Behavioral Ecology and Sociobiology (2023) 77:67
https://doi.org/10.1007/s00265-023-03328-4

ORIGINAL ARTICLE

Behavior automatic analysis for wolf pack hunting: making fast behavior analysis of massive data possible

Dengqing Tang[1] · Cheng Huang[2] · Xiaojia Xiang[1] · Han Zhou[1] · Shaohao Zhu[2] · Tianjiang Hu[2]

Received: 24 August 2022 / Revised: 17 April 2023 / Accepted: 20 April 2023 / Published online: 5 June 2023

Abstract

Automatic wolf pack hunting behavior analysis contributes to the identification of behavioral mechanism and the development of bio-swarm intelligence engineering. However, wolf pack hunting commonly involves a complicated background and high-speed motion, where the frequent interactions with each other lead to frequent visual occlusion of the individual wolves. These difficulties make the automatic analysis of group hunting behavior significantly more challenging. Hence, we develop an automatic wolf pack hunting behavior detection scheme appropriate for videos, comprising a four-level individual feature map (frame, video, space, and semantic) and a group feature map. We propose a temporal–spatial feature fusion-based motion state recognition algorithm appropriate for scale-varied and occlusion-frequent individuals to obtain a robust semantic-level feature. Based on this individual feature map, we propose a video-based wolf pack hunting automatic behavior detection method. The developed scheme is validated on our Wolf2022 dataset, while the proposed motion state recognition and group behavior detection algorithms are further tested through ablation experiments. The results revealed that the motion state recognition accuracy reaches 88%, correctly detecting 15 out of 17 group behavior video clips.

Significance statement

It is difficult to fast extract quantitative analysis results of wolf pack hunting behavior from video data. Our research focused on the research of the automatic analysis method for wolf pack hunting. Using the proposed method, the statical individual and group behavior attributes can be automatically generated from video data, which contributes to the building of a conceptual bridge between the wolf pack hunting behavior and bio-swarm intelligence engineering.

Keywords Wolf pack hunting behavior · Automatic analysis · Behavior video detection · Motion state recognition

Introduction

Wolves are currently one of the species with the highest success rate of group hunting (Mech et al. 2015). Research on the useful, quantifiable, robust descriptions and models of wolf pack hunting behavior contributes to the identification of behavioral mechanism (Cassidy et al. 2015; Dickie et al. 2016; Schlagel et al. 2017) and provides inspirations such as collaborative perception, communication, and decision (Zhao et al. 2011; Strandburg-Peshkin et al. 2015), thereby promoting the development of swarm intelligence models. The description, model, and application of wolf pack hunting behavior have attracted significant research interest (Escobedo et al. 2014; Duan et al. 2019a, b; Xie et al. 2021). Aiming to obtain swarm decision knowledge from wolf pack hunting, Duan et al. (2019ab) proposed a target allocation method based on the wolf behavior mechanism, and it effectively solves the problem of unmanned aerial vehicle swarm collaborative target allocation. MacNulty et al. (2007) proposed a wolf pack hunting scheme that includes search, watch, approach, attack-group, attack-individual, and capture. Based on this description, Madden et al. (2010, Madden and Arkin 2011) designed a probabilistic graphical model-based group hunting decision mechanism and validated their method using a real group of ground robots.

Communicated by K. Eva Ruckstuhl.

✉ Dengqing Tang
tangdengqing09@nudt.edu.cn

[1] College of Intelligence Science and Technology, National University of Defense Technology, Changsha City, China

[2] Machine Intelligence and Collective Robotics, Sun Yat-Sen University, Guangzhou City, China

However, this scheme required manually discriminating the different hunting behaviors from a larger number of videos to generate the state transition probability table. Hence, this work imposed huge labor and time costs, and the behavioral discrimination accuracy is governed by human subjectivity. Therefore, it is meaningful to explore an automatic hunting behavior recognition method.

Recording and analyzing animal behavior by utilizing modern equipment such as GPS trackers provides adequate data for quantitative behavioral research that comprises animal behavior data acquisition and automatic data analysis. Nevertheless, these two parts are often contradictory (Roian Egnor and Branson 2016) as high-quality data can be easily automatically analyzed by limiting the environmental conditions or adding constraints to the animals, e.g., by constructing a controlled laboratory environment. However, a laboratory environment can only handle salient problems, and therefore it is critical to focus on animal behavior in the field since environmental factors, such as light, physical space, and temperature, have profound influences on behavior. Considering wolf pack hunting behavior, which involves a wide range of physical space and high dynamics, obtaining high-quality data in a laboratory environment is unrealistic. Besides, wolves' motion data can be obtained by wolves wearing collars, but it is dangerous and easily affects their behavior (Hawley et al. 2010). Thus, the videos captured by ecologists or media are the primary form of data currently used for wolf pack hunting behavior research.

With the rapid development of computer vision technology, computer programs and tools have been explored for automatically analyzing and recognizing an individual animal's motion. For instance, the idTracker (Pérez-Escudero et al. 2014) realizes accurate animal tracking in groups within a controlled laboratory, even when humans cannot distinguish some of them as precisely. Using deep learning technology, the toolboxes DeepLabCut (Mathis et al. 2018) and DeepPoseKit (Graving et al. 2019) afford automatic animal pose estimation and achieve appealing results that are comparable to human accuracy. Most of the current methods focus on analyzing the visual characteristics of animal groups, while it is still difficult to automatically understand the group's higher-level behaviors. Through automatic analyses of wolf pack hunting behavior, it is expected to extract statistical data about their behavior and habits, e.g., (1) whether wolves have increased hunting success when using human-created linear features (Dickie et al. 2016) and (2) which individuals in a group may be more likely than others to influence conflicts (Cassidy et al. 2015), which contributes to our knowledge of group behavior.

