European Research Studies Journal
Volume XXVI, Issue 1, 2023

pp. 572-585

On Emotion Detection and Recognition Using a Context-Aware Approach by Social Robots– Modification of Faster R-CNN and YOLO v3 Neural Networks

Submitted 03/02/23, 1st revision 20/02/23, 2nd revision 18/03/23, accepted 30/03/23

Eryka Probierz[1]

Abstract:

Purpose: *This paper points out that it is not sufficient only to analyze the human face, but it is also necessary to know the context. This allows for a more accurate classification of emotions, and thus a more appropriate match between the robot's behavior and the social situation in which it finds itself.*
Design/methodology/approach: *Proper situation assessment through a social robot is a fundamental and necessary skill at this point. In order for such an evaluation to be correct, it is necessary to distinguish certain criteria whose fulfillment can be responsible for the robot's better understanding of human intentions. One such criterion is the identification of the interlocutor's emotions. For the analysis, Emotic image database has been used, whose unique character allows to identify 26 emotions, understood as discrete categories. This database is constructed in such a way that it allows to detect emotions from both the face or posture of a person, as well as from the context that occurs in the picture.*
Findings: *The models chosen to solve the problem are Faster R-CNN and YOLO3 networks. In this paper a two-stage analysis is presented. Originally with no changes in the network structure along with the measurement efficiency. And then, as a next step, modifications to the aforementioned neural networks were proposed by introducing the possibility of an internal classifier that allowed for more satisfactory results.*
Practical Implications: *The analyzed solutions allow implementation in social robots due to the speed of operation, but show some hardware requirements. Nevertheless, they are an important support for social robots in social situations and have a chance to be the next step to their dissemination in everyday life.*
Originality/Value: *Emotion detection and recognition is an essential part of the human-robot relationship. Proper recognition increases the acceptance of robots by humans.*

Keywords: *Emotion recognition, emotion detection, neural network, Faster R-CNN, YOLOv3, social robots*

JEL classification: *D12, D47, D53.*
Paper type: *A research study.*

Funding: *The work was supported in part by the European Union through the European Social Fund as a scholarship under Grant POWR.03.02.00-00-I029.*

[1]*Silesian University of Technology, Gliwice; Łukasiewicz Research Network – Institute of Innovative Technologies EMAG, Katowice; ORCID ID 0000-0002-6588-1975, erykaprobierz@gmail.com;*

1. Introduction

Social robots take part in a variety of social scenarios in which humans have certain needs.A set of assumptions about specific actions, attitudes, displayed emotions, or structure exists in a social context, and this knowledge is referred to as social intelligence (Mileounis *et al.,* 2015). The more a person accurately perceives and transmits a clear message, the more satisfying a particular social setting is for each participant since it allows for comprehension and proper communication.

How much a social robot has created social knowledge will influence how well it is gotten by individuals and how well it cooperates with them (Barchard *et al.,* 2020). When people communicate with one another, they send a variety of signals, including facial expressions, postures, the manner they speak, and what they say or do. The emotions that the interlocutor is experiencing at the time influence the content of the message.

When emotions are recognized correctly, the context of the statement is recognized correctly, and there is less information noise in communication (Qureshi *et al.,* 2016). A feeling is a psychological express that occurs when someone experiences or imagines a circumstance (Hirth *et al.,* 2011). The ability to correctly express and recognize emotions is linked to social intelligence, as well as the environment or culture in which a person was raised (Kim *et al.,* 2008).

Emotions are divided into numerous categories in order to classify and name distinct states. The divide between simple and complicated emotion (Thuseethan *et al.,* 2020) is the most common division. The term ”simple emotion” refers to a group of basic emotions that share a common denominator of universal occurrence (Paiva *et al.,* 2014; Rybka and Janicki, 2013).

Numerous endeavors have been made to discern basic emotions based on images (Tsiourti *et al.,* 2019; Leo *et al.,* 2015; Perez-Gaspar *et al.,* 2016), indicating that the methods used are highly accurate and effective. Composite emotions are variously defined as composites of many simple emotions with the addition of extra continuous characteristics like intensity (Wiem and Lachiri, 2017).

