Published in final edited form as:

Med Image Comput Comput Assist Interv. 2016 October ; 9900: 482–490. doi:
10.1007/978-3-319-46720-7_56.

Process Monitoring in the Intensive Care Unit: Assessing Patient Mobility Through Activity Analysis with a Non-Invasive Mobility Sensor

Austin Reiter¹, Andy Ma¹, Nishi Rawat², Christine Shrock², and Suchi Saria¹

¹The Johns Hopkins University, Baltimore, MD, USA

²Johns Hopkins Medical Institutions, Baltimore, MD, USA

Abstract

Throughout a patient's stay in the Intensive Care Unit (ICU), accurate measurement of patient mobility, as part of routine care, is helpful in understanding the harmful effects of bedrest [1]. However, mobility is typically measured through observation by a trained and dedicated observer, which is extremely limiting. In this work, we present a video-based automated mobility measurement system called *NIMS: Non-Invasive Mobility Sensor*. Our main contributions are: (1) a novel multi-person tracking methodology designed for complex environments with occlusion and pose variations, and (2) an application of human-activity attributes in a clinical setting. We demonstrate NIMS on data collected from an active patient room in an adult ICU and show a high inter-rater reliability using a weighted Kappa statistic of 0.86 for automatic prediction of the highest level of patient mobility as compared to clinical experts.

Keywords

Activity recognition; Tracking; Patient safety

1 Introduction

Monitoring human activities in complex environments are finding an increasing interest [2,3]. Our current investigation is driven by automated hospital surveillance, specifically, for critical care units that house the sickest and most fragile patients. In 2012, the Institute of Medicine released their landmark report [4] on developing digital infrastructures that enable rapid learning health systems; one of their key postulates is the need for improvement technologies for measuring the care environment. Currently, simple measures such as whether the patient has moved in the last 24 h, or whether the patient has gone unattended for several hours require manual observation by a nurse, which is highly impractical to scale. Early mobilization of critically ill patients has been shown to reduce physical impairments and decrease length of stay [5], however the reliance on direct observation limits the amount of data that may be collected [6].

To automate this process, non-invasive low-cost camera systems have begun to show promise [7,8], though current approaches are limited due to the unique challenges common to complex environments. First, though person detection in images is an active research area [9,10], significant occlusions present limitations because the expected appearances of people do not match what is observed in the scene. Part-based deformable methods [11] do somewhat address these issues as well as provide support for articulation, however when combining deformation with occlusion, these too suffer for similar reasons.

This paper presents two main contributions towards addressing challenges common to complex environments. First, using an RGB-D sensor, we demonstrate a novel methodology for human tracking systems which accounts for variations in occlusion and pose. We combine multiple detectors and model their deformable spatial relationship with a temporal consistency so that individual parts may be occluded at any given time, even through articulation. Second, we apply an attribute-based framework to supplement the tracking information in order to recognize activities, such as mobility events in a complex clinical environment. We call this system *NIMS: A Non-Invasive Mobility Sensor*.

1.1 Related Work

Currently, few techniques exist to automatically and accurately monitor ICU patient's mobility. Accelerometry is one method that has been validated [12], but it has limited use in critically ill inpatient populations [6]. Related to multi-person tracking, methods have been introduced to leverage temporal cues [13,14], however hand-annotated regions are typically required at the onset, limiting automation. To avoid manual initializations, techniques such as [15,16] employ a single per-frame detector with temporal constraints. Because single detectors are limited towards appearance variations, [15] proposes to make use of multiple detectors, however this assumes that the spatial configuration between the detectors is fixed, which does not scale to address significant pose variations.

Much activity analysis research has approached action classification with bag-of-words approaches. Typically, spatio-temporal features, such as Dense Trajectories [17], are used with a histogram of dictionary elements or a Fisher Vector encoding [17]. Recent work has applied Convolutional Neural Network (CNN) models to the video domain [18,19] by utilizing both spatial and temporal information within the network topology. Other work uses Recurrent Neural Networks with Long Short Term Memory [20] to model sequences over time. Because the "activities" addressed in this paper are more high-level in nature, traditional spatio-temporal approaches often suffer. Attributes describe high-level properties that have been demonstrated for activities [21], but these tend to ignore contextual information.

