2020
THE JOURNAL ON TECHNOLOGY AND PERSONS WITH DISABILITIES

An Assistive Solution to Assess Incoming Threats for Homes

Shahinur Alam, Md Sultan Mahmud, Mohammed Yeasin
The University of Memphis, TN, USA
salam@memphis.edu; mmahmud@memphis.edu; myeasin@memphis.edu

Abstract

An assistive solution to assess incoming threats (e.g., robbery, burglary, gun violence) for homes will enhance the safety of the people with or without disabilities. This paper presents 'SafeNet'- an integrated assistive system to generate context-oriented image descriptions to assess incoming threats. The key functionality of the system includes the detection and identification of human and generating image descriptions from the real-time video streams obtained from the cameras placed in strategic locations around the house. In this paper, we focus on developing a robust model called "SafeNet" to generate image descriptions. To interact with the system, we implemented a dialog enabled interface for creating a personalized profile from face images or videos of friends/families. To improve computational efficiency, we apply change detection to filter out frames that do not have any activity and use Faster-RCNN to detect the human presence and extract faces using Multitask Cascaded Convolutional Networks (MTCNN). Subsequently, we apply LBP/FaceNet to identify a person. SafeNet sends image descriptions to the users with an MMS containing a person's name if any match found or as "Unknown", scene image, facial description, and contextual information. SafeNet identifies friends/families/caregiver versus intruders/unknown with an average F-score 0.97 and generates image descriptions from 10 classes with an average F-measure 0.97.

Keywords

Assistive Solution, Home Safety, Independent living

Introduction

Assessing incoming threats for homes is challenging for the people with or without disability since it requires continuous monitoring by a human observer. Sulman and colleagues (Sulman, et al, 2008) found in a study that when the number of monitoring displays increases, human performance deteriorates. They reported that a human observer missed 20% of the event while monitoring four surveillance display. However, when they increased the number of the display window to nine, missing rates rose to 60%. The recent years have seen an upsurge of interest in developing automated systems for monitoring homes that would eliminate the necessity of human observers. The available automated commercial security solutions such as ADT (ADT, 2019), Vivint (Vivint, 2019), SimpliSafe (Simplisafe, 2019), Frontpoint (FrontPointSecurity, 2019), Honeywell (Honeywell, 2019) etc. cannot detect unusual activities until an event occurs and are not designed for the people with disability. Designing assistive solutions to assess incoming threats requires consolidated knowledge about image understanding, natural language processing, and system development. In the past, a plethora of research reported on navigation (Gude et al, 2013), expression detection (Anam et al, 2014; Asm et al, 2014), ambient awareness (Ahmed et al, 2016), currency recognition (Looktel, 2009), object recognition (Alam et al, 2015; Kao and others, 1996; Mapelli et al, 1997; Chincha et al, 2011; Bigham et al, 2010) to assist people with disabilities. However, developing an automated system to assess incoming threats to homes did not receive considerable attention from the researchers. Although, the recent advancement in Machine Learning and Computer Vision, especially Convolutional Neural Network (CNN) has made the object detection (Krizhevsky et al, 2012; Girshick et al, 2014; Girshick et al, 2015; Sermanet et al, 2013; He et al, 2017) person recognition (Sun et al, 2015; Nezami et al, 2018) and image captioning (Yagcioglu et al , 2015;

Xu et al, 2015; Ushiku et al, 2015; Vedantam et al, 2015; Verma et al, 2014; Vinyals et al, 2015) task robust and efficient compared to the last decade. However, those technologies have not been used widely to build assistive solutions, especially for assessing incoming threats. To fill the void, we have developed an assistive solution to enhance the safety of people with disabilities (i.e., visually impaired, limited mobility). Our main contributions are: -1) Building a new and robust model called "SafeNet" to generate image descriptions from home monitoring cameras. 2) Collecting and processing training samples to train and evaluate SafeNet 3) Designing a recurrent neural network-based language model to generate semantically meaningful messages. Besides, we have addressed challenges related to accessibility and usability, system design, development, and integration.