When wolf packs are hunting in the field, this typically involves a complicated background and high-speed motion. In addition, their frequent interaction leads to frequent occlusion (i.e., one animal is not visible because it is behind another). These difficulties make the automatic recognition of group hunting behavior from videos very challenging. Focusing on the wolf pack hunting behavior, this paper designed a wolf pack hunting behavior feature map description ranging from individual to group maps and proposes a complete pipeline for feature map extraction and automatic hunting behavior video detection and analyses. The contributions of this paper are summarized as follows:

1. Designing a hunting behavior description containing individual and group feature maps. The multi-level individual feature map indicates the behavior from a single frame-to-frame sequence to spatial and semantic motion states.
2. Proposing a temporal–spatial feature fusion-based individual motion state classification method. By integrating spatial appearance and temporal motion features, our method realizes a robust motion state classification for scale-varied and occlusion-frequent individuals.
3. Developing a wolf pack hunting behavior video detection method that exploits individual multi-level feature maps to detect wolf pack hunting behavior in videos accurately.
4. Creating the Wolf2022 dataset for wolf pack hunting behavior analysis, which contains multi-class manual labels and is available for research on vision detection, tracking, motion estimation, and group behavior video detection.

Materials and methods

Hunting behavior description

As described in Madden et al. (2010, Madden and Arkin 2011), a complete hunting process for a wolf pack involves six different states, including search, approach, watch, individual attack, group attack, and capture. Since the state "capture" is more likely a result of hunting, we employ the remaining five states as the different group hunting behaviors that we aim to recognize automatically. The description of the hunting behaviors is shown in Table 1. It should be

Table 1 The definitions of the five group hunting behaviors

Group hunting behavior	Definition
Search	Traveling without fixating on and moving toward the prey
Approach	Fixating on and traveling toward the prey
Watch	Fixating on the prey while not traveling
Individual attack	Running after/lunge at fleeing individual
Group attack	Running after/lunge at fleeing group

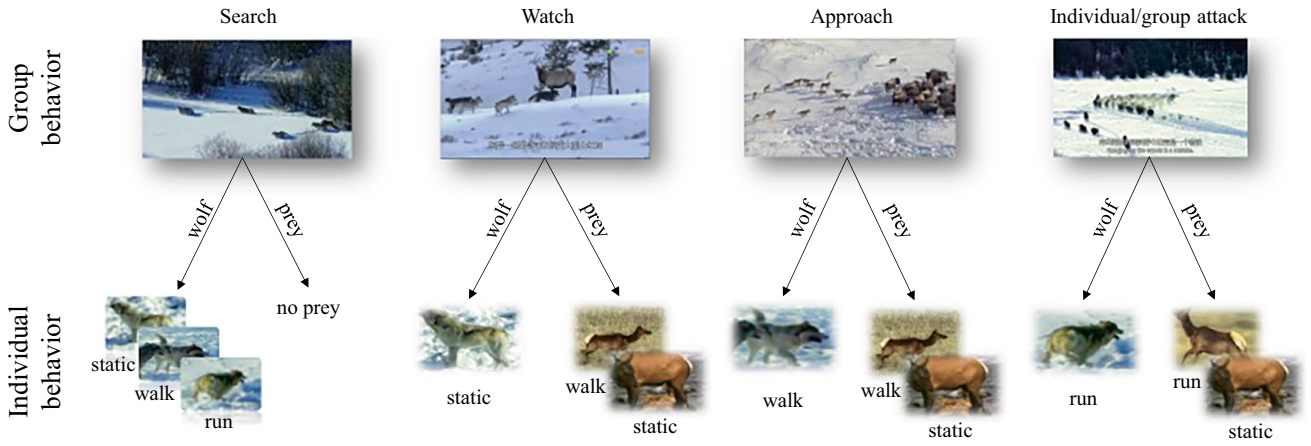

Fig. 1 Behavior decomposition from group to individual. For the search state, the wolves may be static, walking, and running, and no prey appears in video. When the wolf pack watch the prey, the wolves should be static, and the prey may be static or walking. During the approach, the wolves should be walking, and the prey may be static or walking. For the individual/group attack state, the wolves and prey should be running

noted that single blinded method was used when the behavioral data were analyzed since our study involved focal animal in the field.

According to the five group hunting behaviors described in Table 1, these can further be subdivided into individual components, such as static, walk, and run, as illustrated in Fig. 1. Static indicates that the individual remains still, e.g., is standing or lying down. By employing the behavior decomposition model, group behavior can be inferred from a combination of individual sub-behaviors.

As illustrated in Fig. 2, we design a wolf pack hunting automatic behavior detection pipeline comprising multi-level individual behavior descriptions and group behavior inference. We define a 4-level feature map for

