Computer Systems Science & Engineering
DOI:10.32604/csse.2021.015372
Article

Tech Science Press

Leveraging Graph Cut's Energy Function for Context Aware Facial Recognition in Indoor Environments

Kazeem Oyebode[1], Shengzhi Du[2,*] and Barend Jacobus van Wyk[3]

[1]Pan-Atlantic University, Lagos, Nigeria
[2]Department of Electrical Engineering, Tshwane University of Technology, Pretoria, South Africa
[3]Faculty of Engineering, the Built Environment and Technology Nelson Mandela University, South Africa
[*]Corresponding Author: Shengzhi Du. Email: DuS@tut.ac.za
Received: 17 November 2020; Accepted: 14 February 2021

Abstract: Context-aware facial recognition regards the recognition of faces in association with their respective environments. This concept is useful for the domestic robot which interacts with humans when performing specific functions in indoor environments. Deep learning models have been relevant in solving facial and place recognition challenges; however, they require the procurement of training images for optimal performance. Pre-trained models have also been offered to reduce training time significantly. Regardless, for classification tasks, custom data must be acquired to ensure that learning models are developed from other pre-trained models. This paper proposes a place recognition model that is inspired by the graph cut energy function, which is specifically designed for image segmentation. Common objects in the considered environment are identified and thereafter they are passed over to a graph cut inspired model for indoor environment classification. Additionally, faces in the considered environment are extracted and recognised. Finally, the developed model can recognise a face together with its environment. The strength of the proposed model lies in its ability to classify indoor environments without the usual training process(es). This approach differs from what is obtained in traditional deep learning models. The classification capability of the developed model was compared to state-of-the-art models and exhibited promising outcomes.

Keywords: Place recognition; face recognition; deep learning; graph cut

1 Introduction

Facial recognition (FR) systems have been used to resolve a host of challenges ranging from theft [1], development of access control systems [2], online examination management [3], gender recognition [4], facial expression recognition [5,6]; and age estimation [7]. However, there is a growing interest in context-aware FR systems which offer the possibility of recognising faces as well as their immediate environments. With this approach to FR, a domestic robot, for example, can carry out security tasks or indoor services having detected and processed relevant information. To achieve this system, a host of models, including the Principal Component Analysis (PCA) [8–10], have been deployed. More recently,

the deep learning (DL) model has also been used [11]. The latter model works by gathering many training datasets. The training approach fine-tunes the model such that an optimal classification is attained. This method has a downside: when acquiring new faces, the model must be retrained. Recent research has mitigated this challenge by developing an FR model that is trained on millions of faces [12]. For the classification task, vectors of two face images are passed into the model and compared for a match. A threshold value T informs the model's decision; a value greater than T suggests that a match does not exist, while a value less than or equal to T suggests that a match does exist. The model under review has an advantage as it does not require retraining on new faces.

Similarly, there has been growing research interest in place recognition, especially in localising domestic robots. The DL model plays a prominent role in this regard, although traditional models have also been useful. Within these traditional models, global and salient features of indoor environments have been used for indoor classification [13].

Unfortunately, one of the disadvantages of DL models is that they need a considerable number of images to sufficiently learn unique features which distinguish one indoor environment from another. Many studies have, therefore, investigated the use of pre-trained models [14] as they are trained on thousands of images such that they need only two or three custom convolution layers to adapt the model to a given classification problem. These layers are then trained on custom datasets. In Liu et al. [15], a Convolution Neural Network (CNN) feature extraction model is built within a pre-trained model. This model's output is a feature vector of considered environments. Therefore, for the test image, feature vectors of the considered environments are generated and compared; the best match becomes the recognised image. The downside of this model is that more custom data needs to be gathered to train the weights of the added layers [15]. For instance, when indoor environments are unique, acquiring datasets for the environments become a challenge. Fig. 1, for example, depicts a combination of a living room and a dining room. A training dataset for such a scenario is rare because popular place recognition datasets [16] have not considered this uniqueness. This complicates its deployment in real-life scenarios.