Ekman’s (Wiem and Lachiri, 2017) and Plutchik’s (Esau *et al.,* 2007) wheel models are the most popular. Deep neural networks face various difficulties in detecting and recognizing them. However, continuing research shows that emotion identification is extremely accurate (Ebrahimi Kahou *et al.,* 2015; Levi and Hassner, 2015; Jain *et al.,* 2019).

According to the aforementioned studies, there are numerous techniques to effectively detect simple and complicated emotions based on facial photos in the field of emotion detection. However, there are significant downsides to this technique, which become especially important when integrating the given algorithms

into social robots. The first issue is that not every face, or even a portion of one, is available for examination. When there is no face in a snapshot, the algorithm is unable to analyze emotions. The key is to take a gander at both the faceposture and the setting of the photograph (Bendjoudi *et al.,* 2020). This method also allows for the consideration of more than one individual in the snapshot.

Because social robots are expected to participate in a variety of social contexts, it is necessary to consider that the robot will see a variety of interactions, including human-robot, human-human, and robot-robot interactions. Context analysis allows us to account for the links between the actors in a social context, reducing the possibility of artifacts that could lead to misinterpretation (Chen and Whitney, 2021). Social robots are already used in a variety of applications, indicating that further research could lead to improved robot-human interaction and, as a result, increased robot efficiency.

Currently, social robots are being successfully used in rehabilitation, supporting it, pushing people to be more methodical, and facilitating training monitoring (Kellmeyer *et al.,* 2018). Another application is for early diagnosis of depressive symptoms based on the study of patterns and routines of behavior, detection of changes in facial expression or speech, and maintenance of the therapeutic process in the treatment of mental diseases (Cao *et al.,* 2018). They can also help people age gracefully by helping with every day exercises, working with mental and actual wellbeing journals (Broekens *et al.,* 2009).

The use of robots in the development of social skills for people with autism spectrum disorders is a common application, as it allows for the development of certain competences based on a recurring pattern of behaviors and cause-effect correlations (Cabibihan *et al.,* 2013). The prospect of deploying social robots as private tutors is also mentioned in personalized education, allowing for the adaptation of the chosen subject matter to the level of development (e.g., children of immigrants who better their learning of a language that is not their native language) (Heerink *et al.,* 2016).

The work on enhancing the capabilities of social robots can possibly be not simply hypothetical, but also practical, due to the development opportunities mentioned. It is a broad field that provides for growth in many areas (Koziarski and Cyganek, 2018). Emotion discovery and acknowledgment dependent on both the faceposture and the specific situation visible in the image is the problem addressed in this study.

This is to allow for emotion recognition in photographs with multiple people or if faces aren't visible. The EMOTIC database, which comprises images of people in diverse places and scenarios, will be used for this purpose. Two algorithms will be used to examine the photos, allowing for quick calculations. For the use of emotion recognition and practical deployment of social robots, speed of operation is critical.

As a result, the EMOTIC picture database was trained using the YOLOv3 and Faster R-CNN algorithms, with extensive descriptions and results provided below. The results for the same collection of photographs but other algorithms were also analyzed.

2. YOLO v3

Darknet (Redmon and Farhadi, 2018) built the YOLO network, and v3 is the third and enhanced version of the network. YOLO stands for ”You Only Look Once.” This network employs one Darknet variation, consisting of 53 layers that were trained on a batch of ImageNet images. The number of these levels is placed on top of the 53 layers, totaling 106 layers, which is the YOLO v3 fundamental architecture.

The ability to identify at three different scales distinguishes this network from its predecessors. This means that detection can be done using a 1 x 1 kernel, but on three separate feature maps with varied sizes and locations throughout the network. The prediction is made at three scales that reduce the input image by 32, 16, and 8 percent, respectively.

