applied sciences

MDPI

Article

Convolutional Neural Network-Based Automated System for Dog Tracking and Emotion Recognition in Video Surveillance

Huan-Yu Chen [1], Chuen-Horng Lin [1,*], Jyun-Wei Lai [1] and Yung-Kuan Chan [2]

[1] Department of Computer Science and Information Engineering, National Taichung University of Science and Technology, No. 129, Sec. 3, Sanmin Rd., Taichung 404, Taiwan; hychen@nutc.edu.tw (H.-Y.C.)
[2] Department of Management Information Systems, National Chung Hsing University, 145 Xingda Rd., South Dist., Taichung 402, Taiwan
* Correspondence: linch@nutc.edu.tw

Abstract: This paper proposes a multi–convolutional neural network (CNN)-based system for the detection, tracking, and recognition of the emotions of dogs in surveillance videos. This system detects dogs in each frame of a video, tracks the dogs in the video, and recognizes the dogs' emotions. The system uses a YOLOv3 model for dog detection. The dogs are tracked in real time with a deep association metric model (DeepDogTrack), which uses a Kalman filter combined with a CNN for processing. Thereafter, the dogs' emotional behaviors are categorized into three types—angry (or aggressive), happy (or excited), and neutral (or general) behaviors—on the basis of manual judgments made by veterinary experts and custom dog breeders. The system extracts sub-images from videos of dogs, determines whether the images are sufficient to recognize the dogs' emotions, and uses the long short-term deep features of dog memory networks model (LDFDMN) to identify the dog's emotions. The dog detection experiments were conducted using two image datasets to verify the model's effectiveness, and the detection accuracy rates were 97.59% and 94.62%, respectively. Detection errors occurred when the dog's facial features were obscured, when the dog was of a special breed, when the dog's body was covered, or when the dog region was incomplete. The dog-tracking experiments were conducted using three video datasets, each containing one or more dogs. The highest tracking accuracy rate (93.02%) was achieved when only one dog was in the video, and the highest tracking rate achieved for a video containing multiple dogs was 86.45%. Tracking errors occurred when the region covered by a dog's body increased as the dog entered or left the screen, resulting in tracking loss. The dog emotion recognition experiments were conducted using two video datasets. The emotion recognition accuracy rates were 81.73% and 76.02%, respectively. Recognition errors occurred when the background of the image was removed, resulting in the dog region being unclear and the incorrect emotion being recognized. Of the three emotions, anger was the most prominently represented; therefore, the recognition rates for angry emotions were higher than those for happy or neutral emotions. Emotion recognition errors occurred when the dog's movements were too subtle or too fast, the image was blurred, the shooting angle was suboptimal, or the video resolution was too low. Nevertheless, the current experiments revealed that the proposed system can correctly recognize the emotions of dogs in videos. The accuracy of the proposed system can be dramatically increased by using more images and videos for training the detection, tracking, and emotional recognition models. The system can then be applied in real-world situations to assist in the early identification of dogs that may exhibit aggressive behavior.

Keywords: convolutional neural networks; dog detection; dog tracking; dog emotion recognition; long short-term memory

check for updates

Citation: Chen, H.-Y.; Lin, C.-H.; Lai, J.-W.; Chan, Y.-K. Convolutional Neural Network-Based Automated System for Dog Tracking and Emotion Recognition in Video Surveillance. *Appl. Sci.* **2023**, *13*, 4596. https://doi.org/10.3390/app13074596

Academic Editor: Zhengjun Qiu

Received: 6 February 2023
Revised: 26 March 2023
Accepted: 3 April 2023
Published: 5 April 2023

1. Introduction

Keeping pets has become increasingly popular in recent years, leading to a surge in stray dogs due to abandonment, loss, and breeding. This has resulted in numerous issues, such as disease spread, attacks on humans, the disruption of urban cleanliness, and traffic accidents. Although the government uses TNvR and precise capture, addressing dog attacks is time-consuming and labor-intensive. In recent years, many surveillance cameras have been installed in essential areas, such as roads, intersections, transfer stations, and public places. However, these surveillance cameras cannot provide immediate warning messages before incidents occur. Nevertheless, recent computer vision technology can analyze camera footage and replace human reporting by sending alerts to emergency services when one or more stray dogs are detected as being about to attack. Therefore, computer vision has also been widely used for object identification. Integrating these technologies to detect and analyze dog behavior can save time and processing power, and facilitate the real-time collection of dog information and issue immediate warning alerts.

From 2014 to 2022, researchers used animal motion tracking and gesture recognition to study animal emotions and improve their emotional well-being. Sofia et al. used computer vision technology to assess animal emotions and pain recognition through a comprehensive analysis of facial and body behavior [1]. Identifying animal emotional behaviors is challenging because they express internal emotional states subjectively [2]. Researchers traditionally observe or record videos of animal behavior to analyze their behaviors. However, automatic facial and body pose analysis enables the extensive annotation of human emotional states. Fewer studies have focused on the mechanical behavior of non-human animals. Animal tracking studies include pose estimation, canine behavior analysis, and animal identification and tracking techniques using deep learning methods. Analyzing facial expressions and body behaviors to understand animal emotions presents many challenges. Techniques for recognizing animal emotional states and pain are more complex than those for tracking movement.

Recently, researchers have used computer vision and deep learning techniques for canine emotion recognition. Zhu used indoor static cameras to record dogs' behavior during locomotion, and their architecture combined pose and raw RGB streams to identify pain in dogs [3]. Franzoni et al. and Boneh et al. used images of dogs in experiments that elicited emotional states, and the main target was the detection of emotion on the dog's face [4,5]. Ferres et al. recognized dog emotions from body poses, using 23 regions on the body and face as critical points [6]. The imaging dataset for these studies was limited to a single dog, and high-resolution, clear images of faces and limbs were necessary. Research on dog emotion recognition using computer vision and deep learning has mainly focused on high-resolution, clear facial images of a single dog. These studies have generally used surveillance cameras, and the emotional state of animals has been primarily based on physical behavior due to distance and low-resolution videos. Past research on human emotion recognition has used text, audio, or video data and various models to achieve high accuracy, with facial expressions or body language analysis used for emotion recognition. However, no studies investigate dog tracking and emotion recognition due to the complexity of dog behavior and a lack of readily available imaging data.

Numerous studies on object detection have been conducted [7–12]. In object detection, colors, textures, edges, shapes, spatial relationships, and other features are extracted from data, and machine learning methods are used to classify objects according to these features. Dalal and Triggs used the histogram of an oriented gradient image feature extractor and a support vector machine (SVM) classifier to achieve human detection [7]. With the development of deep learning in artificial intelligence, convolutional neural networks (CNNs) have been applied in various deep learning technologies. Deep learning is now commonly used in computer vision, mainly because of the 2012 ImageNet Large-Scale Visual Recognition Challenge [13]. AlexNet, the deep learning network architecture proposed by Alex Krizhevsky [14], heralded the era of the CNN model. Subsequently, VGG, GoogleNet, and ResNet architectures, all of which are commonly used in innovative technologies, were developed [15–17].

Object tracking refers to the tracking of objects in continuous images; after the objects in each image are detected, they are tracked to determine and analyze their movement trajectory. Pedestrians and cars have been the objects most commonly tracked in previous studies [18–22], and the MeanShift tracking method, Kalman filter method, particle filter method, local steering kernel object texture descriptors method, CamShift method, and optical flow method have been commonly used for tracking [12,18–22]. Several methods have been developed for CNN-based feature extraction and object tracking in video. For example, simple online and real-time tracking with a deep association metric (DeepSORT) combines information regarding an object's position and appearance to achieve high tracking accuracy [23].

In most previous studies on human emotion recognition, human emotions have been classified using traditional methods involving feature extractors and classifiers. Some recent studies have explored using CNN models to extract human features. In 2010, Mikolov et al. proposed recurrent neural networks (RNNs) to deal effectively with time series problems [24]. Regarding research on human emotion recognition, Ojala et al. and Gu et al. used the local binary pattern method [25,26] and the Gabor wavelet transform method, respectively, to recognize facial expressions [27]. Oyedotun et al. proposed a facial expression recognition CNN model that receives RGB data and depth maps as input [28]. Donahue et al. introduced long-term recurrent convolutional networks, which combine CNNs and long short-term memory (LSTM) models to recognize people in videos [29].

Animals have basic emotions that result in different emotional states and neural structures in their brains [30]. However, the lack of large datasets makes assessing canine emotional states more challenging than humans. Nevertheless, we can evaluate a dog's physiology, behavior, and cognitive mood [31]. Facial expressions, blink rate, twitching, and yawning are among the essential sources of information for assessing animal stress and emotional states [1,32]. In addition to facial behavior, body posture and movement are associated with affective states and pain-related behaviors [33,34]. Open spaces, novel objects, elevated plus mazes, and qualitative behavioral assessments evaluate animals' pain, discomfort, and emotional mood [35,36]. In recent years, physical and postural behavior has also been utilized to assess affective emotions in dogs and horses [1,37,38].

The present study focused on the recognition of the emotions of dogs in videos to identify potentially aggressive dogs and relay warning messages in real time. The proposed system first uses YOLOv3 architecture to detect dogs and their positions in the input videos. To track the dogs, we modified the sizes of the images input into the DeepSORT model, improved the feature extraction model, trained the model on the dog dataset, and modified each final tracking position to the position of each tracked dog. The modified model is called real-time dog tracking with a deep association metric (DeepDogTrack). Finally, the system categorizes the dogs' emotional behaviors into three types—angry (or aggressive), happy (or excited), and neutral (or general emotional) behaviors—based on manual judgments made by veterinary experts and custom dog breeders. The dog emotion recognition model proposed in this study is called the long short-term deep features of dog memory networks (LDFDMN) model. This model uses ResNet to extract the features of the dog region that are tracked in the continuous images, which are then input into the LSTM model. The LSTM model is then used for emotion recognition.

The contributions of this study are as follows:

1. An automated system that integrates an LSTM model with surveillance camera footage is proposed for monitoring dogs' emotions.
2. A new model for dog tracking (DeepDogTrack) is developed.
3. A new model for dog emotion recognition (LDFDMN) is proposed.
4. The proposed system is evaluated according to the results of experiments conducted using various training data, methods, and types of models.

2. Related Work

2.1. The Processing of the SORT

The overall SORT process involves the detection, estimation, data association, and creation and deletion of tracked identities.

Detection: First, Faster-RCNN is used for detection and feature extraction. Because the detection objects in this study are objects, other objects are ignored, and only objects that are more than 50% likely to be a object are considered.

