https://doi.org/10.2352/ISSN.2470-1173.2021.8.IMAWM-269

Bed Exit Detection Network (BED Net) for Patients Bed-Exit Monitoring Based on Color Camera Images

Fan Bu[1]***, Qian Lin***[2]***, and Jan Allebach***[1]
[1]***School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN***
[2]***HP Labs, HP Inc., Palo Alto, CA***

Abstract

Among hospitalized patients, getting up from bed can lead to fall injuries, 20% of which are severe cases such as broken bones or head injuries. To monitor patients' bed-side status, we propose a deep neural network model, Bed Exit Detection Network (BED Net), for bed-exit behavior recognition. The BED Net consists of two sub-networks: a Posture Detection Network (Pose Net), and an Action Recognition Network (AR Net). The Pose Net leverages state-of-the-art neural-network-based keypoint detection algorithms to detect human postures from color camera images. The output sequences from Pose Net are passed to the AR Net for bed-exit behavior recognition. By formatting a pre-trained model as an intermediary, we train the proposed network using a newly collected small dataset, HP-BED-Dataset. We will show the results of our proposed BED Net.

Introduction

Among hospitalized patients, getting up from bed can lead to fall injuries [1], 20% of which are severe cases such as broken bones or head injuries [2, 3]. Assistance from health care workers can help reduce the chance of a patient falling. However, nurses are usually so busy that they have to leave some patients' care needs unattended. Setting up a bed-exit detection system [1, 4] in hospitals can help the nursing staff assist the patients and reduce the risk of falling during bed-exit.

Contact methods, such as pressure mats, can be used to detect patients' weight shift [5]. However, false alarms often happen regardless of the detection algorithms. Each time visiting the wards has a high time cost. Hence, to quickly verify the alarm, the nursing staff prefers camera-based detection algorithms rather than other set-ups so that they can monitor the live stream remotely.

This paper proposes a camera-based algorithm, which leverages deep neural network models to detect human bed-exiting activities from color camera images. The experimental results show that our method achieves promising performances.

Related works

Non vision-based methods

Among the non-vision-based methods, one approach is to leverage a pressure-sensitive system, such as pressure mats [5, 6]. However, pressure mats need to be regularly checked to ensure correct functionality and positioning, as they are constantly under physical stress and are movable with the movement of the human body. In addition, pressure mats are not disposable and frequently require cleaning and disinfection in environments such as hospitals and care centers, where infections and body fluid leakage are common.

Another solution has included wearable sensors such as accelerometers [7, 8] and radio-frequency identification sensors [9]. Wearable sensors usually require the patient to be wired, or are too bulky to apply to older people or patients. For wireless sensors, people can easily forget to wear such sensors, and those battery-based sensors require frequent recharging.

Vision-based methods

Depth cameras such as Microsoft's Kinect have been used for research to detect the bed position or to monitor a patient's condition. Banerjee et al. estimate the bed based on the height difference from the bed to the floor [10]. When a patient falls, the patient will be disconnected from the bed surface. Hence, they proposed a method that detects patient's fall by observing the shape of the bed's surface. Bauer et al. locates the bed by fitting the pointclouds to a predefined two-rectangle-shape model, and classifies the patient's state with machine learned shape models and a decision-tree [11]. Chen et al. detected congestion in the neural network in the low-resolution depth image to detect the patient's bed-exit posture [12]. However, depth cameras emit sound or light to all people in the scene. The red light emitted from depth cameras may be regarded as a problem by hospital officials or may interfere with other sensors in the room. Deep-learning-based models are a popular solution for recognizing bed-exiting behavior from depth sensor images [12]. Because of the limitation of deep learning models, the detection accuracy is highly based on the training data.

Apart from depth cameras, RGB cameras can be a safer choice for medical use because most of the depth cameras emit sound or light to the environment. Inoue et al. proposed a deep-learning-based model to detect human sitting-up positions from RGB camera images [13]. However, not all people sit up before they exit their beds. A counterexample would be patients directly falling from the bed. We proposed a novel end-to-end method to address this issue and showed that our method outperforms theirs on both test accuracy and generalization ability.

