching this value was registered when evaluating the BEiT-large v.2 model trained on 16 thousand frames (see Fig. 10). As a result, this number of frames was considered optimal for the current case of classification based temporal action localisation in rabbits.

Conducting the abovementioned steps of artificial neural network models training and evaluation leads to the following conclusions. The task of temporal action localisation in a video can be solved with F1 Score above 0.95 via transfer training of the artificial neural network model BEiT-large v.2 with a two-class video frame classifier on an imbalanced training dataset of 16 thousand frames of actor behaviour data. These conclusions can be summarised in the form of a scheme (see Fig. 11).

Step 5. Model evaluation under the data drift conditions

Temporal change is an important factor in research. We addressed the issue of possible data drift, in particular, possible changes in behavioural patterns with time, by including additional data consisting of 2 video files of behaviour of Actor 1 after 10 months of experimental work as well as manual annotations of these experiments. In line with previous analysis steps, we divided one video file into frames and used it for the model training while the other video file was saved for the model performance evaluation while being completely unknown to the neural network at the training stage.

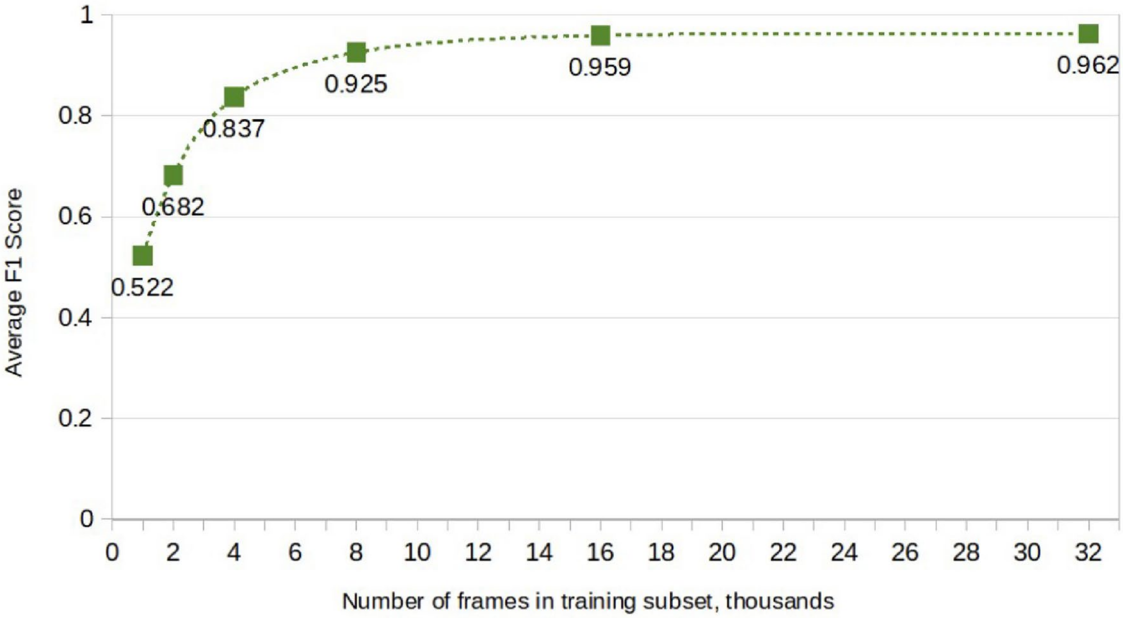

Fig. 8. Average F1 score when using training datasets of different sizes.

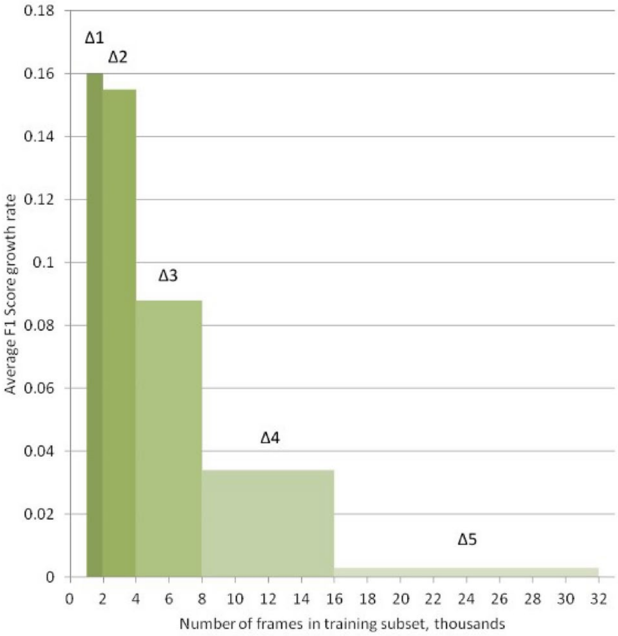

Fig. 9. Increase of average F1 score as a function of increase of training dataset size.

