3D Point Cloud-Based Visual Prediction of ICU Mobility Care Activities

Bingbin Liu*[1] bingbin@stanford.edu
Michelle Guo*[1] mguo95@stanford.edu
Edward Chou[1] ejchou@stanford.edu
Rishab Mehra[1] rishab@stanford.edu
Serena Yeung[1] syyeung@cs.stanford.edu
N. Lance Downing[2] ldowning@stanford.edu
Francesca Salipur[2] fsalipur@stanford.edu
Jeffrey Jopling[2] jjopling@stanford.edu
Brandi Campbell[3] Brandi.Campbell@imail.org
Kayla Deru[3] Kayla.Deru@imail.org
William Beninati[3] bill.beninati@imail.org
Arnold Milstein[2] amilstein@stanford.edu
Li Fei-Fei [1] feifeili@cs.stanford.edu

** The two authors contributed equally to the paper.*

[1] *Department of Computer Science, Stanford University, United States*

[2] *Clinical Excellence Research Center, School of Medicine, Stanford University, United States*

[3] *TeleCritical Care, Intermountain Healthcare, United States*

Abstract

Intensive Care Units (ICUs) are some of the highest intensity areas of patient care activities in hospitals, yet documentation and understanding of the occurrence of these activities remain sub-optimal due in part to already-demanding patient care workloads of nursing staff. Recently, computer vision based methods operating over color and depth data collected from passive mounted sensors have been developed for automated activity recognition, but have been limited to coarse or simple activities due to the complex environments in ICUs, where fast-changing activities and severe occlusion occur. In this work, we introduce an approach for tackling more challenging activities in ICUs by combining depth data from multiple sensors to form a single 3D point cloud representation, and using a neural network-based model to reason over this 3D representation. We demonstrate the effectiveness of this approach using a dataset of mobility-related patient care activities collected in a clinician-guided simulation setting.

1. Introduction

In high-intensity hospital environments such as Intensive Care Units (ICUs), the ability to record the occurrence of patient care activities at a per-patient level is key to both monitoring protocol adherence and studying the correlation of care activities with patient outcomes (Schweickert et al., 2009; Investigators et al., 2015). While it is impractical for nurses or staff to constantly record activities at a sufficient level of detail, computer vision has been proposed as a potential solution for monitoring and automatically recognizing clinical activities in hospitals. Passive sensors placed in the environment can capture visual data on a perpetual basis, and machine learning algorithms can reason over this data to identify clinical activities. Recently, success has been shown for using computer vision to recognize surgical workflow in operating rooms (Twinanda et al., 2015), to assess patient mobility status in ICUs (Ma et al., 2017), and to monitor hand hygiene in hospital corridors (Haque et al., 2017). These studies have used either RGB-D sensors (Ma et al., 2017; Twinanda et al., 2015), or depth sensors only (Haque et al., 2017), such that the entire system is privacy-preserving.

While these early studies have shown success in recognizing coarse or limited activities, an important challenge is extending this recognition to a wider range of patient care activities. Here a number of difficulties arise, due to both the fine-grained nature of subtly different activities, as well as the complexity of environments like ICUs. One common challenge is occlusion, where important visual information is physically obfuscated by objects in the camera's line of sight. This is particularly prevalent with 2-dimensional data from a single viewpoint, especially in complex environments which contain a large number of objects that can occlude the camera. For example, a patient may be occluded by a caretaker or a curtain, which is especially common in an ICU setting where care activities often require multiple caretakers, and are performed under an environment with complex equipment setups. Moreover, 2-dimensional data from a single camera is viewpoint variant, which undesirably requires extra computation such as training one model for each viewpoint or incorporating a spatial transformer network (Jaderberg et al., 2015) to normalize different views, and is difficult to generalize.

To address the challenges of difficult viewpoints and to obtain a more complete understanding of an ICU environment, we propose that in contrast to reasoning over image streams from individual RGB or depth sensors as in (Haque et al., 2017; Ma et al., 2017), a more powerful approach towards robust recognition of activities is to reason over integrated information from multiple depth sensors in 3D. In particular, we collect inputs from two depth sensors capturing complementary viewpoints covering the front and side view of a room, which we then use to build point clouds over the room. Point clouds provide benefits over other representations, namely their ability to model scenes at a shared, global level (Goesele et al., 2010). Additionally, due to their 3D nature, point clouds provide geometric invariance such as viewpoint independence (Henry et al., 2012). These benefits are useful for our task of understanding activities in ICUs, which have inherently complex scenes that could be better reasoned over using global representations.