The remainder of this paper is as follows: first, we describe our multi-person tracking framework followed by our attributes and motivate their use in the clinical setting to predict mobility. We then describe our data collection protocol and experimental results and conclude with discussions and future directions.

2 Methods

Figure 1 shows an overview of our NIMS system. People are localized, tracked, and identified using an RGB-D sensor. We predict the pose of the patient and identify nearby objects to serve as context. Finally, we analyze in-place motion and train a classifier to determine the highest level of patient mobility.

2.1 Multi-person Tracking by Fusing Multiple Detectors

Our tracking method works by formulating an energy functional comprising of spatial and temporal consistency over multiple part-based detectors (see Fig. 2). We model the relationship between detectors within a single frame using a deformable spatial model and then track in an online setting.

Modeling Deformable Spatial Configurations—For objects that exhibit deformation, such as humans, there is an expected spatial structure between regions of interest (ROIs) (e.g., head, hands, etc.) across pose variations. Within each pose (e.g. lying, sitting, or standing), we can speculate about an ROI (e.g. head) based on other ROIs (e.g. full-body). To model such relationships, we assume that there is a projection matrix $A_{ll'}^c$ which maps the location of ROI l to that of l' for a given pose c . With a training dataset, C types of poses are determined automatically by clustering location features [10], and projection matrix $A_{ll'}^c$ can be learnt by solving a regularized least-square optimization problem.

To derive the energy function of our deformable model, we denote the number of persons in the t -th frame as M^t . For the m -th person, the set of corresponding bounding boxes from L ROIs is defined by $X^t = \{X_1^t(m), \dots, X_L^t(m)\}$. For any two proposed bounding boxes $X_{l'}^t(m)$ and $X_l^t(m)$ at frame t for individual m , the deviation from the expected spatial configuration is quantified as the error between the expected location of the bounding box for the second ROI conditioned on the first. The total cost is computed by summing, for each of the M^t individuals, the minimum cost for each of the C subcategories:

$$E_{spa}(X^t, M^t) = \sum_{m=1}^{M^t} \min_{1 \leq c \leq C} \sum_{l \neq l'} \|A_{ll'}^c X_l^t(m) - X_{l'}^t(m)\|^2 \quad (1)$$

Grouping Multiple Detectors—Next we wish to automate the process of detecting people to track using a combination of multiple part-based detectors. A collection of existing detection methods [11] can be employed to train K detectors; each detector is geared towards detecting an ROI. Let us consider two bounding boxes $D_k^t(n)$ and $D_{k'}^t(n')$ from any two detectors k and k' , respectively. If these are from the same person, the overlapped region is large when they are projected to the same ROI using our projection matrix. In this case, the average depths in these two bounding boxes are similar. We calculate the probability that these are from the same person as:

$$p = ap_{over} + (1-a)p_{depth} \quad (2)$$

where a is a positive weight, p_{over} and p_{depth} measure the overlapping ratio and depth similarity between two bounding boxes, respectively. These scores are:

$$p_{over} = \max \left(\frac{|A_{\ell(k)\ell(k')}^c D_k^t(n) \cap D_{k'}^t(n')|}{\min(|A_{\ell(k)\ell(k')}^c D_k^t(n)|, |D_{k'}^t(n')|)}, \frac{|D_k^t(n) \cap A_{\ell(k')\ell(k)}^c D_{k'}^t(n')|}{\min(|D_k^t(n)|, |A_{\ell(k')\ell(k)}^c D_{k'}^t(n')|)} \right) \quad (3)$$

$$p_{depth} = \frac{1}{2} e^{\frac{-(v_k^t(n) - v_{k'}^t(n'))^2}{2\sigma_k^t(n)^2}} + \frac{1}{2} e^{\frac{-(v_k^t(n) - v_{k'}^t(n'))^2}{2v_{k'}^t(n')^2}} \quad (4)$$

where ℓ maps the k -th detector to the l -th region-of-interest, v and σ denote the mean and standard deviation of the depth inside a bounding box, respectively.

By the proximity measure given by (2), we group the detection outputs into N^t sets of bounding boxes. In each group $G^t(n)$, the bounding boxes are likely from the same person. Then, we define a cost function that represents the matching relationships between the true positions of our tracker and the candidate locations suggested by the individual detectors as:

$$E_{det}(X^t, M^t) = \sum_{n=1}^{N^t} \min_{1 \leq m \leq M^t} \sum_{D_k^t(n') \in G^t(n)} w_k^t(n) \|D_k^t(n') - X_{\ell(k)}^t(m)\|^2 \quad (5)$$

where $w_k^t(n)$ is the detection score as a penalty for each detected bounding box.