Method

Bed exit is a type of activity with body movements. To detect these movements, one may not need all the details of the human body. Recently, researchers have developed robust human body keypoints detection algorithms, such as OpenPose [14], Mask R-CNN [15], etc. A typical keypoints detection result is shown on the black-background picture in Fig. 2. We can simplify human motions with the help of these skeleton-like keypoints and then recognize the person's activity.

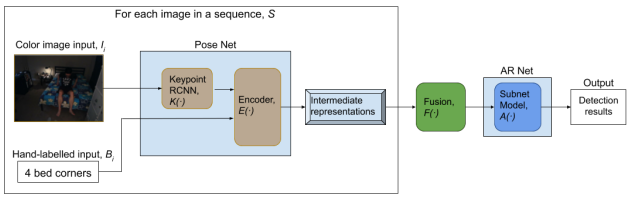


Figure 1: The overall flowchart of our proposed bed-exit detection algorithm based on RGB camera images.

In this paper, we proposed a camera-based bed-exit detection algorithm using deep learning models. The proposed BED Net consists of two sub-networks: a posture detection network (Pose Net), and an action recognition network (AR Net). The Pose Net leverages the state-of-the-art neural-network-based keypoint detection algorithms $K(\cdot)$ to detect human postures from RGB camera images I_i in a sequence S. The outputs of the Pose Net $K(I_i)$, and the four bed corners B_i, are formatted to image-like intermediate representations using encoder $E(\cdot)$. An example of the intermediate representation is shown in Fig. 2. A sequence of those representations are fused by the fusion block $F(\cdot)$, and are then passed to the AR Net $A(\cdot)$ for bed-exit behavior recognition. Sections *Pose Net* and *AR Net* describe the two sub-nets in detail. Fig. 1 shows the overall flowchart of our proposed network. In summary, the output Y of our models is defined by:

$$Y = A\left(F\left(\sum_{i=1}^{S} E\left(K(I_i), B_i\right)\right)\right)$$

Pose Net(Posture Detection Network)

The input to our proposed method is video data collected from monocular cameras for 3 seconds with a sampling rate of $5Hz$. We sub-sample the video by extracting one image every 0.2 seconds and end up with a sequence of $S = 16$ images. All 16 images are forwarded into Pose Net one by one, and hence can generate 16 image-like intermediate representations, as shown in Fig. 1.

In the Pose Net, input RGB images are first resized to 1024×1024, and are passed through a pre-trained* Mask R-CNN network [15]. We chose Mask R-CNN network as it can detect the location of 17 body key-points of each person on the input image. Each key-point comes with a confidence score that reflects how much the Pose Net trusts its detection. Apart from keypoints, Pose Net requests the location of the bed in each image to form the output. Since beds are barely moved in practical use, we simplify beds as a quadrilateral shape on each image and manually label the four corners of each bed. As is shown in Fig. 1, key-point detection results and the bed corners were encoded to an intermediate representation in the encoder $E(\cdot)$.

The output of Pose Net are image-like intermediate representations as shown in Fig. 2. Each image-like representation is a 2-d image with the same height and width of the input camera image, where the background is colored in black, and the bed area is colored in white. Key-points detection results are drawn on the 2-d image as color skeletons. Due to the model limitation of Mask R-CNN, the output of Pose Net is not always perfect. For example, a blue line is drawn on the right side of the black-background image in Fig. 2. The Pose Net draws a blue line to represent a part of the human body, which is caused by noise pixels from the input camera image.

*The pre-trained model and the skeleton color maps are available at `https://github.com/facebookresearch/detectron2`.

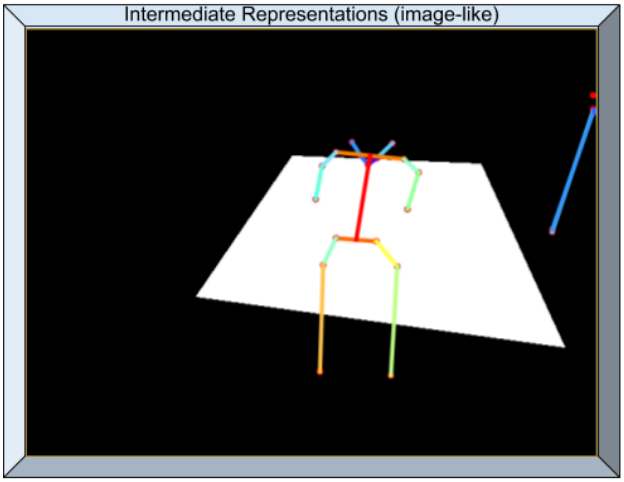


Figure 2: An example of intermediate representation (image-like).

