A Multiview Multimodal System for Monitoring Patient Sleep

Carlos Torres, *Student Member, IEEE,* Jeffrey C. Fried, *Fellow, AMA*
Kenneth Rose, *Fellow, IEEE* and B. S. Manjunath, *Fellow, IEEE*

***Abstract*—Clinical observations indicate that during critical care at the hospitals, a patient's sleep positioning and motion have a significant effect on recovery rate. Unfortunately, there is no formal medical protocol to record, quantify, and analyze motion of patients. There are very few clinical studies that use manual analysis of sleep poses and motion recordings to support medical benefits of patient positioning and motion monitoring. Manual processes do not scale, are prone to human errors, and put strain on an already taxed healthcare workforce. This study introduces Multimodal, Multiview Motion Analysis and Summarization for healthcare (MASH). MASH is an autonomous system, which addresses these issues by monitoring healthcare environments and enabling the recording and analysis of patient sleep-pose patterns. MASH uses three RGB-D cameras to monitor patients in a medical Intensive Care Unit (ICU) room. The proposed algorithms estimate pose direction at different temporal resolutions and use keyframes to efficiently represent pose transition dynamics. MASH combines deep features computed from the data with a modified version of Hidden Markov Model (HMM) to flexibly model pose duration and summarize patient motion. The performance is evaluated in ideal (BC: Bright and Clear/occlusion-free) and natural (DO: Dark and Occluded) scenarios at two motion resolutions and in two environments: a mock-up and a medical ICU. The usage of deep features is evaluated and their performance compared with engineered features. Experimental results using deep features in DO scenes increases performance from 86.7% to 93.6%, while matching the classification performance of engineered features in BC scenes. The performance of MASH is compared with HMM and C3D. The overall over-time tracing and summarization error rate across all methods increased when transitioning from the mock-up to the the medical ICU data. The proposed keyframe estimation helps achieve a 78% transition classification accuracy.**

***Keywords*—Healthcare, sleep poses, multimodal sensor network, ICU monitoring, patient motion analysis, summarization, hidden markov models, time-series motion interference, M.A.S.H.**

I. Introduction

While receiving care in hospitals, ICU patients are continuously monitored by healthcare staff; however, there are no clinical procedures to reliably analyze and understand pose variations from observations (e.g., videos) or the effects of time-based pose patterns on patient health. The recovery of ICU patients varies largely and often inexplicably [8], even for patients with similar initial health conditions. A small number of clinical studies [21] suggests that patient therapies based on body positioning and controlled motion can enhance patient recovery, while inadequate positioning can have negative effects and aggravate patient health. This study attempts to address this crucial healthcare deficiency by introducing MASH's algorithms and multimodal multiview (*mm*) camera network. MASH combines keyframes extracted from *mm* video data with Hidden Semi-Markov Models (HSMM) to represent poses, analyze motion patterns, and model pose duration. The MASH summarization methods enable the following healthcare applications: (1) methods to estimate rate and range of motion to aid the analysis and prevention of bed sores (long periods of time); (2) tools for the analysis of erratic and distressed motion (short periods of time) that can be used to prevent patients from, for example, falling off the bed or accidentally removing intra-venous lines; and (3) historical summarization of pose sequences (short and long periods time) to unobtrusively evaluate sleep hygiene.

The MASH architecture analyzes input videos from multiviews and modalities to deal with variable scene conditions from a purely observation approach. Motion quantization is performed to remove depth's sensor noise and threshold observable levels of detectable motion. After noise and motion thresholding, features are extracted to represent the various poses and pseudo or transitory poses (deep and/or engineered features). MASH uses keyframes because collecting, storing, and processing video data from the six sources becomes a hefty task on its own. This problem is more manageable using keyframes across all views and modalities, which can be considered as the frames that are informative and discriminant (i.e., pose and pseudo-pose centroids). Pose patterns and pose transitions can span seconds, minutes, or hours, so we use a modified HMM to flexibly model state or pose duration. Finally, the summary can tell us whether the observation was a sequence of poses seen over an extended period of time (i.e., hours) or the same sequence of poses a transition (i.e., seconds). With these considerations, the workflow shown in Fig 1 consists of six major blocks: (1) data collection regarding sleep poses and pose transitions; (2) motion thresholding, which uses optic flow vectors to remove noise from the depth cameras and subtlety distinguish between small and large movements; (3) features extracted from the last layer of the Inception architecture [34] to represent body configurations as a numerical vector, (4) keyframe extraction to identify pseudo poses that best represent a transition; (5) time-series analysis via HSMM to identify the most likely sequence and model pose duration; and (6) output summary.