Tracking Framework—We initialize our tracker at time $t = 1$ by aggregating the spatial (Eq. 1) and detection matching (Eq. 5) cost functions. To determine the best bounding box locations at time t conditioned on the inferred bounding box locations at time $t - 1$, we extend the temporal trajectory E_{dyn} and appearance E_{app} energy functions from [16] and solve the joint optimization (definition for E_{exc} , E_{reg} , E_{dyn} , E_{app} left out for space considerations) as:

$$\min_{X^t, M^t} \lambda_{det} E_{det} + \lambda_{spa} E_{spa} + \lambda_{exc} E_{exc} + \lambda_{reg} E_{reg} + \lambda_{dyn} E_{dyn} + \lambda_{app} E_{app} \quad (6)$$

We refer the interested reader to [22] for more details on our tracking framework.

2.2 Activity Analysis by Contextual Attributes

We describe the remaining steps for our NIMS system here.

Patient Identification—We fine-tune a pre-trained CNN [24] based on the architecture in [25], which is initially trained on ImageNet (<http://image-net.org/>). From our RGB-D sensor, we use the color images to classify images of people into one of the following categories: *patient*, *caregiver*, or *family-member*. Given each track from our multi-person tracker, we extract a small image according to the tracked bounding box to be classified. By understanding the *role* of each person, we can tune our activity analysis to focus on the patient as the primary “actor” in the scene and utilize the caregivers into supplementary roles.

Patient Pose Classification and Context Detection—Next, we seek to estimate the pose of the patient, and so we fine-tune a pre-trained network to classify our *depth* images into one of the following categories: *lying-down*, *sitting*, or *standing*. We choose depth over color as this is a geometric decision. To supplement our final representation, we apply a real-time object detector [24] to localize important objects that supplement the state of the patient, such as: *bed upright*, *bed down*, and *chair*. By combining bounding boxes identified as people with bounding boxes of objects, the NIMS may better ascertain if a patient is, for example, “lying-down in a bed down” or “sitting in a chair”.

Motion Analysis—Finally, we compute in-place body motion. For example, if a patient is lying in-bed for a significant period of time, clinicians are interested in how much exercise in-bed occurs [23]. To achieve this, we compute the mean magnitude of motion with a dense optical flow field within the bounding box of the tracked patient between successive frames in the sequence. This statistic indicates how much frame-to-frame in-place motion the patient is exhibiting.

Mobility Classification—[23] describes a clinically-accepted 11-point mobility scale (**ICU Mobility Scale**), as shown in Table 1 on the right. We collapsed this into our **Sensor Scale** (left) into 4 discrete categories. The motivation for this collapse was that when a patient walks, this is often performed outside the room where our sensors cannot see.

By aggregating the different sources of information described in the preceding steps, we construct our attribute feature F_t with:

1. Was a patient detected in the image? (0 for *no*; 1 for *yes*)
2. What was the patient’s pose? (0 for *sitting*; 1 for *standing*; 2 for *lying-down*; 3 for *no patient found*)
3. Was a chair found? (0 for *no*; 1 for *yes*)
4. Was the patient in a bed? (0 for *no*; 1 for *yes*)
5. Was the patient in a chair? (0 for *no*; 1 for *yes*)
6. Average patient motion value
7. Number of caregivers present in the scene

We chose these attributes because their combination describes the “state” of the activity. Given a video segment of length T , all attributes $\mathbf{F} = [F_1, F_2, \dots, F_7]$ are extracted and the

mean $\mathbf{F}_\mu = \sum_{t=1}^T \mathbf{F}_t / T$ is used to represent the overall video segment (the mean is used to account for spurious errors that may occur). We then train a Support Vector Machine (SVM) to automatically map each $\mathbf{F}_\mu$ to the corresponding Sensor Scale mobility level from Table 1.

3 Experiments and Discussions

Video data was collected from a surgical ICU at a large tertiary care hospital. All ICU staff and patients were consented to participate in our IRB-approved study. A Kinect sensor was mounted on the wall of a private patient room and was connected to a dedicated encrypted computer where data was de-identified and encrypted. We recorded 362 h of video and manually curated 109 video segments covering 8 patients. Of these 8 patients, we use 3 of them to serve as training data for the NIMS components (Sect. 2), and the remaining 5 to evaluate.