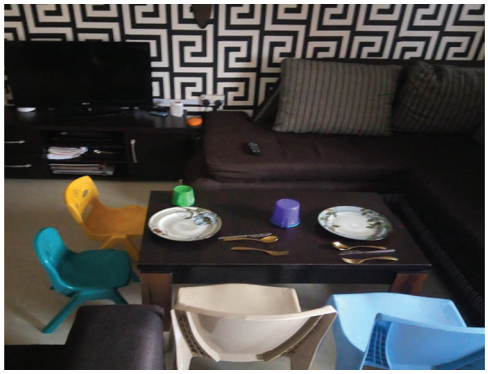

Figure 1: Indoor environment – living room merged with dining table

Fig. 2 depicts how the suggested model works. The names of everyday objects and concepts are passed on to the model through the image, whereafter it makes a prediction regarding the environment. Furthermore, the face in the input image is extracted and recognised using the model proffered in Parkhiet al. [12]. Ultimately, a face together with the immediate environment is recognised.

For an indoor environment, this might be useful as a lightweight domestic robot can perform functions such as cooperating with identified and recognised faces in a defined environment. In the case where a face does not match a given (environment), an alarm or anomaly event can be raised. The contextual face

recognition model can also be extended to other systems where safety is prioritised. For instance, road accidents caused by unlicensed drivers are one of the leading causes of death in America [17]. The model can thus shut down a train/bus/car when an unauthorised operator is observed.

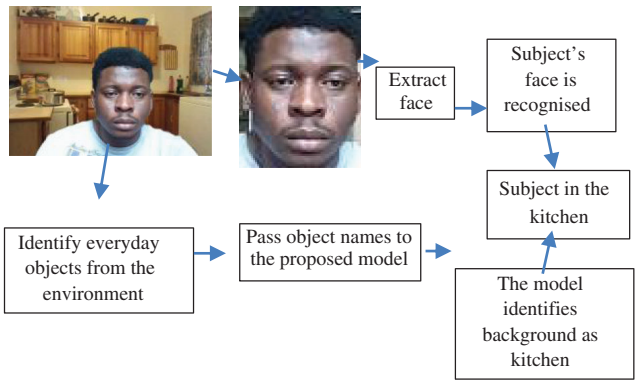

Figure 2: Proposed model

The model put forward in this work takes advantage of a free, DL model for object and concept detection [18] alongside a modified graph cut. It is highly customisable, which means that everyday objects in a given environment can be gathered and used for place recognition without the need for training.

To the best of the researcher's knowledge, no work has been carried out concerning localised FR in an indoor environment. In Davis et al. [19], an FR model is offered whose accuracy is enhanced by incorporating environment metadata. The major disadvantage of this model is that when faces move around, the environment metadata may be unhelpful. Another indoor place recognition design has been put forward using the Bayesian model [20]. The model works by assigning probabilities to everyday objects found in indoor environments. Thereafter, these probabilities are then passed over to a Bayesian model for indoor environment prediction. One downside of this model is that objects detected from a given scene rely on a probability value before place recognition is carried out; therefore, this study does not make use of the approach. Rather, a modified graph cut energy function is used for indoor environment recognition and an existing FR.

Section 2 of this work discusses everyday objects or concepts in the considered indoor environment and it introduces the graph cut energy function as well as the suggested model. The proffered model is then evaluated in Section 3, whilst Sections 4 and 5 advance the discussion and conclusion of the study.

2 Theory

For context-aware FR, the names of everyday objects and concepts for the considered environments must be gathered. This research focuses on indoor environments in homes. Hence, five indoor environments have been considered; namely the bedroom, dining room, kitchen, living room and bathroom. Commonly associated objects and concepts for these five indoor environments are listed below.

(1) Kitchen – microwave, kitchenware, stove, cabinet, counter, oven, toaster, faucet, sink, refrigerator, cupboard, stainless, refrigerator, knife, seat, chair, table, cutlery, shelf, stool, wash closet, curtain, window.

(2) Living room – seat, sofa, chair, rug, cushion, couch, pillow, fireplace, chandelier, cosy, flower arrangement, furniture, decoration, cabinet, armchair, table, television, luxury, refrigerator, curtain, window.

(3) Bedroom – bed, pillow, table, seat, blanket, lamp, mirror, television, furniture, rug, chandelier, luxury, curtain, window, hotel, bedroom, residential.

(4) Dining room – dining, table, lamp, seat, chair, banquet, tableware, vase, flower arrangement, chandelier, dining room, shelf, place setting, tablecloth, rug, knife, cutlery, curtain, window.

(5) Bathroom – toilet, bathtub, mirror, carpet, faucet, sink, shower, shelf, cabinet, chair, seat, wash closet, lavatory, furniture, rug, lavatory, table, luxury, lamp, bathroom.