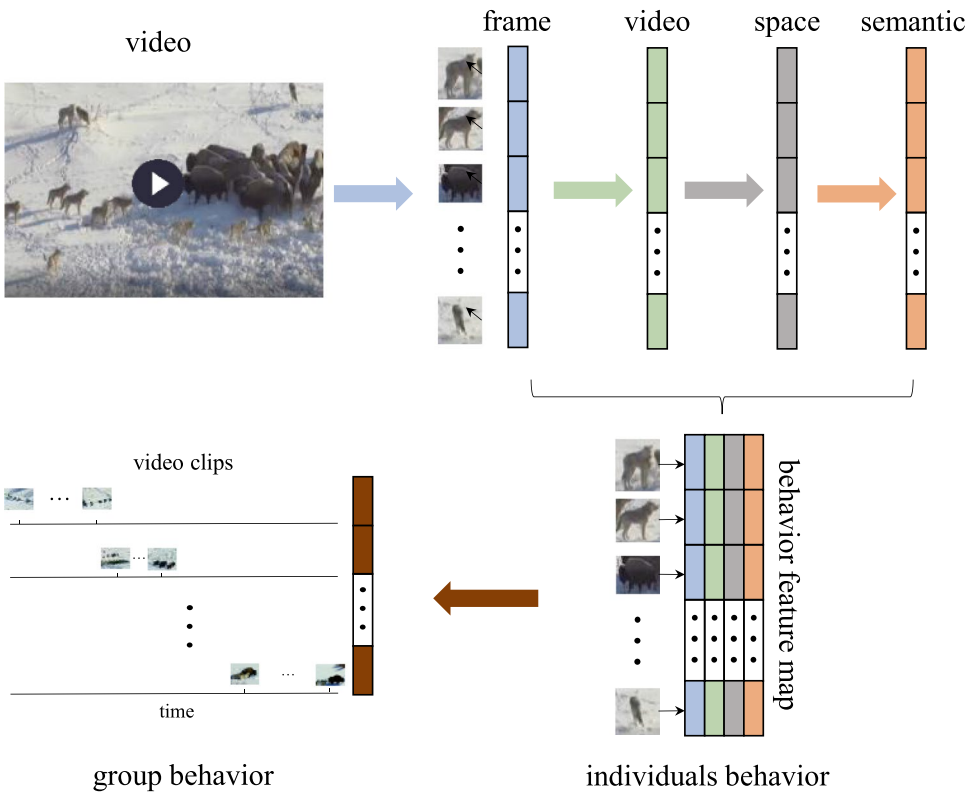

Fig. 2 The wolf pack hunting automatic behavior detection pipeline, from multi-level individual behavior descriptions to group behavior inferences. The individual behavior map includes frame, video, space, and semantic feature levels, and they can derive from the videos in turn. Using the behavior feature maps of all individuals, the group behavior can be inferred, containing the behavior categories and the attributes of video clips. Each video clip should cover a complete group behavior

the multi-level behavior descriptor, including (1) single frame, (2) video, (3) physical space, and (4) semantic description. This 4-level behavior feature map realizes the representations from the two-dimensional image space to time-sequential two-dimensional image space, three-dimensional physical space, and finally semantic space. Based on this feature map, some group behavior video clips and their attributes, such as the numbers of wolves and prey visible in video clips, can be inferred. Next, we describe the behavior feature map generation and the group behavior inference methods.

Individual behavior multi-level feature map

The individual behavior multi-level feature map comprises a 4-level feature setup: (1) frame-level feature F_f, (2) video-level feature F_v, (3) space-level feature F_{sp}, and (4) semantic-level feature F_{se}. For an individual i in frame k, the frame-level feature is defined as

$$\left(B_i, C_i\right)_k \tag{1}$$

$$Box_i = [x, y, w, h]_i \tag{2}$$

where C_i is the category of the individual i and Box_i denotes the image region of individual i. (x, y) is the image position of the regional center, and w and h are the region's pixel width and height, respectively.

The video-level feature involves the identification and motion in each video individually:

$$\left(ID_i, M_i^V\right)_k \tag{3}$$

$$M_i^V = [v_x^V, v_y^V]_i \tag{4}$$

where ID_i is the unique identification of individual i and $M_i^V : [v_x^V, v_y^V]_i$ denotes the velocity vector in the video.

The motion within the video cannot reflect the actual motion in the physical space when the camera is moving. Hence, we employ space-level features to provide motion information in the physical space:

$$\left(ID_i, M_i^P\right)_k \tag{5}$$

$$M_i^P = [v_x^P, v_y^P]_i \tag{6}$$

where $M_i^P : [v_x^P, v_y^P]_i$ is the projection of the actual motion velocity in the three-dimensional physical space on the two-dimensional image plane.

The semantic-level feature gives the sub-behavior category MS_i including static, walk, and run:

$$\left(ID_i, MS_i\right)_k \tag{7}$$

Frame-level feature F_f

In computer vision, object detection can provide the image region and object category within a frame, constituting the frame-level feature F_f. To adapt to complex backgrounds, variable illumination, and individual scale in the wild, we generate F_f by exploiting the neural network-based YOLO-v4 architecture (Bochkovskiy et al. 2020), which affords great robustness and is currently state-of-the-art in object detection fields.

Video-level feature F_v

The identification and motions in the video are obtained using video object tracking methods. Such methods require an initial object region, which in our application relies on the frame-level feature F_f. In this work, we employ the DeepSORT (Wojke et al. 2017) algorithm for robust tracking results that generate the identification ID_i and each object's region in the video. The object's motion velocity $MV\ i$ per frame is inferred by the difference between the object's regional center in adjacent frames.

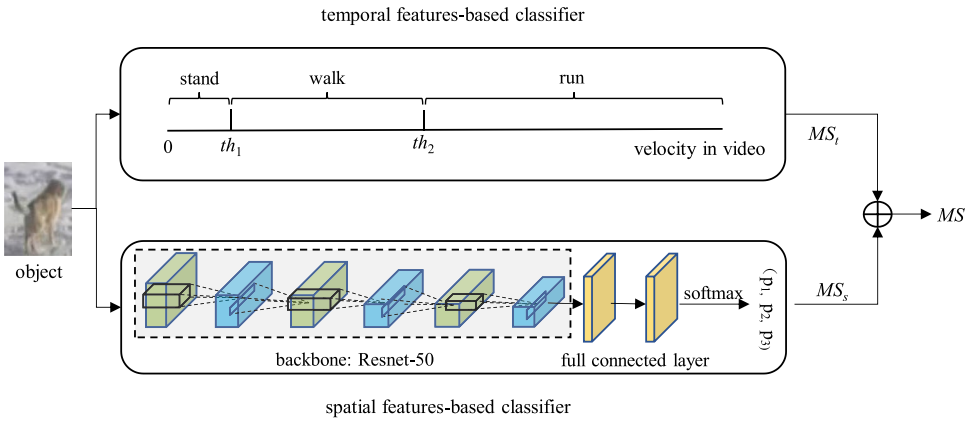

Fig. 3 The proposed temporal–spatial feature fusion-based motion state recognition architecture. A temporal feature-based classifier is designed to extract the temporal feature MS_t, by estimating the velocity. In light of the appearance, a network is used to generate the spatial feature MS_s. The final motion state, MS, is obtained through a weighted sum as formula (9)

Space-level feature F_{sp}

The inference of object motion velocity in the physical space relies on the camera's intrinsic parameters. However, almost all web documentaries hardly provide the corresponding camera parameters. Therefore, we roughly estimate *MP i* by

$$M_i^P = [v_x^V, v_y^V]_i - [v_x^B, v_y^B]_i \tag{8}$$

where $[v_x^B, v_y^B]_i$ denotes the background motion velocity.

