반려견에 초점을 맞춰 추출하는 영상 기반의 행동 탐지 시스템

오스만[1], 이종욱[2], 박대희[2], 정용화[2]
[1]고려대학교 컴퓨터정보학과
[2]고려대학교 컴퓨터융합소프트웨어학과
osuman@korea.ac.kr, eastwest9@korea.ac.kr, dhpark@korea.ac.kr, ychungy@korea.ac.kr

Dog Activities Recognition System using Dog-centered Cropped Images

Othmane Atif[1], Jonguk Lee[2], Daihee Park[2], Yongwha Chung[2]
[1]Dept. of Computer Information Science, Korea University
[2]Dept. of Computer Convergence Software, Korea University

Abstract

In recent years, the growing popularity of dogs due to the benefits they bring their owners has contributed to the increase of the number of dogs raised. For owners, it is their responsibility to ensure their dogs' health and safety. However, it is challenging for them to continuously monitor their dogs' activities, which are important to understand and guarantee their wellbeing. In this work, we introduce a camera-based monitoring system to help owners automatically monitor their dogs' activities. The system receives sequences of RGB images and uses YOLOv7 to detect the dog presence, and then applies post-processing to perform dog-centered image cropping on each input sequence. The optical flow is extracted from each sequence, and both sequences of RGB and flow are input to a two-stream EfficientNet to extract their respective features. Finally, the features are concatenated, and a bi-directional LSTM is utilized to retrieve temporal features and recognize the activity. The experiments prove that our system achieves a good performance with the F-1 score exceeding 0.90 for all activities and reaching 0.963 on average.

1. INTRODUCTION

For centuries, dogs have been loyal and beloved companions to humans and served different purposes such as protection and mental health support, earning them the title of man's best friend. This contributed to their growing popularity, making them lead the pet market in countries like South Korea for example, where the number of dogs raised reached 5.45 million in 2022 [1]. When raising a dog, owners have responsibilities towards it, and these include providing it with basic needs and guaranteeing its safety and health. Although owners make efforts to do so, they cannot continuously watch over them, and one of the main challenges they face is the monitoring of their dog's activities, especially when left alone, which is essential as an initial step to help them understand and ensure their dog's wellbeing. Owing to that, in this study, we propose a camera-based monitoring method to help owners supervise their dog's activities.

As a first step, it is essential to detect the dog in the image sequences received from the camera to confirm its presence before using those images as input data to perform dog activities recognition. In order to achieve that, we looked into recent object detection models and selected the You Only Look Once v7 (YOLOv7) [2], which is currently one of the best real time object detection models. Furthermore, rather than using the full images as input, by cropping them to focus on the dog, we can help reduce the background area, which would allow the recognition model to learn more relevant features [3]. This would help not only improve the recognition performance, but also the system's processing speed as the cropped images would be smaller in size. Since the YOLOv7 model provides bounding boxes that define the area of the dog in the image, we can utilize them to perform image cropping. However, if the bounding boxes are used as is to crop images in an input sequence, the resulting sequence will contain cropped images of different sizes. To prevent that, a simple post processing algorithm is presented to combine the bounding boxes into a single one and use it to crop all the images in a sequence, producing a cropped-images sequence that is forwarded as input to perform activities recognition.

Traditionally, actions and activities recognition methods used machine learning models such as Support Vector Machine (SVM) to classify hand-crafted features extracted from videos [4]. Later, with the improvement of deep learning models, more methods started using two-stream convolutional neural networks (CNN) to extract both appearance features from RGB images and motion features from optical flow [5] to perform recognition. However, when applying such a method, the images in a sequence are processed independently without considering the temporal information in a sequence of images and which are important to recognize activities.

Consequently, in addition to the use of CNN models to learn the appearance and motion features, recent methods [6] utilized the long-short term memory model (LSTM), which delivers good performance with sequential data, to extract temporal features from a sequence of images, in what is referred to as a two-stream CNN-LSTM, achieving by that good recognition performance. Accordingly, to perform dog's activities recognition, we utilize a two-stream CNN-LSTM model. However, unlike previous methods [6] where the full area of the images is fed as input to the CNN-LSTM, in our system, we used the previously mentioned cropping module as a preprocessing step to crop the specific area in the images where the dog is present and use the resulting cropped images as input to ensure that the model delivers good performance.

Therefore, in this paper, we propose a cropped image-based dogs' activities monitoring system. Our work uses the YOLOv7 model to detect the dog in the video stream and utilize the generated bounding boxes to perform cropping on input images through a post processing algorithm before feeding them as a sequence to a two-stream CNN-LSTM for the dog's activities recognition.