It is imperative to mention that objects or concepts attached to these indoor environments are customisable. They can be adjusted to meet specific environmental peculiarity.

2.1 Graph Cut Energy Function

Graph cut is a segmentation method that partitions an image into *foreground* and *background*. In other words, it binarises images [21]. Before segmentation, an image is transformed into a graph G. Image pixels in G are transformed into vertices or nodes (V) while edges (E) connect pixel neighbours. Each node makes a connection with two fundamental nodes – O and B – as depicted in Fig. 3. The figure shows a 2 x 3 image which has been transformed into a graph. Here, *a, b, d, e, f, g* are image pixels that have been transformed into nodes. One can observe that the nodes have two distinct colours which depict the inherent capability of graph cut segmentation as a binary classifier.

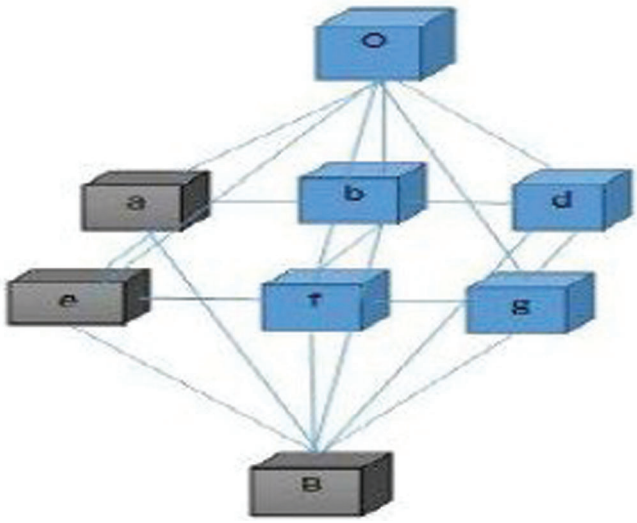

Figure 3: An image represented in the form of a graph

At the heart of this process is the graph cut energy function in Eq. (1), which is used to assign weights to edges of the graph in Fig. 3. For example, edges attached to O and B are given weight values by the graph cut data term (the first term, starting from the left in Eq. (1)). The data term determines the weight using the negative logarithm of the probability, given an observation F, assuming $I(i)$ is attached to nodes O and B. Weights between two neighbouring vertices are assigned by the second term. This term is referred to as the smoothness term. Assuming pixels or vertices $I(i)$and $J(i)$ are neighbours, σ is the pixel similarity variance [21]. The min-cut/max-flow algorithm is generally used [22] to optimally partition an image into foreground and background. The λ in Eq. (1) is a parameter that gives relative importance to the data term at the expense of the smoothness term.

$$E(F) = \lambda \sum_{i=1}^{n} -Log(Prob(I(i)|F)) + \sum_{i,j \in N}^{n} \exp\left(-\frac{|I(i) - I(j)|^2}{2\sigma^2}\right) \tag{1}$$

2.2 Modified Graph Cut Energy Function

One of the limitations of the graph cut segmentation model [23] is that it can only partition an image into foreground and background. This research, therefore, extends the capability of the graph cut energy function from a bi-classification to a multi-classification. Hence, the graph cut energy function is modified to accommodate multiple categorisations or classifications. As observed in Fig. 3, where objects *a, b, d, e, f, g* are pixel objects, they could also be transformed into objects detected in an indoor environment. However, instead of partitioning the objects into binary categories (foreground and background), these objects should be partitioned into five categories – kitchen, living room, dining room, bedroom and bathroom – based on elicited objects. Fig. 4 depicts an envisioned graph cut segmentation for indoor classification. The coloured boxes, in this case, are everyday objects detected from an indoor environment and they are segmented into the five identified environments.

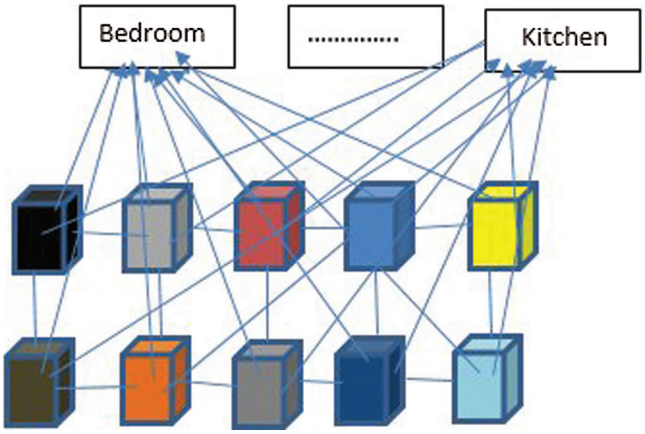

Figure 4: Graph cut seen as a multiclass classification

To achieve the multiple classification goal, Eq. (1) is modified, as shown in Eq. (2). Eqs. (1) and (2) have common attributes. Firstly, they both have data and smoothness terms. However, there is an adjustment in the data term. The modified energy function takes the probability of an object $I(i)$ belonging to a given domain $F_m\{m = 1, 2, \ldots, N\}$ among the considered domains ($N = 5$ for the five domains in the example used). In Eq. (2), n is the total number of objects elicited from a presented scene.

Furthermore, λ is applicable only when a given object is unique to the considered domain or environment. This development is also in line with the graph cut energy function offered in [21] where a particular pixel weight is given a high score, based on an interactive pixel selection. For the smoothness term, when an object $I(i)$ belongs to a domain in question, the value of $I(i)$ is 255. This approach gives the impression that there is cohesiveness; otherwise, $I(i)$ becomes 0. $E(F_m)$ gives the total number of points accumulated for the given domain F_m. In Eq. (2), 60 is assigned to λ while 0.71 is assigned to σ.

$$E(F_m) = \sum_{i=1}^{n} \lambda \, Prob(I(i)|F) + \sum_{i,j \in N}^{n} \exp\left(-\frac{|I(i) - 255|}{2\sigma^2}\right) \tag{2}$$

Therefore, for the considered domain, there would be $E(F_1)$, $E(F_2)$, …, $E(F_5)$. The energy with the highest value from Eq. (3) indicates the identified domain.

$$A = \text{argmax}(\{E(F_m)\}) \tag{3}$$

2.3 Facial Recognition

Ten frontal face images are acquired and then passed into the deep face recognition model [12]. Thereafter, the model generates ten unique vectors V_1, V_2, …, $V_{10}$for these images. It then compares these vectors to the face vector G generated by the model of a face extracted from the presented indoor environment. Eq. (4) is then used to recognise the face G. In this way, the face match with the smallest cosine value becomes the recognised face. A threshold value, T, of 0.4 is set. This means that if G is extracted from a scene and it does not match any stored face in the database, the system gives a *face not recognised* message.

$$B = \min(\{(C(V_1, G), C(V_2, G), \ldots\ldots C(V_{10}, G))\}) \tag{4}$$

Algorithm 1 details how the proposed model works.

Algorithm 1: Localised FR

1. **Inputs**: X and G (X encapsulates the objects elicited from an indoor scene $I(i) \ldots\ldots\ldots I(n)$, n is the total number of objects and I represents everyday objects or concepts. G is the face vector extracted from the indoor environment).
2. **Outputs***: A* and B (A is the recognised indoor environment and B is the recognised face in the indoor environment).
3. **Begin:**
4. **Initialise**
5. List< double > *As*
6. Vector< double > *Fc*
7. *domain* = {Kitchen, Bedroom, Living room, Bathroom, Dining room}
8. **for each** (F in *domain*)
9. pass X into Eq.(2) and return $E(F)$ for F
10. *As.add*($E(F)$)
11. **End for**
12. Use Eq. (3) to perform indoor environment classification on *As* and assign the result to A
13. **for each** (v in $V_{1.....10}$)
14. f = Consine distance (G,v)<= *0.4*
15. *Fc.add(f)*
16. **End for**
17. **If** *Fc is empty*
18. B = 'unknown face'
19. *Else*
20. B = *get the name attached to the face with the minimum value in Fc*
21. ***End for***
22. Return A and B

3 Evaluation

The developed model has been evaluated on 2543 images. These are images of five indoor environments [16], as previously discussed. The Accuracy metric is used to evaluate its performance, as seen in Eq. (5).

$$Accuracy\,(\mathrm{A}) = \frac{total\ number\ of\ correctly\ classified\ Images}{total\ number\ of\ Images\ referenced} \tag{5}$$

4 Results

Tab. 1 shows the performance of the proffered model. The model outperformed others in three categories out of a possible five. It did not perform adequately in the Dining room category because a notable similarity of the objects exists between a living room and dining room. However, when the indoor environment had somewhat unique objects, the developed model performed well (as seen in the Kitchen and Bathroom categories). The CNN feature classification model [15] outperforms the offered model in the Bathroom and Living room categories. The performance of this model may be attributed to the fact that the training and testing images may have been derived from the same source. Hence, there is a possibility that patterns of indoor environment images would have been learned from the training dataset. This approach is different from the proposed model, as training is not required.