For each object, its motion velocity $[v_x^V, v_y^V]_i$ is obtained from its video-level feature. To obtain the actual motion velocity relative to the static background, the background velocity should be inferred. Optical flow represents the instantaneous velocity on the image plane. Hence, first, we extract the first frame's Harris corners (Harris and Stephens 1988) in the background region, which is determined by removing the object regions. Then, we employ the Lucas–Kanade algorithm (Lucas and Kanade 1981) to generate the optical flow vector for these corners in subsequent frames. Next, the outliers of these vectors are eliminated, and finally, the average of the remaining vectors is regarded as the background velocity. Considering that the number of corners changes as some may be out-of-view, the corner should be re-extracted before generating the optical flow vector, if the number of the current corners is below a threshold.

Semantic-level feature F_{se}

Generally, the motion state (static, walk, and run) is directly inferred from the motion velocity in space (temporal feature). However, the motion velocity in the video denotes the projection of the motion velocity from space into the image plane, losing one dimension of information. Therefore, the motion state cannot be accurately inferred only from the video's motion velocity. In addition to the temporal features, the spatial imaging feature can also infer motion states since there are noticeable differences in the animals' posture under three different motion states in a single frame. The imaging of animals leads to one-dimensional information loss, i.e., a projection process from the three-dimensional physical space into a two-dimensional image plane. Moreover, animal body occlusion also reduces posture information. Hence, it is difficult to accurately infer the animal's motion state by solely employing the temporal or spatial appearance features.

Table 2 The behavior map from the sub-behavior of individuals to the group behavior

Prey / Wolf	Static	Walk	Run	Empty
Static	Watch	Watch		Search/watch
Walk	Approach	Approach		Search/approach
Run			Chase	Search/chase

Therefore, we develop a temporal–spatial feature fusion-based motion state recognition architecture to realize a robust motion state recognition. As illustrated in Fig. 3, the proposed pipeline involves two classifiers to generate the motion states relying on temporal and spatial imaging features.

Setting the velocity range (0–th_1, th_1–th_2) of these motion states affords the temporal feature-based classifier to directly categorize the animals' motion state, MS_t, according to their velocity within the video. For the spatial imaging feature-based classifier, we build a network to classify the animals' motion state, MS_s, by feeding it to the network bounding box of each animal. Indeed, the network distinguishes different motion states by extracting the feature of an animal's limb state. Our network's backbone is the ResNet-50 (He et al. 2016) to extract the feature map, followed by two fully connected layers and a softmax classifier. The final motion state, *MS*, is obtained through a weighted sum as follows:

$$MS = r_t * MS_t + r_s * MS_s \tag{9}$$

where r_t and r_s are dynamic weight parameters. For the spatial feature-based module, the details of the appearance feature gradually decrease as the individual scale in images reduces. Therefore, MS_t is theoretically more credible as

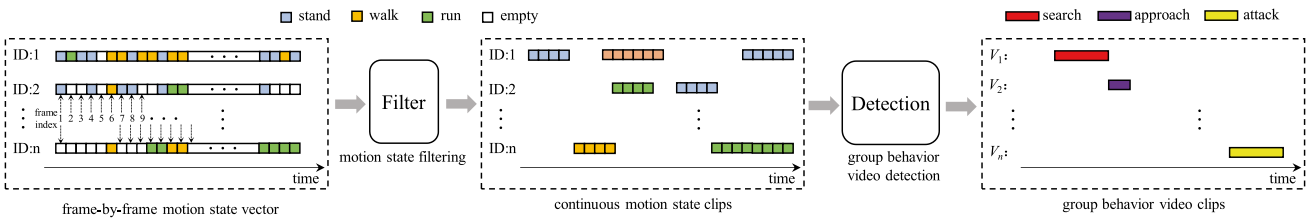

Fig. 4 Diagram of the group behavior inference. A filter is first proposed to extract multiple continuous motion state clips for each individual, and each clip should only contain the same type of motion state. Then, using the continuous motion state clips, a detection module is designed to generate the group behavior video clips

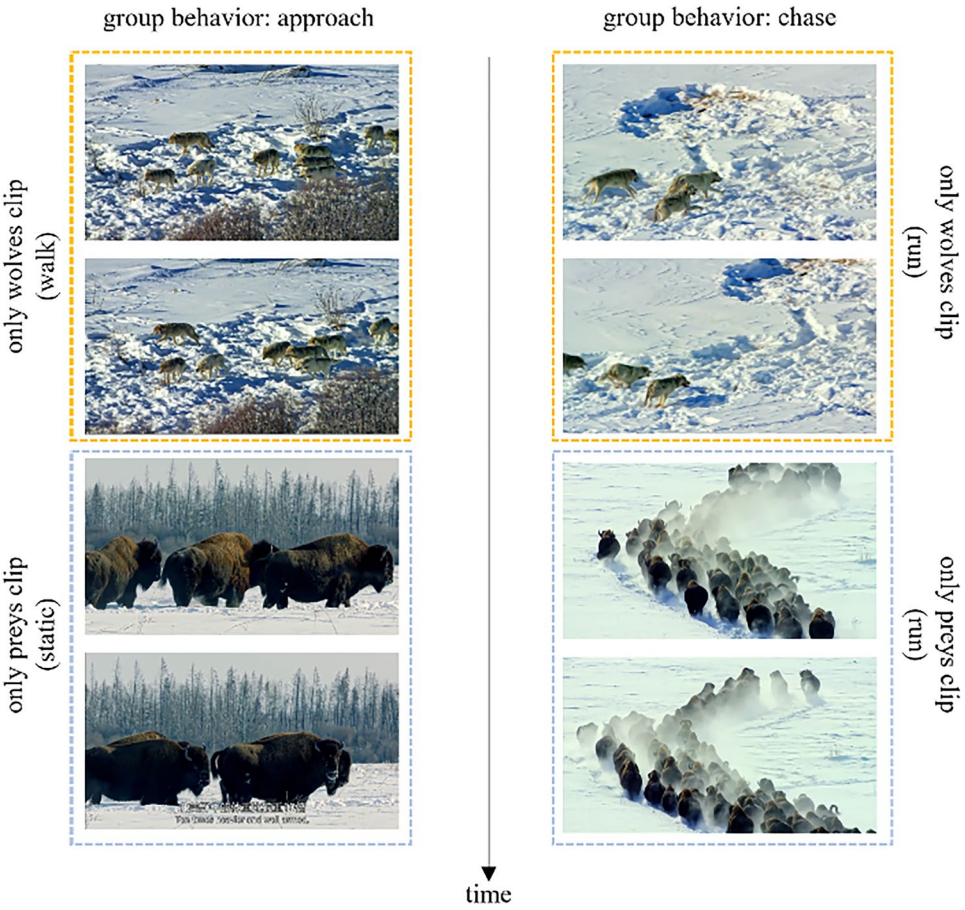

Fig. 5 The particular situation where wolves and prey do not appear in view at the same time from beginning to end of a complete group behavior video clip