We demonstrate the use of a deep learning-based approach reasoning over 3D point cloud representations to recognize a set of patient care activities common in ICU ward settings. We focus in particular on activities related to patient mobility, to encourage early mobility and prevent complications such as ICU acquired weakness. We set up a simulation room of an adult ICU ward, and collect a dataset of clinician-led simulations of these activities using two depth sensors capturing complementary views. We combine data from both sensors to form a point cloud, a view-invariant 3D representation of the room. We then show that a neural network-based model reasoning over the point cloud representation is able to effectively recognize the activities and outperform approaches based on single-sensor data.

2. Related Work

Recognizing and interpreting human activity is a widely studied problem in computer vision (Poppe, 2010; Weinland et al., 2011), across both images and videos. While these have primarily focused in the domain of color internet images and videos (Caba Heilbron

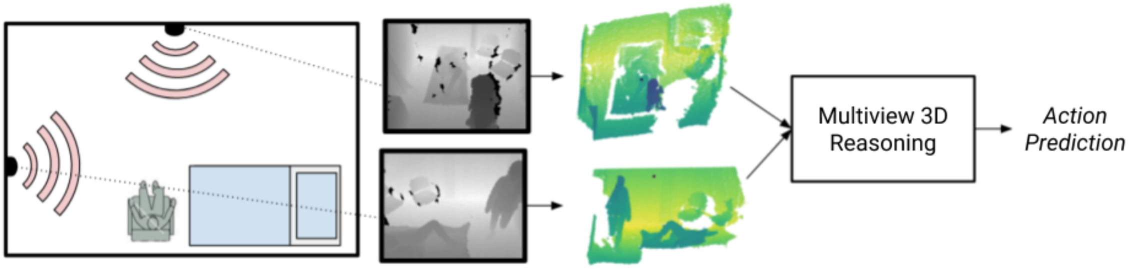


Figure 1: An overview of our proposed approach. Depth data from multiple sensors is combined in a fused point cloud representation. Our model is a neural network-based model operating over this point cloud input that produces the final activity prediction.

et al., 2015; Jiang et al., 2014; Yeung et al., 2015; Ramanathan et al., 2015; Abu-El-Haija et al., 2016; Carreira and Zisserman, 2017), a recent line of work has also explored activity recognition from depth images expressing distances of people and objects from the sensor (Ye et al., 2013).

In contrast to these traditional approaches that reason over 2D image data, an alternative representation of a visual scene is a point cloud, which is a 3D point occupancy-based representation of a given space. Such a 3D representation has several advantages over single viewpoint 2D representation. It has more detailed information of the space, and does not suffer from object occlusion since its construction utilizes information of the whole space rather than a single viewpoint. Recently, neural network-based models have been developed that reason over point cloud inputs (Qi et al., 2017) for the task of object classification. We build upon this work, and reason over point clouds to detect patient care activities in ICUs, in particular a set of activities related to mobility.

In the domain of healthcare, the use of computer vision-based methods to automatically interpret clinically-relevant human activities in healthcare spaces has shown early promises for a number of applications. (Ma et al., 2017) and (Twinanda et al., 2015) both use color video supplemented by depth data (RGB-D), to recognize surgical workflow activities in an operating room and assess patient mobility level in an ICU room, respectively. (Haque et al., 2017) uses a depth sensor-only setup to preserve privacy, and presents a larger-scale study of computer vision for recognizing hand hygiene actions in corridors across a hospital unit.