Training—To train the person, patient, pose, and object detectors we selected 2000 images from the 3 training patients to cover a wide range of appearances. We manually annotated: (1) head and full body bounding boxes; (2) person identification labels; (3) pose labels; and (4) chair, upright, and down beds.

To train the NIMS Mobility classifier, 83 of the 109 video segments covering the 5 left-out patients were selected, each containing 1000 images. For each clip, a senior clinician reviewed and reported the highest level of patient mobility and we trained our mobility classifier through leave-one-out cross validation.

Tracking, Pose, and Identification Evaluation—We quantitatively compared our tracking framework to the current SOTA. We evaluate with the widely used metric MOTA (Multiple Object Tracking Accuracy) [26], which is defined as 100 % minus three types of errors: false positive rate, missed detection rate, and identity switch rate. With our ICU dataset, we achieved a MOTA of 29.14 % compared to −18.88 % with [15] and −15.21 % with [16]. Using a popular RGBD Pedestrian Dataset [27], we achieve a MOTA of 26.91 % compared to 20.20 % [15] and 21.68 % [16]. We believe the difference in improvement here is due to there being many more occlusions in our ICU data compared to [27]. With respect to our person and pose ID, we achieved 99 % and 98 % test accuracy, respectively, over 1052 samples. Our tracking framework requires a runtime of 10 secs/frame (on average), and speeding this up to real-time is a point of future work.

Mobility Evaluation—Table 2 shows a confusion matrix for the 83 video segments to demonstrate the inter-rater reliability between the NIMS and clinician ratings. We evaluated the NIMS using a weighted Kappa statistic with a linear weighting scheme [28]. The strength of agreement for the Kappa score was qualitatively interpreted as: 0.0–0.20 as *slight*, 0.21–0.40 as *fair*, 0.41–0.60 as *moderate*, 0.61–0.80 as *substantial*, 0.81–1.0 as *perfect* [28]. Our weighted Kappa was 0.8616 with a 95 % confidence interval of (0.72, 1.0). To compare to a popular technique, we computed features using Dense Trajectories [17] and trained an SVM (using Fisher Vector encodings with 120 GMMs), achieving a weighted Kappa of 0.645 with a 95 % confidence interval of (0.43, 0.86).

The main source of difference in agreement was contained within differentiating “A” from “B”. This disagreement highlights a major difference between human and machine observation in that the NIMS is a computational method being used to distinguish activities containing motion from those that do not with a quantitative, repeatable approach.

4 Conclusions

In this paper, we demonstrated a video-based activity monitoring system called NIMS. With respect to the main technical contributions, our multi-person tracking methodology addresses a real-world problem of tracking humans in complex environments where occlusions and rapidly-changing visual information occurs. We will continue to develop our attribute-based activity analysis for more general activities as well as work to apply this technology to rooms with multiple patients and explore the possibility of quantifying patient/provider interactions.