Table 1: Indoor environment recognition accuracy

Domain	Model	Image no.	A (%)
Bedroom	Indoor recognition method [13]	662	27.6
	Place recognition model [20]	662	72.4
	Place recognition model [15]	662	**93**
	Proposed model	662	88.5
Bathroom	Indoor recognition method [13]	197	47
	Place recognition model [20]	197	87.3
	Proposed model		**92.4**
Dining room	Indoor recognition method [13]	274	**-**
	Place recognition model [20]	274	41.6
	Proposed model		**70**
Kitchen	Indoor recognition method [13]	734	-
	Place recognition model [20]	734	55
	Place recognition model [15]	734	75
	Proposed model	734	**82.2**
Living room	Indoor recognition method [13]	706	-
	Place recognition model [20]	706	89
	Place recognition model [15]	706	**95**
	Proposed model	706	76.6

It is important to note that the performance of the proffered model hinges on an object detection algorithm. It is observed that the adopted object detection model may not be robust in detecting everyday objects in the images presented. For example, consider the image in Fig. 5. The everyday objects or concepts that were detected are (5 in number, meaning $n = 5$): a window, table, furniture, chair and counter. However, the object detection model omitted objects such as a pot and cooker. Regardless, based on other identified objects, the proposed model can determine that the presented image is a kitchen. The data and smoothness terms for the considered indoor environment were determined as seen below.

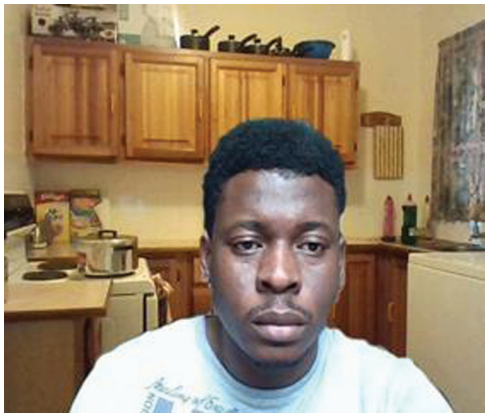

Figure 5: Proposed model can recognise the face together with its environment – kitchen

Data term for Kitchen:

$\sum_{i=1}^{5} \lambda \text{Prob}(I(i)|F)$ = Prob(window|kitchen) + Prob(chair|kitchen) + Prob(table|kitchen) + (λ*Prob(counter|kitchen)) + Prob(furniture|kitchen) + = 4/5+4/5 +5/5 (60*(1/5)) + 0 = 14.6

Smoothness term for Kitchen:

$$\sum_{i,j \in N}^{n} \exp\left(-\frac{|255 - I(i)|}{2 * \sigma^2}\right) = \exp\left(-\frac{|255 - 255|}{2 * 0.71^2}\right) + \exp\left(-\frac{|255 - 255|}{2 * 0.71^2}\right) + \exp\left(-\frac{|255 - 255|}{2 * 0.71^2}\right) + \exp\left(-\frac{|255 - 255|}{2 * 0.71^2}\right) + \exp\left(-\frac{|255 - 0|}{2 * 0.71^2}\right) = 4$$

$$E(\text{kitchen}) = 14.6 + 4 = 18.6$$

Values for others are given as E(Diningroom)- 5.6; E(Bedroom)- 5.4; E(Kitchen)- 18.6; E(Bathroom) - 5.4; E(Livingroom)- 7.2.

Drawing from the calculation, the domain with the highest value was the kitchen with 18.6. Hence, the model classifies the input image as a kitchen. Furthermore, the face in Fig. 5 is extracted and sent to the face recognition model. The model then runs the extracted face G on a collection of registered faces; this returns the least value of 0.308 and a corresponding name. A value of more than 0.4 suggests that a foreign face is identified.

5 Discussion

In this research, the binary graph cut was modified to a multi-classifier as required. A localised FR was developed, which takes advantage of the multi-classifier and an existing FR model. An advantage of the proposed model over others is that it can be adjusted to work in any indoor environment. The simple

process involves gathering everyday objects and concepts particular to the considered environment. This development contrasts the conventional DL model or pre-trained model for place recognition, as they require training.