The performance of MASH is evaluated in ideal (BC:

C. Torres and K. Rose, and B.S. Manjunath are with the Electrical and Computer Engineering Department, University of California, Santa Barbara; J. C. Fried is with the Medical ICU at Santa Barbara Cottage Hospital.

Submitted Aug, 2017; revised Dec, 2017 and Feb 2018

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2829162, IEEE Transactions on Multimedia

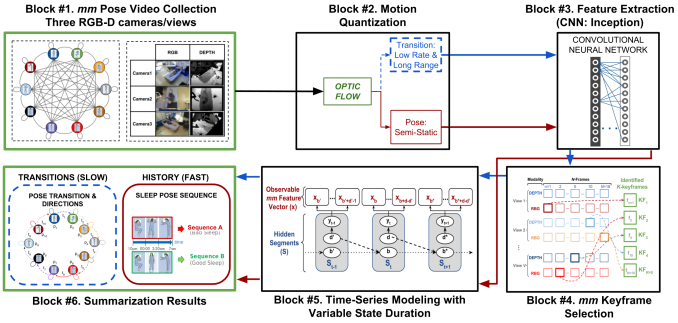


Fig. 1. MASH framework blocks. The process starts with Block #1 and flows clock wise: data collection, motion threshold, deep feature extraction, mm keyframe selection, time-series modeling, and inferred summarized results.

Bright and Clear/occlusion-free) and natural (DO: Dark and Occluded) scenarios in a mock-up and a medical ICU using two motion speeds (fast and slow). Experimental results indicate that using Inception features [34] to represent poses helps MASH match the static classification performance of engineered features in BC scenes, and increases the classification performance in DO scenes from 86.7% using engineered features to 93.6% by using Inception features (an additional 7%) in the mock-up ICU. Pose history summarization shows that the average MASH trace overlap is 83.2% in the mock-up ICU and 80.1% in the medical ICU, which approximately doubles the performance of using conventional HMM. Experimental results indicate that the proposed multimodal multiview keyframe estimation algorithm achieves a mean transition classification accuracy of 78% using five keyframes (or pseudo poses). The keyframe approach avoids using complete videos and provides robustness to variation in transition speeds.

Medical Background. Harvard Medical School reported in August 2016 that monitoring ICUs can save up to $15 billion by saving $20,000 in each of the 750,000 ICU beds in the U.S. by reducing the effect of preventable ICU-related conditions such as poor quality of sleep and decubitus ulcers (DUs) [23]. For instance, ICUs in the U.S. receive about five million patients per year, each with an average stay of 9.3 days and with a mortality rate that ranges from 10 to 30% depending on health conditions. MASH sample applications focus on developing solutions to help understand and address sleep analysis and incidence of DUs. These applications are selected due to their pervasive nature in medical ICUs and the opportunity to improve the quality of care provided to patients. For example, sleep hygiene is correlated to shorter hospital stays, increased recovery rates, and decreased mortality rates. The findings of [3], [13], and [42] correlate sleep positions with sleep hygiene and its various effects on patient health. DUs are preventable, soft tissue wounds that appear on bony areas of the body and are caused by continuous decubitus positions. There is little understanding about the set of poses and pose duration that cause or prevent DUs. MASH enables the inception of required clinical studies that analyze pose duration, rotation frequency and range, and the duration of weight/pressure off-loading, as well as serving as the non-obtrusive measuring tool to collect and analyze pose patterns.