AR Net(Action Recognition Network)

The image-like intermediate representations from the Pose Net are first resized to 112x112. A sequence of these 16 resized images are channel-wise concatenated by the fusion module $F(\cdot)$ to form a 3D tensor, which makes the input of our AR Net.

The AR Net is an action recognition network that has a Res3D architecture following [16]. Fig. 3 shows the detailed architecture of the AR Net. Res3D is based on ResNets, which introduce shortcut connections that bypass a signal from one layer to the next. The connections pass through the gradient flows of networks from later layers to early layers, and ease the training of deep networks. We can tell from Fig. 3 that each skip connection skips two convolutional layers. If we consider these two convolutional layers as a single block, which we will call the ResBlock, then the input to a ResBlock has another path which connects its output. Such connections add the inputs of each ResBlock to its output, giving us the final output for that particular ResBlock. For example, "Tensor 2" is computed by a ResBlock from "Tensor 1". With the help of these ResBlocks, we acquire "Tensor 1" to "Tensor 9".

"Tensor 9" is passed to an average pooling layer to make "Tensor 10"; and the output of the last fully connected layer is the final classification result for human activities ("stay in bed" or "exit bed").

The AR Net is trained with a classification loss. The classification loss computes the cross-entropy between the predicted classification logits and the ground truth classification vectors.

Experiments

baseline methods

Among all similar works, we choose the 2-fully-connected-layer method as our baseline method [13], short as the 14+2FC method in this paper. The baseline method also leverages the

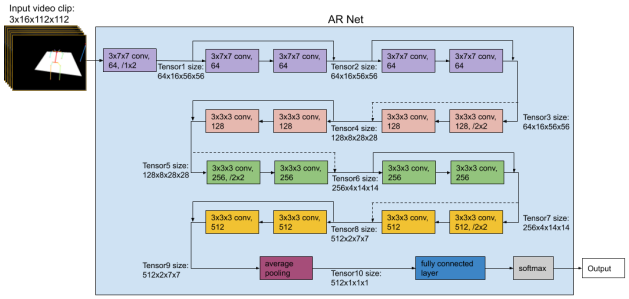


Figure 3: Flowchart of the AR Net.

Key-points detection algorithms to format its intermediate representations. For a fair comparison, we use the same pre-trained model † to detect the keypoints of the person on the input image.

To further interpret different deep-neural-network models' characteristics, we implemented five models and ran them on the same dataset HP-BED-dataset. We show the comparison results in Table 2. The overall structure of all the comparison is smimilar to Fig. 1, which is constructed from two sub-nets: Pose Net and AR Net. All models are trained with the same cross-entropy loss for classification. The inputs, outputs, and the inner architectures of the two sub-nets are different, and are discussed in detail in the following subsections.

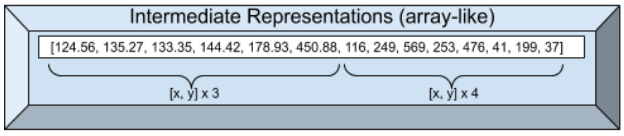


Figure 4: The intermediate representation (array-like, 14 points) for the 14+FC method.

(1) 14+2FC method. 14+2FC method is also our baseline method. Unlike our method, the input to the 14+2FC method is one image instead of a sequence of images. The 14+2FC Pose Net formats an array-like intermediate representation consisting of 14 numbers. The array-like intermediate representation is shown in Fig. 4. The first 6 numbers are the "x position" and the "y position" of the neck, left pelvis, and right pelvis of the person who has the highest confidence score on the input image. The last 8 numbers are the "x position" and "y position" of the four bed-corners in the input image.

The AR Net of the 14+2FC method consists of two fully connected layers as shown in Fig. 5. Both layer sizes are set to 20, following [13].