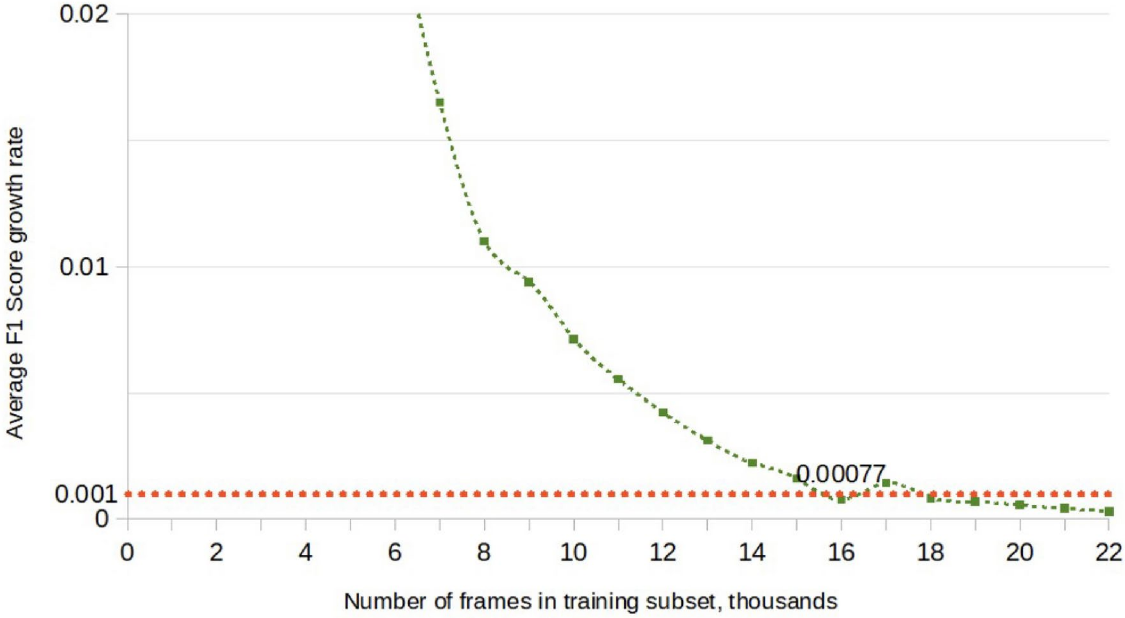

Fig. 10. Optimal size of the training dataset in the context of current case of classification based temporal action localisation in rabbits.

The new experimental setting differed from the previous one in a number of ways:

1. experimental chamber appearance;
2. locations of the feeder and the pedal;
3. appearances of the feeder and the pedal;
4. lightning of the experimental chamber;
5. video camera position and therefore its angle;
6. Internal state and behaviour of the Actor due to a long period between the experiments (10 months).Therefore, we evaluated the model by varying data composition of the training dataset, namely, by including frames from the original and the new experimental settings in different proportions. There were 3 options: only the

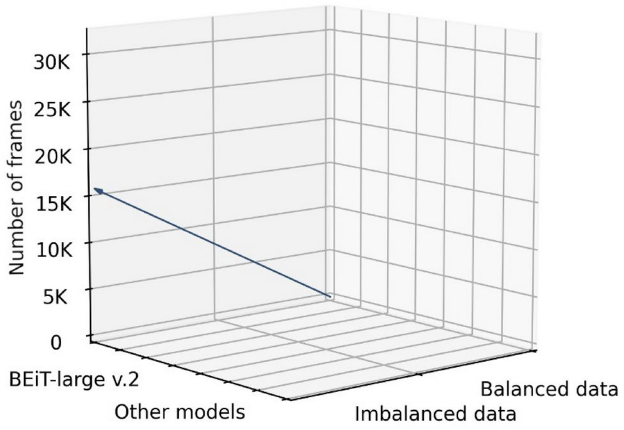

Fig. 11. Selected approach to automatic temporal action localisation in rabbits.