Estimation: The SORT model's estimation model describes the model of the object and enters the movement model of its representation and transmission target in the next frame. First, the Kalman filter is used to predict the target state model (including size and position) of an object detected at time T at time T + 1. An object's state model can be expressed as follows:

$$x = [u, v, s, r, \dot{u}, \dot{v}, \dot{s}]^T \quad (1)$$

where (u, v) represents the coordinates of the object's center at time T; (s, r) represents the region and aspect ratio of the object's bounding box at time T; and $(\dot{u}, \dot{v})$ and $(\dot{s})$, respectively, represent the center point and speed of the object at time T. When the object in the next frame is detected, the object's bounding box $(\dot{u}, \dot{v})$ is used to update the object's status. If no correlations between the objects are detected, the prediction model is not updated.

Data association: The detection result is used to determine the object's target state; that is, the bounding box $(\dot{u}, \dot{v})$ of the object at time T is used to predict the new position of the object at time T + 1. First, the model predicts the bounding box $(\dot{u}^{T+1}, \dot{v}^{T+1})$ of the object at time T and the ith object at time T + 1 (u_i^{T+1}, v_i^{T+1}), and calculates the Mahalanobis distance between them. Thereafter, the model uses the Hungarian algorithm for matching to enable multi-object tracking. When the intersection area (intersection over union [IOU]) is less than the threshold value, the object is regarded as the tracking target.

Creation and deletion of tracked identities: When an object enters or leaves the screen, its identity information must be added or deleted from this system. To prevent erroneous tracking, the model must detect objects to be tracked within a few frames of their entrance to determine whether the object must be newly added to this system. Furthermore, the IOU of the object in each frame and in the next frame is calculated; if its value is less than the threshold value, the object is determined to have left the screen, and the object's identity information is deleted.

2.2. The Processing of the DeepSORT

The overall DeepSORT process involves the detection, estimation, data association, and creation and deletion of tracked identities.

Detection: The DeepSORT model uses YOLOv3 architecture for pedestrian detection. Because the detection objects in this study are pedestrians, other objects are ignored, and only objects that are more than 50% likely to be pedestrians are considered.

Estimation: The pedestrian's description is to enter the motion of its representation and propagation target in the next frame. First, the model uses the Kalman filter to predict the state model (including size and position) of a pedestrian detected at time T at time T + 1. DeepSORT expresses the state model of the pedestrian as eight values $(u, v, r, h, \dot{x}, \dot{y}, \dot{r}, \dot{h})$, as follows:

$$\mathrm{x} = (u, v, r, h, \dot{x}, \dot{y}, \dot{r}, \dot{h})^T \quad (2)$$

where (u, v) and (r, h) are the coordinates of the pedestrian's center and the aspect ratio and height of the bounding box of the pedestrian at time T, respectively. At time T, the Kalman filter is used to predict the pedestrian's position at time T +1. $D_{T+1,1}$, represents the predicted position $(\dot{x}, \dot{y}, \dot{w}, \dot{h})$ of the pedestrian at time T + 1, where $(\dot{x}, \dot{y}, \dot{w}, \dot{h})$ are the coordinates, length, width, and height, respectively, of the pedestrian's center at time T + 1. When a pedestrian is detected, the $(\dot{x}, \dot{y}, \dot{w}, \dot{h})$ values are updated to reflect the target state of the pedestrian. If no pedestrian is detected, the predictive model is not updated.

Pedestrian feature extraction: The trained CNN model, which contains two convolution layers, a max pooling layer, and six residual layers, is used to extract the features of each pedestrian at time $T + 1$, which are output as a 512-dimensional feature vector. The feature vector of the jth pedestrian at time $T + 1$ is expressed as f_j^{T+1}.

Data association: The pedestrian region $(\dot{u}, \dot{v})$ at time T is the predicted new position of the pedestrian at time $T + 1$. Thereafter, the Mahalanobis distance between the pedestrian region at time T $\mathrm{O}(\dot{x}, \dot{y}, \dot{w}, \dot{h})_i^{T+1}$ and the region of the ith pedestrian at time $T + 1$ $\mathrm{O}'(\dot{x}, \dot{y}, \dot{w}, \dot{h})_j^{T+1}$ is calculated as follows:

$$\Delta d_1(i,j) = \min\left[(\mathrm{O'}_i^{T+1} - \mathrm{O}_j^{T+1})^T S_i^{-1} (\mathrm{O'}_i^{T+1} - \mathrm{O}_j^{T+1}),\ i,j = 1,2,\ldots,n \right] \tag{3}$$

First, $(\dot{x}, \dot{y}, \dot{w}, \dot{h})$ is converted into $(\dot{x}, \dot{y}, \dot{r}, \dot{h})$, where $(\dot{x}, \dot{y})$ represents the coordinates of the pedestrian's center, $\dot{r}$ is the aspect ratio of the pedestrian, and $(\dot{h})$ is the height of the pedestrian. $\mathrm{O}'(\dot{x}, \dot{y}, \dot{r}, \dot{h})_i^{T+1}$ represents the new position of the ith pedestrian at time $T + 1$, $\mathrm{O}(\dot{x}, \dot{y}, \dot{r}, \dot{h})_j^{T+1}$ represents the new location of the jth pedestrian at time $T + 1$, S_i^{-1} is the covariance matrix of the ith pedestrian, and n is the total number of pedestrians at time $T + 1$. The detection index based on Mahalanobis distance can be used to obtain the optimal match. The χ^2 distribution and its 95% confidence interval are used as the detection threshold value, which was 9.4877 in the present study.

The Mahalanobis distance is suitable for movement positions that produce low uncertainty regarding the pedestrian's position. The state distribution of a pedestrian is predicted using a frame, and the pedestrian's position in the next frame is obtained using the Kalman filter. This method only provides an approximate position, and the positions of pedestrians that are obstructed or moving quickly will not be correctly predicted. Therefore, the model uses a CNN to extract the feature vector of the pedestrian and calculates the cosine distance between the extracted vector and the feature vector of the pedestrian in this system. The minimum cosine distance is represented as follows:

$$\Delta d_2(i,j) = \min\left\{ \dot{f}_i^{T+1} - f_j^{T+1}, j = 1,2,\ldots,n \right\} \tag{4}$$

Finally, the position and features of the pedestrian are matched and fused. The fused cost matrix $c(i,\ j)$ is expressed as follows:

$$c(i,\ j) = \lambda \Delta d_1(i,\ j) + (1-\lambda)\Delta d_2(i,\ j) \tag{5}$$

where λ is the weight. Because using a nonfixed camera to shoot may cause the image to shake violently, λ should be set to 0. Therefore, λ can also account for the problem of obscured pedestrians and reduce ID switching (IDSW) during tracking.

The creation and deletion of tracked identities is the same as for SORT.

2.3. LSTM Model

In traditional neural networks, each neuron is independent and unaffected by time series. In RNNs, time series data are used as input [24]. Earlier layers of an RNN exert weaker effects than subsequent decisions. When too many series are present in the data, the gradient disappears or explodes. To address this problem, Sepp and Jürgen proposed the LSTM model [39] in 1997. An LSTM model comprises numerous LSTM cells, each having three inputs, three components, and two outputs. The three inputs x_t are the input at time t, the output h_{t-1} at time $t - 1$, and the long-term memory (LTM) c_{t-1} at time $t - 1$. The three components are the input gate i_t, the output gate o_t and the forget gate f_t. The three components all use sigmoid functions as activation functions to obtain an output value between 0 and 1, simulating the opening and closing of a valve. The input gate uses the input x_t at time t and the output h_{t-1} at time $t - 1$ to determine whether the LTM C_t should incorporate the memory $\hat{C}_t$ generated at time t. The output gate determines whether the whether the LTM C_t generated at time t should be output according to the input x_t at time t and the output h_{t-1} at time $t-1$. The forget gate uses the input x_t at time t and the output h_{t-1} at time $t-1$ to determine whether the LTM C_{t-1} at time $t-1$ should be added to the LTM C_t at time t. The two outputs of the LSTM model are the output h_t and the LTM C_t at time t. The LSTM model has one more output (C_t, or LTM) than ordinary RNNs do, which enables it to solve the gradient problem caused by excessive time series in ordinary RNNs.

3. Proposed System

This study automatically detects the dog's movements through surveillance video to predict the dog's emotions. Therefore, this study must first convert the surveillance video into a continuous image, then detect the dogs in each image, track the dogs' position in each image, and make emotional predictions from the dog's movements in the surveillance video.

The proposed system combines CNNs with a deep association metric and RNN technologies to detect, track, and recognize the emotions of dogs. The system process is illustrated in Figure 1. First, dogs in each frame of the input video are detected; then, each dog is tracked; and finally, each dog's behavior is analyzed to determine which emotion is being expressed. The dogs' emotions are categorized into three types: angry (or aggressive), happy (or excited), and neutral (or general). The methods used for dog detection, tracking, and emotion recognition are described in the following sections.

Figure 1. Dog emotion recognition process.

3.1. Dog Detection

The first step of object detection is image feature extraction. Originally, to achieve this end, suitable filters were used to manually extract various features. However, since the rise of deep learning, CNNs have been commonly used to extract features automatically. Experiments have revealed that CNN-based object detection methods are highly accurate. Therefore, the system described herein uses a YOLOv3 CNN-based object detection algorithm [40] for dog detection. In addition to using Darknet53 to extract shadow features, YOLOv3 uses feature pyramid network technology to address the inability of YOLOv2 to detect small objects. The processing method of YOLOv3 involves first dividing the input image into 13 × 13, 26 × 26, and 52 × 52 grid cells. YOLOv3 is pretrained on Microsoft's Common Objects in Context (MSCOCO) image dataset [41], which contains 80 object classes, and generates 13 × 13 + 26 × 26 + 52 × 52 prediction results. Because many overlapping frames may be obtained, the model uses non-maximum suppression (NMS) processing. NMS processing enables the most reliable and unique bounding box to be regarded as the predicted result of object detection. The dog detection process is illustrated in Figure 2.

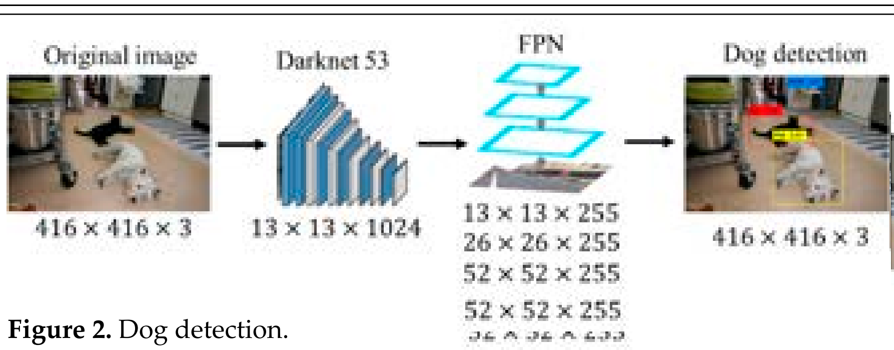


Figure 2. Dog detection.