Technical Background. The analysis of human motion dynamics has captured the attention of researchers in the engineering and health communities. In particular, the ailing healthcare system in the U.S. continues to degrade. This degradation requires that engineers and health professionals join forces to develop new efficient therapies and optimize care techniques and workflows. The latest techniques using convolutional neural network (CNN) architectures achieve impressive classification performance. However, CNN-based techniques require large data sets [2], [4], [39], and [41]. In [28], the authors introduced a CNN-alternative method for action representation via sequential deep trajectory descriptors. The previously cited works recognize actions centered on the camera plane. An exception is the work from [32], which uses stationary cameras and allows off-center actions and is limited to scenes with good illumination that are clear of occlusions (i.e., BC). A supervised method for learning local feature descriptors is introduced in [48]. Best practices for human action and activity recognition are outlined in [44] and [25] with benchmarks described in [16]. The spatio-temporal evolution of features for action recognition is explored in [17] and [15]. A multimodal bilinear method for person detection is explored in [40]. Although these methods are innovative, they tackle conventional activity and action motion dynamics observed, for example, in walking and running, making them inadequate for sleep-pose pattern analysis. Sleep-pose patterns are different; they are subtle, non-continuous, non-sequential, and abrupt. Although effective, the method requires controlled scenes, which are not possible in healthcare. A discriminative multi-instance multitask method to recognize actions in 3D spaces is proposed in [46]. However, the proposed method is unable to distinguish between similar actions, for which their only distinction is their duration. The ICU scenes and bed setting disqualify techniques based on skeletal estimation and tracking [1] and pure RGB data for human body orientation [18]. Although promising, the work described in [19] is limited by partial occlusions and challenging ICU bed configurations, which are tackled using multimodal multiview data.

Analysis of realistic human motion is a challenging problem with intra-class and inter-class variations and similarities that require deep appearance and kinetic analysis [31]. Also, the summarization via camera networks enables systems to represent and analyze environments from multipleviews via hypergraphs [33], motion patterns represented as salient motifs [5] and using graphs [45]. These methods, however, are limited to smooth sequential motion in scenes with relatively good illumination and cannot be applied to the ICU. The work in [30] surveys multimedia methods for large-scale data retrieval and classification using multimedia data. The objective of the survey is to highlight an in-depth understanding of multimedia methods for data analysis and understanding. This will be relevant as more data is collected by MASH. A true multimedia method to summarize events in videos via audio, visual, and textual saliency is introduced in [6], and a multiview method for surveillance video summarization via sparse optimization are presented in [24]. Although interesting, these methods analyze motion dynamics with less subtlety than the motion of patients in the ICU. Also, these studies analyze scenes with better illumination and are not representative of

the ICU environment. In addition, multimedia methods may expect speech or text information as input, which cannot be recorded in the ICU (or hospital space). These infrastructural and privacy limitations thwart the implementation and deployment of the existing methods in healthcare applications. The studies from [11] and [43] use multiview systems and methods for smart environments. Unfortunately, these methods require modifications to existing infrastructure. These studies are limited to ideal scenes because they cannot overcome illumination variations and occlusions. They do not account for subtle motion, which can be non-uniform and non-sequential; therefore, these cannot be deployed in a medical ICU.

The authors from [12] introduced an RGB-pressure system for sleep pose classification. Their technique uses geometric features to represent poses extracted from the pressure array and the static RGB image. However, the system requires complex calibration and a top clear view of the patient's body configuration. Pose classification is also tackled in [38] using RGB, depth, and pressure sensors in simulated healthcare environments. The authors combine RGB, depth, and pressure modalities with room sensors to weight modality reliability. The study in [9] uses bed aligned maps (BAMs) composed of pressure arrays and a single depth camera. Although the BAMs method outperforms previous static sleep pose classification techniques, it does not consider motion. The authors from [36] use convex coupled-constrained least-squares optimization to remove the cumbersome pressure array and create a purely observational system. This latest technique increased the classification accuracy by integrating multimodal sources from multiple views and creating a truly multiview multimodal sleep pose classification system. Unfortunately, no previous method incorporates time to analyze the sequence of poses, pose transition, or pose motion dynamics. The work in [22] tackles a rehabilitation application via pose detection and tracking; however, its applications are limited to ideal scenarios.

Previous Work. In [37] we introduced the time-series representation of sleep-pose patterns using HHMs and deep features to represent sleep poses. Although this improves the static pose classification, the methods are limited by lack of flexibility in modeling state duration and the inability to identify key poses across multiple modalities and views. MASH addresses these limitations by introducing a flexible framework to model state duration using time segments and HMM-modified inference. In addition, MASH introduces a keyframe algorithm to identify discriminant and informative frames (i.e., pseudo-poses), which replaces the conventional K-means method used in [37], and improves the overall summarization performance.

Extensive literature search indicates that MASH and its contributions may be the first of its kind. It analyzes patients' sleep-pose patterns and motion dynamics in a simulated and a medical ICU. Also, it observes the environment from multiple modalities and multiple views to account for challenging natural scene conditions. Two distinctive aspects of MASH include incorporation of variable time information and ability to deal with subtle motion patterns using principled statistics.