Our work builds upon all of these. We use a privacy-preserving depth sensor-only setup similar to (Haque et al., 2017). While they focused on hospital corridors, we study the environment inside an ICU room like (Ma et al., 2017), and in particular focus on the application of encouraging patient mobility. However, while (Ma et al., 2017) classified video segments with the maximum level of patient mobility in each segment (4 levels corresponding to Nothing, In-Bed, Out-of-Bed, and Walking), we focus not on assessing the state of the patient but rather on recognizing the nursing activities involved in a patient's care. Recording these has the potential to assist in ensuring compliance with protocols for patient care activities, as well as correlating performed activities with outcomes towards developing effective care guidelines. In order to recognize such discrete activities, which may be fast-changing and complex, we move from interpreting single-sensor data as in (Haque et al., 2017) and (Ma et al., 2017), to integrating multi-sensor data in a 3D representation and

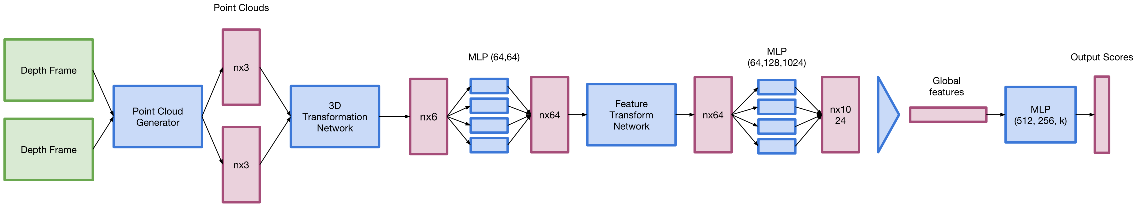


Figure 2: An overview of our model, which takes depth data from two sensors as input to generate and fuse a combined point cloud representation. Green boxes denote individual depth images. Blue boxes denote processing pipelines such as a neural network modules. Purple/pink boxes denote intermediate representations. Data flow is from left to right.

performing 3D reasoning. (Twinanda et al., 2015) also reasons in 3D; however, while they extract 3D features from interest points, we leverage recent advances in deep learning to use a neural network-based approach operating directly over a rich point cloud representation.

3. Method

Our goal is to use computer vision to automatically record and recognize ICU activities. We treat the problem as an activity detection task and train a machine learning model to detect these events. In contrast to recent methods which use 2D methods for temporal activity detection on RGB images, we propose a new approach for temporal activity detection in 3D. Our key insight is that in busy healthcare settings, such as complex environments like the ICU, reasoning on 2D sensor data from a single viewpoint alone is not sufficient to richly cover the entire space. Some downsides and challenges to 2D single-view approaches include viewpoint variance, occlusion, as well as incomplete coverage of entire rooms (Sturm et al., 2012). In contrast, multi-view, 3D monitoring of an ICU patient room can provide a much richer and more comprehensive coverage, both temporally and spatially, of important activities involving both healthcare workers and the patient.

In this work, we propose a viewpoint invariant, multi-view 3D model for temporal activity detection. For a given video clip, the goal is to perform frame-level classification of the activity that is currently occurring in the frame. To construct an approach that is truly 3D and viewpoint invariant, we must address two main challenges.

- First, we note that sensor viewpoints can vary widely across different environments based on the placement and angle of the camera. To work towards viewpoint invariance, we use a 3D spatial transformer network on the input 3D video to normalize multiple views into a common 3D space.

- Second, we must design a method for effectively combining information collected in multiple videos from multiple sensor viewpoints. For this, we process our 3D videos using a PointNet to embed the 3D frames, then perform *late sensor fusion* across the multiple sensors.

Our method is summarized in Figure 2, which consists of 3D data processing, multi-view fusion, and 3D network reasoning.

3.1 Point Cloud Generator

Our data is collected in the form of depth videos. In order to perform 3D video understanding, we first convert depth frames into 3D point clouds. Given an input depth image (video frame) of size $H \times W$, where each pixel (x_{img}, y_{img}) represents the distance from the sensor to the "pixel", we transform the depth image into a point cloud of size $HW \times 3$, where each pixel is represented by a 3D point with coordinate (x, y, z) through the following transformation:

$$\begin{aligned} x &= C_z(x_{img} - 160) \\ y &= -C_z(y_{img} + 120) \\ z &= z_{img} \end{aligned} \tag{1}$$

where C_z, 160 and 120 come from intrinsic camera calibration parameters, and z_{img} is the value of the depth map at the pixel (x_{img}, y_{img}).