3.2. *Dog Feature Extraction*

The model uses a ResNet CNN to extract the features of each dog from the sub-images of all the dogs and Mask R-CNN architecture to remove the backgrounds of the sub-images (Figure 3).

Figure 3. Dog region.

Dog feature extraction: The ResNet uses the shortcut connection method to reinforce the learning bottlenecks of multiple layers. This method involves retaining the input feature map before convolution. After the input feature map is subjected to two layers of convolution, ReLU function, and a third layer of convolution, the output feature map is combined with the retained feature map to preserve the pre-convolution features.

Background removal: The proposed system uses Mask R-CNN architecture to remove the backgrounds of the dog images [42]. Mask R-CNN architecture adds a new output to Faster R-CNN architecture to produce a fully convolutional network that can be used to solve object detection and segmentation problems [43]. Faster R-CNN outputs the classification and coordinate offset of a predicted object. Each pixel in the predicted region is classified as part of the foreground or background, as illustrated in Figure 4.

Figure 4. Background removal.

3.3. *Dog Tracking*

After a dog is detected, it is tracked to determine its movement trajectory. The dog tracking system identifies the position of the same dog in consecutive images and plots these positions to form an action path. The system uses a DeepDogTrack model for d

3.3. Dog Tracking

After a dog is detected, it is tracked to determine its movement trajectory. The dog-tracking system identifies the position of the same dog in consecutive images and plots these positions to form an action path. The system uses a DeepDogTrack model for dog tracking. In addition to using a Kalman filter to predict the dog's position in the next frame, the model also uses a CNN to extract and match the dog's features in consecutive frames to determine the dog's motion status. DeepDogTrack is an improved DeepTrack pedestrian tracking model. The DeepSORT model integrates simple online and real-time tracking (SORT) [44] and CNN technology to extract and match each pedestrian's features and analyze the location and appearance information of each pedestrian to achieve accurate tracking. To reduce the computation time of the system and improve the accuracy of dog tracking, the system adopts our novel DeepDogTrack model, which contains improvements in the processing flow and adjustment of parameters.

3.3.1. SORT and DeepSORT

SORT is a practical multi-object tracking method that can effectively track objects in consecutive frames. The SORT model proposed herein uses Faster-RCNN and a Kalman filter to detect an object's position and to predict the object's position in the next frame, respectively. Thereafter, the model calculates the Mahalanobis distance between an object's location and its predicted location in the next frame and uses the Hungarian algorithm [45] for matching to enable multi-object tracking. Therefore, the overall SORT process involves the detection, estimation, data association, and creation and deletion of tracked identities.

Although SORT is a simple and effective multi-object tracking method, it compares only the size and position of a predicted object and does not consider the object's features. To address this limitation, the proposed system incorporates DeepSORT, which improves upon the detection method of SORT and accounts for the object's features, thus enhancing the accuracy of object tracking. DeepSORT applies SQRT's object tracking to pedestrian tracking. DeepSORT is based on SORT's multiple object tracking (MOT) architecture and uses the Kalman filter to predict a given pedestrian's position in the next frame. The model calculates the Mahalanobis distance between the region of the predicted pedestrian and the region in which other pedestrians may be located. Thereafter, a CNN is used to extract and calculate the minimum cosine distance between the pedestrian's features and the features of all the pedestrians in the next frame. Finally, the Hungarian algorithm is used for matching to enable multi-pedestrian tracking. Accordingly, DeepSORT involves the detection, estimation, feature extraction, data association, and the creation and deletion of tracked identities.

3.3.2. Real-Time Dog Tracking with a Deep Association Metric (DeepDogTrack)

Because DeepSORT is typically used to track pedestrians, and the proportions of the human body are 64×128, the input must be a fixed-size image. Proportion features are extracted using a simple CNN model, and the result predicted using the Kalman filter is used as the tracking region of the object. However, the proportions of dogs are different from those of humans. To adapt DeepSORT for the tracking of dogs and improve the computational efficiency, the DeepDogTrack model takes the detected dog region as input data, and the size of the region is not fixed. To increase the depth of the model and minimize error, a deep residual network (ResNet) is used to extract the dogs' features. The DeepSORT model was retrained using the dog data-set to improve its tracking accuracy. The architecture of the proposed DeepDogSORT dog-tracking model is illustrated in Figure 5. The original and improved results are presented in Figure 6.

Appl. Sci. 2023, 13, x FOR PEER REVIEW 9 of 29

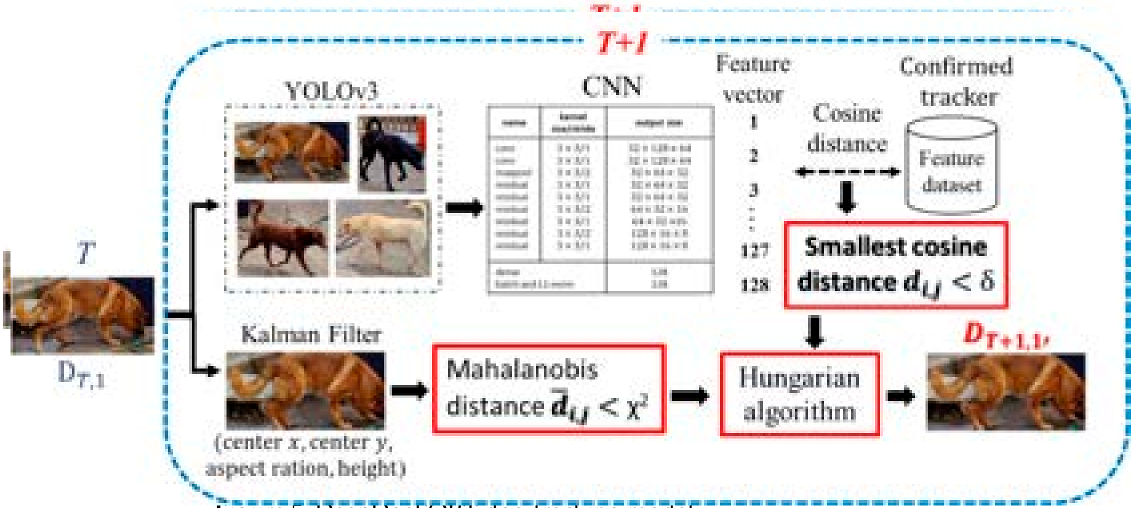


Figure 5. DeepDogSORT dog-tracking model.

(**a**) (**b**)

Figure 6. Dog tracking with DeepSORT and DeepDogTrack models. (**a**) DeepSORT model; (**b**) DeepDogTrack model.

3.4. Dog Emotion Recognition

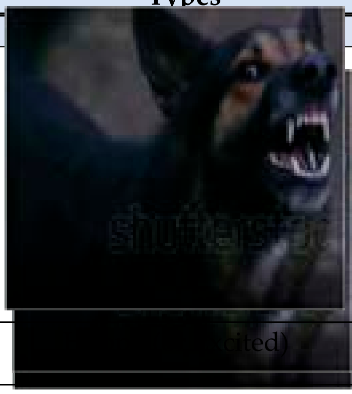
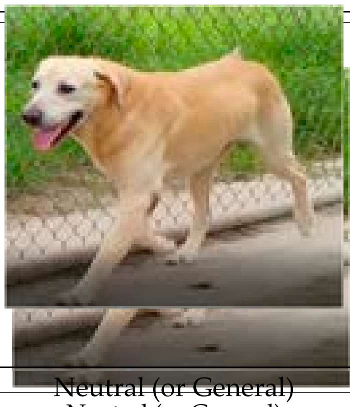
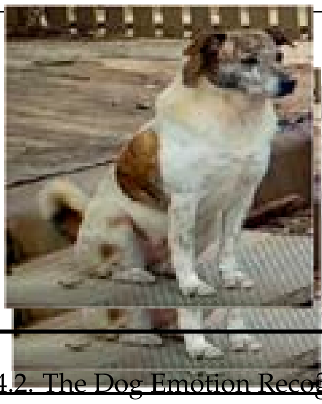
The automatic recognition of dog emotion in this study first defines the emotional type of dogs and then proposes a deep learning technology for predicting dog emotions.

3.4.1. The Emotions of the Dogs

Dogs go through their developmental stages faster than humans and have all the emotional ranges they can reach by four to six months old (depending on how quickly their breed matures). However, the variety of emotions in dogs does not exceed that of humans by two to two and a half years old. Dogs will have all the basic emotions: joy, fear, anger, disgust [46–48], and even love. However, based on current research, dogs do not appear to have more complex emotions such as guilt, pride, and shame [46]. Therefore, we can determine which emotions the dog experiences through the dog's body language. A dog's emotional state is primarily determined by facial and physical behavior, or a combination of the two. However, the data source of this study is surveillance cameras due to their long distance and low-resolution video. Therefore, the dogs' emotional state in this study was generally determined by physical behavior. In addition, since the emotions of fear, anger and disgust need to match the subtle features of the face, these emotions are uniformly assumed to be angry (or aggressive). The proposed model lists the basic human emotions anger (or aggressive) and happiness (or excitement) [49], but these two emotions are relatively extreme behaviors. To strengthen the evaluation of canine emotional types, the third emotion in this study is based on the dog's physical behavior, which is called neutral (or general).

Therefore, the emotions of the dogs in this study are categorized into three types—angry (or aggressive), happy (or excited), and neutral (or general)—according to the manual judgment of veterinary experts and custom dog breeders. The descriptions and characteristics of the three emotional types of dogs are shown in Table 1.

Table 1. The descriptions and characteristics of the three emotional types of dogs.

Types	Characteristics
Anger (or Aggressive) 	For better or worse, dogs' anger is a natural emotion. Protective instincts, health issues, or genetics can cause anger or aggression. It is natural for dogs to be angry occasionally, but we should be aware of situations in which they get angry and avoid them in the future. Dogs will display terrifying postures. The characteristics of anger in dogs include tail wagging, stiffness of the body, trembling, ears pulled back, moving the body weight around, hair standing up, visible sclera, and even defensive aggression such as growling, biting, and sprinting.
Happy (or Excited) 	The happiness of dogs is written all over their faces, and dogs tend to be excited and easily surprised. Dogs are joyful while doing their favorite activities and can lead to some hilarious moments. Dogs will flutter, bounce, and run. When happy, they make a slight whirring; the panting sound is sometimes referred to as laughing. The characteristics of happiness in dogs include lying on the stomach, buttocks wiggling, wild tail wagging, hanging tongue, and relaxed ears, mouth, and body.
Neutral (or General) 	The dog is often classified as a neutral emotional category because it sometimes lacks emotional response or shows indifference, unlike other pets with evident emotions. The characteristics of neutral emotions in dogs include relaxation of the entire body (including the tail, ears, and face), no evident excitement or daze, and observing their environment or sniffing.