2. PROPOSED METHOD

The general proposed architecture of our system is shown below in Figure 1. The system contains two main modules: an *image data collection and cropping module* and a *dog activities recognition module*.

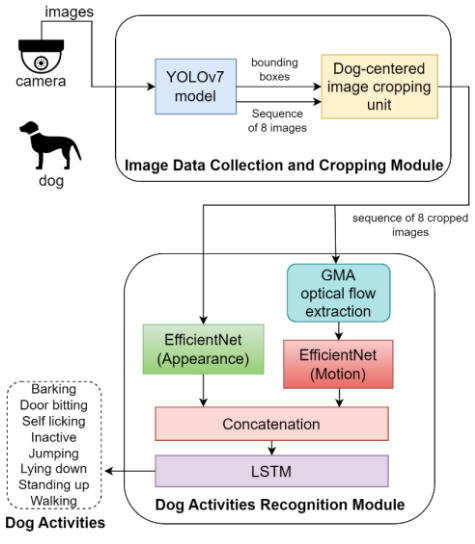

Figure 1: General proposed architecture of the dog activities recognition system

2.1 Image Data Collection and Cropping Module

In this module, images of size 960 × 1080 are collected from a camera and grouped in sequences of 8 images, which represents the input length set to the two-stream CNN-LSTM. Each image from a sequence is fed to the YOLOv7 model to perform object detection and extract dog bounding boxes, and linear interpolation, which is widely used due to its simplicity and efficiency, is used to correct misdetections. If a dog is detected in the sequence of images, the sequence is kept for use as input for activities' recognition. After that, the cropping unit evaluates the 8 bounding boxes to select the smallest minimum x and y and the largest maximum x and y from all the bounding boxes and use them to define a unified bounding box which covers the dog location in the image sequence. At this stage, if each sequence is processed and cropped independently, the size of the images will differ from one sequence to another. To guarantee that all the resulting sequences contain cropped images of the same size, an analysis was performed on the dataset to estimate the largest possible size of the unified box, which gave us a value of 300 × 700. Accordingly, after calculating the coordinates of the united box for each sequence, the width and height of the box are expanded to reach 300 × 700 before using it to crop the 8 images in the input sequence. Finally, this results in a sequence containing 8 images cropped with a focus on the dog, which is passed on to the following module.

2.2 Dog Activities Recognition Module

The sequence of 8 dog-focused images is fed to the recognition module in two separate streams. The first stream takes the RGB images as is and feeds them to their corresponding CNN model which has been pretrained to extract spatial appearance features. On the other hand, the second stream first uses the Global Motion Aggregation (GMA) [7] algorithm, as it a low error rate and guarantees fast processing, to extract the optical flow from the RGB sequence. The resulting 7 optical flow images are then fed to the CNN pretrained to extract motion features. The EfficientNet-B0 [8] was selected as CNN model in this module for both appearance and motion features extraction since it delivers a good performance with fast inference. After that, two features' vectors of shape (8,1280) and (7,1280) are extracted from the RGB and flow streams respectively. Before concatenating the two feature vectors, the first element of the RGB features vector is dropped to guarantee a matching shape with the flow features vector. The concatenated feature vector of shape (7, 2560) is then given to the LSTM model to extract temporal features and use a SoftMax layer to recognize the dog activity. A bi-directional LSTM (Bi-LSTM) is used here as it has a capacity to better define temporal boundaries of activities [9].

3. EXPERIMENTAL RESULTS

3.1 Data Collection Experiments

The data collection was conducted in a safe enclosed area with a camera recording from the ceiling at a rate of 21 fps. Two dog owners volunteered to help us collect video data recordings of the activities of two different small-sized dogs separately. The recordings were limited to 15~20 minutes with breaks to prevent any discomfort for the dogs.

The object detection annotation to train the YOLOv7 model and temporal activities annotation to train the CNN-LSTM model were performed on the collected data using VitBAT [10]. We selected eight classes of activities for which data was collected and prepared: the "*Standing up*" class contained 7928 images, and "*Barking*", "*Door Biting*", "*Self-Licking*", "*Jumping*", "*Lying down*", "*Walking*" and "*Inactive*" (which is when the dog is standing or sitting still) each contained 8000 images, with the total adding up to 63928 images. All the images were organized in sequences of 8 images to fit the size of the CNN-LSTM model's input, where each sequence represents an activity. The dataset was then split into 3 sets for training, validation and testing using the

