영상 기반 강아지의 이상 행동 탐지

오스만*, 이종욱**†, 박대희**, 정용화**
*고려대학교 컴퓨터정보학과, **고려대학교 컴퓨터융합소프트웨어학과
e-mail: osumanatif@gmail.com

Camera-based Dog Unwanted Behavior Detection

Othmane Atif*, Jonguk Lee**†, Daehee Park**, Yongwha Chung**
*Dept. of Computer Information Science, Korea University
**Dept. of Computer Convergence Software, Korea University

Abstract

The recent increase in single-person households and family income has led to an increase in the number of pet owners. However, due to the owners' difficulty to communicate with them for 24 hours, pets, and especially dogs, tend to display unwanted behavior that can be harmful to themselves and their environment when left alone. Therefore, detecting those behaviors when the owner is absent is necessary to suppress them and prevent any damage. In this paper, we propose a camera-based system that detects a set of normal and unwanted behaviors using deep learning algorithms to monitor dogs when left alone at home. The frames collected from the camera are arranged into sequences of RGB frames and their corresponding optical flow sequences, and then features are extracted from each data flow using pre-trained VGG-16 models. The extracted features from each sequence are concatenated and input to a bi-directional LSTM network that classifies the dog action into one of the targeted classes. The experimental results show that our method achieves a good performance exceeding 0.9 in precision, recall and f-1 score.

1. INTRODUCTION

Nowadays, it has become very common for families around the world to raise pets as companions. In South Korea for example, a survey conducted by the Korea Pet Food Association (Kopfa) in 2017 estimated the percentage of Koreans who own a pet to be 20%, with dogs being the most popular pet choice at a percentage of 78.7%. This number increased to 81.3% in the report from the 2018 survey and the number is expected to increase [1]. Dogs, by nature, tend to be very dependent on their owners and when left alone at home, they start displaying some disruptive and undesired behaviors. These behaviors can cause damage to the house, can physically harm dogs resulting in broken teeth and hurt jaw, and are one of the main reasons pushing owners to abandon them [2]. For owners, their little companions' well-being is a serious matter of concern that affects them emotionally and financially. In order to detect and suppress those behaviors, dogs left alone at home need constant monitoring. However, most owners cannot afford to keep observing their dogs' behavior when they are absent. In this study, we propose a method to monitor the behavior of dogs left alone at home to detect the occurrence of unwanted behaviors based on computer vision and deep learning algorithms.

Recently, research that makes use of convergence technology for the understating of domestic animal behavior has been increasing. In some research, monitoring of animals in farms such as pigs and cows was done through image processing for productivity and health management. For example, Nasirahmadi et al [3] presented a work where they used Support Vector Machines (SVM) to automatically classify pigs' postures from RGB images collected through a camera. In the method proposed by Zhang et al [4], a Convolutional Neural Networks (CNN) model used to extract features was combined with a tracking algorithm to perform individual tracking and identification on a group of pigs standing close to each other and overlapping. In addition, by considering the captured videos as time series, other research took advantage of Long-Short Term Memory (LSTM) in order to extract and learn from the temporal features, like in the work done by Brattoli et al [5] to analyze the postures and behavior of rats. This type of approach combining both spatial and temporal features leads to better behavior and action detection results.

In this paper, we extend the use of cameras from livestock monitoring to the field of a dog that single-person household raised and propose a camera-based monitoring system to detect unwanted behavior in the dog left alone at home. Our work is based on a two-stream CNN-LSTM algorithm similar but simpler than the one used by Singh et al [6] to detect human actions while shopping. The system works by using two separate CNN to extract features from both the appearance (RGB) and the motion (optical flow) data before combining their outputs and feeding it to a Recurrent Neural Network of type LSTM to classify the action.

† Corresponding author

2. PROPOSED METHOD

The overall architecture of the proposed system is shown in Figure 1. The system is composed of three modules: *Video data collection module*, *flow extraction and preprocessing module*, and *dog behavior detection module*.

2.1 Video Data Collection Module

This module is the one responsible for collecting data from the camera sensor, trim it into equal sequences of frames and resize them to 224×224 to fit the feature extractor's input requirements.