3.4.2. The Dog Emotion Recognition Model

The dog emotion recognition model proposed herein is the LDFMN model. After a dog is detected, the dog region and the dog's features are extracted using the ResNet model. Thereafter, these continuous and time-series-associated features are sent to the LSTM model for processing, and the time series output results are obtained. Dog emotion recognition is based on dogs' continuous behaviors; analyzing continuous actions is therefore essential to the proposed system, and the RNN and LSTM models for this purpose are described as follows.

LDFMN Model

In the proposed system, a ResNet CNN and DeepDogTrack model are used to extract features from and to track dog regions, respectively. The tracked dog region is converted into an image set, as illustrated in Figure 7. Each image set depicts the continuous

Appl. Sci. **2023**, *13*, x FOR PEER REVIEW 11 of 29

movement of a dog and is used as a data-set for dog emotion recognition. If the image set comprises fewer than 16 images, it is deleted; if the image set exceeds 16 images, the set is trimmed to 16 images. Thereafter, the image set is input into the LDFDMN model, and the dog emotion recognition results are obtained. The architecture of the LDFDMN model is illustrated in Figure 8.

Figure 7. Dog image set.

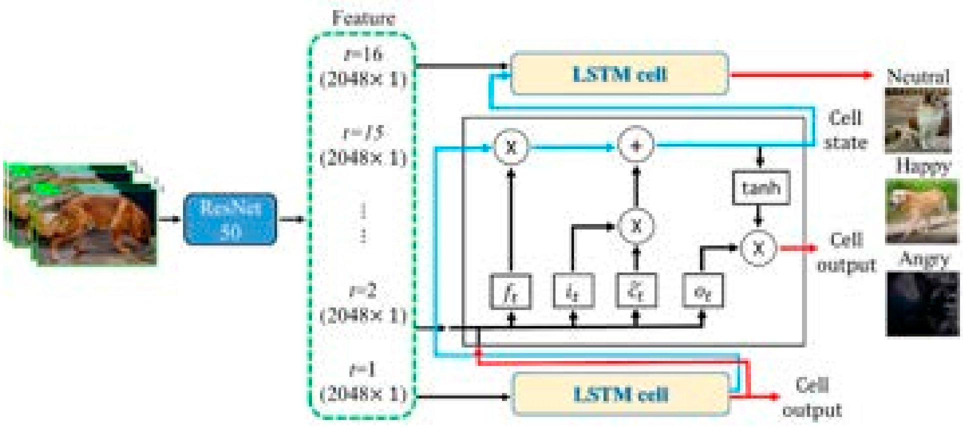


Figure 8. LDFDMN model.

Dog Emotion Recognition after Background Removal

Each of the model's detection regions includes nondog regions, or backgrounds. If the background area is larger than the dog area, the extracted dog features will be affected, resulting in a reduced dog emotion recognition rate. Therefore, the proposed model uses a Mask R-CNN model to remove backgrounds from the image set before the dog tracking and emotion recognition are processed by DeepDogTrack and the LDFDMN model, respectively.

Video Preprocessing

In this study, we trained the LDFDMN model by using videos collected from YouTube, the Folk Stray Dog Shelter, and the Dog Training Center (hereafter, DTC) of the Customs Administration of Taiwan's Ministry of Finance. The input data of the LDFDMN model must be a fixed-length feature vector, but the lengths of the videos collected for this study differed, and multiple dogs may have been present in each video. Therefore, each video was divided into multiple sub-images, each of which was resized to 360 × 360 pixels. Sub-images of the same dog were used to create experimental videos in order to analyze the dog's emotions.

Although the backgrounds of the dog regions are supposed to be removed by the Mask R-CNN before tracking, the sub-images may depict the background instead of the dog because of classification errors, resulting in a set of fewer than 16 continuous sub-images. To address this problem, the Farneback optical flow method is applied [50], and the 16 sub-images in each image set are linearly interpolated according to the optical flow value. The results of the linear interpolation of an image are presented in Figure 9. In this figure, the optical flow information of the image at times $t(0)$ and $t(1)$ is used to produce a linear interpolation of the image at time $\tilde{t}(\frac{1}{2})$.

Appl. Sci. **2023**, *13*, x FOR PEER REVIEW 12 of 29

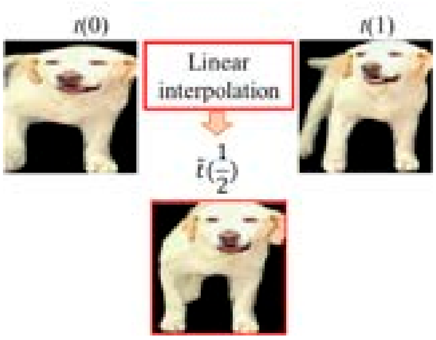


Figure 9. Linear image interpolation.

3.5. Dog Emotion Recognition in Surveillance Videos

The proposed system was tested using three dog-tracking methods (DeepSORT, DeepSORT_retrained [a version of the DeepSORT model retrained using the dog data-set], and DeepDogTrack) and two dog emotion recognition methods (sub-images with and without backgrounds). The methods were combined into six models, as listed in Table 2.

Table 2. Dog emotion recognition model types.

Type	Detection	Tracking	Emotion Recognition
Type_1	YOLOv3	DeepSORT	LDFDMN with background
Type_2		DeepSORT_retrained	
Type_3		DeepDogTrack	
Type_4		DeepSORT	LDFDMN with without background
Type_5		DeepSORT_retrained	
Type_6		DeepDogTrack	

4. Experiments

The performance of the DeepDogTrack and LDFDMN models for dog tracking and emotion recognition, respectively, were evaluated through a series of experiments on dog detection, tracking, and emotion recognition. The hardware and software employed in the experiments, experimental image and video datasets, experimental procedures and evaluation criteria, and model performance evaluation are present in the following relevant information.

4.1. Software and Hardware

The hardware and software systems used in the experiments are listed in Tables 3 and 4. The CNN architecture incorporates Darknet53 and PyTorch [51], both of which use the Python programming language, and a computer vision library (OpenCV for Python) [52].

Table 3. Hardware.

Device	Specification
CPU processor	Intel Core i7-8700 3.2 GHz
GPU processor	NVIDIA GeForce GTX1080Ti 11 G
RAM memory	32 G

Appl. Sci. **2023**, *13*, x FOR PEER REVIEW 13 of 29

Table 4. Software.

	Detection	Tracking	Emotion Recognition
Network architecture	YOLOv3	DeepDogTrack	LDFDMN
System		Windows 10 Pro	
Programming language		Python 3.5.4	
Neural network framework	Darknet	PyTorch 0.4.1	PyTorch 0.4.1
Computer vision library		OpenCV-python 3.4.4	

4.2. Image Data-Sets

Experiments were conducted to evaluate the dog detection, tracking, and emotion recognition models and the proposed system overall. In each set of experiments, different image datasets were used for training and testing. There may be more than two dogs in one image.

4.2.1. Data-Set for Dog Detection Experiments

The proposed model used a YOLOv3 model for dog detection, and the MSCOCO image set was used to train the YOLOv3 model. The image set contained 80 classes of objects and a total of 118,287 images, as shown in Figure 10. The test images were divided into two image databases in the dog detection experiment. The first (TestSet1) is the image database established by Columbia University and the University of Maryland [53], which contains images from ImageNet, Google, and Flickr. The database contains 8351 images of 133 dog breeds, as shown in Figure 11. The second (TestSet2) is the image database established by Stanford University [54], which contains images from ImageNet. The database contains 20,580 images of 120 dog breeds, as shown in Figure 12.

Figure 10. Some images of the MSCOCO dataset.

Appl. Sci. 2023, 13, x FOR PEER REVIEW 14 of 29

Figure 11. Some images of the TestSet1.

Figure 12. Some images of the TestSet2.

4.2.2. Data-Set for Dog-Tracking Experiments

The CNN in the DeepSORT model used two pedestrian reidentification data-sets, Market-1501 and MARS, which contain images of 1501 and 1261 pedestrians [55,56], respectively. The training data-set used by the ResNet CNN in the DeepDogTrack model proposed in this study contains data from YouTube, the Folk Stray Dog Shelter, and the DTC, accounting for a total of 40 dogs. Three test videos from the Folk Stray Dog Shelter and DTC, containing a total of 5 dogs, were used in the experiment. The data-set information is presented in Table 5.

Appl. Sci. 2023, 13, x FOR PEER REVIEW 16 of 32

Table 5. Test videos used in dog-tracking experiment.

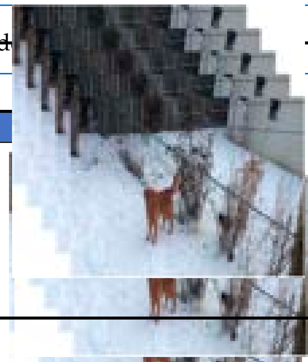
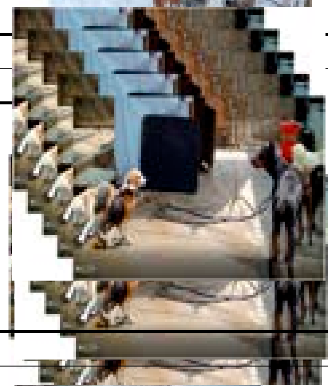
Source	Video	Dog Number	Image Number
 DTC	IMG_0043_5	1	240
	IMG_0041_1	1	180
 Folk Stray Dog Shelter	IMG_0014	4	371

Note: DTC, Dog Training Center of the Customs Administration of Taiwan's Ministry of Finance.

4.2.3. Data-Set for Dog Emotion Recognition Experiments

Since dogs' emotional states in this study were considered in terms of physical behaviors and considering the generality of future applications, few existing emotional video datasets exist. Therefore, in addition to collecting some videos from YouTube, this study set up several surveillance cameras at the Folk Stray Dog Shelter and the DTC to capture the emotional stages of dogs, as shown in Table 6. Since dog emotions are easily disturbed by external things, the collected videos included neutral (general), happy (excited), and angry (aggressive) dogs; therefore, we removed the videos one by one. After screening, 813 sub-videos from YouTube were divided into training and testing sub-video groups, each with 572 and 241 sub-videos; the training video was divided into two groups of training sub-videos, TrainSet4_1 and TrainSet4_2, each with 294 and 278 sub-videos; the test video was divided into two groups of training sub-movies, TestSet4_1 and TestSet4_2, each with 121 and 120 sub-movies, as shown in Table 6. In addition, in this study, static surveillance cameras were set up outdoors at the Folk Stray Dog Shelter and the DTC. Dogs were allowed to move freely. We focused our attention on physical behavior indicators. The determination of the dogs' behaviors was purely based on observed behaviors, without considering human-induced behaviors. In total, 246 sub-videos were selected from the videos of the Folk Stray Dog Shelter, which were divided into training and testing sub-video groups, each with 176 and 70 sub-videos; the training video was split into two groups of training sub-videos, TrainSet4_1 and TrainSet4_2, each of which included 88 and 88 sub-videos; the test video was divided into two groups of training sub-movies, TestSet4_1 and TestSet4_2, each with 35 and 35 sub-movies. After screening, 278 sub-movies from the DTC movies were divided into training and testing sub-movies, each with 196 and 82 sub-movies; the training movie was divided into two groups of training sub-movies, TrainSet4_1 and TrainSet4_2, each with 98 and 98 sub-movies; the test video was divided into two groups of training sub-movies, TestSet4_1 and TestSet4_2, with 41 and 41 sub-movies, respectively.

Table 6. The data-set for dog emotion recognition model.

Dataset	Source	Videos
TrainSet4_1 TrainSet4_2 TestSet4_1 TestSet4_2	YouTube	
	Folk Stray Dog Shelter	
	DTC	

Note: DTC, Dog Training Center of the Customs Administration of Taiwan's Ministry of Finance.

In the experiment, the training data-set was divided into TrainSet4_1 and TrainSet4_2. The information of the datasets is presented in Table 7. In both datasets, image sets containing fewer than 16 images were deleted. If an image set contained more than 16 images, it was equally divided into subsets of 16 images. Each image was resized to 360 × 360 pixels, and sets of images of the same dog were used as training data-sets for the dog-tracking model. To create TrainSet4_2, a Mask R-CNN was used to remove the backgrounds from 16 images of the same dog.

The videos in the test dataset for the dog emotion recognition experiment were obtained from YouTube, the Folk Stray Dog Shelter, and the DTC. TestSet4_1 contained 197 preprocessed videos, each of which consisted of more than 16 sub-images. TestSet4_2 contained 196 preprocessed videos, and the background of each sub-image of each video was removed using the Mask R-CNN. If an image set contained fewer than 16 sub-images, the sub-images were interpolated linearly. The test dataset information is presented in Table 8.

Table 7. Training data-set for dog emotion recognition model.

Dataset	Emotion Type	Source	Video Number		Total Video Number
TrainSet4_1	Neutral/General	YouTube	116	206	480
		Folk Stray Dog Shelter	63		
		DTC	27		
	Happy/Excited	YouTube	30	124	
		Folk Stray Dog Shelter	23		
		DTC	71		
	Angry/Aggressive	YouTube	148	150	
		Folk Stray Dog Shelter	2		
		DTC	0		
TrainSet4_2	Neutral/General	YouTube	108	198	464
		Folk Stray Dog Shelter	63		
		DTC	27		
	Happy/Excited	YouTube	30	124	
		Folk Stray Dog Shelter	23		
		DTC	71		
	Angry/Aggressive	YouTube	140	142	
		Folk Stray Dog Shelter	2		
		DTC	0		

Note: DTC, Dog Training Center of the Customs Administration of Taiwan's Ministry of Finance.

Table 8. Test data-set for the dog emotion recognition experiment.

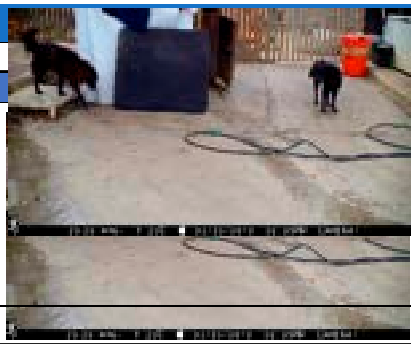
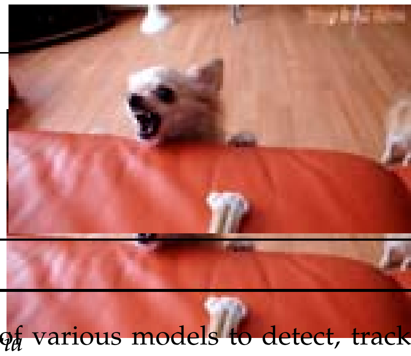
Dataset	Emotion Type	Source	Video Number		Total Video Number
TestSet4_1	Neutral/General	YouTube	48	85	197
		Folk Stray Dog Shelter	26		
		DTC	11		
	Happy/Excited	YouTube	11	50	
		Folk Stray Dog Shelter	9		
		DTC	30		
	Angry/Aggressive	YouTube	62	62	
		Folk Stray Dog Shelter	0		
		DTC	0		
TestSet4_2	Neutral/General	YouTube	47	84	196
		Folk Stray Dog Shelter	26		
		DTC	11		
	Happy/Excited	YouTube	11	50	
		Folk Stray Dog Shelter	9		
		DTC	30		
	Angry/Aggressive	YouTube	62	62	
		Folk Stray Dog Shelter	0		
		DTC	0		

Note: DTC, Dog Training Center of the Customs Administration of Taiwan's Ministry of Finance.

4.2.4. Test Data-Set of the Integrated System

The integrated system proposed herein was tested using two videos, the information of which is presented in Table 9. The IMG_0033 video, taken from the Folk Stray Dog Shelter, contains two dogs with similar appearances. The dogs' emotions are mostly neutral but seem happy at a few points in the video, and one dog moves more frequently than the other. The AngryDogs video, taken from YouTube, depicts only one dog. The dog appears angry, although its emotions seem neutral at a few points in the video.

Table 9. Test data-set of the integrated system.

Video	Total Image Number	Number of Dog	Emotion Type	Image
IMG_0033	400	2	Neutral/Happy	
AngryDogs	400	1	Neutral/Angry	

4.3. Model Training Parameters and Evaluation Criteria

This paper proposes and explains the training of various models to detect, track, extract the features of, and recognize the emotions of dogs in videos. This paper also aimed to verify the accuracy of the models in terms of dog detection, tracking and emotion recognition. Various evaluation criteria were used for different tasks.

4.3.1. Model Training Parameters

In the proposed system, the YOLOv3 and the DeepDogTrack models were used for dog detection and tracking, respectively. The ResNet50 and Mask R-CNN models, combined with the LSTM model, were used for dog emotion recognition. In this experiment, to train the LSTM model, the Mask R-CNN model and ResNet50 models were used to remove the image backgrounds and extract each dog's features, respectively. The model parameters were those of ImageNet. The LSTM model used the feature vectors from ResNet50 as input data, and its training parameters are presented in Table 10.

Table 10. Training parameters of LSTM model.

	Parameters
Input size	16 × 2048
Feature length	16
Learning rate	0.0001
Dropout	0.4
Batch size	2
Activation function	tanh
Epoch	50

4.3.2. Model Evaluation Criteria

In the dog detection, tracking, and emotion recognition experiments, various evaluation criteria were used to examine the performance of the models.

Evaluation Criteria for Dog Detection

The dog detection performance of the proposed system was evaluated according to the rate of correct predictions (vs. the ground truth region). This experiment used three evaluation criteria, the first of which is Recall. Recall represents the number of predicted ground truth pixels and is calculated as follows:

$$\text{Recall} = \frac{1}{N}\sum_{i=1}^{N}\frac{Gt_i \cap P_i}{Gt_i} \tag{6}$$

where Gt_i represents the ground truth region of the ith dog, P_i represents the predicted region of the ith dog, N is the total number of dogs, and $Gt_i \cap P_i$ represents the intersection between the ground truth and predicted regions.

The second evaluation criterion used was Precision. Precision represents the number of correctly predicted pixels and is calculated as follows:

$$\text{Precision} = \frac{1}{N}\sum_{i=1}^{N}\frac{Gt_i \cap P_i}{P_i} \tag{7}$$

The third evaluation criterion used was the mean IOU (mIOU), that is, the average number of pixels detected correctly in the ground truth and predicted regions. It is calculated as follows:

$$\text{mIOU} = \frac{1}{N}\sum_{i=1}^{N}\frac{Gt_i \cap P_i}{Gt_i \cup P_i} \tag{8}$$

where $Gt_i \cup P_i$ represents the union of the ground truth region Gt_i and the predicted region P_i.

The fourth evaluation criterion used was the detection rate. The detection rate is considered satisfactory if the Recall, Precision, or mIOU value is $\geq$0.5.

Evaluation Criteria for Dog Tracking

In the dog tracking experiment, the models were evaluated in terms of MOT accuracy (MOTA), as defined by the MOT Challenge [57]. MOTA is calculated as follows:

$$\text{MOTA} = 1 - \frac{\sum_t (FN_i + FP_i + IDSW_i)}{\sum_i GT_i} \tag{9}$$

where GT_i is the ground truth region of the dog in the ith image, FN_i (false negative) is the number of dogs that are not tracked in the ith image, and FP_i (false positive) is the number of tracked dogs in the ith image for which the tracked region is incorrect. Incorrectly tracked regions are those for which the IOU between the tracked region and the ground truth region is less than 50%. $IDSW_i$ (ID Switch) represents the number of dogs tracked as other dogs in the ith image. Therefore, larger MOTA values indicate higher MOTA.

Evaluation Criteria for Dog Emotion Recognition

Dog emotion recognition was evaluated by comparing the predicted results with the ground truth results and is presented herein in terms of identification accuracy ACC, which is calculated as follows:

$$ACC = \sum_{i=1}^{N_T} P_i \text{ and } P_i = \frac{NT_i}{N_i} \tag{10}$$

where P_i is the identification rate of the ith category of emotions, N_T represents the total number of images, NT_i represents the number of correct recognitions in the ith category, and N_i represents the total number of dogs in the ith category.

4.4. Performance Analysis

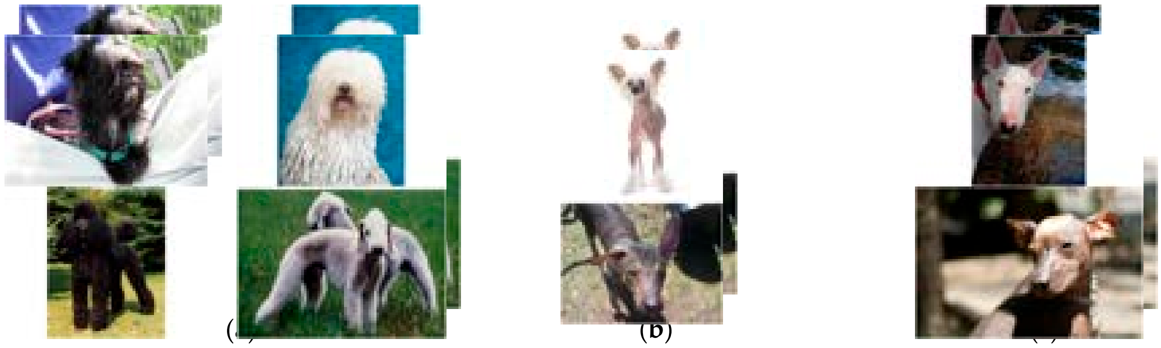
An analysis of the performance of the proposed system according to the results of the dog detection, tracking, and emotion recognition experiments is presented in the following sections.