the individual scale increases. On the flipside, the temporal feature-based module has more errors as the individual scale increases due to the extensive occlusion between individuals. Considering the above characteristics, the weight parameters are defined as

$$r_s = \begin{cases} 0 & s <= s_{\min} \\ \frac{s - s_{\min}}{s_{\max} - s_{\min}} * r_m & s_{\min} < s < s_{\max} \\ \frac{s - s_{\max}}{s_{im} - s_{\max}} * (1 - r_m) + r_m & s_{\max} <= s \end{cases} \tag{10}$$

where s is the pixel area of each animal's individual bounding box, and s_{im} is the pixel area of the entire image. The $s_{\max}$ and $s_{\min}$ indicate the upper and lower bounds of the effective pixel area, respectively, and are set to 1500 and 10,000 for the images with a 1920×1080 resolution, based on experience. Finally, r_{m} denotes the maximal weight when $s_{\min} < s < s_{\max}$ and is set to 0.9.

Group behavior feature map

Definition and diagram

The group behavior is the group state that the wolf pack maintains over a period. For a video, the group behavior feature map comprises several video clips, where each indicates a complete group behavior. Therefore, we define the group behavior feature map as

$$\{V_i | i = 1, ..., n\} \tag{11}$$

$$V_i = (B_i^G, T_i^s, T_i^e, N_i^w, N_i^p) \tag{12}$$

where V_i represents the attribute of the group behavior in video clip i, n is the number of video clips, B_i^G is the group behavior category, and the start and end time of the video clip i is denoted by $\left(T_i^s, T_i^e\right)$. N_i^w and N_i^p are the numbers of the wolf and prey, respectively.

The group behavior inference diagram is displayed in Fig. 4. First, we build each animal's frame-by-frame motion state vector (indicated by ID in the figure). Then, for each animal, several continuous motion states and their start and end timestamps in the video are separated from the vector using a filter (the detail of the filter is described in "Algorithm" section). Finally, from the motion state vectors, we generate the group behavior video clips $\{V_i | i = 1, ..., n\}$, utilizing the group behavior video detection model.

Table 3 Object tracking results for the two parts of the test set. "Small," "medium," and "big" represent the number of individuals that are "< 5," "5 ~ 20," and "> 20"

Test video		Environment	Group size	Prey	Field-of-view	Precision
Part 1	Video-1	Beach	Medium		Aerial	84.2%
	Video-2	Forest	Big, medium	Cervidae	Ground	96.0%
	Video-3	Tundra	Small		Aerial	84.2%
	Video-4	Tundra	Medium	Buffalo	Aerial	79.1%
	Video-5	Swamps	Big	Cervidae	Ground	90.7%
Part 2	Video-6	Tundra	Big	Buffalo	Ground	79.9%
	Video-7	Swamps	Big, medium	Cervidae	Ground	85.8%
	Video-8	Grassland	Big		Ground	86.5%

Algorithm

Considering the motion state continuity in the video, the animals should keep one motion state for several continuous frames. Therefore, we build a filter to correct the misrecognition of motion states by extracting the independent clips that satisfy two conditions: containing at least m ($m = 20$) frames, and the motion state of all frames is the same. Then, we apply a one-dimensional dilation process on each clip, i.e., each clip expands p ($p = 2$) frames forward and backward. Finally, if the clips overlap and have the same motion state, we merge them after they undergo a dilation process. The clips after merging are used for the subsequent detection. After the dilation and merging, the outliers of the continuous motion state should be filtered out.

In our group behavior video detection step, we first segment multiple video clips satisfying the condition that each must be containing at least one complete motion state clip of the wolf. Then, we count the number of wolves and prey in three different motion states for each clip and consider the motion state with the largest number as the motion state of the wolf pack and its prey. If the clip does not contain a prey's motion state, it is defined as empty. Besides, if there are two motion states with the largest and same number, priority is given to run, walk, and static. Finally, according to the inference model illustrated in Fig. 1, the behavior map from the sub-behavior of the individuals to the group's behavior is reported in Table 2.

Generally, the group behavior should be further researched when the prey's motion state is empty, i.e., the clip contains only wolves. Sometimes, the field of view must be narrowed to capture more details, prohibiting wolves and prey from appearing in the video simultaneously. In this case, almost all videos usually stitch the clips containing only wolves and only prey, indicating that they are captured in the same scene. As shown in Fig. 5, the images are sampled at equal intervals on the video time axis, where the left case displays an approach group behavior containing two independent motion state clips. The first one only presents the wolves, and the second only the prey. To deal with such situations, when the prey's current clip motion state is empty, we regard the prey motion state based on the neighboring clips. If the neighboring clips do not involve prey, the group behavior is identified as a search.