3.2 3D Transformation Network

Our goal is to design a method that is capable of processing raw 3D point cloud videos captured from a large variety of viewpoints. The main technical challenge at hand is therefore the problem of viewpoint variance across different sensor viewpoints, which we address by normalizing different viewpoints using a spatial transformer. Having generated point clouds using the intrinsics of the camera, we attempt to address variations in camera extrinsics (position, angle, viewpoint, etc.) in the real world by constructing a network that learns to predict an affine transformation matrix for each given example. This transformation network effectively normalizes 3D coordinate spaces across different 3D inputs with varying viewpoints.

3.3 Deep Learning on Point Sets

Given normalized point clouds across different viewpoints, we now perform classification on these point clouds to predict the currently occurring activity at a given time. Traditionally, classification on depth images is performed using a CNN using 2D reasoning. In this work, our aim is to reason over the 3D space using point clouds. To this end, we seek to perform classification over point clouds using a neural network architecture with the following properties:

1. **Order invariance.** Point clouds of size $(H \times W, 3)$ can contain its points in any order. This network should be invariant to $(HW)!$ permutations of the input point set.

2. **Local point interaction.** Rather than treating points as being isolated, the network should be capable of performing spatial reasoning in 3D space as a CNN would in 2D space. More concretely, the network should capture local structures of points neighboring each other.

3. **Viewpoint invariance.** This is the most crucial property in a desired model for processing 3D inputs in our setting. We aim to have a network that treats the input point set as a geometric object, and the learned 3D representation by the network should be invariant to rigid transformations on the point set.

Following Qi et al. (2017), we employ a PointNet architecture to process our point clouds. Given n points representing a point cloud, our network computes learned transformations between input and intermediary features, followed by feature aggregation using max pooling. The final step of the network is a linear layer which outputs k classification scores for the k patient care activities.

4. Data

To evaluate our approach, we collect a dataset of patient care activities in a simulated ICU room. We collect data from two depth sensors of three patient care activities related to mobility: getting out of bed, getting in chair, and turning in bed. The two depth sensors capture complementary views, namely the front view facing directly towards the bed, and the side view that centers around the chair and covers the head of the bed. The simulation was guided by a clinician, who helped make sure the activities are reflective of real-world scenarios and cover enough variations to encompass the diverse scenarios that can occur in ICU wards. For example, depending on the condition of the patient, "getting patient out of bed" may involve different numbers of caretakers, and may vary significantly in duration. "Turning in bed" could be performed in different ways to include different practices, with different types of back support and equipment used to turn the patient. After data collection, the data was examined by the clinician to ensure the activities were conducted following the correct protocol.

In total, 16853 seconds of videos are collected from 10 actors under two pairs of viewpoint settings, comprising of 316 activity instances, and a difficult negative portion of background classes. In particular, only 3144 seconds out of the 16853 seconds, or 9455 out of 14075 frames are non-background. Table 1 and Fig 3 show statistics of the simulation data. We note that activities involving chairs are particularly short, with an average duration of 3.0 seconds. In contrast, bed-related activities have a considerably longer average duration of more than 10 seconds, as well as a greater variability compared to other activities, due to the wide-ranging efforts that can be needed to get a patient with little mobility and/or significant weakness out of bed.

Depth sensors are used to collect de-identified depth images out of privacy concern of the patients. The sensors produce 240 by 320 depth images with a recording speed of 30 frames per second. The images are subsampled to 3 frames per second, which helps saving storage and computation time while retaining enough information to obtain a decent performance.

5. Experiments

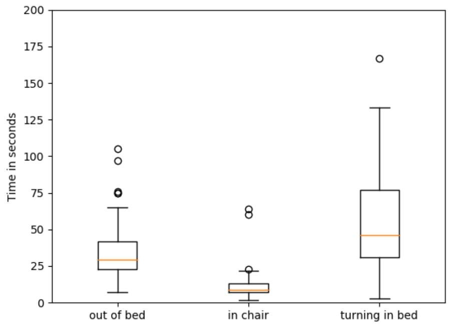
In this section we present quantitative and qualitative results of our point cloud-based neural network approach (based on (Qi et al., 2017) and referred to as PointNet) for recognizing mobility-related care activities from depth images captured by multiple sensors.