4.4.1. Performance for Dog Detection

The results of the dog detection experiment are listed in Table 11. Since the experimental images were taken from the video on the camera, there may be more than two dogs in one picture. Therefore, the number of images in the table will be less than the number of dogs. The detection rate of the TestSet1 data-set was 97.62%; in total, 199 dogs were undetected. The reasons for the detection errors were the obstruction of the facial features of the dog, the breed of the dog, and the obstruction or cropping of the body of the dog (Figure 13). Another factor contributing to the detection error rate may have been the training data-set, which accounted for too many object categories and contained too few dog samples. The detection rate of the TestSet2 data-set was 98.39%; in total, 357 dogs were undetected. In addition to the aforementioned factors contributing to the detection error rate, some detection errors in the experiment conducted using the TestSet2 data-set were attributable to incomplete dog regions, as illustrated in Figure 14. In the future, training data-sets that contain higher numbers of dog images and that account for the types of detection errors identified in this study should be used to improve the detection rate of the proposed system.

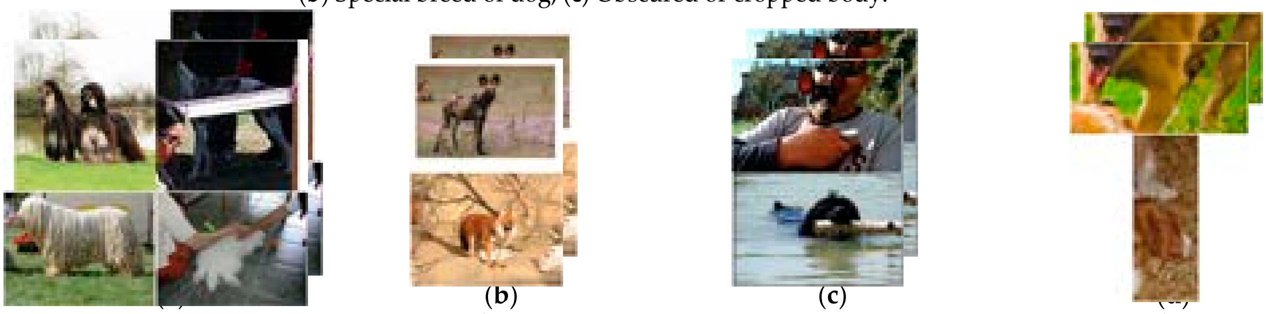
Table 11. Results of dog detection experiments.

Datasets	Image Number	Dog Number	Detection Rate	Precision	Recall	mIOU
TestSet 1	8351	8371	97.62%	93.49%	83.72%	80.27%
TestSet 2	20580	22126	98.39%	88.87%	85.67%	80.48%



(a) **(b)** **(c)**

Figure 13. Reason for detection errors in TestSet1 data-set experiment. (**a**) Obscured facial features; (**b**) Special breed of dog; (**c**) Obscured or cropped body.



(a) **(b)** **(c)** **(d)**

Figure 14. Reasons for detection errors in TestSet2 data-set experiment. (**a**) Obscured facial features; (**b**) Special breed of dog; (**c**) Obscured or cropped body; (**d**) Incomplete dog region.

4.4.2. Performance for Dog Tracking

Dog tracking is an experiment with a single dog after detection. The DeepSORT, DeepSORT_retrained, and DeepDogTrack models (Models 1, 2, and 3, respectively) were used in the dog-tracking experiment. The experimental results for the IMG_0043_5 data-set are presented in Table 12. The MOTA values of Model 1 and Models 2 and 3 were 81.1% (false negatives [FNs]: 33, false positives [FPs]: 9) and 83.88% (FNs: 33, FPs: 1), respectively. The MOTA values of Models 2 and 3 were higher than that of Model 1 because the prediction regions of these two models use YOLOv3 detection. Two reasons for tracking failure were identified: the obstruction of the dog's body in many regions (Figure 15)

Appl. Sci. 2023, 13, x FOR PEER REVIEW 21 of 29

are presented in Table 12. The MOTA values of Model 1 and of Models 2 and 3 were 81.1% (false negatives [FNs]: 33, false positives [FPs]: 9) and 83.88% (FNs: 33, FPs: 1), respectively. The MOTA values of Models 2 and 3 were higher than that of Model 1 because the prediction regions of these two models use YOLOv3 detection. Two reasons for tracking failure were identified: the obstruction of the dog's body in many regions (Figure 15) and the dog's back being turned to the camera (Figure 16). The YOLOv3 model was not trained using images of dogs' backs, which differ considerably from those taken from front or side views; consequently, the tracked region in such images is incorrect. If the Kalman prediction region is used as the dog region, the IOU between the ground truth and predicted region is less than 20%. This is illustrated in Figure 17, in which blue and red boxes are the predicted and ground truth regions, respectively; the IOUs in images 47 and 48 are 0.46 and 0.43, respectively.

Table 12. Results of dog tracking experiments conducted using IMG_0043_5 data-set.

Methods	Number of Dog	Total Image Number	Number of Dogs Tracked	FN	FP	IDSW	MOTA
Model 1			169	33	9	0	81.1%
Model 2	1	240	177	33	1	0	83.88%
Model 3			177	33	1	0	83.88%

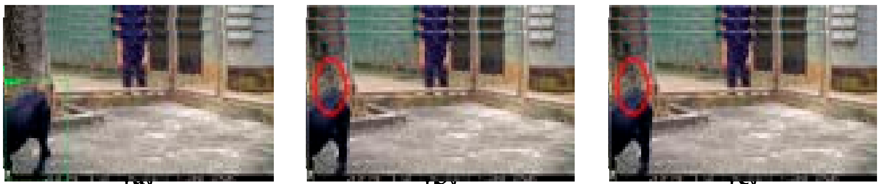

(a) (b) (c)

Figure 15. Dog not tracked in images 211 and 212. (a) Image 210; (b) Image 211; (c) Image 212.


(a) (b) (c)

Figure 16. Dog not tracked in image 40 to 42. (a) Image 40; (b) Image 41; (c) Image 42.

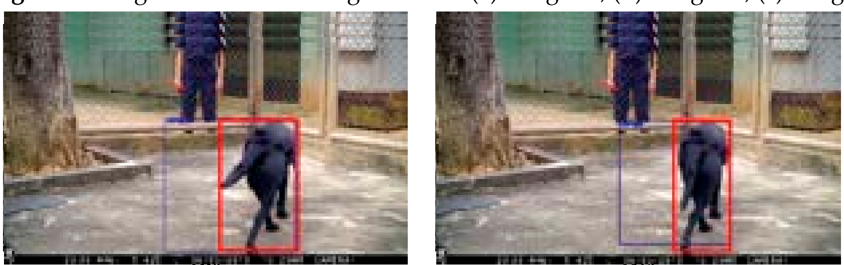

(a) (b)

Figure 17. Dogs with incorrect tracked regions in images 47 and 48. (a) Image 47; (b) Image 48.

The experimental results for the IMG_0041_1 data-set are presented in Table 13. The MOTA values of Model 1 and of Models 2 and 3 were 92.24% (FNs: 8, FPs: 2) and 93.02% (FNs: 8, FPs: 1), respectively. The MOTA values of Models 2 and 3 were again higher than that of Model 1 because the prediction regions of these two models use YOLOv3 detection. The main reason for tracking failure was the obstruction of the vital part (body and head) of the dog, as illustrated in Figure 18.

Table 13. Results of dog-tracking experiments conducted using IMG_0041_1 data-set.

Methods	Number of Dogs	Total Image Number	Number of Dogs Tracked	FN	FP	IDSW	MOTA
Model 1	1	180	119	8	2	0	92.24%
Model 2			120	8	1	0	93.02%
Model 3			120	8	1	0	93.02%

(a) (b) (c)

Figure 18. Dogs not tracked in images 164 and 165. (a) Image 163; (b) Image 164; (c) Image 165.

The IMG_0014 data-set used in the dog-tracking experiment contained sub-images of four different dogs (ID 1–4). The experimental results for the data-set are presented in Table 14. In the experiments involving ID 1, each of the three models achieved a MOTA value of 98.32%; the tracked regions were all correct. In those involving ID 2, Models 2 and 3 achieved a MOTA value of 69.26%, which was higher than that achieved using Model 1, but considerably lower than that achieved in the experiments involving other dogs. This is partially because several dogs were tracked as the same dog. In the experiments involving ID 3, Models 2 and 3 achieved a MOTA value of 87.50%, which was far higher than that achieved using Model 1 and second only to that obtained in the experiments involving ID 1. In the experiment involving ID 4, Models 1 and 3 achieved a MOTA value of 82.50%, and the tracked regions in both models were all correct; however, Model 2 achieved a lower MOTA value (81.25%) because several dogs were tracked as the same dog.

Table 14. Results of dog tracking experiments conducted using IMG_0014 data-set.

ID	Number of Dog	Total Image Number	Number of Dogs Tracked	FN	FP	IDSW	MOTA
1	Model 1	357	351	6	0	0	98.32%
	Model 2	357	351	6	0	0	98.32%
	Model 3	357	351	6	0	0	98.32%
2	Model 1	231	159	70	1	1	68.83%
	Model 2	231	160	70	0	1	69.26%
	Model 3	231	160	70	0	1	69.26%
3	Model 1	48	31	6	11	0	64.58%
	Model 2	48	42	6	0	0	87.50%
	Model 3	48	42	6	0	0	87.50%
4	Model 1	80	66	14	0	0	82.50%
	Model 2	80	66	14	0	1	81.25%
	Model 3	80	66	14	0	0	82.50%

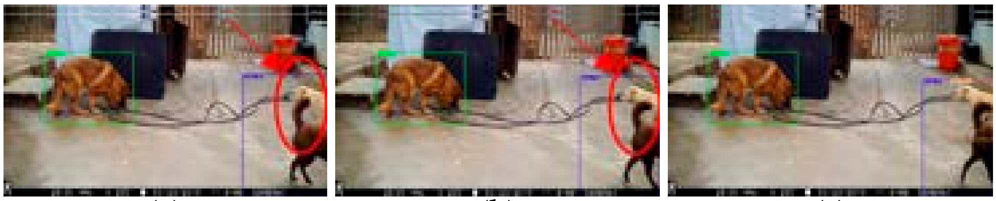
The numbers of FNs obtained for IDs 2 and 4 were higher than those obtained for IDs 1 and 3. Examples of images resulting in FNs for IDs 2 and 4 are presented in Figures 19 and 20, respectively. ID 2 corresponds to a black dog far from the camera. In images 266 to 274, the dog is obscured, leading to tracking failure. ID 4 corresponds to a white dog that entered the frame during recording. In images 302 and 303, the dog has not yet completely entered the frame, resulting in tracking failure.