System Overview

In order to use the SafeNet system, the house needs to be equipped with cameras covering the critical points such as front door, back door, driveway, off-street, etc. We assumed that people with disabilities would receive help from sighted people in installing cameras. A raspberry pi connected to the home monitoring cameras captures and sends video frames to the image descriptions generator (see "Image Description generator"). The data transmission between cameras and raspberry pi can be done using intra-home Wi-Fi or wired connection based on coverage area and distance. The image descriptions generator identifies persons from the video frames by matching faces with a personal profile and generates a short description. According to the guideline from UC Berkley police department (UCPD, 2019), a suspect can be described by information about the weapon, shirt, pants, color, eyeglasses, hair, facial hair, etc. To limit the scope of work, we have included identified person's name if any match found, location around house, whether he/she has a gun at hand or talking over the phone or wearing

masks and information about facial features such as beard, mustache, eyeglasses, bald head, in the image descriptions. Here are two sample image descriptions: 1) “John at the entrance talking over the phone”; 2) “An unknown person with a gun who has beard, mustache, hair, and no-eyeglass at the back door.” The system sends the image descriptions and a scene image via Multimedia Messaging (MMS) to the listed users. People with visual impairment can use the smartphone's screen reader to read out notifications. However, to make it more convenient for users’ system makes a phone call and reads out the generated image descriptions. Besides, the system records the videos and image descriptions in the persistent storage based on user preferences and allows them to query summarized history. The end-to-end process is shown in figure 1.

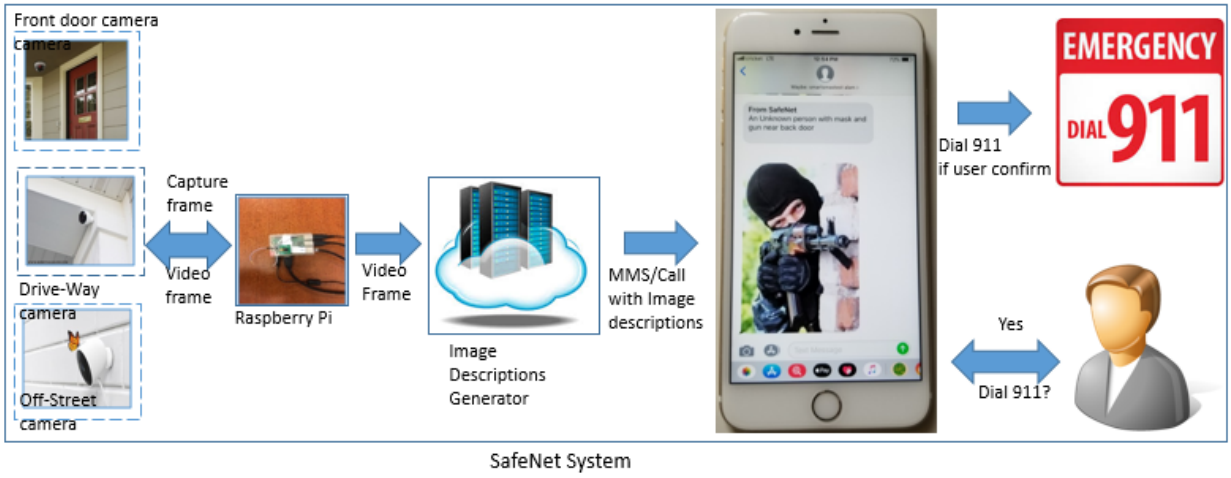

Fig. 1. System architecture and process flow.

System Design

Designing an assistive system is very challenging because it requires a comprehensive understanding of user's needs depending on the severity and type of disability, their technical adaptability, cognitive ability, and financial affordability. To design an effective system, we included a visually impaired individual who lost his vision after surgery in the development loop.

The functional requirements of the system were collected through participatory design with a group of sighted and visually impaired individuals. Then, we refined and validated collected requirements by conducting a survey via Amazon Mechanical Turk (Mechanical Turk, 2019) with a set of questions (see survey questions) where 30 people with disabilities participated. To address accessibility and usability issues we incorporated design thinking concept and followed "Adding Accessibility Features to Apps" (Google, 2017) technical guidelines to design a user interface. SafeNet system has been compartmentalized into three modules: - 1. Personal Profile Module, 2. Image Description Generator Module, 3. Feedbacks Module.

Personal Profile: To identify a person, first, the recognition model needs to be trained with face images of friends/families. We have developed a smartphone app with four utility options (Add Person, Add Views, Delete Person, and Readout Summary) to enlist a person to the profile with voice-over interaction. The personal profile contains demographic information (Name, Email, Contact) and face images of friends/families with various expression (Joy, Sad, Surprise, Fear, Contempt, Disgust), orientation and poses so that system can recognize them robustly from different view angle, position, and distances. The app allows the user to collect face images from the photo gallery/video clip/camera preview. In camera preview mode, when a subject/friend/family stands in front of the camera, the system automatically detects face and read out the position to make sure the collected images do not have a cropped face. To capture face images from different view system guides user to rotate smartphone around the face from left to right or right to left. The captured pictures become blurry when the rotational speed is high. In order to prevent it, we provide a feedback "too fast" when the rotational speed exceeds 20 degrees per second. The collected images are sent to a **Deep Webservice** which is responsible for the training/re-training recognition model, versioning trained model, and data. We have

developed this REST (RESTWeb, 2019) Deep Webservice to handle query requests robustly and seamlessly.