Scenario	Percent of the frames from the original setting	Percent of the frames from the new setting	F1 score for annotation of the video file recorded in the original setting	F1 score for annotation of the video file recorded in the new setting
1	100	0	0.973	0.701
2	75	25	0.944	0.794
3	50	50	0.947	0.838

Table 3. Comparison of the networks trained on datasets which included video data recorded in the new experimental setting.

frames from the original setting; the frames from the original setting and the frames from the new setting in proportion 3:1; the frames from the original setting and the frames from the new setting in proportion 1:1. The analysis results are presented in the Table 3.

These results underline the ability of the model developed at the previous stage to analyse new data. Using mixed training dataset results in a small reduction in temporal action localisation quality for the data recorded in the original setting but allows to increase the quality of the analysis of the data recorded in the new setting. Nevertheless, progress in this direction depends on many variables such as the nature of behavioural patterns in question, changes of the experimental setting, quantity and quality of the collected data, dataset balancing strategy and timely data preprocessing procedures revision with respect to its characteristics. As there can be a lot of possible strategies in working with this issue, further detailed discussion is required but at the moment we believe it to be outside the scope of this work.

Discussion

As a discussion we present an evaluation of the concordance between the automatic algorithm and a human expert in terms of overall event detection and temporal precision of the event onsets and offsets detection. Concordance evaluation and further statistical analysis was performed on a dataset consisting of frames taken from a video independently annotated by the algorithm and a human expert. In order to mitigate the possible issue of expert subjectivity influence on the algorithm performance we took the following steps: 1) Expert annotation was performed using very strict, pre-defined and literature-based criteria for behavioural pattern (rear) onset and offset identification in video frames. 2) In the course of the study manual annotations were performed by several experts, whose temporal concordance was found to be no less than 90%. A concordance as high as this can be considered as another proof that our criteria of rear onset and offset annotation were effective in decreasing expert subjectivity in video data analysis.

As discussed to a greater extent above, for a number of research tasks in biology a precise determination of times of onsets and offsets of behavioural patterns of interest is of critical importance. The majority of works pertaining to automatic video labelling focused on detection of a single behavioural pattern[39] and calculating its overall duration[40,41]. Automatic algorithm performance was evaluated in accordance with the goals of the aforementioned studies. First of all, we evaluated the quality of the model performance, namely, the quality of detection of the behavioural pattern of interest (a rear) executed by the Actor. In order to do that we calculated the standard metrics. The values are given in the Table 4. As all the metrics values were found to be above 0.97 so we can conclude that the proposed model achieved a very high level of performance in annotating rears in video data.

Nevertheless, in our work we considered the main approach to the algorithm performance evaluation to be calculation of the temporal precision with which it detected the behavioural event of interest—a rear of the

Metric	Value
Kappa coefficient	0.97
Precision	0.97
Recall	0.99
F1-score	0.98
Correlation coefficient phi	0.97
Cosine similarity	0.98

Table 4. Statistical metrics measuring the quality of the pattern of interest (a rear) detection by the model.

actor (rabbit). Taking this approach, we firstly converted the automatic frame labels to the form of annotation representing the onsets and offsets of detected events (rears). As behavioural events occupy a certain amount of time due to the limitations of animal physiology, we based the criteria of onset and offset frames selection on the duration of a normal rear. Minimum time required for an animal to perform a rear is reported to be 0.5 s[42], and in our experience minimal rear duration was shorter, around 0.3 s. The difference may be due to the effects of prolonged training making it an automatic reflexive action. We also recognise that there is a minimum possible interval between two subsequent events, which in our case is 0.9 s, and events separated by lesser intervals should be excluded as well.

Therefore, we formulated criteria for event onset and offset, and applied them to the array of the automatic annotation. The frame pertaining to the onset of the event had to meet the following criteria:

1. It should be marked as containing the behavioural event by the algorithm;
2. The previous frame should be marked as not containing the event;
3. At least 10 subsequent frames (0.3 s) should be marked as containing the event.The frame representing the offset of an event had to meet the following criteria:

1. It should be marked as containing the behavioural event by the algorithm;
2. The subsequent frame should be marked as not containing the event;
3. At least 10 preceding frames should be marked as containing the event.All frames marked as containing the event which did not meet the aforementioned criteria were excluded. The remaining pairs of onsets and offsets were considered to represent detected events and were subjected to further analysis.

After selecting onsets and offsets of the events from the automatic labelling, we could move to evaluating temporal precision of the rabbit rear annotation in comparison with the annotation done by a human expert. Such an evaluation is best done in a pairwise manner, that is, by calculating for every event a delta between its onset time determined by an automatic algorithm and its onset time determined by a human expert. The same is applicable to the event offsets.