The IMG_0014 data-set used in the dog-tracking experiment contained sub-images of four different dogs (ID 1–4). The experimental results for the data-set are presented in Table 14. In the experiments involving ID 1, each of the three models achieved a MOTA value of 98.32%; the tracked regions were all correct. In those involving ID 2, Models 2 and 3 achieved a MOTA value of 69.26%, which was higher than that achieved using Model 1, but considerably lower than that achieved in the experiments involving other dogs. This is partially because several dogs were tracked as the same dog. In the experiments involving ID 3, Models 2 and 3 achieved a MOTA value of 87.50%, which was far higher than that achieved using Model 1 and second only to that obtained in the experiments involving ID 1. In the experiment involving ID 4, Models 1 and 3 achieved a MOTA value of 82.50%, and the tracked regions in both models were all correct; however, Model 2 achieved a lower MOTA value (81.25%) because several dogs were tracked as the same dog.

The numbers of FNs obtained for IDs 2 and 4 were higher than those obtained for IDs 1 and 3. Examples of images resulting in FNs for IDs 2 and 4 are presented in Figures 19 and 20, respectively. ID 2 corresponds to a black dog far from the camera. In images 266 to 274, the dog is obscured, leading to tracking failure. ID 4 corresponds to a white dog that entered the frame during recording. In images 302 and 303, the dog has not yet completely entered the frame, resulting in tracking failure.

(a) (b) (c)

Figure 19. Dog with ID 2 not tracked from image 266 to 274. (a) Image 265; (b) Image 266; (c) Image 274.



(a) (b) (c)

Figure 20. Dog with ID 4 not tracked in images 302 and 303. (a) Image 302; (b) Image 303; (c) Image 304.

4.4.3. Performance for Dog Emotion Recognition

The LDFMN model and the TestSet4_1 and TestSet4_2 data-sets were used for the emotion recognition experiments. In the experiment conducted using TestSet4_1, 16 images were selected as prediction targets, and the ResNet50 model incorporated into the LDFMN model was trained using ImageNet parameters. In the experiment conducted using TestSet4_2, 16 images were selected as prediction targets, and the Mask R-CNN and ResNet50 models incorporated into the LDFMN model were both trained using ImageNet parameters.

The results of the emotion recognition experiments are presented in Table 15. In the experiments conducted using the TestSet4_1 data-set, the average identification accuracy of the LDFMN model was 81.73%, which is higher than that obtained using Convolutional 3D (C3D) [58] architecture (71.07%). The identification accuracy for anger/aggression (96.77%) was the highest among those for the three emotions. In the experiments conducted using the TestSet4_2 data-set, the average identification accuracy of the LDFMN model was 76.02%, higher than that obtained using C3D architecture (66.84%). Again, the identification accuracy for anger/aggression (88.70%) was the highest. The identification accuracy achieved using the TestSet4_1 data-set was higher than that achieved using the TestSet4_2 data-set, indicating that background removal did not contribute to higher dog emotion recognition. However, the identification accuracy for happiness achieved using the TestSet4_2 data-set was higher than that achieved using the TestSet4_1 data-set, indicating that background removal may be conducive to the recognition of happiness in dogs. Nevertheless, as illustrated in Figure 21, background removal can cause the loss of a dog's features, resulting in dog emotion recognition errors.

Table 15. Results of dog emotion recognition experiments.

Dataset	Methods	ACC of the Emotion		Average ACC
		Emotion Type	ACC	
TestSet4_1	LDFMN	Neutral/General	77.65%	81.73%
		Happy/Excited	70.00%	
		Angry/Aggressive	96.77%	
	C3D (Tran et al., 2015 [58])	Neutral/General	74.11%	71.07%
		Happy/Excited	66.00%	
		Angry/Aggressive	70.96%	
TestSet4_2	LDFMN	Neutral/General	66.66%	76.02%
		Happy/Excited	76.00%	
		Angry/Aggressive	88.70%	
	C3D (Tran et al., 2015 [58])	Neutral/General	68.51%	66.84%
		Happy/Excited	68.00%	
		Angry/Aggressive	64.52%	

(**a**) (**b**)

Figure 21. Images before and after background removal. (**a**) Original image; (**b**) After background removal.

The reasons for emotion recognition errors, illustrated in Table 16, can be classified into the following four cases:

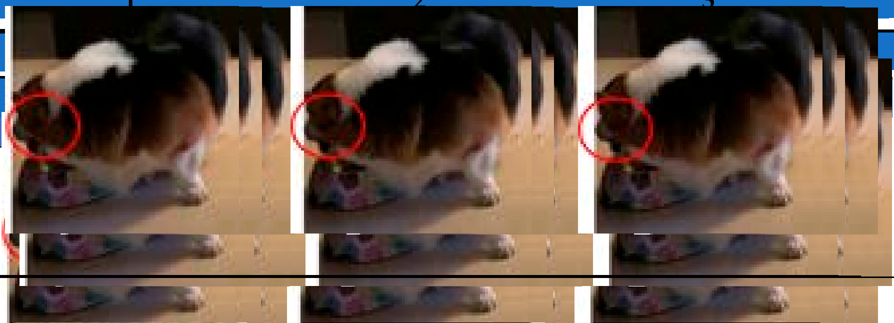
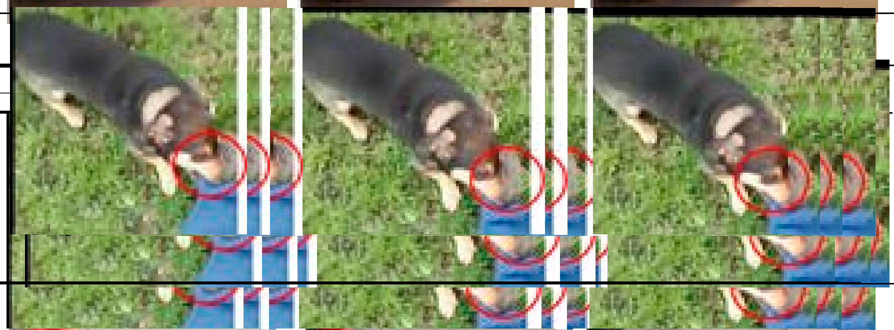
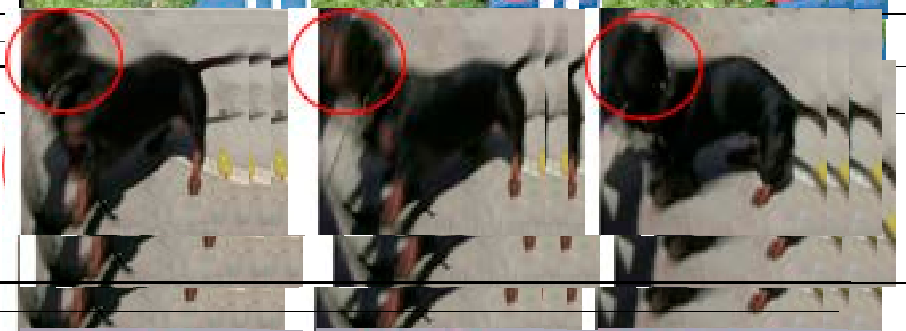
Case 1: An angry or aggressive dog is categorized as being happy or excited. For example, in the image in Table 15, the dog's mouth is only slightly open, and the dog's movements are too subtle.

Case 2: The shooting angle is suboptimal.

Case 3: The dog moves too quickly, resulting in blurry images.

Case 4: The resolution of the image is too low.

Table 16. Emotion recognition errors.

Image Cases	1	2	3
Case 1			
Case 2			
Case 3			
Case 4			

4.4.4. Performance for Dog Emotion Recognition in Surveillance Videos

The models used in the dog detection, tracking, and emotion recognition experiments

Table 17. Identification accuracy of model types in experiments conducted using IMG_0033 data-set.

Type of the Processing	ACC of the Dog Emotion
Type_1	75.45%
Type_2	76.36%
Type_3	76.36%
Type_4	63.89%
Type_5	63.89%
Type_6	62.46%

Table 18. Identification accuracy of model types in experiments conducted using AngryDogs data-set.

Type of the Processing	ACC of the Dog Emotion
Type_1	76.36%
Type_2	76.36%
Type_3	76.36%
Type_4	53.24%
Type_5	53.24%
Type_6	53.76%

(a) (b)

Figure 22. Dogs with similar emotions. (a) Neutral (or general); (b) Happy (or excited).

In the experiment conducted using the AngryDogs data-set, the Type_1, Type_2, and Type_3 models achieved the highest identification accuracy (76.36%), and Type_4 and Type_5 achieved the lowest (53.24%). This indicates that, as with the IMG_0033 data-set, the models that removed the image backgrounds did not effectively recognize the dogs' emotions. Because the dogs in this data-set remain mostly still over the course of the video, the tracking results and identification accuracy values of the Type_1, Type_2, and Type_3 models were the same.

5. Conclusions

The primary purpose of this study was to develop a multi-CNN model for dog detection, tracking, and emotion recognition. The dog detection model was trained using the MSCOCO data-set, and dog tracking and emotion recognition models were trained using videos collected from YouTube, the Folk Stray Dog Shelter, and the DTC. In the dog detection experiment, the detection rates for the TestSet1 and TestSet2 data-sets were 97.59% and 95.93%, respectively. The reasons for detection errors were obscured facial features, special breeds of dogs, obscured or cropped bodies, and incomplete regions. The effects of these factors can be minimized by reducing the number of object types, increasing the sample size of dogs in the training data-set and making the ground truth region more apparent. In the dog-tracking experiment, the MOTA values for videos of a single dog and for multiple dogs were as high as 93.02% and 86.45%, respectively. The tracking failures occurred in cases where large parts of the dog's body were obscured. In the dog emotion recognition experiments, the identification accuracy rates for the two data-sets were 81.73%, and 76.02%, respectively. The results of the emotion recognition experiment indicate that removing the backgrounds of dog images negatively affects the identification accuracy. Furthermore, happy and neutral emotions are similar and therefore difficult to distinguish. In other cases, the dog's movements may not be apparent, the image may be blurred, the shooting angle may be suboptimal, or the image resolution may be too low. Nevertheless, the results of the experiments indicate that the method proposed in this paper can correctly recognize the emotions of dogs in videos. The accuracy of the proposed system can be further increased by using more images and videos to train the detection, tracking, and emotion recognition models presented herein. The system can then be applied in real-world contexts to assist in the early identification of dogs that exhibit aggressive behavior.