References

1. Brower R. Consequences of bed rest. *Crit Care Med.* 2009; 37(10):S422–S428. [PubMed: 20046130]
2. Corchado J, Bajo J, De Paz Y, Tapia D. Intelligent environment for monitoring Alzheimer patients, agent technology for health care. *Decis Support Syst.* 2008; 44(2):382–396.
3. Hwang J, Kang J, Jang Y, Kim H. Development of novel algorithm and real-time monitoring ambulatory system using bluetooth module for fall detection in the elderly. *IEEE EMBS.* 2004
4. Smith, M., Saunders, R., Stuckhardt, K., McGinnis, J. *Best Care at Lower Cost: the Path to Continuously Learning Health Care in America.* National Academies Press; Washington, DC: 2013.
5. Hashem M, Nelliott A, Needham D. Early mobilization and rehabilitation in the intensive care unit: moving back to the future. *Respir Care.* 2016; 61:971–979. [PubMed: 27094396]
6. Berney S, Rose J, Bernhardt J, Denehy L. Prospective observation of physical activity in critically ill patients who were intubated for more than 48 hours. *J Crit Care.* 2015; 30(4):658–663. [PubMed: 25813549]
7. Chakraborty, I., Elgammal, A., Burd, R. Video based activity recognition in trauma resuscitation. *International Conference on Automatic Face and Gesture Recognition;* 2013.
8. Lea, C., Facker, J., Hager, G., et al. 3D sensing algorithms towards building an intelligent intensive care unit. *AMIA Joint Summits Translational Science Proceedings;* 2013.
9. Dalal N, Triggs B. Histograms of oriented gradients for human detection. *IEEE CVPR.* 2005
10. Chen X, Mottaghi R, Liu X, et al. Detect what you can: detecting and representing objects using holistic models and body parts. *IEEE CVPR.* 2014
11. Felzenszwalb P, Girshick R, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. *PAMI.* 2010; 32(9):1627–1645.
12. Verceles A, Hager E. Use of accelerometry to monitor physical activity in critically ill subjects: a systematic review. *Respir Care.* 2015; 60(9):1330–1336. [PubMed: 25852167]
13. Babenko D, Yang M, Sivic S. Robust object tracking with online multiple instance learning. *PAMI.* 2011; 33(8):1619–1632.
14. Lu Y, Wu T, Zhu S. Online object tracking, learning and parsing with and-or graphs. *IEEE CVPR.* 2014
15. Choi W, Pantofaru C, Savarese S. A general framework for tracking multiple people from a moving camera. *PAMI.* 2013; 35(7):1577–1591.
16. Milan A, Roth S, Schindler K. Continuous energy minimization for multi-target tracking. *TPAMI.* 2014; 36(1):58–72.
17. Wang H, Schmid C. Action recognition with improved trajectories. *IEEE ICCV.* 2013

18. Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks. IEEE CVPR. 2014
19. Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. NIPS. 2014
20. Wu Z, Wang X, Jiang Y, Ye H, Xue X. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. ACMMM. 2015
21. Liu J, Kuipers B, Savarese S. Recognizing human actions by attributes. IEEE CVPR. 2011
22. Ma AJ, Yuen PC, Saria S. Deformable distributed multiple detector fusion for multi-person tracking. 2015 arXiv:1512.05990 [cs.CV].
23. Hodgson C, Needham D, Haines K, et al. Feasibility and inter-rater reliability of the ICU mobility scale. Heart Lung. 2014; 43(1):19–24. [PubMed: 24373338]
24. Girshick R. Fast R-CNN. 2015 arXiv:1504.08083.
25. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. NIPS. 2012
26. Keni B, Rainer S. Evaluating multiple object tracking performance: the CLEAR MOT metrics. EURASIP J Image Video Proces. 2008; 2008:1–10.
27. Spinello L, Arras KO. People detection in RGB-D data. IROS. 2011
28. McHugh M. Interrater reliability: the Kappa statistic. Biochemia Med. 2012; 22(3):276–282.

Fig. 1.

Flowchart of our mobility prediction framework. Our system tracks people in the patient's room, identifies the "role" of each ("patient", "caregiver", or "family member"), relevant objects, and builds attribute features for mobility classification.

Fig. 2.

Full-body (red) and Head (green) detectors trained by [11]. The head detector may fail with (a) proximity or (d) distance. The full-body detector may also struggle with proximity [(b) and (c)]. (To protect privacy, all images are blurred). (Color figure online)

Table 1

Table comparing our **Sensor Scale**, containing the 4 discrete levels of mobility that the NIMS is trained to categorize from a video clip of a patient in the ICU, to the standardized **ICU Mobility Scale** [23], used by clinicians in practice today.

Sensor Scale	ICU Mobility Scale
A. Nothing in bed	0. Nothing (lying in bed)
B. In-bed activity	1. Sitting in bed, exercises in bed
C. Out-of-bed activity	2. Passively moved to chair (no standing)
	3. Sitting over edge of bed
	4. Standing
	5. Transferring bed to chair (with standing)
	6. Marching in place (at bedside) for short duration
D. Walking	7. Walking with assistance of 2 or more people
	8. Walking with assistance of 1 person
	9. Walking independently with a gait aid
	10. Walking independently without a gait aid

Table 2

Confusion matrix demonstrating clinician and sensor agreement.

	A. Nothing	B. In-Bed	C. Out-of-Bed	D. Walking
A. Nothing	18	4	0	0
B. In-Bed	3	25	2	0
C. Out-of-Bed	0	1	25	1
D. Walking	0	0	0	4

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript