Sentiment analysis of pets using deep learning technologies in artificial intelligence of things system

Ming-Fong Tsai1 · Jhao-Yang Huang1

Accepted: 12 July 2021 / Published online: 5 August 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
This research paper proposes sentiment analysis of pets using deep learning technologies in the artificial intelligence of things system. Mask R-CNN is used to detect image objects and generate contour mask maps, pose analysis algorithms are used to obtain object posture information, and at the same time object sound signals are converted into spectrograms as features, using deep learning image recognition technology to obtain object emotion information. By using the fusion of object posture and emotional characteristics as the basis for pet emotion identification and analysis, the detected specific pet behaviour states will be actively notified to the owner for processing. Compared with traditional speech recognition, which uses mel-frequency cepstral coefficients for feature extraction, coupled with a Gaussian mixture model-hidden Markov model for voice recognition, the experimental method of this research paper effectively improves the accuracy by 70%. Prior work on the implementation of smart pet surveillance systems has used the pet’s tail and mouth as features, and has combined these with sound features to analyse the pet’s emotions. This research paper proposes a new method of sentiment analysis in pets, and our method is compared with previous related work. Experimental results show that our approach increases the accuracy rate by 70%.

Keywords Sentiment analysis · Deep learning · Artificial intelligence of things

1 Introduction

Pet sentiment analysis can be used to analyse whether a pet is suffering from anxiety, hypothetical disorders and other mental illnesses. As the number of pets increases, the demand for pet sentiment analysis will increase. Pet sentiment analysis can also be used to obtain more subtle information where pets hide the emotion. For example, when pets are in a vigilant mood, the hidden subtle messages are the response to strangers or strange objects. Traditional sentiment analysis uses voice recognition to analyse emotions, and this uses mel-frequency cepstral coefficients to extract features of the input audio. The main process of using MFCC for feature extraction is obtained by performing the following eight steps on the input source. The first step is pre-emphasis, used to highlight the high-frequency formant, and the second step is frame blocking, which combines x sampling points into a sound frame, where x is usually 256 or 512, and the third step is the Hamming window, which multiplies each sound frame by the Hamming window to increase the continuity between the left and right ends of the sound frame; the fourth step is fast Fourier transform, the fifth step is triangular bandpass filters, and the sixth step is discrete cosine transform; the seventh step is log energy, and the eighth step is the delta cepstrum. After obtaining the features through the above eight steps, a Gaussian mixture model–Hidden Markov model is finally used for speech recognition analysis. Because the MFCC is based on the human ear which can accurately distinguish human speech, it simulates the operation of the human ear in an artificial way, and the main sensitive frequency range of the human ear is 200–5000 Hz, so mel-cepstral coefficients are not suitable for processing sounds other than from human. In traditional sentiment analysis, the target object of the analysis is human beings. The characteristic media used in sentiment analysis can be divided into two main categories:
the first involves image characteristics (Bartlett et al. 2005; Cohn 2007; Khattak et al. 2021; Sebe et al. 2007; Itti-chaichareon et al. 2012; Hasan et al. 2004; Gales and Young 2007; Xie et al. 2019; Rabiner 1989; Schuller et al. 2003), while the second involves sound characteristics (Muda et al. 2010; Chen and Bilmes 2007; Benba et al. 2016; Dave 2013; Brigham 1988; Ahmed et al. 1974; Deng et al. 2021, 2020; Cai et al. 2021; Kalarani and Brunda 2019; Lee and Narayanan 2005). In the first method, most of the image features are based on the human face as the target area of emotion analysis. However, it is impractical to rely solely on either image or sound characteristics for emotion analysis.