Research on automatic face and emotion recognition technology has developed rapidly and matured, and many data-sets have been collected. However, because dogs are not easy to control, there are few datasets for dog tracking and emotion recognition. Therefore, to improve the accuracy of tracking and emotion recognition, it is necessary to further collect many dog-tracking and emotion recognition data-sets in the future.

Author Contributions: Conceptualization, Y.-K.C.; Methodology, H.-Y.C. and C.-H.L.; Software, J.-W.L.; Validation, H.-Y.C. and C.-H.L.; Investigation, J.-W.L.; Resources, Y.-K.C.; Data curation, J.-W.L.; Writing—original draft, C.-H.L. and J.-W.L.; Supervision, H.-Y.C., C.-H.L. and Y.-K.C.; Funding acquisition, Y.-K.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Agricultural Technology Research Institute of Taiwan, R.O.C. grant number Grant No. 109 Agriculture-6.1.1-Nomadic-U2.

Institutional Review Board Statement: The Agricultural Technology Research Institute of Taiwan, R.O.C, approved the study protocol.

Informed Consent Statement: Not applicable.

Data Availability Statement: (1) TestSet1 is the image database established by Columbia University and the University of Maryland (Liu, J.; Kanazawa, A.; Jacobs, D.; Belhumeur, P.), which contains images from ImageNet, Google, and Flickr. (2) TestSet2 is the image database established by Stanford University (Khosla, A.; Jayadevaprakash, N.; Yao, B.; Li, F.F), which contains images from ImageNet. (3) Two pedestrian reidentification data sets, Market-1501 and MARS, which contain images of 1501 and 1261 pedestrians (Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. and Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; Tian, Q.). (4) The data set for dog tracking and emotion recognition contains data from YouTube, the Folk Stray Dog Shelter, and the DTC.

Acknowledgments: This work was supported in part by Agricultural Technology Research Institute of Taiwan under Grant No. 109 Agriculture-6.1.1-Nomadic-U2.

Conflicts of Interest: Chuen-Horng Lin reports financial support was provided by Agricultural Technology Research Institute.

References

1. Broome, S.; Feighelstein, M.; Zamansky, A.; Carreira Lencioni, G.; Haubro Andersen, P.; Pessanha, F.; Mahmoud, M.; Kjellström, H.; Salah, A.A. Going Deeper than Tracking: A Survey of Computer-Vision Based Recognition of Animal Pain and Emotions. *Int. J. Comput. Vis.* **2022**, *131*, 572–590. [CrossRef]
2. Anderson, D.J.; Adolphs, R. A framework for studying emotions across species. *Cell* **2014**, *157*, 187–200. [CrossRef] [PubMed]
3. Zhu, H. Video-Based Dog Pain Recognition via Posture Pattern Analysis. Master's Thesis, Utrecht University, Utrecht, The Netherlands, 2022.
4. Franzoni, V.; Milani, A.; Biondi, G.; Micheli, F. A preliminary work on dog emotion recognition. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume, Thessaloniki, Greece, 14–17 October 2019; pp. 91–96.
5. Boneh-Shitrit, T.; Amir, S.; Bremhorst, A.; Riemer, S.; Wurbel, H.; Mills, D.; Zamansky, A. Deep learning models for classification of canine emotional states. *Comput. Vis. Pattern Recognit.* **2022**, arXiv:2206.05619.
6. Ferres, K.; Schloesser, T.; Gloor, P.A. Predicting dog emotions based on posture analysis using deeplabcut. *Future Internet* **2022**, *14*, 97. [CrossRef]
7. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.

8. Yeo, B.C.; Lim, W.S.; Lim, H.S. Scalable-width temporal edge detection for recursive background recovery in adaptive background modeling. *Appl. Soft Comput.* **2013**, *13*, 1583–1591. [CrossRef]
9. Rakibe, R.S.; Patil, B.D. Background subtraction algorithm based human motion detection. *Int. J. Sci. Res. Publ.* **2013**, *3*, 2250–3153.
10. Mashak, S.V.; Hosseini, B.; Mokji, M.; Abu-Bakar, S.A.R. Background subtraction for object detection under varying environments. In Proceedings of the 2010 International Conference of Soft Computing and Pattern Recognition, Paris, France, 7–10 December 2010; pp. 123–126.
11. Li, H.; Achim, A.; Bull, D.R. GMM-based efficient foreground detection with adaptive region update. In Proceedings of the 2009 16th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, 7–10 November 2009; pp. 3181–3184.
12. Horn, B.K.; Schunck, B.G. Determining optical flow. In Proceedings of the Techniques and Applications of Image Understanding, International Society for Optics and Photonics, Washington, DC, USA, 12 November 1981; Volume 281, pp. 319–331.
13. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FA, USA, 20–25 June 2009; pp. 248–255.
14. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Nice, France, 2012; pp. 1097–1105.
15. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
16. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 1–9.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
18. Comaniciu, D.; Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 603–619. [CrossRef]
19. Kalman, R.E. A new approach to linear filtering and prediction problems. *J. Basic Eng.* **1960**, *82*, 35–45. [CrossRef]
20. Bazzani, L.; Cristani, M.; Murino, V. Decentralized particle filter for joint individual-group tracking. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1886–1893.
21. Zoidi, O.; Tefas, A.; Pitas, I. Visual object tracking based on local steering kernels and color histograms. *IEEE Trans. Circuits Syst. Video Technol.* **2012**, *23*, 870–882. [CrossRef]
22. Bradski, G.R. Computer vision face tracking for use in a perceptual user interface. *Intel Technol. J.* **1998**, *3*, 49–54.
23. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649.
24. Mikolov, T.; Karafiát, M.; Burget, L.; Černocký, J.; Khudanpur, S. Recurrent neural network-based language model. In Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, Makuhari, Japan, 26–30 September 2010.
25. Ojala, T.; Pietikainen, M.; Harwood, D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In Proceedings of the 12th International Conference on Pattern Recognition, Jerusalem, Israel, 9–13 October 1994; Volume 1, pp. 582–585.
26. Gu, W.; Xiang, C.; Venkatesh, Y.V.; Huang, D.; Lin, H. Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. *Pattern Recognit.* **2012**, *45*, 80–91. [CrossRef]
27. Shan, C.; Gong, S.; McOwan, P.W. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vis. Comput.* **2009**, *27*, 803–816. [CrossRef]
28. Oyedotun, O.K.; Demisse, G.; El Rahman Shabayek, A.; Aouada, D.; Ottersten, B. Facial expression recognition via joint deep learning of rgb-depth map latent representations. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 3161–3168.
29. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
30. Panksepp, J. Affective neuroscience of the emotional BrainMind: Evolutionary perspectives and implications for understanding depression. *Dialogues Clin. Neurosci.* **2010**, *12*, 533–545. [CrossRef] [PubMed]
31. Kret, M.E.; Massen, J.J.; de Waal, F. My fear is not, and never will be, your fear: On emotions and feelings in animals. *Affect. Sci.* **2022**, *3*, 182–189.
32. Descovich, K.A.; Wathan, J.; Leach, M.C.; Buchanan-Smith, H.M.; Flecknell, P.; Framingham, D.; Vick, S.-J. Facial expression: An underutilised tool for the assessment of welfare in mammals. *Altex* **2017**, *34*, 409–429. [CrossRef]
33. Briefer, E.F.; Tettamanti, F.; McElligott, A.G. Emotions in goats: Mapping physiological, behavioural and vocal profiles. *Anim. Behav.* **2015**, *99*, 131–143.
34. Walsh, J.; Eccleston, C.; Keogh, E. Pain communication through body posture: The development and validation of a stimulus set. *PAIN®* **2014**, *155*, 2282–2290. [CrossRef]
35. Lecorps, B.; Rödel, H.G.; Féron, C. Assessment of anxiety in open field and elevated plus maze using infrared thermography. *Physiol. Behav.* **2016**, *157*, 209–216. [CrossRef]

36. Kremer, L.; Holkenborg, S.K.; Reimert, I.; Bolhuis, J.; Webb, L. The nuts and bolts of animal emotion. *Neurosci. Biobehav. Rev.* **2020**, *113*, 273–286. [CrossRef]
37. Rashid, M.; Silventoinen, A.; Gleerup, K.B.; Andersen, P.H. Equine facial action coding system for determination of pain-related facial responses in videos of horses. *PLoS ONE* **2020**, *15*, 0231608. [CrossRef]
38. Lundblad, J.; Rashid, M.; Rhodin, M.; Haubro Andersen, P. Effect of transportation and social isolation on facial expressions of healthy horses. *PLoS ONE* **2021**, *16*, 0241532. [CrossRef]
39. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
40. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
41. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 740–755.
42. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
43. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; NIPS: Montreal, QC, Canada, 2015; Volume 28, pp. pp. 91–99.
44. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468.
45. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1995**, *2*, 83–97. [CrossRef]
46. Carolyn Steber, 11 Emotions You Didn't Realize Dogs Could Feel, Bustle. Available online: https://www.bustle.com/p/11-emotions-you-didnt-realize-dogs-could-feel-15644499 (accessed on 25 May 2022).
47. Stanley Coren, Which Emotions Do Dogs Actually Experience? ModernDog. Available online: https://moderndogmagazine.com/articles/which-emotions-do-dogs-actually-experience/32883 (accessed on 25 May 2022).
48. PetFinder, Do Dogs Have Feelings? PetFinder. Available online: https://www.petfinder.com/dogs/dog-training/do-dogs-have-feelings/ (accessed on 25 May 2022).
49. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
50. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. In Proceedings of the Scandinavian Conference on Image Analysis, Halmstad, Sweden, 2–29 June 2003; Springer: Berlin, Heidelberg, 2003; pp. 363–370.
51. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In *Advances in Neural Information Processing Systems Workshop on Autodiff*; NIPS: Long Beach, CA, USA, 9 December 2017.
52. Bradski, G.R.; Kaehler, A. OpenCV. *Dr. Dobb's J. Softw. Tools* **2000**, *120*, 122–125.
53. Liu, J.; Kanazawa, A.; Jacobs, D.; Belhumeur, P. Dog breed classification using part localization. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Springer: Berlin, Heidelberg; pp. 172–185.
54. Khosla, A.; Jayadevaprakash, N.; Yao, B.; Li, F.F. Novel dataset for fine-grained image categorization: Stanford dogs. In Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization (FGVC), Online, 25 June 2011; Volume 2.
55. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1116–1124.
56. Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; Tian, Q. Mars: A video benchmark for large-scale person re-identification. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland; pp. 868–884.
57. Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; Schindler, K. MOT16: A benchmark for multi-object tracking. *arXiv* **2016**, arXiv:1603.00831.
58. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.