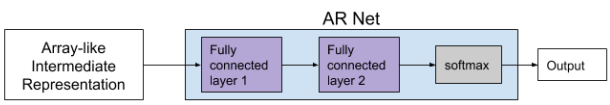


Figure 5: AR Net structure of the 14+2FC method.

(2) 59+2FC method. The 59+2FC method is very similar to the 14+2FC method. The Pose Net of 59+2FC method formats an array-like intermediate representation consists of 59 numbers. An example is shown in Fig. 6. The AR Net structure of the 59+2FC method also consists of two fully connected layers as shown in Fig. 5. Both layer sizes are set to 59 because the array-like intermediate representation contains more information than the 14+2FC method.

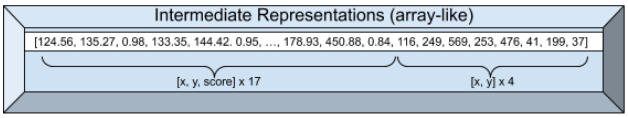


Figure 6: The intermediate representation (array-like, 59 points) for the 16x59+LSTM method.

(3) 16x59+LSTM method. We build this model to test the impact of sequential data. Similar to our proposed method, the input of the 16x59+LSTM Pose Net is a sequence of 16 images using the same sampling rate of $5Hz$. Each image is passed through the Key-points detection algorithm and formatted to an array-like intermediate representation consists of 59 numbers. Similar to the 59+2FC method, the array-like intermediate representation is shown in Fig. 6. A fusion function concatenates the array-like intermediate representations to a 2D tensor of size 16x59, which makes the input to its AR Net.

The AR Net structure of the 16x59+LSTM method is shown in Fig. 7. The sub-net model is a LSTM[17] Recurrent Neural Network following by a fully connected layer. The input size and the hidden size of the LSTM are 59, and the number of recurrent layers is 2.

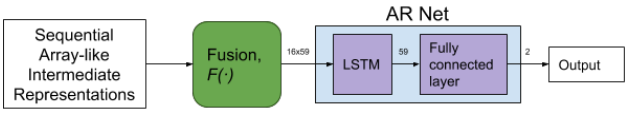


Figure 7: AR Net structure of the 16x59+LSTM method.

(4) C3D method. The input of this model is the same video data as our proposed model, shown in Fig. 1. We collect 16 images with the same sub-sampling rate and hence the Pose Net generates the same image-like intermediate representations as shown in Fig. 2.

Fig. 8 shows the AR Net architecture of the C3D model. It is a 3D convolutional network following [18]. We test this model to compare the performance of different 3D convolutional networks. The 16 image-like intermediate representations are resized to 112x112 and concatenated in temporal order to form a 3d tensor and are passed into the AR Net. The convolutional layer sizes are shown in Fig. 8, and the layer sizes for the last two fully connected layers are 8192 and 4096.

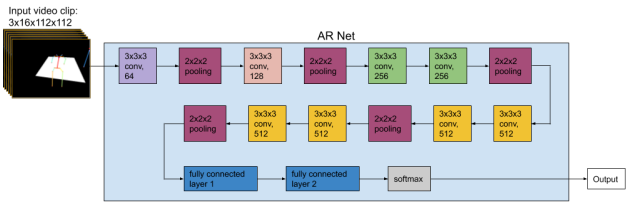


Figure 8: AR Net structure of the C3D method.

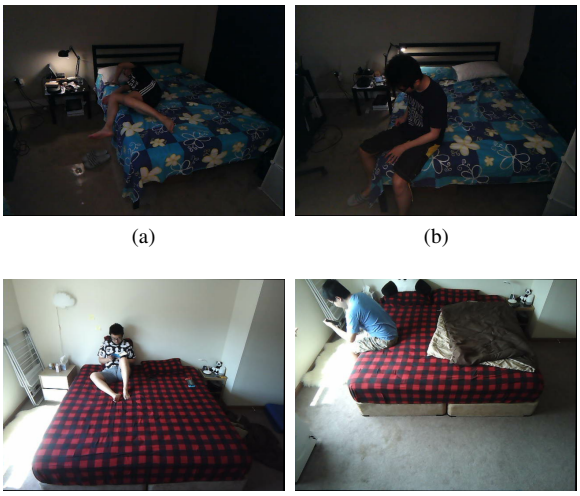
HP-BED-Dataset

We collected a dataset called HP-BED-Dataset for the ablation study. The dataset is collected in two bedrooms with a color camera mounted at a height range from 1.5 to 1.8 meters facing the bed.