In order to determine whether the algorithm detected all the event onsets and offsets correctly, we introduced a maximum delta of 0.5 s which is allowed to exist between the onset/offset time detected by the algorithm and the onset/offset time detected by the human expert. That is, if the event onset/offset time determined by the algorithm differed from the time determined by a human expert by more than 0.5 s, this event was considered to be detected incorrectly by the algorithm. If the delta was less than 0.5 s, the onsets were considered agreeing between the algorithm and a human expert. Then, we calculated the percentage of agreeing event onsets and offsets, and used the number as a quantitative measure of temporal concordance between the algorithm and a human expert. The formula for percentage of temporal concordance is given in Formula (2). The analysis showed that 100% of event onsets and 100% event offsets detected by the algorithm were in agreement with a human expert.

$$R_0 = \frac{SRB}{RB}100\%, \quad R_1 = \frac{SRE}{RE}100\% \tag{2}$$

where R_0—the percentage of agreement of rear onset localisations, *SRB*—the number of agreeing event onsets; *RB*—the total number of detected event onsets, R_1—the percentage of agreement of event offset localisations, *SRE*—the number of agreeing event offsets; *RE*—the total number of detected event offsets.

In order to further explore the temporal precision of the algorithm and evaluate the concordance between a human expert and the algorithm, we visualised it in the form of a histogram of the distributions of deltas between onset times of the same events detected by the algorithm and a human expert (Fig. 12A). The same was performed for the matching offsets (Fig. 12B). For a statistical evaluation, we then analysed the distributions by calculating the median, the first and the third quartiles (Fig. 12).

From the first histogram we can conclude that the automatic algorithm tends to detect the onsets of some events several frames earlier than a human expert. This tendency is more pronounced in the case of event onset detection, which may be due to its greater visual ambiguity in comparison to the event offset: a rabbit can start a rear from slightly different locations in the experimental chamber, and the moment when the front paws are lifted from the floor cannot always be precisely detected. However, the distribution demonstrates that for the

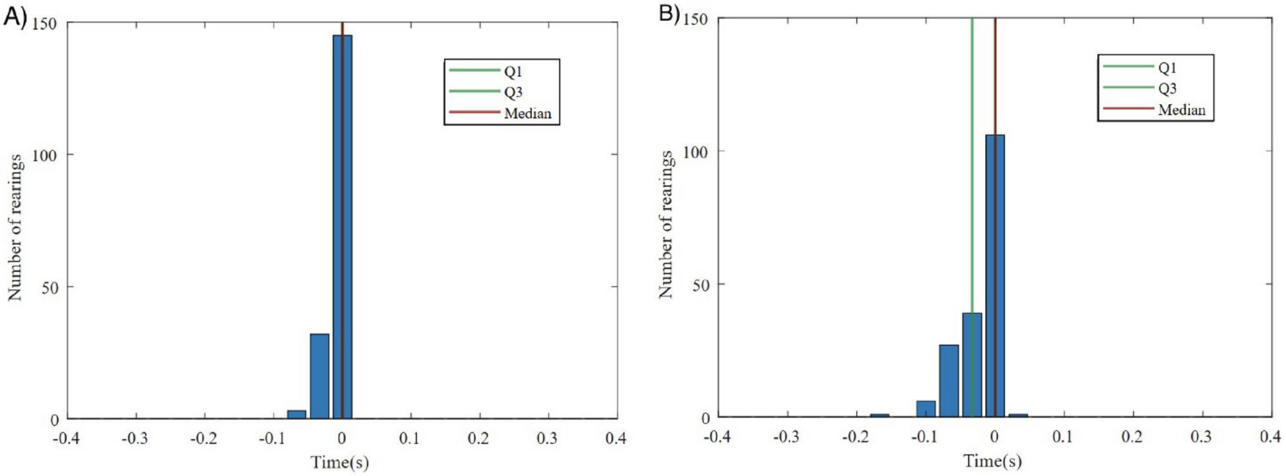

Fig. 12. (**A**) The histogram of the distribution of the event onset deltas between the algorithm and a human expert (median = 0; the first quartile = − 0.033; the third quartile = 0); (**B**) The histogram of the distribution of the event offset deltas between the algorithm and a human expert (median = 0; the first quartile = 0; the third quartile = 0).

majority of the events the times of event onsets and offsets are detected by a human expert and the automatic algorithm with a single frame precision.

This, as well as the high percentage of concordance calculated above (100% for the test data), allows us to conclude that the developed algorithm performed very well and showed a high level of concordance with a human expert in detecting event onsets and offsets.

We then decided to address yet another issue. As was mentioned above, 'human factor' is believed to exert a significant influence on the quality of video annotation especially if there is a substantial amount of data. In particular, ambiguity while interpreting behavioural pattern annotation criteria, fatigue and carelessness can contribute to a low level of concordance in event time detection between two human experts[40]. One possible solution to this problem is to automate the annotation of behavioural patterns allowing to increase consistency of the result due to lower ambiguity of classification. In our research we encountered this issue as well, and we were able to resolve it using automatic labelling. In order to illustrate this point, we propose an evaluation of concordance of event onset and offset times detection between 2 human experts. We implemented the same approaches as before while evaluating concordance between the automatic algorithm and a human expert. We found that two human experts annotating one video demonstrated a lower concordance for both onsets and offsets (94% each).

The distributions of onset and offset deltas between two human experts for one video were also calculated, and the histograms are shown in Fig. 13A and B, respectively. The widths of distributions and median and quartiles shifts from zero indicate that temporal concordance in event onset and offset times detection between two human experts is lower than between a human expert and an algorithm (Fig. 12). This suggests that implementation of an automatic algorithm may be a useful tool in annotating huge video datasets without losing the quality or temporal concordance in rear detection.

Conclusion

In this study we implemented a system for automatic temporal action localisation in video data based on two-class classification of frames. The classification was based on presence or absence of the action of interest in the frame for a single-actor scenario. While designing the temporal action localisation system we took into consideration both temporal and spatial boundaries of the action execution process. Special attention was paid to the peculiarities associated with automatic localisation of action onset and offset times, as well as to the distinctive features of manual and automatic localisation of these times. The actors executing actions were animals in controlled experimental conditions.

The task of automatic classification of frames in video data was solved by using artificial neural networks improved by transfer learning. The training consisted of 5 steps: selecting the most efficient neural network model, picking training dataset balancing strategy, choosing the optimal number of different actors in the training dataset, and determining the optimal training dataset size. As a result, after training we achieved the accuracy level of 0.993 and F1 Score of 0.978 in an automatic temporal action localisation with neural network model for the case of an actor not presented in the training dataset. Temporal precision in determining onset and offset of the action of interest by the neural network reached 100%. This result allows the algorithm to be used in research requiring highly accurate temporal localisation of animal actions, for example, while studying fast behaviours[43–45] or perception of various animal actions and facial microexpressions in primates at the neural level[2,3,46].

Apart from that we managed to substantially reduce time costs for video data labeling due to automation of the process enabled by the developed system. While the manual labeling of a single video file of a given duration

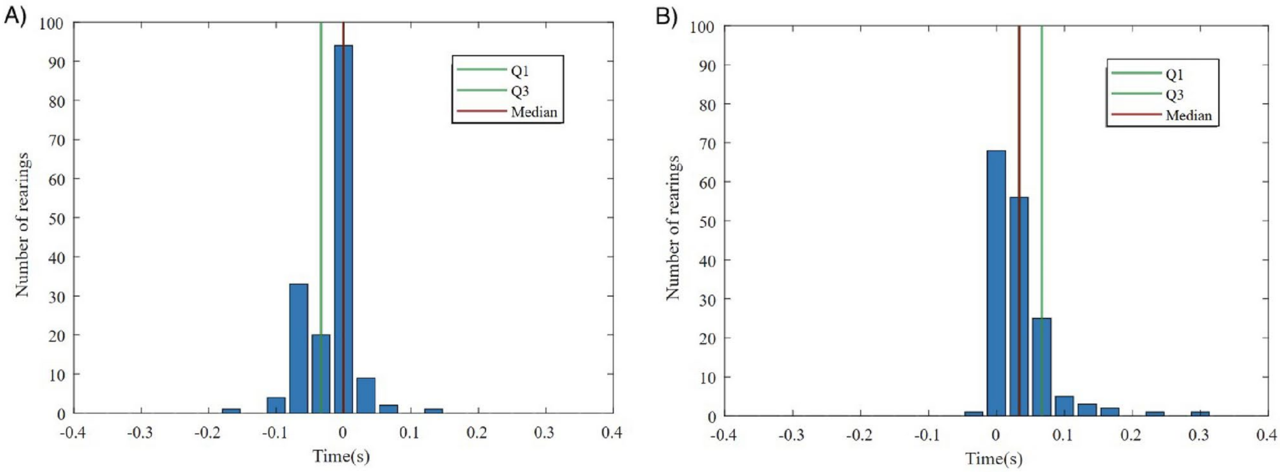

Fig. 13. (**A**) The histogram of the distribution of the event onset deltas between the algorithm and a human expert (median = 0; 0.25 quartile = − 0.033; 0.75 quartile = 0); (**B**) the histogram of the distribution of the event offset deltas between the algorithm and a human expert (median = 0; 0.25 quartile = 0; 0.75 quartile = 0).

required about an hour-long human expert work, the automatic algorithm finished it in 40 min (2462.7 s for the evaluation video), thus reducing analysis time by one third. It should also be noted that with better hardware the algorithm will perform even quicker. Another advantage is an improvement in temporal precision reflected in the increase in the percentage of temporal fits of onsets and offsets of actions from 94% between two human experts to 100 between a human expert and the algorithm.

One of the advantages of the presented approach to automatic temporal action localisation in videos is its ability to identify the same behavioural pattern in different animals. The evaluation dataset was taken from a video of an animal whose behaviour was not included in the training dataset. Successful temporal localisation of one behaviour in different animals may present difficulties due to the diversity in its execution by individuals of the same species (e.g., a rear in different rabbits may be of different height, with or without support etc.).

We are planning to improve the performance of the system by expanding the scope of behavioural patterns it is capable of identifying in the same actor as well as to increase the number of possible actors. Another possible future direction is to broaden the range of available environments allowing us to study behaviour in different experimental conditions and also move towards studying animals in their natural habitats.

Our results suggest further study and development of the approach in use to be highly promising in several directions. On the one hand, an investigation of features and peculiarities of the approach is necessary. Firstly, we would like to conduct a comparison with the analogues. Secondly, we are interested in the peculiarities of implementation of our system with increased number of actors in the training dataset, and in describing dependency of the optimal training dataset size on the number of actors in it. On the other hand, there is a promising task of providing opportunities for optimising the processes involved in practical application of the system, including its embedding in pipelines, deployment, and subsequent support. In particular, it is of interest to work on determining optimal repertoire of behavioural events that need to be included in the training dataset to ensure better quality of automatic temporal action localisation.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 22 August 2024; Accepted: 6 February 2025
Published online: 17 February 2025

References

1. Kingsbury, L. et al. Correlated neural activity and encoding of behavior across brains of socially interacting animals. *Cell* **178**, 317–330. https://doi.org/10.1007/s12110-009-9068-2 (2009).
2. Zhang, M., Zhao, K., Qu, F., Li, K. & Fu, X. Brain activation in contrasts of microexpression following emotional contexts. *Front. Neurosci.* **14**, 329. https://doi.org/10.3389/fnins.2020.00329 (2020).
3. Tombaz, T. et al. Action representation in the mouse parieto-frontal network. *Sci Rep.* **10**, 5559. https://doi.org/10.1038/s41598-020-62089-6 (2020).
4. Grundemann, J. *et al.* Amygdala ensembles encode behavioral states. *Science* **364**, eaav8736, https://doi.org/10.1126/science.aav8736 (2019).
5. Mazuski, C. & OḰeefe, J. Representation of ethological events by basolateral amygdala neurons. *Cell Reports* **39**, 10921, https://doi.org/10.1016/j.celrep.2022.110921 (2022).
6. Egnor, K., S. E. end Branson. Computational analysis of behavior. *Annu Rev Neurosci* **39**, 217—236, https://doi.org/10.1146/annurev-neuro-070815-013845 (2016).