Subsequent methods of sentiment analysis were developed based on a combination of the characteristics of images and sounds (Zeng et al. 2006). In this research work, pets are the object of sentiment analysis. This is a novel method in which we use sentiment analysis to obtain subtle information about the hidden emotions of pets. For example, when a pet is wary, there is hidden subtle information based on its reaction to strangers or strange objects. The method used in this research paper is based on the framework of human emotion analysis, which is adapted for use with pets. We analyse the emotions of pets by combining image and sound features. Since pets do not have facial muscles that have developed in the same way as humans, the posture of the pet will be used as an indicator for sentiment analysis based on image features. In terms of sound features, a sentiment analysis is carried out using a spectrogram in the same way as in image analysis, and this is used as an analysis index. Overall, the aim of our method is to simulate the behaviour of humans, who judge the emotions of a pet based on visual cues. In prior related work, a Smart Pet Surveillance System Implementation (SPSSI) framework has been proposed for sentiment analysis in pets (Tsai et al. 2020). This framework combines image and sound characteristics to analyse the emotions of pets. In terms of image features, the pet’s mouth and tail are used as analysis indicators. However, this previously developed approach is not able to accurately analyse emotions when the image does not contain clear features relating to the mouth and tail. Hence, in this research paper, we use the pet’s pose as an indicator for sentiment analysis. The posture of the animal can be obtained at the same time as an image is detected. The approach proposed in this paper can obtain more accurate results for pet sentiment analysis than previous related schemes that require a clear image of the pet’s mouth and tail.

This research paper proposes sentiment analysis of pets using deep learning technologies in the artificial intelligence of things system, using Mask R-CNN (He et al. 2017) to detect and recognise object tags and generate corresponding contour masks to obtain posture features, and uses object sound signals to convert into spectrograms for recognition and analysis to obtain emotions features in order to realise pet emotion analysis through a non-contact smart Intelligence of Things system. The second chapter explains the pet sentiment analysis system architecture, system process and algorithm of the smart Intelligence of Things system. The third chapter explains the experimental environment setting and performance analysis of pet sentiment analysis of the smart Internet of Things system. The fourth chapter is the conclusion and recommends future work.

2 Sentiment analysis of pets in artificial intelligence of things system

2.1 System overview

The overview of the pet sentiment analysis system of the Smart Internet of Things system is shown in Fig. 1. A smart web camera is used to capture pet video and audio information, pet posture analysis for continuous images, and pet sentiment analysis for sound. Pet emotion analysis and recognition is performed based on the posture and emotion information of the above-mentioned deep learning image recognition. When the specific emotional state of the pet is determined, it will be notified to the owner in real time through the communication software for processing.

2.2 System architecture

The structure of the pet sentiment analysis system of the smart IoT system is shown in Fig. 2 and consists of three parts: the hardware layer, the software layer and the application layer. The hardware layer mainly uses network cameras for video and audio capture and computing core platforms for data analysis and calculation. The software layer mainly uses the Tensorflow framework as a machine learning development environment platform, OpenCV-Python for image display and storage functions, PyEmail for email processing kit tools, PyAudio for sound file processing kit tools, and MoviePy for sound file extraction and storage. The application layer has functions of deep learning emotion recognition, deep learning gesture recognition and pet specific state notification.

The pet emotion analysis process of the smart Internet of Things system is shown in Fig. 3, which includes three parts: the user side, the hardware side, and the software side. The pet body on the user side is the target of the analysed emotional state, and the owner’s smart handheld device is the carrier for receiving notifications of pet state analysis. The hardware side is a smart webcam and a
computing core platform. The smart webcam is used to capture pet video and audio files, and the pet video and audio information is input to the computing core platform for analysis, identification and notification. The software side is the environment and package tools of the computing core platform. Pet audio files are extracted and stored through MoviePy, and sound analysis preprocessing is performed through PyAudio, the Tensorflow deep learning image recognition framework is used for emotion analysis and recognition, and then Mask R-CNN is used for mask generation and OpenCV-Python for pose analysis and recognition. The above emotion and posture analysis recognition results are used to determine the specific behaviour state of the pet. The specific behaviour state result is stored in the database and the owner is notified by the email package PyEmail for subsequent processing. The pet sentiment analysis network architecture of the smart Internet of Things system is shown in Fig. 4, including data preprocessing, the Faster R-CNN neural network (Ren et al. 2017), Mask R-CNN neural network, and specific behaviour state analysis. The data preprocessing part performs image framing for the images recorded by the webcam, as well as having the function of extracting sound files and generating spectrograms. After dividing the image into frames, the Mask R-CNN neural network is performed to generate the contour mask map and the pose analysis algorithm is used to obtain the pose analysis result. The spectrogram uses Faster R-CNN neural network image recognition to obtain the sentiment analysis results. According to the above-mentioned posture and emotion analysis results, the pet’s specific behaviour state is determined.