Proposed Approach. MASH is a new multimodal multiview framework to monitor patients in healthcare environments independent of motion rate and range. Its elements include a multimodal multiview (*mm*) data collection camera network, a *mm* keyframe extraction algorithm, and a *mm* time-series analysis algorithm to model variable pose duration and distinguish between sleep poses and transitory (or pseudo) poses. The views and modalities are shown in Fig. 1 Block #1 with sample motion summaries shown in Block #6. The two resolutions are based on two of the most common ICU conditions: sleep hygiene and DU analysis. Pose history summarization is the coarser resolution. It provides a pictorial representation of poses over time. The applications of the pose history include prevention and analysis of DUs and analysis of sleep-pose effects on quality of sleep. The pose transition summarization is the finer resolution. MASH looks at the pseudo-poses that occur while a patient transitions between two poses. Applications of pose transition summarization include analyzing and quantifying physical therapy and distressed sleep motion quantification and analysis.

Contributions. The technical contributions of MASH are:

1) An adaptive framework capable of monitoring patient motion at various resolutions.
2) A non-disruptive and non-obtrusive monitoring system robust to natural healthcare scenarios and conditions such as variable illumination and partial occlusions.
3) An algorithm that effectively compresses sleep pose transitions using a subset of the most informative and most discriminative keyframes.
4) A fusion technique to incorporate observations from multiple modalities and views (complementary data) into emission probabilities to estimate intermediate poses and pose transitions over time.

Organization of the Manuscript. The MASH system components are described in Section II, which includes modalities, views, and temporal characteristics. Section III describes the protocols for data collection and feature extraction and selection. The description of the problem including the temporal analysis, inference, and keyframe extraction procedures are discussed in Section IV. Thorough experimental results regarding the historical summarization of poses (coarser motion resolution) and the rate and range of motion during pose transitions (finer motion resolution) are shown in Section V. Conclusion and future work are discussed in Section VI. Supplemental materials including: larger figures, datasets, and deployment details can be found online at vision.ece.ucsb.edu.

II. The MASH System

The MASH system is composed of three nodes. They are battery powered, enclosed by aluminum cases, controlled by Raspberry Pi3 [35] ARM-computers running Ubuntu 16.04 (to record video using a Carmine RGB-D sensor), and synchronized using TCP/IP communication, which are shown in Fig. 2.

Multiple Modalities (Multimodal). Multimodal studies use complementary modalities to classify static sleep poses in natural ICU scenes with large variations in illumination and occlusions. MASH leverages the findings from [36] and [27] as evidence of the benefits of multimodal systems. The RGB and Depth views are shown in Fig. 1 Block #1.

Fig. 2. Elements in one MASH node: Raspberry Pi3 B+, Carmine RGB-Depth sensor, 24000 mAh battery, and aluminum enclosure used to deploy MASH in the mock-up and the medical ICU rooms.

a) RGB (R): Standard video data provides information to represent and classify human sleep poses in scenes with relatively ideal conditions. However, most people sleep in imperfectly illuminated scenarios, using sheets, blankets, and pillows that block and disturb sensor measurements. The system collects RGB frames of dimensions 640×480 pixels. Pose appearance features representing human body configurations are extracted from these videos in BC and DO scenes.

b) Depth (D): Infrared depth cameras are resilient to illumination changes. The MASH sensor network uses Primesense's Carmine devices to collect depth data. The devices acquire images of dimensions 640×480 and use 16 bits to represent pixel intensity values, which correspond to the distance from the sensor to a point in the scene. Their operating distance range is 0.8 m to 3.5 m; and their spatial resolution for scenes 2.0 m away is 3.5 mm for the horizontal (x) and vertical (y) axes, and 30 mm along the depth (z) axis. The system uses the depth images to represent the 3-dimensional shape of the poses. However, depth information alone is not sufficient since it requires depth contrast, which is negatively affected by the deformation properties of mattresses, pillows, and blankets in the ICU. The work in [47] surveys methods using depth cameras for semi-controlled scenarios.