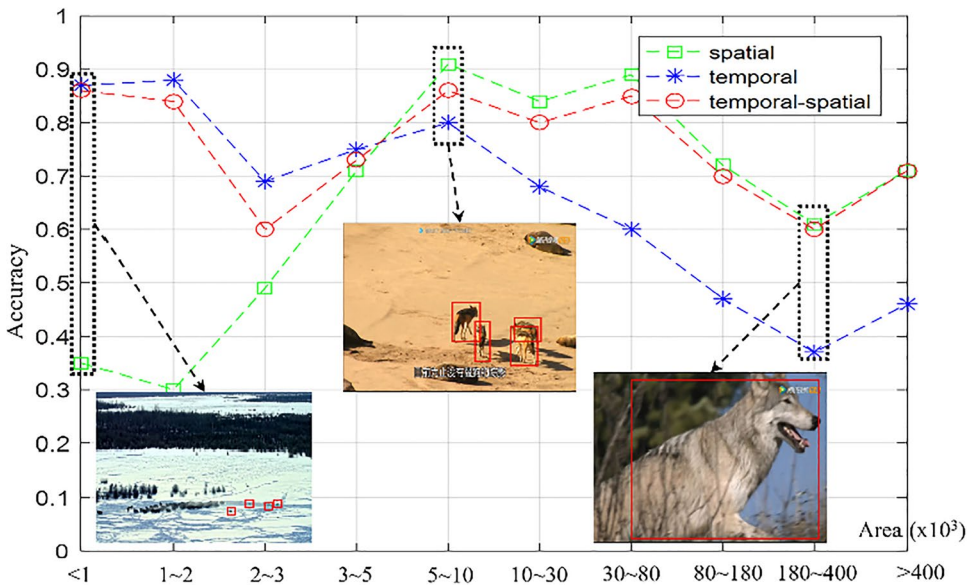

Fig. 6 The motion state recognition accuracy of spatial, temporal, and temporal–spatial features-based methods for the animals in different scale intervals. The horizontal axis indicates the pixel area (pixel^2) of the animal image region

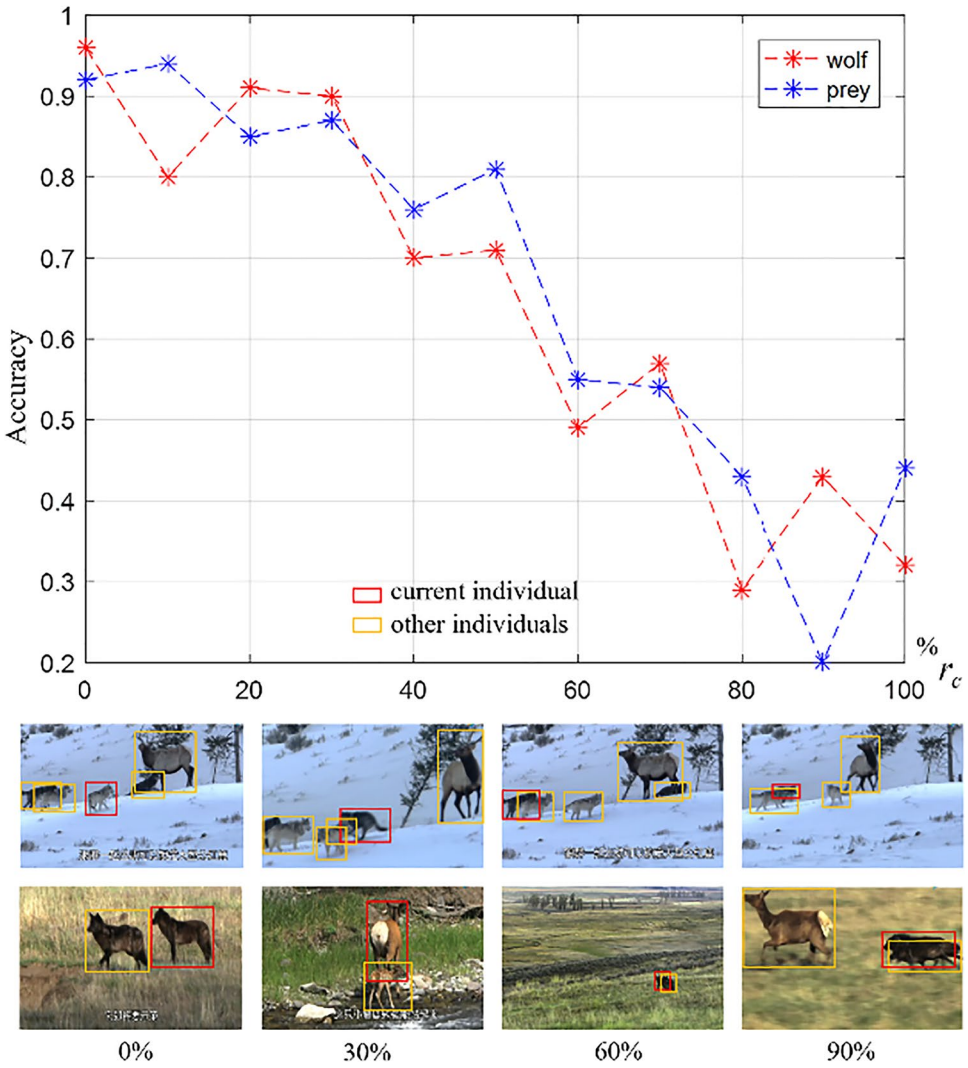

Fig. 7 The motion state recognition accuracy of wolf and grey under different occlusion rates, defined as formula (13). Sample frames with 0%, 30%, 60%, and 90% occlusion rates are given

Data compilation of Wolf2022 dataset

The *Wolf2022* dataset comprises 13,100 frames, sampled at equal intervals from 13 videos (total 670 min) collected from the internet involving wolf pack hunting, and the source of these videos could be found in the data availability statement at the end of this paper. The dataset involves five scenarios/habitats, including grassland, swamps, forest, sand beach, and tundra, and six hunting states, including search, watch, approach, attack on an individual, attack on a group, and capture of prey. It provides two types of manual annotation for the tasks of group behavior analysis and individual behavior analysis, respectively. For the group behavior, the annotation comprising the behavior category, start and end time, wolf group size, prey group size, illumination, and shooting perspective is provided for each hunting behavior video clip. For the individual behavior, category, image index, image region (bounding box), and motion state are given for all individuals, in each frame, even if a frame contains a large number of individuals. The individual annotation enables the research of the object detection, tracking, and motion analysis. In addition, the video's attribute, i.e., location, environment, group size, prey, and field-of-view, is also provided for each video.