2.3 System functions

The main functional flow of the system is shown in Fig. 5. It uses a web camera to capture the video and audio of the pet body, and the core computing platform performs framed image and sound files to analyse the pet’s posture and emotional information. When the pet’s emotional analysis indicates a specific state such as alert, the owner will be notified for processing. The system randomly samples the framed images of the pet body for Mask R-CNN object detection and obtains the contour mask map
Fig. 3 System flowchart

Fig. 4 Identification network architecture
and then uses the posture analysis algorithm to obtain the pet posture information. The system converts sound files into spectrograms and uses Faster R-CNN for emotion recognition to obtain pet emotion information. After the system has successfully obtained the pet’s posture and emotion information, it makes specific state association judgments to notify the owner of subsequent processing.

2.3.1 Mask R-CNN contour mask

The system is based on Mask R-CNN object detection to identify pets and generate contour masks. The sample set of contour masks in Fig. 6 includes posture categories such as pets standing, sitting, and lying. The system sets the label category as background and pet. Two types are used for deep learning recognition model training to generate weight files for contour mask recognition.

2.3.2 Faster R-CNN identification analysis

Figure 7 shows spectrograms for the pet’s barking in different moods. The left picture is the angry spectrogram, the middle picture is the sad spectrogram, and the right picture is the normal barking spectrogram. The system is based on the Faster R-CNN network architecture to recognise the spectrogram of the pet’s emotional bark, and uses the deep learning recognition model to train and generate the weight file for emotion recognition.

2.4 System algorithm

2.4.1 Posture analysis algorithm

The system posture analysis algorithm is shown in Fig. 8. It performs image framing for the recorded video, and randomly selects $\beta$ pieces of framed images as posture analysis samples, where $\beta$ is less than or equal to the number of frames. When the selected framed image is judged to have a pet based on Mask R-CNN, the pose is judged to be empty; otherwise, a pet contour mask map of the framed image will be generated. The position of the pet in the image is found by using the contour mask. We use $(x_{\text{min}}, y_{\text{min}})$ to represent the coordinate value of the upper left corner of the object box, and use $(x_{\text{max}}, y_{\text{max}})$ to represent the coordinate value of the lower right corner of the object box. According to formula (1), we calculate the row position value of the most pet-rich area (white value) in the object frame and set it to the $\max_x$ value. According to formula (2), we calculate the position value of the column with the most pet areas (white value) in the object box and set it to the $\max_y$ value. We use $\text{IMG}$ to represent the contour mask image array. The white value is 255, and the black value is 0.

\[
\max_x := \max_j \left( \sum_{i=x_{\text{min}}}^{x_{\text{max}}} \text{img}_{i,j} \right) 
\]

\[
\max_y := \max_i \left( \sum_{j=y_{\text{min}}}^{y_{\text{max}}} \text{img}_{i,j} \right) 
\]

\[
\max_x - x_{\text{min}} \leq \max_x - x_{\text{max}} \]