Because theimage's fine granularity is preserved, this property enables for the detection of small objects. This network supports cross-error based class prediction, i.e., utilizing logistic regression, as opposed to the prior version's quadratic errors-based approach. A Softmax layer is not present in this network since its presence would suggest that each image must be assigned to only one class (Ju *et al*., 2019). Because this layer is missing, many labels can be applied to a single image; the greater the class score predicted using logistic regression, the higher the ranking.

3. Faster R-CNN

In 2015, the Faster R-CNN network was created as a follow-up to the R-CNN family of networks (Ren *et al.,* 2015). A region creation technique that locates an object on a map, feature generation for objects, classification of objects into classes, and a regression layer that allows for more accurate coordinates for specific objects are all common elements in all networks. The region proposal technique is what sets Faster R-CNN apart from the other networks in this family.

A selective search algorithm was previously proposed, but it required CPU calculation and was slower. RPN or convolutional network is utilized in Faster R-CNN to produce regions that will be evaluated later. This reduces analysis time while also allowing for the sharing of a given region between layers, which improves feature representation. The underlying network is made up of an RPN (region proposal algorithm) and a Fast R-CNN (detector). The network is anchor-based in both this framework and YOLO v3, and regression is used again. The most significant distinction is in the manner in which it operates.

YOLO v3 does regression classification of the bounding boxes simultaneously, whereas Faster R-CNN does it in two steps. This means that utilizing the YOLO v3 algorithm (Mao *et al.,* 2019) has no drawbacks.

Faster R-CNN is difficult to implement in real-time due to the two-step structure. On the other hand, the speed of the aforementioned algorithm is sufficient to be deemed a social robot functionality that does not require real-time. It was decided to conduct the investigation using the Faster R-CNN implementation (Ho *et al.,* 2019), which allows this model to be applied in real-time.

4. Experimental Results

4.1 EMOTIC Database

The EMOTIC database has 23,571 photos representing 34,320 people (Kosti *et al.,* 2019). It is a compilation of data from the COCO (Lin *et al.,* 2014) and Ade20K (Zhou *et al.,* 2017) databases. This database is unusual in that it comprises images of individuals in a variety of scenarios and scenarios, in a variety of settings, allowing for a diversified collection in terms of activities, social circumstances, and quantity of individuals.

The database has been tagged, allowing for the assignment of several emotions from 26 distinct categories to each individual. Unlike traditional emotion detection databases, this database contains not only photographs of faces, but the entire picture of a scene. The database's creators identified the 26 emotions based on a review of dictionaries and psychiatric texts, allowing for the formation of clusters of terms with comparable meanings.

From these clusters, 26 categories were chosen, taking into account the element of emotion distinguishability based on the photo, as well as six basic emotions that have been studied previously. The level of agreement for the tagged photos was also evaluated so that the information gathered could be used as a valid reference source.

The categories were: affection, anger, annoyance, anticipation, aversion, confidence, disapproval, disconnection, disquietment, doubt/confusion, embarrassment, engagement, esteem, excitement, fatigue, fear, happiness, pain, peace, pleasure, sadness, sensitivity, suffering, surprise, sympathy, and yearning.

The developers of the database (Kosti *et al.,* 2019) provided detailed definitions for the different categories. Among the database's fundamental statistics, it should be mentioned that 66% of the characters are male and 34% are female. The majority of the characters are adults (83%) although there are also youngsters (10%) and teenagers (7%) (Kosti *et al.,* 2019).

Figure 1. *Example of a image from EMOTIC database*

Source: *Kosti et al., 2019.*

4.2 Simulation Results on YOLOv3

YOLO v3 was the model used. TensorFlow 2 (Lin *et al.*, 2014) does this, allowing the network to be trained on a GPU. The annotations for the EMOTIC database were converted using convert2Yolo (Zhou *et al.*, 2017), which creates a text file with class assignment information and image dimensions. The text files were placed in a label folder, while the processed photos were placed in a directory.

After that, a 70:30 split was used to construct a training and test set. The executable was constructed by first creating a file with the names of 26 emotion categories, and then using the dataset that had been constructed. The batch parameter was set to 64, and each network layer's line classes were adjusted to correspond to the 26 emotions. The filter values were also changed as a result.