†The pre-trained model is available at `https://github.com/facebookresearch/detectron2` with a configure file of COCO-InstanceSegmentation mask_rcnn_R_50_FPN_3x

The training set and the validation set contain 120 videos with each video length between 10 to 30 seconds, 30 frames per second. The resolution of all images is 640x480. Each video sequence is choreographed to include periods when the subject is motionless in the bed, and periods when the subject is moving. In some sequences the subject enters the field of view and lies down on the bed. All videos in the training set and the validation set are taken in a bedroom with the same bed and the same person with varied clothing. The camera position may vary among all videos, but are all placed closer to the end of the bed and facing the head of the bed. Example images are shown in Figs 9a and 9b. The test set contains 38 videos with similar video length and image resolution, but are collected from a different bedroom of another participant with a different bed set-up. The participant in the test set is different from the participant in the training set. Example images are shown in Figs 9c and 9d.

Based on the participant's position with respect to the bed, all images are labeled with a class, either "stay" or "exit," describing the participant's status. To balance the number of the two classes, we collected videos of 9 scenarios. As it is defined in the previous sections, some of the baseline models read one image at a time, while others read a sequence of images as their data input. When a model reads sequential images, the video data is clipped to a 3-second length and sub-sampled at a sampling rate of $5Hz$, hence ending up with a sequence of $S = 16$ images. The label for each sequence is decided by the label of the last image in that sequence. For example, if a participant sits on the bed edge for the first few seconds, and in the last image frame, he/she leaves the bed or lifts his/her pelvis up from the bed surface, then the whole sequence of 16 images will be labeled as "exit". The details of the training and testing sequences are shown in Table 1. As is shown in the table, there are 38 videos in the test set. If a model input is sequential data, the model cannot generate results for the first 3 seconds of that video, because the model has to collect all 16 images before it starts the inference. For a fair comparison against all models, the results of the first 3 seconds of the 38 videos are not valid in the evaluation. Therefore, the frame numbers in Table 1 counts only the number of valid frames.



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

Figure 9: Example images from the dataset.

Evaluation Metrics

We evaluated the performance of all models using the Accuracy metric defined as below:

$$ACC = \frac{tp+tn}{tp+fp+tn+fn}$$

Our task is a binary classification problem, the images with the ground-truth label "exit" are called the positive samples, and the others with the ground-truth label "stay" are called the negative samples. Accuracy of the test result can be calculated from the true positive (tp), true negative (tn), false positive (fp), and the false negative (fn) numbers. As described in the HP-BED-dataset section, some of the models leverage temporal information, so we ignore the first 3 seconds of each video when evaluating the accuracy of each model. Table 2 shows the comparison results between all models. For each scenario, we calculated the test accuracy of each model, and their tp, tn, fp, fn numbers. For example, under the scenario of "Stay in bed" with blanket, the "14+2FC" model gains a test accuracy of 97.0%. Table 1 shows that all 1109 test cases under the same scenario are negative samples. Thus, the 33 wrong detection are marked as false positives (fp) in Table 2.

Results Comparison

As we've illustrated in the *baseline methods* section, model "14+2FC" and model "59+2FC" inference one image at a time, while other models, including our proposed method, leverage temporal information when making predictions.

The last row of Table 2 shows the overall detection accuracy of our proposed method against other ablation models, and we observe that our proposed method outperforms the others.

The "Stay in bed" row of Table 2 shows that all models generate robust detection when the participant stays in bed, especially when the participant is not covered with any blanket. The main reason is that when most of the body parts are exposed directly to the camera, all body keypoints can be easily detected by the Mask R-CNN network [15]. Another reason is that "Stay in bed" scenario requires less knowledge of the motions. Therefore, models like "14+2FC" and "59+2FC" that are not leveraging temporal information can also achieve test accuracies of 97% and 100%, respectively.