Fig. 5 Main function of system

Fig. 6 Pose category of contour mask sample set
According to the above values of \( x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}}, x_{\text{max}} \), max \in \mathbb{Z}^+, \) the head direction of the pet is judged by the distance between the object’s head and the left and right borders of the object. If the condition of formula (3) is met, the head of the pet in the framed image faces to the left; otherwise, it faces to the right. The posture is judged by judging the distance between the head of the object and the upper and lower boundaries of the object. If the framed image with the pet’s head facing to the left meets the condition of formula (4), it is judged to be standing. If the standing posture is not met, the ratio of the pet area to the background area of the framed image is judged. If the condition of formula (5) is met, it is judged as the prone posture; otherwise, it is the sitting posture. The framed image with the pet’s head facing to the right still uses formula (4) to determine the standing posture conditions. If the standing posture is not met, formula (6) is used to determine the lying posture conditions. If none of them meets formula (4) and formula (6), the condition is sitting. In this research paper, the variable \( \alpha \) represents the threshold for judging whether the animal has a standing posture, whereas the variable \( \kappa \) represents the threshold for judging whether the posture is prone. \( \alpha \) and \( \kappa \) are adjusted using empirical methods. This research paper designs \( \alpha \) and \( \kappa \) to be a real number of 1.2 and a real number of 0.38.

### 2.4.2 Sentiment analysis algorithm

The system sentiment analysis algorithm is shown in Fig. 9. It extracts sound information from video files for sentiment analysis. Using the spectrogram as a sentiment analysis feature, the horizontal coordinate of the spectrogram is time, the vertical coordinate is frequency, and the coordinate point value is the speech data energy, as shown in Fig. 7. The system defines sentiment analysis categories as angry, sad, and normal, and is based on the Faster R-CNN network architecture to train and recognize the spectrogram model. Faster R-CNN arranges the identified possibility results from high to low reliability to form a one-dimensional array. From the identified one-dimensional array results, the top five emotion results are obtained as the voting results of the emotion analysis, and finally the emotion with the highest number of votes is used as the final emotion analysis result of the voice.

### 2.4.3 Specific state analysis algorithm

The system specific state analysis algorithm is shown in Fig. 10, which makes association judgments based on the results of posture and emotion analysis. As shown in Fig. 11, the system determines that the alert state is defined as a pet standing and making angry sounds. If the above specific status occurs, an email is sent to the owner’s smart handheld device application as a reminder notification.
3 Evaluation and results

3.1 Experimental platform and environment

The experimental platform information is shown in Table 1. It uses Logitech Webcam C925 as the network camera, the core computing platform is embedded systems, the system is written using Python programming language with a library for the Tensorflow deep learning development environment, and Pycocotools uses COCO library, OpenCV-Python image processing library, PyAudio speech processing library, MoviePy video editing library, and PyEmail email library.

3.2 Mask R-CNN mask training

The system implemented Mask R-CNN network architecture recognition model training with 475 pet images and trained 60,000 steps to generate model weight files for contour mask recognition. The success rate of generating the contour mask map with the training sample image set is 100% and the average cosine similarity accuracy is 96.78%. We add 10%, 30%, 50%, and 70% of the salt and pepper noise to the training sample image set to generate the contour mask. The respective success rates are 72.94%, 51.29%, 37.87%, and 5.84%, and the average cosine is similar. The accuracy is 92.90%, 88.89%, 81.87%, and 62.23%, respectively. The values are shown in Table 2. Figure 12 shows the similarity percentage data of the