Results and discussion

Individual behavior

For the frame, video, and space-level feature extractions, we directly employ existing algorithms for their extraction (YOLO-v4 for frame-level, DeepSort for video-level, and optical flow for space-level), and they all are contributing to the semantic-level feature inference. Therefore, this part mainly focuses on the performance validation and analysis

Table 4 The group behavior detection results (in boldface, the misrecognized motion/behavior). "Number" shows the total number of prey and wolves. "GT" indicates the ground truth, and "Ours" means the results inferred by our method

Index	Number	Environment	Motion state				Time IOU	Group behavior	
			Wolves		Prey				
			GT	Ours	GT	Ours		GT	Ours
Clip 1	3	Tundra	Static	Static	Static	Static	96%	Watch	Watch
Clip 2	3	Tundra	Run	Run	Run	Run	93%	Chase	Chase
Clip 3	6	Grassland	Walk	Walk	Static	Static	99%	Approach	Approach
Clip 4	2	Grassland	Walk	Walk	Walk	Walk	95%	Approach	Approach
Clip 5	2	Grassland	Run	Run	Run	Run	96%	Chase	Chase
Clip 6	> 20	Tundra	Walk	Walk	Empty	Empty	96%	Search	Search
Clip 7	> 20	Tundra	**Walk**	**Static**	Static	Static	97%	Approach	Approach
Clip 8	> 20	Tundra	Run	Run	Run	Run	99%	Chase	Chase
Clip 9	> 20	Tundra	**Run**	**Walk**	**Run**	**Walk**	93%	**Chase**	**Approach**
Clip 10	4	Forest	Run	Run	Run	Run	94%	Chase	Chase
Clip 11	8	Forest	Run	Run	Run	Run	94%	Chase	Chase
Clip 12	8	Swamp	Walk	Walk	Empty	Empty	98%	Search	Search
Clip 13	8	Swamp	Static	Static	Walk	Walk	98%	Approach	Approach
Clip 14	7	Swamp	Static	Static	**Static**	**Empty**	**37%**	**Watch**	**Search**
Clip 15	7	Swamp	Run	Run	Run	Run	96%	Chase	Chase
Clip 16	14	Tundra	Empty	Empty	Static	Static	92%	Search	Search
Clip 17	> 20	Tundra	Walk	Walk	Static	Static	96%	Approach	Approach

for the temporal-spatial feature fusion-based semantic-level feature inference module.

Video-level feature

Since the video-level feature extraction plays a vital role in semantic-level feature inference, we first briefly present and discuss the object tracking precision. The video-level feature is generated using DeepSORT, while YOLO-v4 is used for object detection. We train YOLO-v4 employing about 30% of the images from video-1 to video-5, constituting the training and validation set. The test set comprises two parts: the rest of images of video-1 to video-5 and the images from the three other videos. According to the tracking accuracy shown in Table 3, the average tracking precision of the two parts of the test set reaches 86.8% and 84.1%, respectively. Although the training set does not contain the images from video-6 to video-8, the tracking precision of the second testing part is not significantly reduced compared with the first part test. Since the individuals are numerous and dense, about 90% of mis-tracking is due to occlusion between individuals.

Semantic-level feature

Spatial appearance feature-based motion state recognition essentially relies on the differences in the animal's limb state under different motion states. However, when the animal scale in the images is too small, the animal limb parts are difficult to distinguish, significantly reducing recognition accuracy. Therefore, we introduce the temporal feature to improve robustness of the animals' scale. To validate this concept, we set up multiple intervals of object image scale and compute the recognition accuracy for all scale intervals. For an independent analysis of the proposed motion state recognition algorithm, we employ the object tracking

Fig. 8 The sample frames of clip 9. The average occlusion rate between individuals reaches about 58%, and flying snow and surrounding plants often occlude the animals' limbs

Fig. 9 The frame sequence in clip 14. After the wolf scene, the following scene is not about prey. A commentary is inserted between the wolf and prey scenes

ground truth for validating the motion state recognition. Figure 6 depicts the motion state recognition accuracy of the three feature types and the image demos under multiple animal image scales. For the spatial feature-based results, for an animal image area less than $(5 \sim 10) \times 10^3$, the larger the image scale, the higher the recognition accuracy. However, the accuracy gradually increases when the area exceeds 10×10^3 pixel^2 although the image features become more remarkable with an increasing scale. This is due to the more frequent occlusion between animals as the scale increases. According to the temporal feature-based method, accuracy is mainly affected by mis-tracking. Since the occlusion between the individuals is more frequent as the individual scale increases, the tracking precision gradually decreases as the scale increases. Hence, recognition accuracy gradually decreases as the scale increases (the blue line in Fig. 6). Although none of the spatial and temporal feature-based modules have achieved more than 70% recognition accuracy, employing the dynamic weighting (formula (9)) increases the recognition accuracy to 88%.

To validate that occlusion is the leading cause of motion state misrecognition, we define the individual occlusion rate r_c:

$$r_c = \frac{s_{in}}{s} \tag{13}$$

where s denotes the individual pixel area and s_{in} is the sum of the overlapping area of all individual bounding boxes. As Fig. 7 illustrates, the recognition accuracy shows a remarkable downtrend with increased occlusion rate r_c. For the individuals without occlusion, we realize a recognition accuracy exceeding 90%, while for individuals with occlusion of more than 60%, the recognition accuracy drops below 50%, indicating that the occlusion between individuals is one of the critical factors causing recognition errors.