The "Sit on bed edge" row of Table 2 reveals the advantages of temporal-involved models. The "14+2FC" model and the "59+2FC" model show an accuracy less than 70%. An intuitive explanation is that when a person is sitting on the bed edge, without further information, one can hardly tell if he/she is leaving the bed or staying in the bed. On the contrary, if we have beforehand information that a participant sits on the bed edge in the last 3 seconds, most likely he/she will keep sitting on the edge because he/she is not showing any body movements that indicates a bed-exit. We observed from the last column that our proposed model outperforms the "14+2FC" model by approximately 50% accuracy, and outperforms the "59+2FC" model by 27% accuracy. Similar tendencies are found with the "16x59+LSTM" model and the "C3D" model. This result confirms our conjecture that "sitting on bed edge" cannot be accurately deduced from a single image.

The "Stay ←transit→ Exit" scenario contains videos of the participant switching his/her status, namely leave the bed or enter-

Table 1: Details of each scenario in HP-BED-dataset.

Dataset	No. of video sequences	Scenario	With blanket	No. of sequences	No. of frames	No. of "Stay" sequences	No. of "Exit" sequences
training & validation	120	Stay in bed	Yes	15	9811	9811	0
			No	18	10944	10944	0
		Sit on bed edge	Yes	7	5512	5512	0
			No	6	4604	4604	0
		Stay <—transit—>Exit	Yes	8	4801	3489	1312
			No	23	7302	3693	3609
		Fall when exiting bed	Yes	6	4270	1652	2618
			No	15	6324	2278	4046
		Always not in bed	N/A	19	10683	0	10683
test	38	Stay in bed	Yes	2	1109	1109	0
			No	2	1228	1228	0
		Sit on bed edge	Yes	2	933	944	0
			No	3	1309	1309	0
		Stay <—transit—>Exit	Yes	5	2426	1388	1038
			No	8	2937	1378	1559
		Fall when exiting bed	Yes	3	649	260	389
			No	4	981	514	467
		Out of bed	N/A	9	3870	0	3870

Table 2: Test accuracy for all ablation models.

Scenario	With blanket	Model name Accuracy tp/fp/tn/fn				
		14+2FC	59+2FC	16x59+LSTM	C3D	Res3D (Ours)
Stay in bed	Yes	97.0 0/33/1076/0	97.4 0/28/1081/0	**100.0** **0/0/1109/0**	**100.0** **0/0/1109/0**	**100.0** **0/0/1109/0**
	No	**100.0** **0/0/1228/0**	**100.0** **0/0/1228/0**	**100.0** **0/0/1228/0**	**100.0** **0/0/1228/0**	**100.0** **0/0/1228/0**
Sit on bed edge	Yes	46.3 0/505/436/0	68.9 0/293/651/0	82.5 0/165/779/0	95.2 0/44/882/0	**96.1** **0/36/908/0**
	No	30.2 0/910/394/0	26.2 0/966/343/0	79.9 0/262/1047/0	73.3 0/347/956/0	**80.9** **0/250/1059/0**
Stay <-transit-> Exit	Yes	79.7 688/142/1245/350	80.4 696/132/1256/342	86.7 838/122/1266/200	76.0 967/512/876/70	**88.8** **833/65/1323/205**
	No	81.8 1131/105/1272/427	78.4 1060/133/1245/499	83.3 1278/207/1171/281	67.9 1559/940/437/0	**84.9** **1331/213/1165/228**
Fall when exit bed	Yes	62.5 151/5/254/237	69.9 206/12/248/183	**80.1** **260/0/260/129**	66 389/220/39/0	68.2 255/72/188/134
	No	77.1 285/43/471/181	69.7 225/55/459/242	**77.3** **302/57/457/165**	76.7 467/227/283/0	**77.3** **253/8/506/214**
Out of bed	N/A	53.3 2060/0/0/1803	55 2130/0/0/1740	80.5 3116/0/0/754	**100** **3870/0/0/0**	87.9 3402/0/0/468
Overall Rate		69.2 27.9/11.2/41.3/19.4	70 27.9/10.4/42.1/19.4	84.8 37.4/5.2/47.3/9.8	84.6 47.0/14.8/37.6/0.4	**87.7** **39.3/4.1/48.4/8.0**

ing the bed. We can see from Table 2 that our model outperforms others under this unpredictable scenario.