Multiple Views (Multiview). The studies from [36] and [27] show that analyzing actions from multiple views and multiple orientations greatly improves detection. These studies indicate that the analysis of multiple views yield algorithms, which are independent of view and orientation. The positions of the cameras in the medical ICU are shown in Fig. 4. (see Fig. **??** in Appendix **??** for the mock-up ICU views and node locations).

Time Analysis. ICU patients move subtly and slowly, very different from active motions like jumping or walking, which are easier to detect. MASH effectively monitors subtle and abrupt patient motion by breaking the motion cues into segments to flexibly model pose and pseudo-pose duration. The variable pose duration is modeled via HSMM, which uses segments and is derived from conventional HMM.

Motion Quantization The optic flow estimation is computed using the OpenCV [14] implementations of Lucas-Kanade [20] and Farneback [7]. Implementation and experimental results indicate that Lucas-Kanade led to faster results, while Farneback's led to higher accuracy in the detection of the most subtle pose transitions. Such pose transition is observed when transitioning from the left-yearner to the left-log positions without rotating. The two poses and their transition are shown on the bottom row of Fig. 6 in green. From left to right, the second and third pose are yearner-left and log-left.

Inception CNN Feature Extraction Deep feature extraction of using Google's Inception architecture required sizing the frames the appropriate image dimensions of 224×224 pixels. The offline analysis and approach uses Inception features due to the infrastructure restrictions, which prohibit the use large computation equipment. The deployed RPi3-based system cannot compute Inception features. Instead, the deployed system uses the online feature extraction method from [36].

III. The MASH Dataset

The MASH dataset is collected from two environments: the mock-up ICU with views shown in Fig. 1 Block #1 and the medical ICU with views shown in Fig. 4. The fully annotated dataset will be available online to researchers. The real patient data is not controlled and only annotated after the fact. Fig. 3 shows the observed counts of poses in number of minutes. Fig.5 shows the counts of pose transitions observed in the medical ICU room. The cell colors indicate the transition is not applicable (marked N/A), the transition includes no rotation (gray), includes left (orange), or includes right (green) rotation.

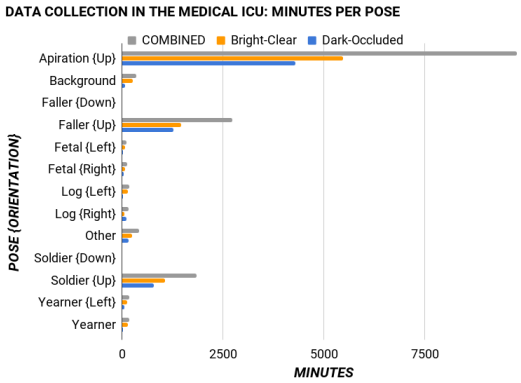


Fig. 3. Number of minutes for poses recorded in the medical ICU.

The mock-up ICU. This room allows researchers to collect static and dynamic data, design and test algorithms, and evaluate and refine the MASH system and algorithms.

1) Poses Static Data: The mock-up sequence is set at random. All actors in the mock-up ICU are asked to assume and hold each of the poses while videos are recorded. The combination of two separate recording sessions of six actors (three female and three male) yield a total of 24 sessions: 12 for BC and 12 for DO scene conditions. Each pose is recorded for one minute, which makes each session 10 minutes long.

2) Pose Transitions Data: The actors start in the initial pose and transition towards a final pose by rotating left or right. This processes is repeated for all initial poses and until all possible combinations between initial and final poses are covered. The combination of ten poses, with two possible transition rotations each generates a set of 20 sequences for each initial pose. Each recording session includes ten initial poses and ten final poses; therefore, each transition recording session generates 200 sequence pairs. A sample transition sequence with left and right rotation directions is shown in Fig. 6. The initial and final poses are Faller Up (*falU*) and Fetal Left (*fetL*), respectively. The top sequence (orange) shows the left rotation and the bottom sequence shows the right (green) rotation. A

small ($\leq 180^o$) rotation or a large ($> 180^o$) rotation are the possible transitions between the poses.

The medical ICU. The battery operated MASH network is currently deployed in a local community hospital where it is used to collect ICU data. The ICU patient dataset is thoroughly anonymized to protect the privacy of patients and medical staff. The dataset includes the video recordings of five consenting patients from periods of time that range from one to five days.