Group behavior

The test dataset involves eight videos and contains 17 group behavior clips. Each clip contains only one group behavior and lasts no less than 10 s. The behavior detection result is deemed correct when it satisfies two conditions: the category of the group behavior is correct, and the intersection over union (IOU) of the start and end time in the video is not less than 90%. According to the group behavior detection results reported in Table 4, using the individual feature map generated by the proposed methods, two clips have the wrong group behavior category, and one has a time IOU below 90%. For clip 9, the group behavior is misclassified because the motion states of the wolves and prey are misrecognized (the motion state of most individuals is classified as "walk," while it should be "run"). Figure 8 displays multiple typical frames of clip 9, highlighting that the high occlusion rate between individuals (up to 58%) directly causes a significant motion state misclassification. Besides direct occlusion, the flying snow and surrounding plants also occlude the animals' limbs. For clip 14, our method identifies the prey's motion state as empty, which means no prey is detected. Figure 9 reveals that after the wolf scene is played, the prey scene is not played continuously, and a commentary is inserted between the wolf and prey scenes. Thus, our method does not consider them to be the same scene. Although clip 7 is correctly detected, the wolf pack's motion state is misrecognized, as most wolves are pacing, while part of them are misidentified as "static" by our method.

Conclusion

This paper presents a new wolf pack hunting behavior description method comprising multi-level individual feature maps and group feature maps, and develops a pipeline that automatically detects the hunting behavior from video clips. Moreover, we propose a temporal–spatial feature fusion-based motion state recognition method for a robust individual semantic-level feature extraction. The experimental results demonstrate our method's robustness and accuracy in motion state recognition for scale-varied and occlusion-frequent individuals with complicated backgrounds. This paper validated our pipeline's feasibility and accuracy in hunting behavior video detection. Future work will focus on constructing a toolbox for standardized, diverse, and convenient group behavior detection.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00265-023-03328-4.

Acknowledgements The authors thank Chao Yan, Zhen Lan, Zixing Li, Xingyu Liu, Yixin Huang, Yihao Sun, Qin Tan, and Fan Yang for their contribution on the dataset Wolf2022 construction and the method development. The authors also thank the reviewers for their valuable and constructive suggestions and comments.

Author contribution DT: conceptualization, methodology, writing original draft. CH: dataset construction. XX, HZ: investigation, supervision. SZ: software, validation. TH: supervision, writing reviewing and editing.

Data availability The dataset (Wolf2022) analyzed during the current study are available in the repository: https://drive.google.com/drive/folders/1e0W6Eu0YOJa1LdJBdVAMzPnNwBqNiPam?usp=sharing.

Declarations

Ethical approval All applicable international, national, and/or institutional guidelines for the use of animals were followed. Our study does not involve experimental activities such as animal photography, and directly uses publicly available documentary online videos as data sources. Therefore, ethical approval is not required.

Conflict of interest The authors declare no competing interests.

References

Bochkovskiy A, Wang CY, Liao H (2020) YOLOv4: optimal speed and accuracy of object detection. arXiv:2004.10934

Cassidy KA, MacNulty DR, Stahler DR, Smith DW, Mech LD (2015) Group composition effects on aggressive interpack interactions of gray wolves in Yellowstone National Park. Behav Ecol 26:1352–1360

Dickie M, Serrouya R, McNay RS, Boutin S (2016) Faster and farther: wolf movement on linear features and implications for hunting behaviour. J Appl Ecol 54:253–263

Duan H, Yang Q, Deng Y, Li P, Qiu H, Zhang T, Zhang D, Huo M, Shen Y (2019a) Unmanned aerial systems coordinate target allocation based on wolf behaviors. Sci China 62:014201

Duan H, Zhang D, Fan Y, Deng Y (2019b) From intelligence of wolves to collaborative decision make of UAV swarm. Sci China 49:112–118

Escobedo R, Muro C, Spector L, Coppinger P (2014) Group size, individual role differentiation and effectiveness of cooperation in a homogeneous group of hunters. J R Soc Interface 11:20140204

Graving JM, Chae D, Naik H, Li L, Koger B, Costelloe BR, Couzin ID (2019) DeepPoseKit: a software toolkit for fast and robust animal pose estimation using deep learning. eLife 8:47994

Hawley JE, Gehring TM, Schultz RN, Rossler ST, Wydeven AP (2010) Assessment of shock collars as nonlethal management for wolves in Wisconsin. J Wildlife Manag 73:518–525

Harris C, Stephens M (1988) A combined corner and edge detector. Alvey Vision Conference, Manchester, pp 147–151

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778

Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. International Joint Conference on Artificial Intelligence, Vancouver, pp 674–679

Madden JD, Arkin RC, Macnulty DR (2010) Multi-robot system based on model of wolf hunting behavior to emulate wolf and elk interactions. IEEE International Conference on Robotics and Biomimetics, Tianjin, China, pp 1043–1050

Madden JD, Arkin RC (2011) Modeling the effects of mass and age variation in wolves to explore the effects of heterogeneity in robot team composition. IEEE International Conference on Robotics and Biomimetics, Karon Beach, Thailand, pp 663–670

Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, Bethge M (2018) DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci 21:1281–1289

MacNulty DR, Mech LD, Smith DW (2007) A proposed ethogram of large-carnivore predatory behavior, exemplified by the wolf. J Mammal 88:595–605

Mech LD, Smith DW, MacNulty DR (2015) Wolves on the hunt: the behavior of wolves hunting wild prey. The University of Chicago Press, Chicago

Pérez-Escudero A, Vicente-Page J, Hinz RC, Arganda S, Polavieja GG (2014) IdTracker: tracking individuals in a group by automatic identification of unmarked animals. Nat Methods 11:743–748

RoianEgnor SE, Branson K (2016) Computational analysis of behavior. Annu Rev Neurosci 39:217–236

Strandburg-Peshkin A, Farine DR, Couzin ID, Crofoot MC (2015) Shared decision-making drives collective movement in wild baboons. Science 348:1358–1361

Schlagel UE, Merrill EH, Lewis MA (2017) Territory surveillance and prey management: wolves keep track of space and time. Ecol Evol 7:8388–8405

Wojke N, Bewley A, Paulus D (2017) Simple online and realtime tracking with a deep association metric. International Conference on Image Processing, pp 3645–3649

Xie Y, Han L, Dong X, Li Q, Ren Z (2021) Bio-inspired adaptive formation tracking control for swarm systems with application to UAV swarm systems. Neurocomputing 453:272–285

Zhao L, Yang G, Wang W, Chen Y, Huang JP, Ohashi H, Stanley HE (2011) Herd behavior in a complex adaptive system. P Natl Acad Sci USA 108:15058–15063