Image Description Generator: The system generates image descriptions based on the detection and recognition outcomes. The process flow of generating image descriptions is shown in figure 2. First, we use Faster-RCNN (Ren et al, 2015) to detect the human presence and extract faces using Multitask Cascaded Convolutional Networks (MTCNN) (Zhang et al, 2016). The reason for including a person detection model (Faster R-CNN) is to notify users about human presence even if there is no face found, especially when the front view of a person is not visible to the cameras. Second, we use LBP (Ahonen et al, 2004)/FaceNet (Schroff et al, 2015) to identify a person and groups by matching extracted faces with the profile (see Personal Profile). Third, face parts are extracted from detected faces using algorithm 1 (see figure 7). Fourth, a facial and contextual description is generated by classifying individual face parts and images using SafeNet (see "Model Development"). The color of the hair has been determined by calculating the intensity histogram from the cropped head patch. The location of the detected person is obtained based on the source camera of the video frames. Finally, a semantically and syntactically meaningful image description is generated from obtained words by applying rule-based grammar to create initial sequences first and then refined with Long-Short Term Memory (LSTM) (LSTM, 2019) language model trained with possible sequences.

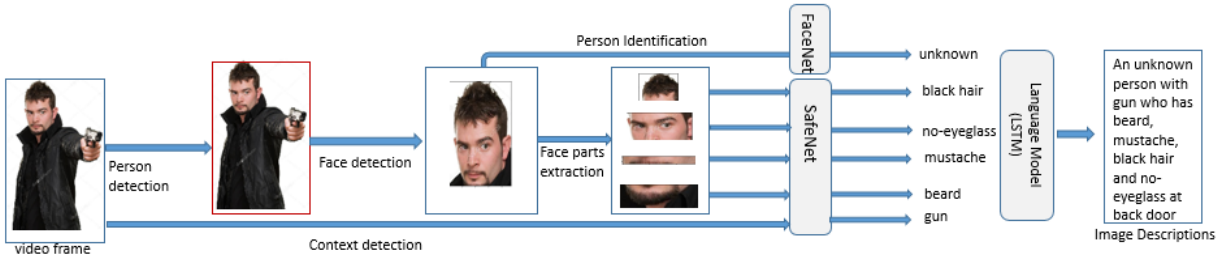

Fig. 2. Workflow for generating image descriptions.

Feedback Module: The primary task of the feedback module is sending notifications to users with image descriptions. Although smartphones are easily accessible nowadays, lots of people with a disability do not know how to utilize accessibility features such as TalkBack, Siri, etc. properly. Hence, designing an effective feedback system for people with disabilities is very challenging. Considering the technical adaptability of the users, we have included four types of feedback modes such as MMS, alert message, email, and phone call. The feedback mode can be set based on user choice. We have developed a communication API using SMTP (Simple Mail Transfer Protocol) server to send feedback messages to the users via their phone operator. To make a phone call, we are using Twilio (Twilio, 2019) 3rd party service.

Model development

We have built a Convolutional Neural Network (CNN) based robust model called "SafeNet" to generate image descriptions. SafeNet is built with one input layer of dimension 320x256x3, 14 convolution layers, seven dense layers, and 5 MaxPooling layers. The output of each activation is normalized by a batch normalization layers. BatchNormalization (Ioffe et al, 2015) helps to prevent covariance shift and model overfitting. Since pooling layers reduce network dimensions very cheaply by discarding lots of spatial information, we used only five MaxPooling layers, which help to reduce the number of parameters and required computational resources. The network architecture is shown in figure 3, and the loss curve in figure 4. We have come up with this architecture considering some key factors such as the size of the training dataset, overfitting vs. an underfitting problem with network depth and complexity, co-adaptation, feature learning, and predictive ability of individual neurons. The standard network such as VGG16 (Simonyan et al , 2014), ResNet50 (He et al, 2016), MobileNet (Howard et al, 2017), etc. do not perform well with/without transfer learning (see Quantitative Evaluations) for this dataset because simple

networks do not learn all distinguishing features and very complex network suffers from over-fitting problem.