The entire network is based on a Darknet 53 structure consisting of 53 convolutional network layers (Lee *et al.*, 2020). Anchors are employed in this model to identify areas, and the entire network is based on a Darknet 53 structure consisting of 53 convolutional network layers.

Table 1*. Performance scores (AP) for emotic dataset*

Labels	CNN (Kosti *et al.*, 2019)	GCN (Zhang *et al.*, 2019)	EmotiCon (Mittal *et al.*, 2020)	YOLO v3	Faster R-CNN
Affection	26,01	**46,89**	45.23	31.12	29.47
Anger	11,29	10,87	**15.46**	14.74	11.29
Annoyance	16,39	11,27	**21.92**	18.41	21.16
Anticipation	58,99	62,64	**72.12**	67.25	71.18
Aversion	9,56	5,93	17.81	**18.81**	14.39
Confidence	81,09	72,49	68.65	77.53	**86.84**
Disapproval	16,28	11,28	19.82	**20.54**	18.46

Disconnection	21,25	26,91	**43.12**	33.02	27.56
Disquietment	20,13	16,94	18.73	17.42	**23.21**
Doubt/Confusion	33,57	18,68	35.12	31.89	**35.47**
Embarrasement	3,08	1,94	14.37	**17.22**	6.04
Engagement	86,27	88,56	**91.12**	89.11	87.26
Esteem	18,58	13,33	23.62	**25.55**	22.73
Excitement	78,54	71,89	93.26	**95.62**	92.19
Fatigue	10,31	13,26	16.23	18.41	**19.47**
Fear	16,44	4,21	**23.65**	19.92	17.52
Happiness	55,21	73,26	74.71	76.70	**77.41**
Pain	10,00	6,52	13.21	11.29	**14.12**
Peace	22,94	32,85	34.27	30.12	**36.87**
Pleasure	48,65	57,46	**65.53**	29.99	24.36
Sadness	19,29	25,42	23.41	24.20	**26.07**
Sensitivity	**8,94**	5,99	8.32	8.46	6.71
Suffering	17,60	23,39	26.39	**27.44**	25.81
Surprise	21,96	9,02	17.37	16.74	**24.12**
Sympathy	15,25	17,53	34.28.	31.85	**36.44**
Yearning	9,01	10,55	**14.29**	12.81	9.47
mean	28,33	28,42	**35.48**	33.31	33.29

Source: *Own study.*

4.3 Simulation Results on Faster R-CNN

Initial assumptions were required by the model (Ho *et al.*, 2019). Along with cuDNN, a CUDA package was installed, enabling for GPU processing. This made it possible to train the Faster R-CNN network on a GPU with 8GB of RAM (using cuDNN). The network uses the default anchor solution, which consists of 256 image channels and nine potential windows (anchors) produced by multiplying three areas by their coefficients.

This corresponds to three areas, 1252, 2562, and 5122, respectively, for coefficients 1:1, 1:2, and 1:3. The anchor placements and parameters can be determined using a mixture of classification and regression layers. For the region proposal network, determining these is critical (RPN). A pre-trained network (Gao *et al.*, 2018) is used in the model to detect features in the image.

Because the initial layers differentiate edges or color patches that are universal independent of the implemented set, this is achievable. The computed regression coefficients for the envelope were used to change the generated anchors (i.e., windows). This enables us to generate foreground and background windows that will be used by the region proposal network (RPN).

The threshold value distinguishes between background and foreground anchors; if the value is higher, it is a foreground anchor; if the value is lower, it is a background anchor. The default settings specified by the authors were used to train the model,

namely a threshold of foreground anchors above 0.7, background anchors below 0.3, and batchsize for RPN 256. The next step was to examine the above layer's Region of Interest (ROI) to determine the foreground and background ROIs.