For quantitative experiments, we perform frame-level evaluation as opposed to instance-level, similar to Yeung et al. (2015), as our objective is to accurately predict and obtain the total time of every activity for richer information useful to document, instead of merely occurrences of the activities. We use the standard average precision (AP) metric for evaluation.

Table 1: Statistics of care activities in our dataset. # refers to the number of simulation instances. The chair activities are the shortest in both average and total time, whereas “turning in bed" is the longest since turning an immobile patient requires following specific protocols.

Activity	#	Total Time (s)	Total Frames
Getting out of bed	92	1103	3309
Getting in chair	133	414	1244
Turning in bed	91	1627	4882
All activities	316	3144	9455

Figure 3: Box plot illustrating the variability in the duration of activities.



Split	Getting out of bed	Getting in chair	Turning in bed	Total
Train	1829 (57.7%)	989 (67.3%)	2896 (60.2%)	5714
Val	529 (16.7%)	171 (11.6%)	877 (15.8%)	787
Test	812 (25.6%)	310 (21.1%)	1109 (24.0%)	2231
Total	3170	1470	4882	8732

Table 2: Number of frames with activities and their percentage of total in train/val/test splits.

5.1 Data preparation

Train, Val, Test splits. The data collected from two depth sensors are first temporally aligned, then divided into training, validation, testing set, roughly following a 60%, 15%, 25% split with no temporal overlap between the splits. Table 2 shows detailed statistics of each split.

PointCloud Subsampling The direct output from depth sensors comprises of 76800 (i.e. 320 by 240) points, which would be computationally expensive for downstream models. We therefore follow the suggestion in Qi et al. (2017) to uniformly subsample 1024 points out of the 76800. Our experiments show that training with subsampled data achieves comparable results or even outperforms the full point clouds on certain activities, which may be related to having fewer noises than the original constructed point clouds, and the fact that the randomness in subsampling effectively augments the data by adding variations which help prevent overfitting. In addition, subsampled points are more than 50 times more memory efficient, and help shorten the computation time significantly.

5.2 Backbone model

Baseline: CNN As a baseline, we implemented a CNN following ResNet-18. The model is trained from scratch. We experimented with initializing the model using weights pretrained on ImageNet (Russakovsky et al., 2015) but did not see improvement, which may be caused by the discrepancy between depth and RGB data. Training is performed using SGD with a

Model	Sensor	Out of Bed	In Chair	Turning in Bed	Mean AP
CNN	A	32%	**28%**	80%	47%
CNN	B	20%	12%	77%	36%
PointNet	A	49%	26%	79%	51%
PointNet	B	44%	24%	83%	50%
PointNet	A or B	47%	22%	81%	50%
PointNet+STN	A or B	48%	24%	83%	51%
CNN	AB	48%	20%	81%	45%
PointNet	AB	49%	25%	**85%**	53%
PointNet+STN	AB	**52%**	27%	84%	**54%**

Table 3: Frame level mean average precision (mAP) comparison of the 2D CNN model against the 3D point cloud model, using different combinations of sensors. A and B correspond to the two sensors in a room. Our PointNet approach is able to significantly outperform the CNN on the Out of Bed and Turning in Bed activities as well as on the overall mAP, especially when the two sensors are combined. It is able to perform comparably on the Getting in Chair activity.

learning rate of 1e-2 and a weight decay of 1e-6 on batches of 64 frames, and runs for 10 epochs.

5.3 Importance of 3D

The first 6 lines of Table 3 show results using inputs from a single camera. In particular, the first four lines show the comparison between CNN and PointNet by only taking inputs from one of the two sensors. The PointNet method consistently outperforms CNN across all activities and cameras for two reasons. First, processing the scene in 3D space allows better understanding of the relative positions. Secondly and perhaps more importantly, the spatial transformer removes discrepancies caused by the camera positioning and angle which is an important step towards better scene understanding. The reason is that the 3D point cloud is a viewpoint-invariant representation of the room, and this viewpoint normalization helps enhance the model's robustness and generalizability (Jaderberg et al., 2015). As a comparison, PointNet is quite robust under viewpoint change, which is demonstrated by row 5 and 6 where there is only a marginal gain by adding an STN.

5.4 Importance of Multi-View

Table 3 also compares results from single or multi-view inputs, for both CNN and PointNet. The performance gain proves that multi-view inputs are able to provide a more informative representation of the scene than the single view.