6 Conclusion

This paper developed a modified graph cut energy function for place classification. The model was combined with facial recognition to deliver a localised FR system. An experiment carried out on 2543 images showed encouraging prospects.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] P. Wang, W. H. Lin, K. Chao and C. C. Lo, "A face-recognition approach using deep reinforcement learning approach for user authentication," in *IEEE 14th Int. Conf. on e-Business Engineering*, Shanghai, China, pp. 183–188, 2017.

[2] A. Derbel, D. Vivet and B. Emile, "Access control based on gait analysis and face recognition," *Electronics Letters*, vol. 51, no. 10, pp. 751–752, 2015.

[3] P. Mahes and K. Selvajyothi, "Impersonation detection in online examinations," in *IEEE Int. Conf. on Signal Processing, Informatics, Communication and Energy Systems*, Kollam, India, pp. 1–5, 2017.

[4] G. Azzopardi, A. Greco, A. Saggese and M. Vento, "Fast gender recognition in videos using a novel descriptor based on the gradient magnitudes of facial landmarks," in *14th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance)*, Lecce, Italy, pp. 1–6, 2017.

[5] Y. Ding, Q. Zhao, B. Li and X. Yuan, "Facial expression recognition from image sequence based on LBP and taylor expansion," *IEEE Access*, vol. 5, pp. 19409–19419, 2017.

[6] J. H. Mosquera, H. Loaiza, S. E. Nope and A. D. Restrepo, "Identifying facial gestures to emulate a mouse: navigation application on Facebook," *IEEE Latin America Transactions*, vol. 15, no. 1, pp. 121–128, 2017.

[7] E. Eidinger, R. Enbar and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[8] R. Kaur and E. Himanshi, "Face recognition using principal component analysis," in *IEEE Int. Advance Computing Conf. (IACC)*, Banglore, India, pp. 585–589, 2015.

[9] M. P. R. Kumar, S. R. Keerthi and K. M. Aishwarya, "Artificial neural networks for face recognition using PCA and BPNN," in *TENCON IEEE Region 10 Conf.*, Macao, China, pp. 1–6, 2015.

[10] W. Deng, J. Hu, J. Lu and J. Guo, "Transform-invariant PCA: A unified approach to fully automatic face alignment, representation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1275–1284, 2014.

[11] Y. Liu and J. Yang, "Face recognition method based on convolutional neural network," in *Int. Conf. in Communications, Signal Processing and Systems, Singapore*, pp. 1925–1929, 2019.

[12] O. M. Parkhi, A. Vedaldi and A. Zisserman, "Deep face recognition," in *British Machine Vision Conf., Britain*, pp. 1–12, 2015.

[13] J. Niu, X. Bu, K. Qian and Z. Li, "An indoor scene recognition method combining global and saliency region features," *Jiqiren/Robot*, vol. 37, no. 1, pp. 122–128, 2015.

[14] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba and A. Oliva, "Places: An image database for deep scene understanding," *Journal of Vision*, vol. 17, no. 10, pp. 1–12, 2016.

[15] S. Liu and G. Tian, "An indoor scene classification method for service robot based on CNN feature," *Journal of Robots*, vol. 2019, no. 1, pp. 1–12, 2019.

[16] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Miami, USA, pp. 413–420, 2009.

[17] R. Subramanian, *Motor Vehicle Traffic Crashes as a Leading Cause of Death in the United States*. USA, 2007. [Online]. Available: https://trid.trb.org/view/809424.

[18] Clarifai, *Computer Vision and Artificial Intelligence for All*. New York, USA, 2013. [Online]. Available: https://www.clarifai.com/.

[19] M. Davis, M. Smith, J. Canny, N. Good, S. King *et al.,* "Towards context-aware face recognition," in *13th Annual ACM Int. Conf. on Multimedia*, Singapore, pp. 483–486, 2005.

[20] K. Oyebode, S. Du, B. J. Van Wyk and K. Djouani, "A sample-free bayesian-like model for indoor environment recognition," *IEEE Access*, vol. 7, pp. 79783– 79790, 2019.

[21] Y. Boykov and J. Marie-Pierre, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," in *Int. Conf. on Computer Vision*, Vancouver, Canada, pp. 105–112, 2001.

[22] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.

[23] R. Carsten, K. Vladimir and B. Andrew, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 309–314, 2004.