Foreground ROIs are utilized above 0.5, whereas background ROIs are used between 0.5 and 0.1, and the batchsize for ROIs is 128. Furthermore, a fraction factor of 0.25 is employed, implying that the number of foreground ROIs cannot exceed batchsize fraction. The final stage is a classification layer, which allows a class to be assigned to a certain window. A detection threshold is set that permits as many envelopes to be drawn in the image as there are classes assigned to a window. Articles by the developers and modified versions of Faster R-CNN (Jiang and Learned-Miller, 2017; Chen *et al.*, 2018; Fan *et al.*, 2016) provide a more detailed discussion of the architecture.

Similar findings were achieved in all of the tests. The average score for both networks was 33. This isn't the finest score in the literature, but it's close to the 35.48 (Mittal *et al.*, 2020) score. The emotions produced the best outcomes for the Faster R-CNN network: confidence, disquietment, doubt/confusion, fatigue, happiness, pain, peace, sadness, surprise and sympathy. To be used on the YOLO v3 network. The emotions that produced the best results were aversion, disapproval, embarrassment, esteem, enthusiasm, and suffering. For YOLO v3, the results were similar for both networks. Faster R-CNN outperformed Faster R-CNN in 14 areas.

Table 2. *Performance scores (AP) for emotic dataset on modified neural networks*

Labels	CNN (Kosti *et al.*, 2019)	GCN (Zhang *et al.*, 2019)	EmotiCon (Mittal *et al.*, 2020)	YOLO v3	Faster R-CNN
Affection	**45.23**	31.12	29.47	30.14	41.41
Anger	**15.46**	14.74	11.29	12.12	13.88
Annoyance	21.92	18.41	21.16	16.48	**27.48**
Anticipation	72.12	67.25	71.18	69.44	**82.04**
Aversion	17.81	**18.81**	14.39	18.44	14.02
Confidence	68.65	77.53	86.84	71.79	**88.44**
Disapproval	19.82	20.54	18.46	**21.88**	19.43
Disconnection	**43.12**	33.02	27.56	31.94	38.25
Disquietment	18.73	17.42	23.21	15.47	**27.81**
Doubt / Confusion	35.12	31.89	35.47	36.44	**39.44**
Embarrasement	14.37	**17.22**	6.04	17.02	14.70
Engagement	**91.12**	89.11	87.26	90.78	89.38
Esteem	23.62	25.55	22.73	**26.77**	25.31
Excitement	93.26	**95.62**	92.19	91.45	94.15
Fatigue	16.23	18.41	19.47	19.44	**22.66**
Fear	**23.65**	19.92	17.52	18.40	19.12
Happiness	74.71	76.70	77.41	72.88	**79.49**
Pain	13.21	11.29	14.12	10.61	**16.77**
Peace	34.27	30.12	36.87	29.11	**45.63**

Pleasure	**65.53**	29.99	24.36	31.25	40.47
Sadness	23.41	24.20	26.07	22.22	**28.08**
Sensitivity	8.32	8.46	6.71	6.55	**9.09**
Suffering	26.39	27.44	25.81	26.05	**29.12**
Surprise	17.37	16.74	24.12	22.18	**31.71**
Sympathy	34.28.	31.85	36.44	27.44	**38.43**
Yearning	**14.29**	12.81	9.47	11.03	14.22
mean	35.48	33.31	33.29	32.59	**38.09**

Source: *Own study.*

4.4 Comparison

It was chosen to compare the collected results to three previously published solutions. The authors of the EMOTIC database (Ho *et al.,* 2019) proposed the classic CNN, as well as graphical modification (Zhang *et al.,* 2019) and the EmotiCon solution (Mittal *et al.,* 2020). CNN (Kosti *et al.,* 2019) is made up of two feature extractions.

The first contains feature extractions from highlighted text, while the second has feature extractions from the entire image. After they've been distinguished, they're joined using the Fusion network, and their emotion category scores are calculated. The graph network (Zhang *et al.,* 2019) has two branches, one of which is a convolutional network and the other of which employs RPN to highlight image windows. Then, taking into account the interrelationships between the highlighted windows, an emotional graph is built for them.

This information is then combined with the convolutional network, and emotion categories are highlighted.. The last answer achieved the best result in the comparison. Each of these networks uses the strategy of identifying and classifying images based on their collation of selected windows.