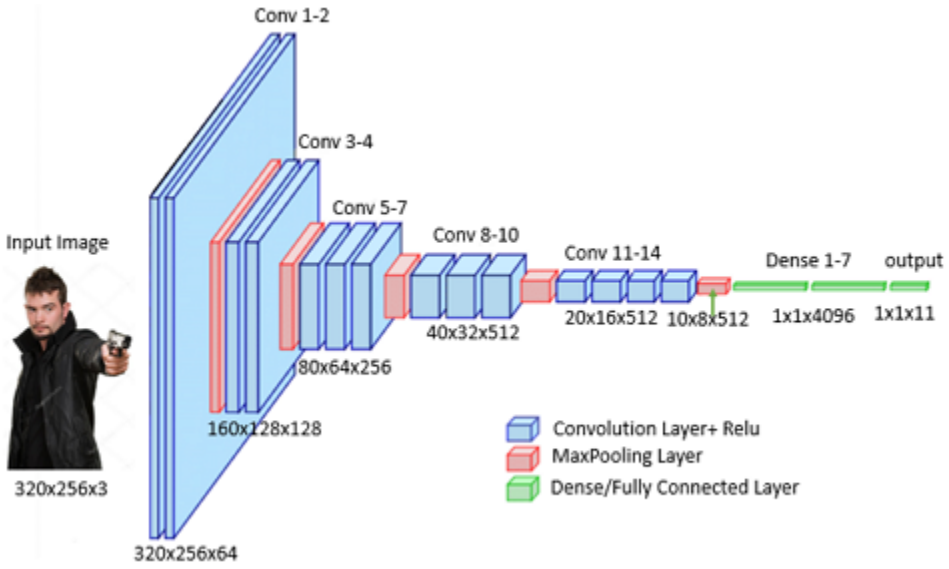

Fig. 3. SafeNet network architecture.

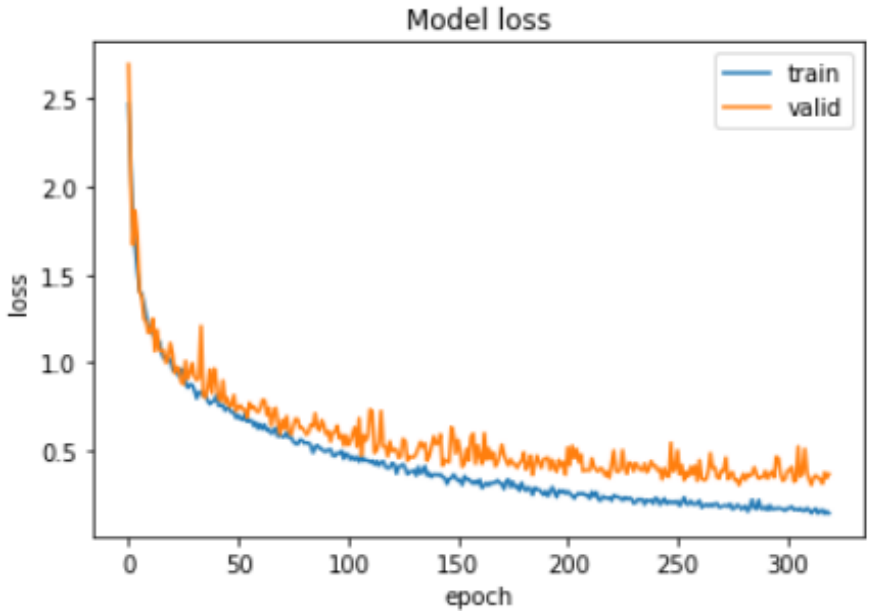

Fig. 4. Loss curve.

SafeNet has been trained for 320 epochs with a batch size of 24 using Stochastic gradient descent (SGD) optimizer. Finding optimal values for the key parameters (learning rate,

momentum) of the SGD optimizer manually/iteratively is exhaustive. In addition, a low learning rate may cause slow convergence, while a high learning rate may miss the global optimum. To address this issue, Bayesian Optimization [10], a method to search hyperparameters for finding the optimal value of an unknown function has been used. The optimal values obtained from Bayesian optimizer are 0.00101 and 0.7605 respectively from a wide search space of learning rate [0.00001, 0.1] and momentum [0.5, 0.9]. The SafeNet network learned total 285,634,121 parameters. There was some criticism in the past regarding the feature learning and their interpretability in the neural network. It is indispensable to have a clear understanding of why any model performs so well, what kind of features that model learned, which part of an image played a significant role in the final classification, whether two independent neurons learned different features and all neurons have the predictive capacity, etc. In order to understand and explain those scenarios, we have visualized filters, weights, and activity of the network layer by layer (see figure 5.) in input pixels space. We can see from figure 5 that the bottom-layers learned edges, blobs, and textures, while upper-level layers learned higher-level abstract. All layers learned distinct features and contributed to the final classification.