The "Fall when exit bed" row of Table 2 shows that the "16x59+LSTM" model outperforms our proposed method. We replayed the corresponding videos and found that the Mask R-CNN detects noisy skeletal keypoints under this scenario. In some cases, Mask R-CNN can not locate the correct position of the participant, especially when the participant is laying horizontally in the images. Before further processing, the "14+2FC" model, the "59+2FC" model and the "16x59+LSTM" model filter the noisy keypoints from Mask R-CNN by selecting the keypoints with the highest confidence score on the input image. Hence, the "C3D" model and our proposed behave unstable when most images in the sequence contain noise. Future works should be focusing on providing higher quality keypoint inputs.

Though it occurs less frequently, similar noises can be found in the "Out of bed" scenario. The second last row of Table 2 shows that the "C3D" model and our proposed model are robust to such noises.

Beyond the previous models, a variant of the 3D convolu-

tional network[19] is developed to decompose the spatiotemporal convolution process into a (2+1)D convolution. Following the idea of [19], we reproduce our Res3D model with a R(2+1)D implementation, which uses a single type of spatiotemporal residual block without bottlenecks. The test result is very similar to our proposed method; and it runs even slower in Pytorch. Therefore, we did not list it as an ablation model in Table 2.

Implementation details

We test the inference time of our method using the Pytorch framework. The input image has a resolution of 640x480, and an Nvidia GeForce GTX 1070 graphic card is used to speed up the inference time. Our method takes 0.2 second/image on average to compute the detection result. It is also the reason that we subsample the video input every 0.2 second.

Conclusion

When one analyzes human activities with Convolutional Neural Networks, a straightforward solution would be to collect a huge amount of data that contains a large variance of people with different bed settings. However, it is unrealistic to collect extensive data with all types of bedroom set-ups. In this paper, we proposed BED Net, a novel end-to-end method that leverages state-of-the-art neural-network-based keypoint detection algorithms and uses its output as an intermediary. It helps to transfer the large variance of the RGB images to a narrower domain so that our method achieves high detection accuracy with a relatively small dataset. By evaluating the ablation methods on the same dataset, we showed that our method achieves accurate performances and outperforms other methods on both test accuracy and generalization ability.

Acknowledgments

This research project was sponsored by HP Labs, HP Inc., Palo Alto, CA.

References

[1] H. Madokoro, N. Shimoi, and K. Sato, "Unrestrained multiple-sensor system for bed-leaving detection and prediction," *Nursing and Health*, vol. 3, no. 3, pp. 58–68, 2015.

[2] D. Oliver, F. Healey, and T. P. Haines, "Preventing falls and fall-related injuries in hospitals," *Clinics in Geriatric Medicine*, vol. 26, no. 4, pp. 645–692, 2010.

[3] E. D. Bouldin, E. M. Andresen, N. E. Dunton, M. Simon, T. M. Waters, M. Liu, M. J. Daniels, L. C. Mion, and R. I. Shorr, "Falls among adult patients hospitalized in the united states: prevalence and trends," *Journal of Patient Safety*, vol. 9, no. 1, p. 13, 2013.

[4] J. Xu, K. D. Kochanek, S. L. Murphy, and B. Tejada-Vera, "Deaths: final data for 2014," 2016.

[5] R. I. Shorr, A. M. Chandler, L. C. Mion, T. M. Waters, M. Liu, M. J. Daniels, L. A. Kessler, and S. T. Miller, "Effects of an intervention to increase bed alarm use to prevent falls in hospitalized patients: a cluster randomized trial," *Annals of Internal Medicine*, vol. 157, no. 10, pp. 692–699, 2012.

[6] O. Sahota, A. Drummond, D. Kendrick, M. J. Grainge, C. Vass, T. Sach, J. Gladman, and M. Avis, "Refine (reducing falls in in-patient elderly) using bed and bedside chair pressure sensors linked to radio-pagers in acute hospital care: a randomised controlled trial," *Age and Ageing*, vol. 43, no. 2, pp. 247–253, 2014.

[7] A. Härmä, W. ten Kate, and J. Espina, "Bed exit prediction based on movement and posture data," in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2014, pp. 165–168.

[8] R. L. S. Torres, Q. Shi, A. van den Hengel, and D. C. Ranasinghe, "A hierarchical model for recognizing alarming states in a battery-less sensor alarm intervention for preventing falls in older people," *Pervasive and Mobile Computing*, vol. 40, pp. 1–16, 2017.

[9] A. Wickramasinghe, D. C. Ranasinghe, C. Fumeaux, K. D. Hill, and R. Visvanathan, "Sequence learning with passive rfid sensors for real-time bed-egress recognition in older people," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 4, pp. 917–929, 2016.

[10] T. Banerjee, M. Enayati, J. M. Keller, M. Skubic, M. Popescu, and M. Rantz, "Monitoring patients in hospital beds using unobtrusive depth sensors," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014, pp. 5904–5907.

[11] P. Bauer, J. B. Kramer, B. Rush, and L. Sabalka, "Modeling bed exit likelihood in a camera-based automated video monitoring application," in *2017 IEEE International Conference on Electro Information Technology (EIT)*. IEEE, 2017, pp. 056–061.

[12] T.-X. Chen, R.-S. Hsiao, C.-H. Kao, D.-B. Lin, and B.-R. Yang, "Bed-exit prediction based on 3d convolutional neural network," in *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 2018, pp. 1185–1188.

[13] M. Inoue, R. Taguchi, and T. Umezaki, "Bed-exit prediction applying neural network combining bed position detection and patient posture estimation," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 3208–3211.

[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[16] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, "Convnet architecture search for spatiotemporal feature learning," *arXiv preprint arXiv:1708.05038*, 2017.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[19] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

Author Biography

Fan Bu received a bachelor's degree in mechanical engineering from Huazhong University of Science and Technology (2015) and a master's

degree in mechanical engineering from the University of Michigan (2017). He is now a Ph.D. student in electrical and computer engineering at Purdue University, working on research projects in communications, networking, signal and image processing.

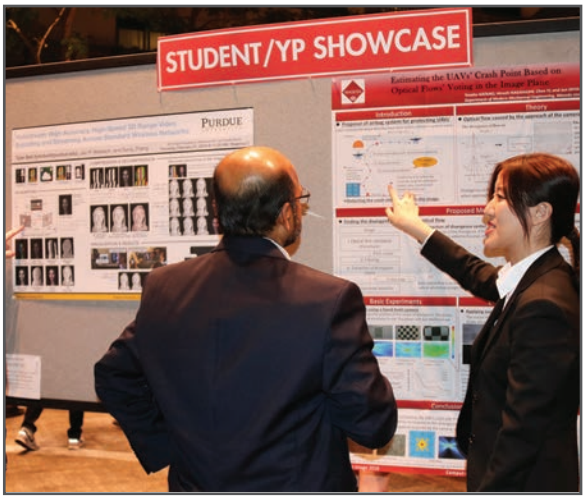
JOIN US AT THE NEXT EI!

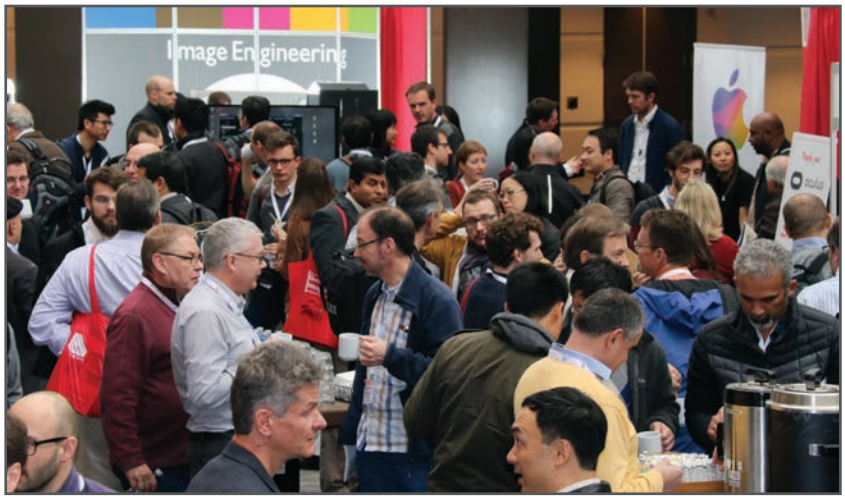
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org