The network presented by the base authors had the best sensitivity result; it is also worth mentioning that the results of Faster R-CNN were comparable to CNN. For affection, the graph network based on the convolutional network produced the best results. The Emoti-Con network had the greatest performance in this category: anger, annoyance, anticipation, disconnection, engagement, fear, pleasure and yearning.

4.5 Simulation based on Modified Faster R-CNN with Classifier

Based on the literature analysis, it was decided to modify the Faster R-CNN network with the approach proposed for vehicle detection (Nguyen, 2019). In this approach, a soft non-maximal suppression (NMS) algorithm is used at the anchor generation stage of the RPN, allowing better results for repeated positions. The use

of a soft solution is important here, since a classical one could lead to complete item removal.

The soft solution allows the ROI to be updated with neighbor proposals and the winning proposal based on objectivity evaluations. In practice, this means that when, for example, two context elements or two face elements are superimposed on each other in a photo, the algorithm will distinguish them, then evaluate which one is the neighbor and which one is the winning item with an assumed cross-validation factor of 0.5 (Hu *et al.,* 2018).

Another element that modifies the network is the use of context-dependent ROI fields (Nguyen, 2019). Using this mechanism allows the proposal size to be adjusted to a fixed size, but without the occurrence of inaccurate representations in forward propagation and the occurrence of increased errors in backward propagation.

That is, if the proposal size is larger than the predetermined feature map size, it will be reduced to the item size from the fixed size using the output feature map. If the proposal image is smaller then a deconvolution operation will be applied, allowing the proposal size to grow to a fixed value. The final element of the change is the application of a classifier allowing the extracted features to be used for classification. This allows the classification of propositions into an emotion class and an emotion class for context (Nguyen, 2019).

4.6 Simulation based on Modified YOLOv3 - YOLO-tiny

In order to improve the results for the YOLO3 network, it was decided to use a simplified model of this network called YOLO-tiny (Lestari *et al.,* 2019). This model in contrast to its larger counterpart contains only 7 convolutional layers (not 53) and 6 max-pooling layers. This network uses a CBL or Convolutional Batch Normalization Leaky Relu model which allows to combine several functions, i.e., convolutional layer, batch normalization layer and Leaky Relu (Zhang *et al.,* 2020).

This allows for a significant speed up of the model, however, the risk of inappropriate feature extraction is indicated, which may lead to inaccurate classification.

The results obtained indicate that the use of YOLO-tiny did not achieve higher performance than YOLO3, or other models tested. The lack of accurate feature extraction, due to the reduced size of the model, is probably indicated as the reason for the results obtained. In contrast, the modified Faster R-CNN model obtained satisfactory results, allowing for higher performance in detecting the indicated emotions not only compared to the classic Faster R-CNN, but also to the model presenting the highest performance to date (Mittal *et al.,* 2020).

Importantly, the modified network no longer exhibits the biases seen in the classical solution. That is, it has increased both those scores that, in similar network arrangements, also came out quite high, but most importantly it has managed to obtain higher recognition performance for other emotions that previously had a much lower recognition rate by the Faster R-CNN network.

5. Discussion and Conclusion

This research examines the use of neural networks to recognize emotions in social robots as a means of assisting in the right interpretation of social situations and its participants. Two networks were chosen for this purpose, YOLO v3 and Faster R-CNN, both of which can be implemented in real time. EMOTIC images were used to train the architects, which allowed them to detect and assign 26 different emotions.

While no average greater than the EmotiCon network was obtained, the results are promising. Therefore, it was decided to modify both networks and undertake recalculations for their modified versions. While the YOLO-tiny network did not obtain satisfactory results, the modified Faster-R-CNN network allowed higher scores for the detected emotions. Nonetheless, this outcome is satisfactory for several emotion categories. Others will need to put in more time studying and training the network.

However, the gathered data can provide a crucial component for a social robot, which might combine the results of picture and text analysis and make judgements about the emotions displayed in a given social context based on the two aspects.