2.2 Optical Flow Extraction and Preprocessing Module

The sequence of resized frames goes into two parallel flows of data. One flow is used for optical flow extraction to obtain motion data between every two consecutive frames and the other one is kept as RGB frames for appearance data. The optical flow extraction unit makes use of PWC (Pyramid, Warping, and Cost volume)-net, which is based on CNNs and uses pyramid, warping and cost volume, and gives a good trade-off between accuracy and speed [7]. The optical flow data is then color-coded. We apply the data preprocessing required for the feature extractor on both flows of data before forwarding them to the next module.

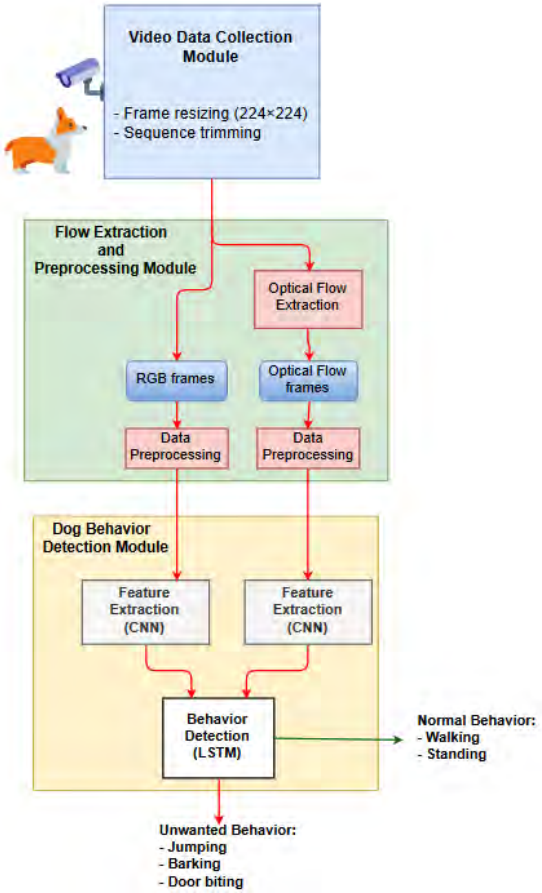


Figure 1: General architecture of the dog-unwanted behavior detection system

2.3 Dog Behavior Detection Module

Both sequences of RGB and optical flow data are fed to two independent convolutional neural networks based on the VGG-16 architecture. The VGG-16 is a popular CNN model introduced by Simonyan et al [8] that outperformed earlier models and ranked first in different competitions. Although outperformed by more recent models, it remains a popular choice as a CNN feature extractor due to its simple sequential nature and its Area Under ROC Curve (AUC) value which is better than some other models. This guarantees more capability of distinguishing between classes.

Each network was fine-tuned on its corresponding type of data between RGB and optical flow and the top block was removed so that it can be used as a feature extractor. The output is concatenated and input to the LSTM network to detect the corresponding behavior and notify the user when the behaviors displayed are unwanted.