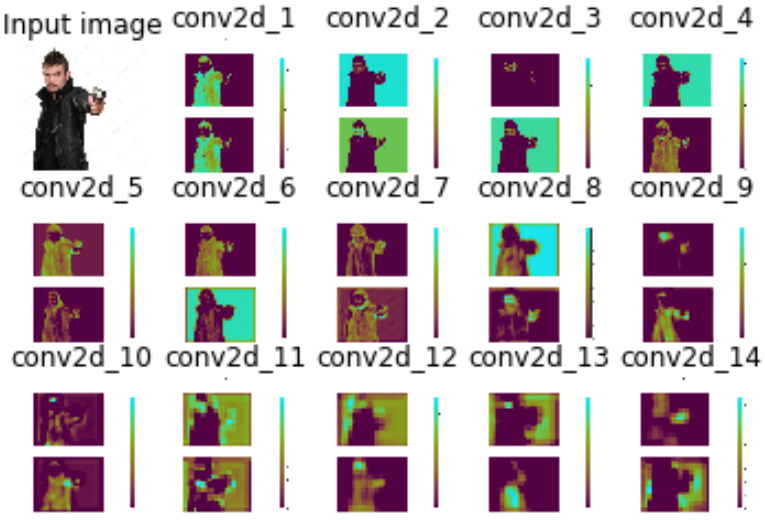

Fig. 5. Two activation map/filters layer by layer.

Data collection and Model Training

We have included ten classes (cellphone, gun, eyeglass, mask, beard, nobeard, mustache, nomustache, baldhead, hair) to generate image descriptions and collected 8128 image samples from ImageNet (Deng et al, 2009), RGB-D (Lai et al, 2011) and web. Then the collected images are sorted out based on the availability of faces with/without a beard, eyeglasses, mustache, and hair. We have developed a simple and computationally efficient algorithm (see figure 7) to crop different parts of a face from collected images by finding facial landmarks (Kazemi et al, 2014) (see figure 6.a) and grouped similar parts together (see two sample groups in figure 6.b). Then, three-domain experts examined every single face-part and filtered out parts that are too small (less than 20x20). To make the model affine (rotation, translation, shear, scale) invariant within a certain range, we have applied data augmentations so that it can recognize face parts robustly with various orientation and head poses. Data augmentation is a very useful technique to significantly increase the diversity of the data by padding, flipping, rotating, scaling, etc. We have generated a total of 50112 samples for training and validation. Person detection and recognition model has been trained with PASCAL-VOC2012 (Everingham et al, 2015) dataset and a profile with 180 images captured from 16 people.

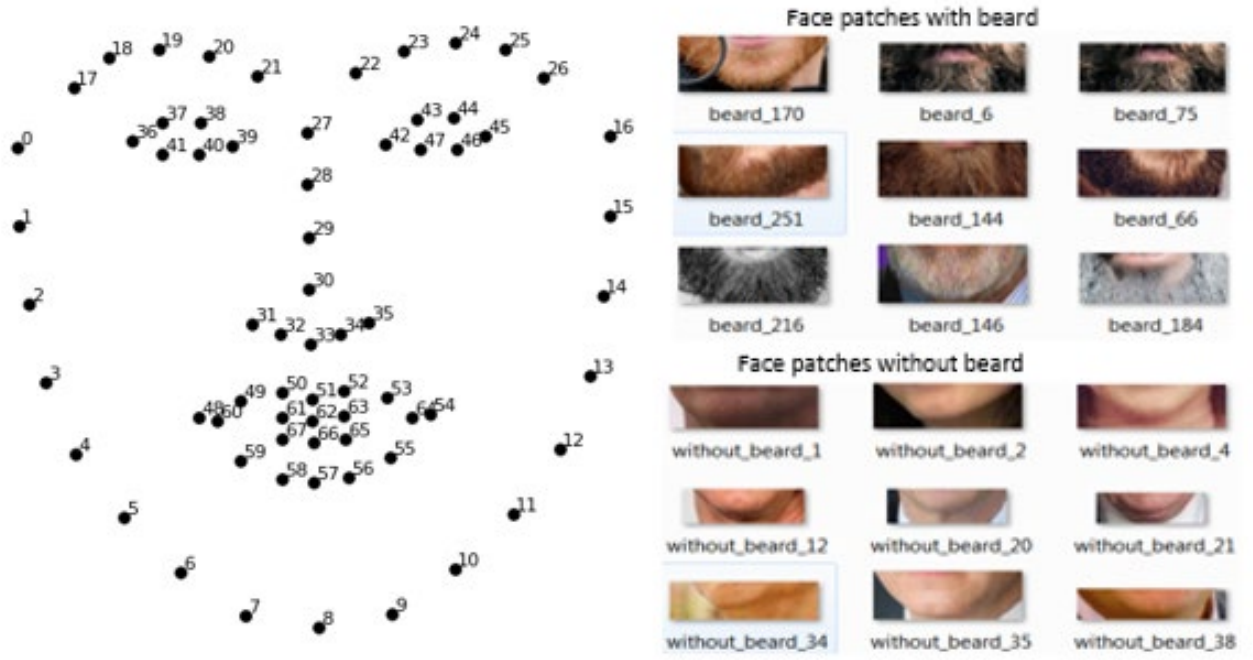

Fig. 6. a) facial landmarks b) cropped samples from two groups/class.

Require: **Input:** $face\_image$
Output: $face\_patches$
$l \leftarrow find\_face\_landmarks(face\_image)$
$(x, y, w, h) \leftarrow find\_face\_bound\_rect(face\_image)$
$(x1, y1) \leftarrow l[0]$
$(x2, y2) \leftarrow l[16]$
$(x3, y3) \leftarrow l[29]$
$(x4, y4) \leftarrow l[19]$
$(x5, y5) \leftarrow l[24]$
$(x6, y6) \leftarrow l[4]$
$(x7, y7) \leftarrow l[12]$
$(x8, y8) \leftarrow l[8]$
$(x9, y9) \leftarrow l[30]$
$(x10, y10) \leftarrow l[33]$
$(x11, y11) \leftarrow l[31]$
$(x12, y12) \leftarrow l[57]$
$offset\_up \leftarrow integer((y8 - y1)/3)$
$offset\_down \leftarrow integer((y11 - y9)/3)$
$offset\_left \leftarrow integer((x1 - x)/2)$
$ep \leftarrow face\_image[y1 - offset\_up :$
$y6 + offset\_down, x1 - offset\_left : x2 + offset\_left$
$offset \leftarrow y8 - y1$
if $y - offset \leq 0$ **then**
$y \leftarrow 0$
else
$y \leftarrow y - offset$
end if
$hp \leftarrow face\_image[y : y5, x : x2]$
$bp \leftarrow face\_image[y6 : y8, x6 : x7]$
$mp \leftarrow face\_image[y10 : y6, x6 : x7]$
$face\_patches \leftarrow (ep, hp, bp, mp)$
$return face\_patches$

Algorithm 1: algorithm takes image as an input and outputs four patches from different area of a face such as eye(ep), beard(bp), mustache(mp), hair(hp). The variable "l" contains facial landmark points and (x,y,w,h) are the bounding box of the detected face. "offset_up","offset_down" and "offset_left" has been used to adjust cropping area around eyes

Fig. 7. Algorithm for croping face-parts.

Results

The person detection and recognition model has been evaluated thoroughly with real-time video captured using Logitech C270 HD webcam placed in-front of the door from the different time points of a day to check how the system performs at different lighting conditions. The

average F-measures of person identification is 0.97. SafeNet model has been evaluated with four standard datasets; Caltech (Caltech, 2019), UTK (UTKFace, 2019), CelebA (CelebA, 2019), Yale (yaleface, 2019) and image samples collected from the web, which contains people with guns, mask, and cellphone. SafeNet generates image descriptions from 10 classes with an average F2-measure 0.97, which outperformed VGG16, ResNet50, and MobileNet for this dataset.

Table 1: Average F-measure of generating image descriptions from 10 classes.

Model and Dataset	**Precision**	**Recall**	**F2-Measure**	**Average F-measure**
SafeNet: Caltech (faces)	0.99	0.99	0.99	**0.97**
SafeNet: UTK	0.99	0.97	0.97	
SafeNet: CelebA	0.97	0.96	0.96	
SafeNet: Yale	0.96	0.96	0.96	
ResNet50: Caltech (faces)	0.87	0.83	0.84	0.85
ResNet50: UTK	0.92	0.91	0.91	
ResNet50: CelebA	0.94	0.91	0.92	
ResNet50: Yale	0.86	0.69	0.72	
VGG16: Caltech (faces)	0.96	0.96	0.96	0.86
VGG16: UTK	0.97	0.95	0.95	
VGG16: CelebA	0.98	0.97	0.97	
VGG16: Yale	0.74	0.53	0.56	
MobileNet: Caltech (faces)	0.96	0.85	0.87	0.86
MobileNet: UTK	0.96	0.94	0.94	
MobileNet: CelebA	0.97	0.91	0.92	
MobileNet: Yale	0.82	0.67	0.7	

Discussion

We started developing this system with participatory design and conducted a survey with a set of 15 questions. Among them, twelve questions were asked to refine the system's functional needs and to find out user's preferences for different functional modes (feedback modes, user interaction modes with a smartphone). Thirty people (26 males, four females) with disabilities participated in that survey where five persons were paralyzed/partially paralyzed, and eight were visually impaired, six persons with hearing disability, and 12 persons had other disabilities. We asked seven questions to understand the effectiveness and impact of our work and survey results showed (see figure 8) that our system will increase the safety and comfort of the people with disabilities. In response to a question, "how secure do you feel at home without any smart system to assess incoming threats," a visually impaired participant, who was an instructor for "technology & apps-uses" at Clovernook Center for the Blind and Visually Impaired, Memphis, TN, said:

"*Nowadays, I am just afraid of coming to the door and opening it without knowing who is there. This system will increase my comfort if I can know who is entering my house. Moreover, people will be able to replace their 250$ doorbell and intercom system if you choose a camera that has both audio input-output and find a way to talk to the incoming person on doorstep. People like me, who is blind and retired depends on SSI (Social Security Income), cannot afford 250$ doorbell. You wouldn't believe how many blind people like me would buy your system if it cost 150$ or less*."