References:

Barchard, K.A., Lapping-Carr, L., Westfall, R.S., Fink-Armold, A., Banisetty, S.B., Feil-Seifer, D. 2020. Measuring the perceived social intelligence of robots, ACM Transac- tions on Human-Robot Interaction (THRI), 9(4), 1-29.

Bendjoudi, I., Vanderhaegen, F., Hamad, D., Dornaika, F. 2020. Multi-label, multi-task cnn approach for context-based emotion recognition. Information Fusion.

Broekens, J., Heerink, M., Rosendal, H. 2009. Assistive social robots in elderly care: a review. Gerontechnology, 8(2), 94-103.

Cabibihan, J.J., Javed, H., Ang, M., Aljunied, S.M. 2013. Why robots? a survey on the roles and benefits of social robots in the therapy of children with autism. International journal of social robotics, 5(4), 593-618.

Cao, H.L., Van de Perre, G., Kennedy, J., Senft, E., Esteban, P.G., De Beir, A., Simut, R., Belpaeme, T., Lefeber, D., Vanderborght, B. 2018. A personalized and platform-independent behavior control system for social robots in therapy: development and applications. IEEE Transactions on Cognitive and Developmental Systems, 11(3), 334-346.

Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L. 2018. Domain adaptive faster r-cnn for object detection in the wild. Proceedings of the IEEE conference on computer vision and pattern recognition, 3339-3348.

Chen, Z., Whitney, D. 2021. Inferential affective tracking reveals the remarkable speed of context-based emotion perception. Cognition 208, 104549.

Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., Pal, C. 2015. Recurrent neural networks for emotion recognition in video. Proceedings of the 2015 ACM on international conference on multimodal interaction, 467-474.

Esau, N., Wetzel, E., Kleinjohann, L., Kleinjohann, B. 2007. Real-time facial expression recognition using a fuzzy emotion model. 2007 IEEE international fuzzy systems conference, IEEE, pp. 1-6.

Fan, Q., Brown, L., Smith, J. 2016. A closer look at faster r-cnn for vehicle detection. 2016 IEEE intelligent vehicles symposium (IV), IEEE, pp. 124-129.

Gao, C., Zhou, J., Wong, W.K., Gao, T. 2018. Woven fabric defect detection based on convolutional neural network for binary classification. International Conference on Artificial Intelligence on Textile and Apparel, Springer, pp. 307-313.

Heerink, M., Vanderborght, B., Broekens, J., Albo´-Canals, J. 2016. New friends: social robots in therapy and education.

Hirth, J., Schmitz, N., Berns, K. 2011. Towards social robots: Designing an emotion-based architecture. International Journal of Social Robotics, 3(3), 273-290.

Ho, M.J., Lin, Y.C., Hsu, H.C. and Sun, T.Y. 2019. An efficient recognition method for watermelon using faster r-cnn with post-processing. 2019 8th International Conference on Innovation, Communication and Engineering (ICICE), IEEE, 86-89.

Xu, X., Xiao, Y., Chen, H., He, S., Qin, J., Heng, P.A. 2018. Sinet: A scale-insensitive convolutional neural network for fast vehicle detection. IEEE transactions on intelligent transportation systems, 20(3), 1010-1019.

Jain, D.K., Shamsolmoali, P., Sehdev, P. 2019. Extended deep neural network for facial emotion recognition. Pattern Recognition Letters, 120, 69-74.

Jiang, H., Learned-Miller, E. 2017. Face detection with the faster r-cnn. 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition, (FG 2017), IEEE, pp. 650-657.

Ju, M., Luo, H., Wang, Z., Hui, B., Chang, Z. 2019. The application of improved yolo v3 in multi-scale target detection. Applied Sciences, 9(18), 3775.

Kellmeyer, P., Mueller, O., Feingold-Polak, R., Levy-Tzedek, S. 2018. Social robots in rehabilitation: A question of trust. Sci. Robot, 3(21).

Kim, H.S., Sherman, D.K., Taylor, S.E. 2008. Culture and social support. American psychologist, 63(6), 518.

Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A. 2019. Context based emotion recognition using emotic dataset. IEEE transactions on pattern analysis and machine intelligence, 42(11), 2755-2766.