7. Datta, S. R., Anderson, D. J., Branson, K., Perona, P. & Leifer, A. Computational neuroethology: A call to action. *Neuron* **104**, 11–24. https://doi.org/10.1016/j.neuron.2019.09.038 (2019).
8. Pereira, T. D., Shaevitz, J. W. & Murthy, M. Quantifying behavior to understand the brain. *Nat Neurosci* **23**, 1537–1549. https://doi.org/10.1038/s41593-020-00734-z (2020).
9. von Ziegler, L., Sturman, O. & Bohacek, J. Big behavior: challenges and opportunities in a new era of deep behavior profiling. *Neuropsychopharmacol* **46**, 33–44. https://doi.org/10.1038/s41386-020-0751-7 (2021).
10. Sturman, O., Germain, P. L. & Bohacek, J. Exploratory rearing: a context- and stress-sensitive behavior recorded in the open-field test. *Stress* **21**, 443–452. https://doi.org/10.1080/10253890.2018.1438405 (2018).
11. Tinbergen, N. *The study of instinct* (Clarendon PressOxford University Press, 1951).
12. Kruuk, H. *Niko's nature: The life Niko Tinbergen and his science of animal behaviour* (Oxford University Press, 2003).
13. Stern, U., He, R. & Yang, C. H. Analyzing animal behavior via classifying each video frame using convolutional neural networks. *Scientific Reports* **5**, 1–13. https://doi.org/10.1038/srep14351 (2015). Last accessed 28 February (2024).
14. Geuther, B. Q. *et al.* Action detection using a neural network elucidates the genetics of mouse grooming behavior. *eLife* **10**, 1—32, https://doi.org/10.7554/eLife.63207 (2021). Last accessed 01 March 2024.
15. Ipek, N., Van Damme, L. G. W., Tuyttens, F. A. M. & Verwaeren, J. Quantifying agonistic interactions between group-housed animals to derive social hierarchies using computer vision: A case study with commercially group-housed rabbits. *Scientific Reports* **13**, 1–14. https://doi.org/10.1038/s41598-023-41104-6 (2023). Last accessed 02 March (2024).
16. Schindler, F., Steinhage, V., van Beeck Calkoen, S. T. S. & Heurich, M. Action detection for wildlife monitoring with camera traps based on segmentation with filtering of tracklets (SWIFT) and mask-guided action recognition (MAROON). *Applied Sciences* **54**, 1—17, https://doi.org/10.3390/app14020514 (2024). Last accessed 02 March 2024.
17. Nath, T., Mathis, A. & Chen, A. C. Using DeepLabCut for 3d markerless pose estimation across species and behaviors. *Nat Protoc* **14**, 2152–2176. https://doi.org/10.1038/s41596-019-0176-0 (2019).
18. Wiltschko, A. B. et al. Mapping sub-second structure in mouse behavior. *Neuron* **88**, 1121–1135. https://doi.org/10.1016/j.neuron.2015.11.031 (2015).
19. Berman, G. J. Measuring behavior across scales. *BMC Biol* **16**, https://doi.org/10.1186/s12915-018-0494-7 (2019).
20. Hsu, A. I. & Yttri, E. A. B-SOiD, an open-source unsupervised algorithm for identification and fast prediction of behaviors. *Nature Communications* **12**, 1—13 (2021). Last accessed 03 March 2024.
21. Tillmann, J. F., Hsu, A. I., Schwarz, M. K. & Yttri, E. A. A-SOiD, an active learning platform for expert-guided, data efficient discovery of behavior. *Nature Methods* 1—28, https://doi.org/10.1038/s41592-024-02200-1 (2024). Last accessed 26 February 2024.
22. Bohnslav, J. P. *et al.* DeepEthogram, a machine learning pipeline for supervised behavior classification from raw pixels. *eLife* **10**, 1—39, https://doi.org/10.7554/eLife.63377 (2021). Last accessed 28 February 2024.
23. Harris, C., Finn, K. R., Kieseler, M. L., Maechler, M. R. & Tse, P. U. DeepAction: a matlab toolbox for automated classification of animal behavior in video. *Scientific Reports* **13**, 1–19. https://doi.org/10.1038/s41598-023-29574-0 (2023). Last accessed 02 March (2024).
24. Walf, A. A. & Frye, C. A. The use of the elevated plus maze as an assay of anxiety-related behavior in rodents. *Nat Protoc.* **2**, 322–328. https://doi.org/10.1038/nprot.2007.44 (2007).
25. Rebik, A. et al. Audiogenic seizures and social deficits: No aggravation found in Krushinsky-Molodkina rats. *Biomedicines* **11**, 2566. https://doi.org/10.3390/biomedicines11092566 (2023).
26. Lever, C., Burton, S. & OḰeefe, J. Rearing on hind legs, environmental novelty, and the hippocampal formation. *Reviews in the Neurosciences* **17**, https://doi.org/10.1515/revneuro.2006.17.1-2.111 (2006).