---

Fig. 8 Posture analyze algorithm

1. INPUT \( \beta \) images
2. INPUT posture[\( \beta \)]
3. FOR i=0 to \( \beta -1 \) DO
4. IF Image has dogs THEN
5. Generate the contour mask of the i-th image
6. Find the position of the i-th image dog
7. IF The head of the dog in the i-th image is facing left THEN
8. Analyze the posture of the dog in the i-th image
9. IF Posture is standing THEN
10. posture[i] set to standing posture
11. ELSE IF Posture is lying THEN
12. posture[i] set to lying posture
13. ELSE
14. posture[i] set to sitting posture
15. ENDIF
16. ELSE
17. Analyze the posture of the dog in the i-th image
18. IF Posture is standing THEN
19. posture[i] set to standing posture
20. ELSE IF Posture is lying THEN
21. posture[i] set to lying posture
22. ELSE
23. posture[i] set to sitting posture
24. ENDIF
25. ENDIF
26. ELSE
27. posture[i] set to NULL
28. ENDIF
29. ENDFOR
30. Determine the most posture in posture[\( \beta \)] as the posture information of the video
contour mask image generated by the standing, sitting and lying postures at different levels of noise. The similarity of the standing posture, sitting posture and lying posture without noise interference values is 96.45%, 97.01% and 97.27%, respectively. The similarity of the standing, sitting and lying postures of the framed image under 10% noise interference is 92.63%, 93.81% and 93.00%, respectively. The similarity of the standing, sitting and lying postures of the framed image under 30% noise interference is 89.98%, 87.73% and 87.77%, respectively. The similarity of the standing, sitting and lying postures of the framed image under 50% noise interference is 85.60%, 74.12% and 85.68%, respectively. The similarity of the standing, sitting and lying postures of the framed image under 70% noise interference is 69.81%, 60.33% and 67.06%, respectively.

3.3 Posture analysis

The system is based on the contour mask map generated by Mask R-CNN to perform pose analysis. The results of the algorithm are shown in Fig. 13 as lying, sitting, and standing. The red frame line is the target object position, the green line is the vertical position where the outline mask image contains the most target object information, and the cyan line is the horizontal position where the outline mask image contains the most target object information under 50% noise interference is 85.60%, 74.12% and 85.68%, respectively. The similarity of the standing, sitting and lying postures of the framed image under 70% noise interference is 69.81%, 60.33% and 67.06%, respectively.

Table 1  Experimental platform

| Camera                        | Logitech webcam C925 |
|-------------------------------|----------------------|
| Platform                      | Computer             |
| Programming language          | Python3.7             |
| Main library                  | Tensorflow = 1.14.0   |
|                               | Pycocotools = 2.0     |
|                               | Opencv-Python = 3.4.3.18 |
|                               | PyAudio = 0.2.11      |
|                               | Moviepy = 1.0.3       |
|                               | PyEmail               |

Fig. 9  Sentiment analysis algorithm

Fig. 10  Specific state analysed algorithm

1. INPUT Posture result
2. INPUT Emotional result
3. Infer the state of the dog based on the correlation between the result of the posture and the result of the emotion
4. IF The status result of the dog is the specified status THEN
5. Notify users of analyzed information by E-Mail
6. Perform the next canine status analysis
7. ELSE
8. Perform the next canine status analysis
9. ENDIF

Fig. 11  Definition of alert status

Alert := Posture : Standing ∧ Emotion : Angry
information. The system uses 269 contour mask images for the pose analysis accuracy test and its accuracy is 90.7%.

### 3.4 Sentiment analysis

The system uses 30 voice files to perform emotion analysis accuracy experiments, including three emotional states, angry, sad, and normal, and each emotional state has ten data. When the one-dimensional array identified by Faster R-CNN uses the TOP-1 result as the basis for emotion voting, the emotion with the highest vote is the final emotion result. The analysis accuracy of angry, sad, and normal states is 80%, 60%, and 90%, respectively. The average accuracy is 76.6%, as shown in the histogram on the left of Fig. 14. When the one-dimensional array identified by Faster R-CNN uses the TOP-3 result as the basis for emotion voting, the analysis accuracy of angry, sad, and normal states is 80%, 60%, and 70%, respectively. The average accuracy is 70%. The value is shown in the middle histogram in Fig. 14. When the one-dimensional array identified by Faster R-CNN uses the TOP-5 result as the basis for emotion voting, the analysis accuracy of angry, sad, and normal states are 80%, 90%, and 90%, respectively. The average accuracy is 86.6%, as shown in the histogram on the right side of Fig. 14. Using the TOP-5 results as the basis for emotion voting in the system, the