A major advantage of having multiple viewpoints is the reduction of occlusion-related ambiguities by combining complementary views. It is worth noting that even though occlusion may be caused by poor camera settings, the problem of occlusion is not likely to be entirely eliminated by a better single camera setup for several reasons. Firstly, the complex setup of ICUs places constraints on where the cameras can be installed, severely reducing the likeliness of having a full coverage of a room with a single camera alone. Secondly, since the ICU patients are usually in a poor condition, multiple caretakers are often involved in each

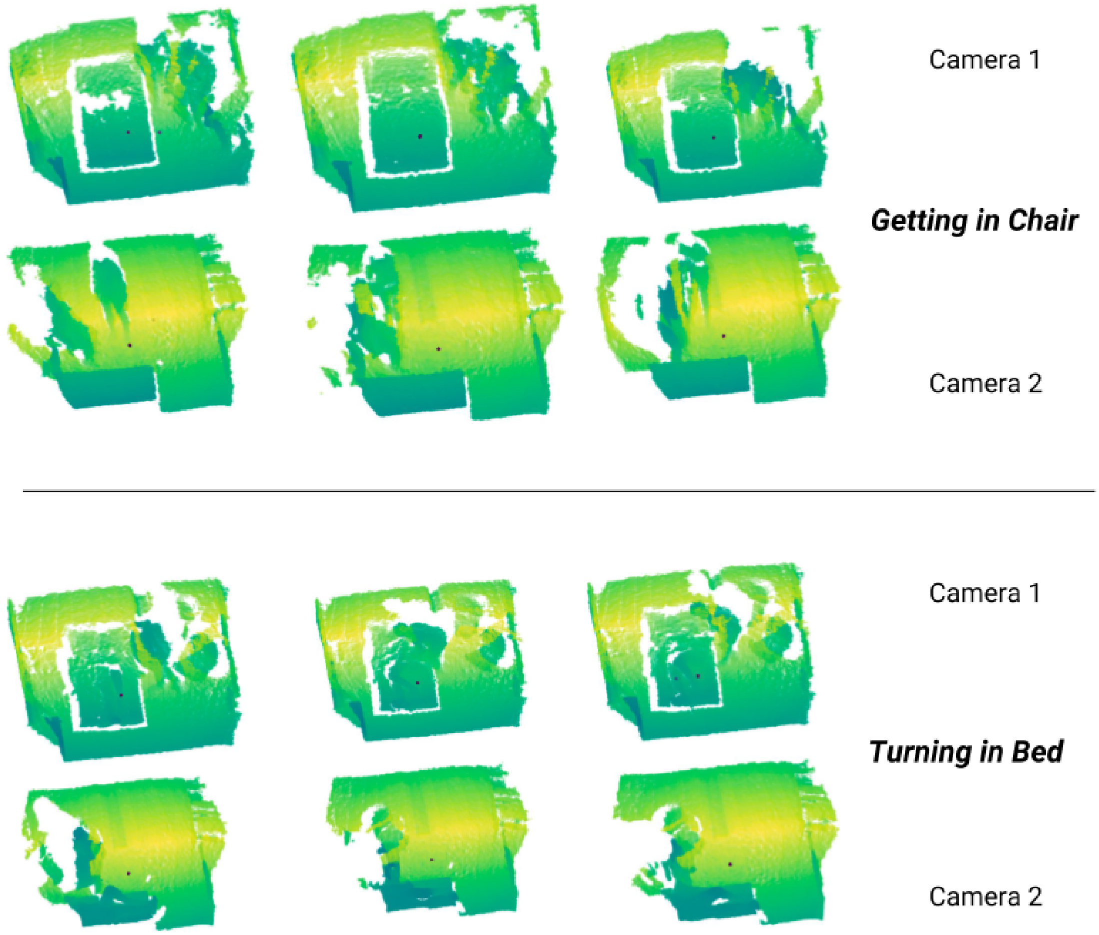


Figure 4: Example correct predictions by our model. The fused point-cloud representations over a sequence of time are shown for each example.

activity in close proximity of the patients, leading to body-to-body occlusion. Thirdly, due to the dynamically changing environment, there is usually no fixed angle at which occlusion may happen. In general, there is a large improvement on handling occlusions by introducing even only one extra viewpoint, which intuitively helps provide a similar level of information that a human annotator would have access to. We also see from experiments that combining results obtained from both cameras can usually help to compensate for the negative effect of occlusion.