3. EXPERIMENTAL RESULTS

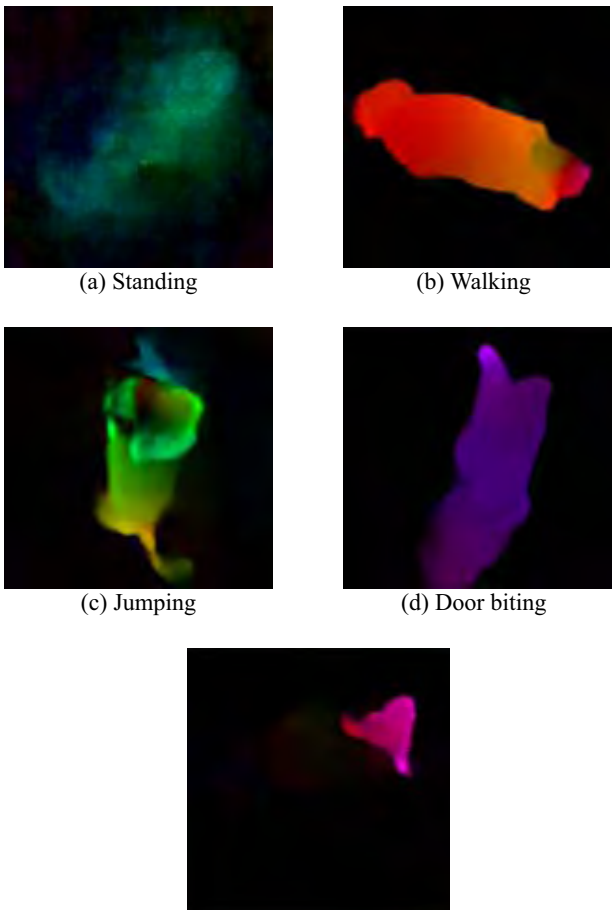
3.1 Data Collection Experiments

The dog recruited for the experiment is a rescued one owned by a local who volunteered to participate in our data collection. The owner lived alone with the dog and reported some unwanted behaviors that confirm that the pet has separation anxiety issues, which made it a suitable subject for our study. The data collection experiments were conducted in a closed and safe area and were recorded using a CCTV camera recording at a frame rate of 23 fps. In order to ensure the safety and comfort of the subject, the experiments were limited to a short duration of time with breaks in between and participants were constantly monitoring it in real time inside the same room. Figure 2 shows an example of a frame captured by the camera during the experiment.

Figure 2: Data collection for experiments

In our experiments, the camera was mounted on the ceiling and adjusted to provide a top-down angle view. The videos were trimmed using VideoPad Video Editor to create short clips containing our targeted actions and then labelled using ViTBAT [9]. We focused on five actions including two normal behaviors, "*Walking*" and "*Standing*", and three unwanted actions, namely "*Jumping*", "*Barking*" and "*Door biting*". Each class of the dataset contains a total of 1860 frames arranged in sequences of 10 frames. This means that every 10 RGB frames represent a continuous sequence of the same action and the reason for this organization is to simplify the optical flow extraction and to maintain a fixed length of input for the LSTM network. For each sequence of 10 RGB frames, the optical flow was extracted using PWC-net resulting in a corresponding sequence of 9 optical flow color-coded frames. After the extraction, we obtain a dataset of optical flow frames containing 1674 frames per class, and each of the two datasets

is used for fine-tuning its feature extractor. Figure 3 below shows examples of color-coded results for each class.



(a) Standing (b) Walking

(c) Jumping (d) Door biting

(e) Barking

Figure 3: Actions' optical flow extracted with PWC-net.

3.2 Feature extraction using VGG-16

The action detection architecture used in the proposed system is similar to the one used to detect fine-grained action in the MERL Shopping Dataset by Singh et al. [6]. However, instead of using a multi-stream network, we only used a two-stream one. The reason for this is that the nature of the actions in the dataset they used is different from our targeted dataset. Unlike the subjects shopping, the dog moves around in the designated area and our targeted actions tend to happen in specific areas of the room. Hence, using two more streams with a cropped figure to remove the background is not necessary in our case. In addition, the nature of the dog's movements inside the room makes it more challenging to get a stable tracking and can negatively affect the detection.

Each of the RGB and optical flow datasets were separately used to fine tune CNNs based on the VGG-16 architecture. Only the layers of the three lower convolutional blocks were frozen and the two remaining ones, containing 4 layers each, were set to trainable in order to fine-tune them on our datasets.

3.3 Bi-directional LSTM for Action Detection

After fine-tuning them the two networks, we disregard their classifiers and use the bottleneck features extracted as vectors containing 512 features. When inputting RGB sequences, we disregard the first frame of each sequence to match the size of the optical flow sequence. The means of the extracted features from each network are then concatenated and the result serves as input to a bi-directional LSTM as shown in Figure 4. The bi-directional LSTM has proven to give better results [6] since it runs forward and backward on the sequences of data allowing better learning of temporal features. A softmax layer is put on top of the LSTM to classify the actions.

3.4 Implementation Details

The networks fine-tuning and training were conducted in a computer running Windows 10, with an Intel i5-8500 CPU, 16 GB of RAM and a GTX 1080 Ti graphic card.

In order to construct the networks, we used the Keras library with TensorFlow as a back-end. The VGG-16 networks were based on their Keras implementation with ImageNet weights and were fine-tuned using a learning rate of 0.000001 with Rectified Linear Unit as activation function and a Stochastic Gradient Descent (SDG) optimizer. In this step, we use the two separate datasets for training. As mentioned earlier, we disregard 1 frame per RGB sequence to match the length with its equivalent optical flow sequence. This means the total number of RGB frames and optical flow color-coded frames is 8370 frames per each dataset, 80% of it used for training and 20% for validation.