ratio 7:2:1. In order to increase the size of the training dataset, shear transformation of [[1,0,0],[0.2,1,0]], horizontal and vertical flip, and rotation with a factor of 0.10 were applied, which resulted in a training set of 27880 images for "*Standing up*" and 28000 images for each of the other classes. From this training dataset, 24520 images were randomly selected and split into datasets using a ratio of 8:2 to train and test the YOLOv7 model, which was then used with the cropping unit to prepare a train, validation, and testing datasets of cropped images. After that, GMA was used to extract the optical flow of every sequence of 8 images in the 3 datasets to generate 3 corresponding datasets with sequences of 7 color-coded optical flow frames. Each EfficientNet model (appearance and motion) was trained using its corresponding datasets.

3.2 Implementation Details

The experiments and implementation were performed on a computer with an AMD Ryzen 9 5900X CPU, 32 GB of RAM, an RTX 3080Ti graphic card, and a Windows 10 operating system. The YOLOv7 model and GMA followed the official implementations and used the default configurations. The two EfficientNet-B0 and bidirectional LSTM models were built and trained using the TensorFlow library with Adam optimizer. The EfficientNet (appearance) was trained for 100 epochs with a learning rate of 5×10^{-5} while the EfficientNet (motion) was trained for 300 epochs with a learning rate of 1×10^{-6}. The top classifier layers from both models were removed after the training to use the models as feature extractors, with each outputting a feature vector of size 1280. The bi-directional LSTM was built with 2 layers, 60 hidden units and was trained for 20 epochs with a learning rate of 5×10^{-6}, 0.5 of dropout and a recurrent dropout of 0.1.

3.3 Object Detection and Activities Recognition Results

The YOLOv7 object detection results showed a mean average precision of 0.987 in dog detection. On the other hand, to evaluate the activities recognition results, Precision, Recall and F1-score were used as shown in Table 1 below. As seen in the table, the F-1 score results of each class exceeded 0.90 with an average of 0.963 for all activities, confirming that the proposed method attains good recognition results which can help owners monitor their dog's activities.

<Table 1> Dog activities recognition experiment results

Actions	Precision	Recall	F-1 score	Support
Barking	0.980	0.990	0.985	100
Door Biting	0.962	1.000	0.980	100
Self-Licking	0.990	1.000	0.995	100
Inactive	0.956	0.860	0.905	100
Jumping	0.942	0.980	0.961	100
Lying down	0.944	1.000	0.971	100
Standing up	0.978	0.928	0.952	97
Walking	0.960	0.950	0.955	100
Average/ total	**0.964**	**0.964**	**0.963**	797

4. CONCLUSION

In this work, we proposed a system to recognize daily life dog activities to help owners better monitor them. The system uses YOLOv7 with a simple algorithm to crop the input images, and then uses a two stream EfficientNet to extract appearance and motion features before feeding them to a bi-directional LSTM to recognize the activity. The results of the experiments prove that the system that we proposed can deliver a good performance in recognizing dog activities.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1I1A3070835 and NRF-2021R1I1A3049475).

References

[1] N. Jobst, "Estimated number of pet dogs in South Korea from 2010 to 2022," 2023. https://www.statista.com/statistics/661495/south-korea-dog-population/

[2] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," arXiv:2207.02696, 2022.

[3] B. K. Mishra, D. Thakker, S. Mazumdar, D. Neagu, M. Gheorghe, and S. Simpson, "A novel application of deep learning with image cropping: a smart city use case for flood monitoring," Journal of Reliable Intelligent Environments, Vol. 6, No. 1, pp. 51–61, 2020.

[4] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," Computer Vision and Image Understanding, Vol. 150, pp. 109–125, 2016.

[5] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," Proceedings of the Advances in Neural Information Processing Systems (NIPS), Montreal, 2014, pp. 568–576.

[6] O. Atif, J. Lee, D. Park, and Y. Chung, "Camera-based dog unwanted behavior detection," Proceedings of the Korea Information Processing Society (KIPS), Seoul, 2019, pp. 419-422.

[7] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley, "Learning to estimate hidden motions with global motion aggregation," Proceedings of the IEEE International Conference on Computer Vision (ICCV), Montreal, 2021, pp. 9752–9761.

[8] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 36th International Conference on Machine Learning (ICML), Long Beach, 2019, pp. 10691–10700.

[9] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016, pp. 1961–1970.

[10] T. A. Biresaw, T. Nawaz, J. Ferryman, and A. I. Dell, "ViTBAT: Video tracking and behavior annotation tool," IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Colorado Springs, 2016, pp. 295–301.