5.5 Qualitative Results

Figure 4 shows correct predictions by our model. We show that our model is able to simultaneously reason on two point clouds from different sensors, and correctly predict "Getting in chair" and "Turning in bed" activities. We can see from the visualized point clouds that the viewpoints are indeed complementary, and that out model's capability to learn to combine both 3D geometric inputs for activity understanding is powerful.

6. Conclusion

In this work, we presented a method for detecting challenging patient care activities in ICUs by combining depth data from multiple sensors to form a single 3D point cloud representation. By reasoning on this single shared 3D representation, we demonstrated the effectiveness of our approach using a dataset of mobility-related patient care activities. For future work, we would like to examine the use of calibrated point clouds and incorporate temporal information with recurrent neural networks. We would also like to introduce automated annotations using current NLP techniques and video annotation tools. We believe the method presented in this work demonstrates the potential for computer vision to automatically document and monitor clinically relevant activities in complex healthcare environments such as ICUs.

Acknowledgement

We would like to acknowledge members in the Stanford Partnership in AI-Assisted Care (PAC) for helpful discussions and support in data collection. We would also like to thank the anonymous reviewers for their comments which helped improve the paper.

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.

Michael Goesele, Jens Ackermann, Simon Fuhrmann, Carsten Haubold, Ronny Klowsky, Drew Steedly, and Richard Szeliski. Ambient point clouds for view interpolation. In *ACM Transactions on Graphics (TOG)*, volume 29, page 95. ACM, 2010.

Albert Haque, Michelle Guo, Alexandre Alahi, Serena Yeung, Zelun Luo, Alisha Rege, Jeffrey Jopling, Lance Downing, William Beninati, Amit Singh, et al. Towards vision-based smart hospitals: A system for tracking and monitoring hand hygiene compliance. *Machine Learning for Healthcare*, 2017.

Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. *The International Journal of Robotics Research*, 31(5):647–663, 2012.

TEAM Study Investigators et al. Early mobilization and recovery in mechanically ventilated patients in the icu: a bi-national, multi-centre, prospective cohort study. *Critical Care*, 19 (1):81, 2015.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

YG Jiang, J Liu, A Roshan Zamir, G Toderici, I Laptev, M Shah, and R Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2014.

Andy J Ma, Nishi Rawat, Austin Reiter, Christine Shrock, Andong Zhan, Alex Stone, Anahita Rabiee, Stephanie Griffin, Dale M Needham, and Suchi Saria. Measuring patient mobility in the icu using a novel noninvasive sensor. *Critical care medicine*, 45(4):630–636, 2017.

Ronald Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 1(2):4, 2017.

Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Charles Rosenberg, and Li Fei-Fei. Learning semantic relationships for better action retrieval in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1100–1109, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

William D Schweickert, Mark C Pohlman, Anne S Pohlman, Celerina Nigos, Amy J Pawlik, Cheryl L Esbrook, Linda Spears, Megan Miller, Mietka Franczyk, Deanna Deprizio, et al. Early physical and occupational therapy in mechanically ventilated, critically ill patients: a randomised controlled trial. *Lancet (London, England)*, 373(9678):1874–1882, 2009.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Intelligent Robots and Systems (IROS)*, pages 573–580. IEEE, 2012.

Andru P Twinanda, Emre O Alkan, Afshin Gangi, Michel de Mathelin, and Nicolas Padoy. Data-driven spatio-temporal rgbd feature encoding for action recognition in operating rooms. *International journal of computer assisted radiology and surgery*, 10(6):737–747, 2015.

Daniel Weinland, Remi Ronfard, and Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer vision and image understanding*, 115(2):224–241, 2011.

Mao Ye, Qing Zhang, Liang Wang, Jiejie Zhu, Ruigang Yang, and Juergen Gall. A survey on human motion analysis from depth data. In *Time-of-flight and depth imaging. sensors, algorithms, and applications*, pages 149–187. Springer, 2013.

Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, pages 1–15, 2015.