### Table 2 Mask similarity

| Added noise ration | 0%          | 10%         | 30%          | 50%          | 70%          |
|-------------------|-------------|-------------|--------------|--------------|--------------|
| Average similarity| 96.78%      | 92.90%      | 88.89%       | 81.87%       | 62.23%       |
| Success rate of mask generation | 100% | 72.94% | 51.29% | 37.87% | 5.84% |

![Fig. 12](image-url) The noise level corresponds to the similarity of the three postures

![Fig. 13](image-url) Posture analysis chart
average emotion accuracy is the best and the accuracy of each emotion is the most stable. In traditional voice recognition, MFCC plus GMM-HMM is used for voice recognition, and the same 30 voice files are used to perform emotion analysis accuracy experiments, including ten data for each of the three emotional states of angry, sad, and normal. The accuracy of the analysis in angry, sad, and normal states is 10%, 80%, and 10%, respectively. The average accuracy is 33.3%. The values are shown in Table 3.

### 3.5 State recognition

When the system uses the pet specific state analysis algorithm to determine that it is alert, it will immediately send an email to notify the owner, as shown in Fig. 15. The system uses seven audiovisual files to determine the alert state, and its accuracy is 85.71%. If the sentiment analysis uses MFCC plus GMM-HMM for voice recognition, the accuracy of judging the alert state is 14.29%.

### 3.6 Comparison of results for execution time and accuracy

For comparison, we use the results from a prior scheme for smart pet surveillance system called SPSSI, as shown in Fig. 16. The experiment used 14 test images, representing the category of alarm. The average execution time for the proposed method was 1.32 min. In terms of the average execution time, our method was better than SPSSI, as shown in Fig. 16. The accuracy of the proposed method was 85.71%, while the accuracy of SPSSI was 14.29%, as shown in Table 4. This experiment proves that when clear information on the positions of the mouth and tail is not available from a video, the SPSSI method cannot be used to accurately analyse the pet’s emotions. However, the method presented in this paper combines the pet’s posture

![Sentiment analysis accuracy](image1)

**Table 3** MFCC + GMM-HMM sentiment analysis accuracy

| Emotion | Angry | Sad | Normal |
|---------|-------|-----|--------|
| Accuracy | 10%   | 80% | 10%    |
| Average  | 33.33%|     |        |
and voice characteristics to conduct a sentiment analysis that is more reliable than the SPSSI method.

4 Conclusions

This research paper proposes a method of pet sentiment analysis that is based on the artificial Intelligence of Things system; a Mask R-CNN deep learning approach is used for pet object detection and the generation of contour masks, and Faster R-CNN deep learning is used for the recognition and classification of the animal’s emotions. Our sentiment analysis method combines object posture and emotional characteristics, and uses these as a basis for the identification and analysis of the animal’s emotions. Our pet sentiment identification method has an accuracy of 85.71% for successful recognition in a non-contact manner, and informs the owner for processing. The approach proposed in this paper is compared with related work on smart pet surveillance systems, and an implementation of our method is shown to improve the accuracy by 70%. Moreover, the experiments presented here prove that when the object of emotion analysis is a pet, the accuracy of emotion analysis based on a combination of the posture and sound features will be better than an analysis of the pet’s body part features (for example, using the mouth and tail characteristics) combined with sound features. The main challenge of this approach is that when the distance between the pet and the focal distance of camera is too small, most of the captured images will contain the pet’s body or other parts of the pet, which makes it impossible to analyse the characteristics of the pet’s posture. In this case, the method presented here will rely only on sound features for pet emotion analysis, which will affect the accuracy of the results.

Author’s contributions M-FT was involved in supervision. M-FT and J-YH were involved in writing—original draft. All authors have read and agreed to the published version of the manuscript.

Funding This research was funded by National United University, Taiwan.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.
Ethical standards This article does not contain any studies with human participants or animals performed by any of the authors.