27. Meijsser, F. M., Kersten, A. M. P., Wiepkema, P. R. & Metz, J. H. M. An analysis of the open-field performance of sub-adult rabbits. *Applied Animal Behaviour Science* **24**, 147–155. https://doi.org/10.1515/revneuro.2006.17.1-2.111 (1989).
28. Schneider, A. et al. 3d pose estimation enables virtual head fixation in freely moving rats. *Neuron* **110**, 2080-2093.e10. https://doi.org/10.1016/j.neuron.2022.04.019 (2022).
29. Welcome to fastai (2024). Last accessed 19 June 2024.
30. Kaggle (2024). Last accessed 19 June 2024.
31. pytorch image models (2024). Last accessed 19 June 2024.
32. Peng, Z., Dong, H., L.and Bao, Ye, Q. & Wei, F. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *ArXiv* 1–15, https://doi.org/10.48550/arXiv.2208.06366 (2022). Last accessed 22 March 2024.
33. Liu, Z. *et al.* A convnet for the 2020s. *ArXiv* 1–15, https://doi.org/10.48550/arXiv.2201.03545 (2022). Last accessed 22 March 2024.
34. Touvron, H., Cord, M. & Jégou, H. Deit iii: Revenge of the vit. *ArXiv* 1–27, https://doi.org/10.48550/arXiv.2204.07118 (2022). Last accessed 22 March 2024.
35. Li, Y. *et al.* Efficientformer: Vision transformers at mobilenet speed. *ArXiv* 1–19, https://doi.org/10.48550/arXiv.2206.01191 (2022). Last accessed 22 March 2024.
36. Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K. & Dollár, P. Designing network design spaces. *ArXiv* 1–12, https://doi.org/10.48550/arXiv.2003.13678 (2020). Last accessed 22 March 2024.
37. Ridnik, T. *et al.* Tresnet: High performance gpu-dedicated architecture. *ArXiv* 1–12, https://doi.org/10.48550/arXiv.2003.13630 (2020). Last accessed 24 March 2024.
38. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv* 1–22, https://doi.org/10.48550/arXiv.2010.11929 (2021). Last accessed 24 March 2024.
39. Negrete, S. B., Arai, H., Natsume, K. & Shibata, T. Multi-view image-based behavior classification of wet-dog shake in kainate rat model. *Front. Behav. Neurosci.* **17**, 1148549. https://doi.org/10.3389/fnbeh.2023.1148549 (2023).
40. Heredia-Lopez, F. J., May-Tuyub, R. M., Bata-García, J. L., Góngora-Alfaro, J. L. & Álvarez Cervera, F. J. A system for automatic recording and analysis of motor activity in rats. *Behav Res.* **45**, 183-190, https://doi.org/10.3758/s13428-012-0221-1 (2013).
41. Bordes, J. et al. Automatically annotated motion tracking identifies a distinct social behavioral profile following chronic social defeat stress. *Nat Commun.* **14**, 4319. https://doi.org/10.1038/s41467-023-40040-3 (2023).
42. Stone, B. T., Lin, J. Y., Mahmood, A., Sanford, A. J. & Katz, D. B. Licl-induced sickness modulates rat gustatory cortical responses. *PLoS Biol* **20**, e3001537. https://doi.org/10.1371/journal.pbio.3001537 (2022). Last accessed 24 March (2024).
43. Portugal, S. J., Murn, C. P., Sparkes, E. L. & Daley, M. A. The fast and forceful kicking strike of the secretary bird. *Current Biology* **26**, R58–R59. https://doi.org/10.1016/j.cub.2015.12.004 (2016).
44. Whitford, M. D., Freymiller, G. A. & Clark, R. W. Avoiding the serpent's tooth: predator-prey interactions between free-ranging sidewinder rattlesnakes and desert kangaroo rats. *Animal Behaviour* **130**, 73–78. https://doi.org/10.1016/j.anbehav.2017.06.004 (2017).
45. Rossoni, S. & Niven, J. E. Prey speed influences the speed and structure of the raptorial strike of a "sit-and-wait" predator. *Biol. Lett.* **16**, 20200098. https://doi.org/10.1098/rsbl.2020.0098 (2020).
46. Mosher, C. P., Zimmerman, P. E. & Gothard, K. M. Videos of conspecifics elicit interactive looking patterns and facial expressions in monkeys. *Behavioral Neuroscience.* **125**, 639–652. https://doi.org/10.1037/a00242648 (2011).

Author contributions

Conceptualization on organization of calculations, Y.S.; methodology, data analysis and software, S.I.; validation, I.Sh. and J.B; formal analysis, I.B.; statistical analysis, J.B.; writing-original draft preparation S.I., I.Sh., J.B; writing-review and editing, Y.S.; visualization, J.B.; supervision, Y.S.; project administration, I.B.; funding acquisition, Y.S. All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.S.

Reprints and permissions information is available at www.nature.com/reprints.