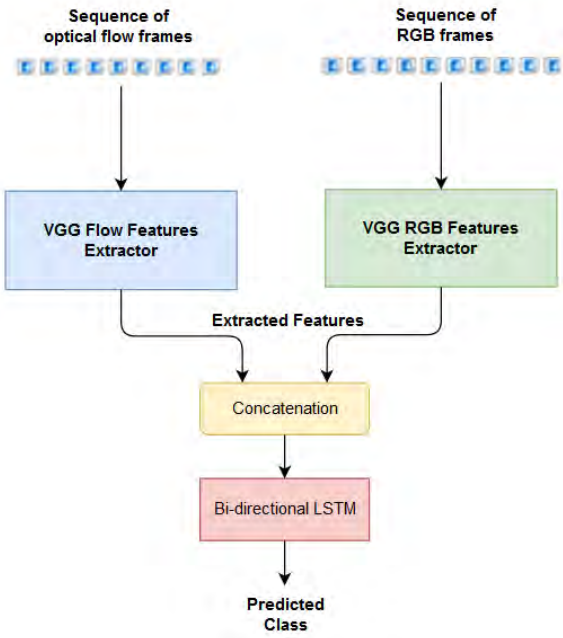


Figure 4: Two-stream bi-directional LSTM network

The bi-directional LSTM had only one layer since a deeper network does not learn well due to the dataset relatively small size. We used 60 hidden units and it was trained with an SDG optimizer, a learning rate of 0.00001, a decay of 10^{-6} and momentum of 0.9.

3.5 Action Detection Results

A summary of the results of the dog actions classification using our proposed method is shown in table 1 below. As seen in the table, we used three different metrics: Precision, Recall and, f1-score (The closer the value is to 1, the better the model performance is). The performance results for each of the

targeted actions exceeded 0.9 and reached 0.94 on average, confirming that the proposed method achieves good action classification results that can help monitor dogs to detect unwanted behaviors.

Table 1: Proposed method experiment results

Actions	Precision	Recall	F-1 score
Barking	0.90	0.93	0.92
Door Biting	0.99	0.96	0.98
Jumping	0.96	0.93	0.95
Standing	0.91	0.95	0.93
Walking	0.92	0.91	0.92
Average/ total	**0.94**	**0.94**	**0.94**

4. CONCLUSION

In this paper, we proposed a system to classify some basic dog actions in order to detect the occurrence of unwanted behavior that can harm it when left alone. We used two parallel CNNs to extract features from both RGB and optical flow frames and fed them to a bi-directional LSTM to classify the action. We also show through experimental results that the proposed method can accurately classify the different targeted actions.

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-2018R1D1A3B07044938).

References

[1] Korea Pet Food Association (Kopfa), “2017 & 2018 반려동물 보유 현황 및 국민 인식 조사 보고서,” [http://kopfa.kr/index.php?mid=pds]

[2] S. Cannas, D. Frank, M. Minero, A. Aspesi, R. Benedetti, and C. Palestrini, **“**Video analysis of dogs suffering from anxiety when left home alone and treated with clomipramine,” Journal of Veterinary Behavior , Vol. 9, No. 2, pp. 50-57, 2014.

[3] A. Nasirahmadi, B. Sturm, A.C. Olsson, K.H. Jeppsson, S. Müller, S. Edwards, and O. Hensel, “Automatic scoring of lateral and sternal lying posture in grouped pigs using image processing and Support Vector Machine,” Computers and Electronics in Agriculture, Vol. 156, pp. 475-481, 2019.

[4] L. Zhang, H. Gray, X. Ye, L. Collins, and N. Allinson, “Automatic individual pig detection and tracking in surveillance videos,” Sensors, Vol.19, 2019

[5] B. Brattoli, U. Buchler, A.S. Wahl, M.E. Schwab, and B. Ommer, “LSTM self-supervision for detailed behavior analysis,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6466-6475, 2017

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

[7] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. “Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8934-8943, 2018.

[8] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” ArXiv, 1409.1556, 2014.

[9] T.A. Biresaw, T. Nawaz, J. Ferryman, and A. Dell, "ViTBAT: Video Tracking and Behavior Annotation Tool," IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 295–301, 2016.