References

Ahmed N, Nasir T, Rao K (1974) Discrete Cosine Transform. IEEE Trans Comput C–23(1):90–93
Bartlett M, Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J (2005) Recognizing facial expression: machine learning and application to spontaneous behavior. In: IEEE Computer society conference on computer vision and pattern recognition, pp 568–573
Benba A, Abdelilah J, Hammouch A (2016) Discriminating between patients with Parkinson’s and neurological diseases using cepstral analysis. IEEE Trans Neural Syst Rehabil Eng 24(10):1100–1108
Brigham E (1988) The fast Fourier transform and its applications. Prentice Hall, Englewood Cliffs, New Jersey
Cai X, Zhao H, Shang S, Zhou Y, Deng W, Chen H, Deng W (2021) An improved quantum-inspired cooperative co-evolution algorithm with multi-strategy and its application. Elsevier Exp Syst Appl 171:114629–114642
Chen C, Bilmes A (2007) MVA processing of speech features. IEEE Trans Audio Speech Lang Process 15(1):257–270
Cohn J (2007) Foundations of human computing: facial expression and emotion, lecture notes in computer science artificial intelligence for human computing, pp 1–16
Dave N (2013) Feature extraction methods LPC PLP and MFCC in speech recognition. Int J Adv Res Eng Technol 1(6):1–5
Deng W, Shang S, Cai X, Zhao H, Song Y, Xu J (2021) An improved differential evolution algorithm and its application in optimization problem. Springer Soft Comput 25:5277–5298
Deng W, Xu J, Gao X, Zhao H (2020) An enhanced MSIQDE algorithm with novel multiple strategies for global optimization problems. IEEE Trans Syst Man Cybern Syst. https://doi.org/10.1109/TSMC.2020.3030792
Gales M, Young S (2007) The application of hidden Markov models in speech recognition. Found Trends Signal Process 1(3):195–304
Hasan M, Jamil M, Rabbani G, Rashidul M (2004) Speaker identification using mel frequency cepstral coefficients. In: International conference on electrical and computer engineering, pp 565–568
He K, Gkioxari G, Dollar P, Girshick R (2017) Mask R-CNN. In: IEEE conference on computer vision and pattern recognition, pp 1–12
Ittichaichareon C, Sukri S, Yingthawornsuk T (2012) Speech recognition using MFCC. In: International conference on computer graphics, simulation and modeling, pp 135–138
Kalarani P, Brunda S (2019) Sentiment analysis by POS and joint sentiment topic features using SVM and ANN. Springer Soft Comput 23:7067–7079
Khattak A, Habib A, Asghar M, Subhan F, Razzaq I, Habib A (2021) Applying deep neural networks for user intention identification. Springer Soft Comput 25:2191–2220
Lee C, Narayanan S (2005) Toward detecting emotions in spontaneous dialogs. IEEE Trans Speech Audio Process 13(2):293–303
Muda L, Begam M, Elamvazuthi I (2010) Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. J Comput 2(3):138–143
Rabiner L (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 36(9):1137–1149
Schuller B, Rigoll G, Lang M (2003) Hidden Markov model-based speech emotion recognition. In: IEEE international conference on acoustics, speech, and signal processing, pp 401–404
Sehe N, Lew M, Sun Y, Cohen I, Gevers T, Huang T (2007) Authentic facial expression analysis. Elsevier Image vis Comput 25:1856–1863
Tsai M, Lin P, Huang Z, Lin C (2020) Multiple feature dependency detection for deep learning technology - smart pet surveillance system implementation. MDPI Electronics 9(9):1387–1403
Xie X, Ge S, Hu F, Xie M, Jiang N (2019) An improved algorithm for sentiment analysis based on maximum entropy. Springer Soft Comput 23:599–611
Zeng Z, Hu Y, Fu Y, Huang T, Roisman G, Wen Z (2006) Audio-visual emotion recognition in adult attachment interview. In: ACM international conference on multimodal interfaces, pp 139–145

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.