1. Do you feel robbery/burglary/theft may happen to your house?
2. Would it increase your comfort and home safety if a system notifies you who is entering/leaving home?
3. Would you prefer to receive a message which contains information about appearance, facial features (beard, eyeglasses etc.), estimated age and height of a person when the identity of the person at doorstep cannot be reliably ascertained?
4. Would it help you to assess incoming threat if a system notifies you when an unknown person with/without mask/gun/baseball bat/knife/iron rod etc. trespass your house/property?
5. Do you want the system to call emergency service/911/police with your confirmation when a threat is detected?
6. Do you feel insecure opening door when someone unknown comes to your house?
7. Are you willing to use this system/service if it cost 150$

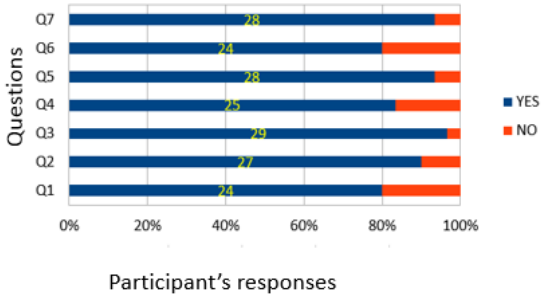

Fig.8. A user survey from 30 participants to find the effectiveness and impact of our system.

The crowd survey revealed that voice-over interaction with a smartphone is preferable than the touch screen. In addition, most of the participants selected "alert message" as a primary feedback mode since it draws more attention compare to MMS/text. Another visually impaired individual suggested us to talk to vendors (Apple/Google) so that we can bypass phone's "Do not Disturb" list for this emergency alert message. He commented:

"In the daytime, I want to receive all alert messages, but in the nighttime, I just want to know about the alert that saves my life."

Conclusion

In this paper, we have presented an interactive assistive solution to assess incoming threats which increase the safety of the people with disability. We have built a novel and robust model to generate image descriptions and collaborating with a group of people with disabilities to improve the efficacy of the system. The extensive quantitative evaluations and initial user study demonstrated that the SafeNet system enhances the safety of people with disabilities. In the

alpha version, the descriptions of images have been generated from 10 categories. The visually impaired participants suggested adding more details (such as the color of shirt/pant, estimated age, and height) about a detected or recognized person in the image descriptions. Moreover, they want the system to detect a few more harmful items such as the knife, baseball bat, and iron rod. In the beta version, we are including new features based on the user's recommendation and training SafeNet model with new items to recognize more harmful materials.

Works Cited

ADT, "Home Security", Date Accessed:01-15-2019, available at: https://security.adt.com/s/d/secureyourhome/

Ahmed, Faruk, et al. "Image captioning for ambient awareness on a sidewalk." 2018 1st International Conference on Data Intelligence and Security (ICDIS). IEEE, 2018.

Ahonen, Timo, Abdenour Hadid, and Matti Pietikäinen. "Face recognition with local binary patterns." European conference on computer vision. Springer, Berlin, Heidelberg, 2004.

Alam, Shahinur, Iftekhar Anam, and Mohammed Yeasin. "O'Map: An assistive solution for identifying and localizing objects in a semi-structured environment." (2015).

Anam, A. S. M., Shahinur Alam, and Mohammed Yeasin. "Expression: A Google Glass based assistive solution for social signal processing." Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility. ACM, 2014.

Asm Iftekhar, Anam, Shahinur Alam, and Mohammed Yeasin. "Expression: A dyadic conversation aid using Google Glass for people who are blind or visually impaired." 6th International Conference on Mobile Computing, Applications and Services. IEEE, 2014.

Bigham, Jeffrey P., et al. "VizWiz:: LocateIt-enabling blind people to locate objects in their environment." 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. IEEE, 2010.