Koziarski, M., Cyganek, B. 2018. Impact of low resolution on image recognition with deep neural networks: An experimental study. International Journal of Applied Mathematics and Computer Science, 28(4).

Lee, K.H., Park, J., Yoo, S.M., Yoon, S.J., Cho, C.S. 2020. An implementation of nnef-darknet neural network model converter. 2020 International Conference on Information and Communication Technology Convergence, IEEE, 1186-1188.

Leo, M., Del Coco, M., Carcagni, P., Distante, C., Bernava, M., Pioggia, G., Palestra, G. 2015. Automatic emotion recognition in robot-children interaction for ASD

treatment. Proceedings of the IEEE International Conference on Computer Vision Workshops, 145-153.

Lestari, D.P., Kosasih, R., Handhika, T., Sari, I., Fahrurozi, A. 2019. Fire hotspots detection system on cctv videos using you only look once (yolo) method and tiny yolo model for high buildings evacuation. 2019 2nd International Conference of Computer and Informatics Engineering (IC2IE), IEEE, 87-92.

Levi, G., Hassner, T. 2015. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. Proceedings of the 2015 ACM on international conference on multimodal interaction, pp. 503-510.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dolla´r, P., Zitnick, C.L. 2014. Microsoft coco: Common objects in context, European conference on computer vision, Springer, pp. 740-755.

Mao, Q.C., Sun, H.M., Liu, Y.B., Jia, R.S. 2019. Mini-yolov3: real-time object detector for embedded applications. IEEE Access 7, 133529-133538.

Mileounis, A., Cuijpers, R.H., Barakova, E.I. 2015. Creating robots with personality: The effect of personality on social intelligence. International Work-Conference on the Interplay Between Natural and Artificial Computation, Springer, pp. 119-132.

Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D. 2020. Emoticon: Context-aware multimodal emotion recognition using frege's principle. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14234-14243.

Nguyen, H. 2019. Improving faster r-cnn framework for fast vehicle detection. Mathematical Problems in Engineering.

Paiva, A., Leite, I., Ribeiro, T. 2014. Emotion modeling for social robots. The Oxford handbook of affective computing, pp. 296-308.

Perez-Gaspar, L.A., Caballero-Morales, S.O., Trujillo-Romero, F. 2016. Multimodal emotion recognition with evolutionary computation for human-robot interaction. Expert Systems with Applications, 66, 42-61.

Qureshi, A.H., Nakamura, Y., Yoshikawa, Y., Ishiguro, H. 2016. Robot gains social intelligence through multimodal deep reinforcement learning. 2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids), IEEE, 745-751.

Redmon, J., Farhadi, A. 2018. Yolov3: An incremental improvement, arXiv preprint arXiv, 1804.02767.

Ren, S., He, K., Girshick, R., Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, arXiv preprint arXiv, 1506.01497.

Rybka, J., Janicki, A. 2013. Comparison of speaker dependent and speaker independent emotion recognition. International Journal of Applied Mathematics and Computer Science, 23(4).

Thuseethan, S., Rajasegarar, S., Yearwood, J. 2020. Complex emotion profiling: An incremental active learning based approach with sparse annotations, IEEE Access, 8, 147711-147727.

Tsiourti, C., Weiss, A., Wac, K., Vincze, M. 2019. Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots. International Journal of Social Robotics, 11(4), 555-573.

Wiem, M.B.H., Lachiri, Z. 2017. Emotion classification in arousal valence model using mahnob-hci database. International Journal of Advanced Computer Science and Applications, 8(3).

Zhang, M., Liang, Y., Ma, H. 2019. Context-aware affective graph reasoning for emotion recognition. 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp. 151-156.

Zhang, S., Cao, J., Zhang, Q., Zhang, Q., Zhang, Y., Wang, Y. 2020. An FPGA-based reconfigurable cnn accelerator for yolo. 2020 IEEE 3rd International Conference on Electronics Technology (ICET), IEEE, 74-78.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A. 2017. Scene parsing through ade20k dataset. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 633-641.