Caltech, "Face Dataset", Date Accessed:09-01-2019 available at: http://www.vision.caltech.edu/Image_Datasets/Caltech_10K_WebFaces/#References

CelebA, "Large-scale CelebFaces Attributes (CelebA) Dataset", Date Accessed: 09-01-2019 available at: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

Chincha, Ricardo, and YingLi Tian. "Finding objects for blind people based on SURF features." 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW). IEEE, 2011

Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

Everingham, Mark, et al. "The pascal visual object classes challenge: A retrospective." *International journal of computer vision* 111.1 (2015): 98-136.

FrontPointSecurity, "Home Security", Date Accessed:01-15-2019 available at: https://www.frontpointsecurity.com/get/july-2019-dm-offer

Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

Girshick, Ross. "Fast r-cnn." Proceedings of the IEEE international conference on computer vision. 2015.

Google, "Adding Accessibility Features", Date Accessed:07-10-2018 available at: https://www.youtube.com/watch?v=1by5J7c5Vz4

Gude, R., M. Østerby, and S. Soltveit. "Blind navigation and object recognition." Laboratory for Computational Stochastics, University of Aarhus, Denmark (2013).

He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017

Honeywell, "Home Security", Date Accessed:07-22-2019 available at: https://www.honeywellstore.com/

Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861 (2017).

Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03167 (2015).

Kao, Gordon, Penny Probert, and David Lee. "Object recognition with fm sonar; an assistive device for blind and visually-impaired people." AAAI Fall Symposium. 1996.

Kazemi, Vahid, and Josephine Sullivan. "One millisecond face alignment with an ensemble of regression trees." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

Lai, Kevin, et al. "A large-scale hierarchical multi-view rgb-d object dataset." 2011 IEEE international conference on robotics and automation. IEEE, 2011.

Looktel, "Currency recognizer", Date Accessed:03-10-2018 available at: http://www.looktel.com/

LSTM, "Language model", Date Accessed:05-05-2019 available at: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Mapelli, Daniela, and Marlene Behrmann. "The role of color in object recognition: Evidence from visual agnosia." Neurocase 3.4 (1997): 237-247.

Mechanical Turk, "Crowd source", Date Accessed:09-01-2019 available at: https://www.mturk.com/

Nezami, Omid Mohamad, et al. "Face-Cap: Image Captioning Using Facial Expression Analysis." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2018.

Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.

RESTWeb, "Web service", Date Accessed:07-10-2018 available at: https://en.wikipedia.org/wiki/Representational_state_transfer

Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Sermanet, Pierre, et al. "Overfeat: Integrated recognition, localization and detection using convolutional networks." arXiv preprint arXiv:1312.6229 (2013).

Sierra, Javier Sánchez, and J. Togores. "Designing mobile apps for visually impaired and blind users." The Fifth International Conference on Advances in Computer-Human Interactions. 2012.

Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

Simplisafe, "Home Security", Date Accessed:07-20-2019 available at: https://simplisafe.com/security?

Snoek, Jasper, Hugo Larochelle, and Ryan P. Adams. "Practical bayesian optimization of machine learning algorithms." Advances in neural information processing systems. 2012.

Sulman, Noah, et al. "How effective is human video surveillance performance?." 2008 19th International Conference on Pattern Recognition. IEEE, 2008.

Sun, Yi, et al. "Deepid3: Face recognition with very deep neural networks." arXiv preprint arXiv:1502.00873 (2015).

Twilio, "Communication API", Date Accessed:06-14-2019 available at: https://www.twilio.com/

UCPD, Berkeley, "Describe a suspect", Date Accessed:08-10-2019 available at: https://ucpd.berkeley.edu/campus-safety/report-crime/describe-suspect

Ushiku, Y., Yamaguchi, M., Mukuta, Y., & Harada, T. (2015). Common subspace for model and similarity: Phrase learning for caption generation from images. In International Conference on Computer Vision.

UTKFace, "Large Scale Face Dataset", Date Accessed: 09-01-2019 available at: https://susanqq.github.io/UTKFace/

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In IEEE Conference on Computer Vision and Pattern Recognition.

Verma, Y., & Jawahar, C. V. (2014). Im2Text and Text2Im: Associating Images and Texts for Cross-Modal Retrieval. In British Machine Vision Conference.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In IEEE Conference on Computer Vision and Pattern Recognition.

Vivint, "Home Security", Date Accessed:07-21-2019 available at: https://www.vivint.com/ppc

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In International Conference on Machine Learning

Yagcioglu, S., Erdem, E., Erdem, A., & Cakici, R. (2015). A Distributed Representation Based Query Expansion Approach for Image Captioning. In Annual Meeting of the Association for Computational Linguistics.

yaleface, "Face dataset", Date Accessed: 09-01-2019 available at: http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html

Zhang, Kaipeng, et al. "Joint face detection and alignment using multitask cascaded convolutional networks." IEEE Signal Processing Letters 23.10 (2016): 1499-1503.