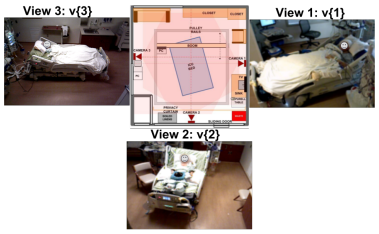


Fig. 4. MASH node locations and views of the patient in the medical ICU.

IN PLACE: NO ROTATION	FINAL POSES																					
LEFT ROTATION	Aspiration		Faller U		Faller D		Fetal L		Fetal R		Log L		Log R			Soldier D		Soldier U		Yearner L		Yearner R	
RIGHT ROTATION	LEFT	RIGHT	LEFT	RIGHT	LEFT	RIGHT	LEFT	RIGHT	LEFT	RIGHT	LEFT	RIGHT	LEFT	RIGHT	OTHER	LEFT	RIGHT	LEFT	RIGHT	LEFT	RIGHT	LEFT	RIGHT
INITIAL POSE: Apiration	N/A		8				2	N/A	N/A	3	2	N/A	N/A	3	2			8	5	N/A	N/A	4	
INITIAL POSE: Faller D	N/A						N/A				N/A							N/A					
INITIAL POSE: Faller U	7		N/A	N/A			4	N/A	N/A	3	3	N/A	N/A	2	4			N/A	N/A	0	N/A	N/A	0
INITIAL POSE: Fetal L	N/A	3	N/A	4			N/A		N/A	2	4	N/A	N/A	0	0			N/A	2	3	N/A	N/A	0
INITIAL POSE: Fetal R	2	N/A	3	N/A			1	N/A		N/A	0	N/A	3		1			0	N/A	0	N/A	5	
INITIAL POSE: Log L	N/A	5	N/A	5			3		N/A	1	2	N/A	N/A	0	2			N/A	3	4	N/A	5	
INITIAL POSE: Log R	2	N/A	4	N/A			5	N/A	2	1	N/A	N/A			0			1	N/A	1	N/A	3	
INITIAL POSE: OTHER	2	4	2	3				N/A							3			3	3				
INITIAL POSE: Soldier D	N/A						N/A											N/A					
INITIAL POSE: Soldier U	6		4					N/A	N/A	2	4	N/A	N/A	2	3			N/A	N/A	0	N/A	N/A	0
INITIAL POSE: Yearner L	N/A	4	N/A	2			6	N/A	0	5	N/A	1	0					N/A	0	N/A	N/A	N/A	0
INITIAL POSE: Yearner R	3	N/A	1	N/A				N/A	5	0	N/A	5	0					0	N/A	0	N/A	N/A	

Fig. 5. Pose transition count recorded by MASH from the medical ICU. The cell colors indicate the transition is not applicable (labeled N/A), the transition has no rotation (gray), left rotation (orange), or right rotation (green).

MASH Feature Extraction and Validation. The methods from [10] are used to calibrate the cameras prior to background subtraction and feature extraction. The background extraction stage detects the bed using the depth modality (i.e., largest square). The four corners of the *depth-bed* serve to estimate the perimeter and surface plane elements, which are then used to crop the camera views and remove the background as in [38]. Camera-based sleep pose classification studies commonly use geometric moments (gMOMs) and histograms of oriented gradients (HOG) to represent poses. Feature extraction of gMOM and HOG features is based on the parameters from [38]. Pose classification results (see Section V) suggest that using Inception [34] outperforms gMOMs, HOG, and VGG [29] features in pose representation and classification.

IV. The MASH Problem

In order to effectively analyze patient motion, the MASH system and algorithms need to properly handle both motion rates (speed) and motion range (rotation angle). The initial assumption for all video frames is that they belong to pose transitions (pseudo-poses), but if the motion rate is identified as slow, these frames can be used to identify true poses, which are needed to identify pose histories (i.e., the sequence of poses). The pose transition involves identifying the set of pseudo poses representing a transition between two poses, and it quantifies the direction of rotation. The first challenge arises because conventional algorithms are unable to model pose duration effectively. The second challenges involves detecting the direction of rotation when transitioning between poses. The last challenge involves representing pseudo poses, for which MASH uses keyframe estimation. The M multimodal cameras are stationed at different locations to obtaining V views of the patients as shown in Fig. 4 and estimate the pose transition dynamics, such as the ones in Fig. 6.

The features extracted from video frames $\mathcal{F} = \{f_t\}$, for $1 \leq t \leq T$ to construct feature vectors $\mathbf{X} = X_{1:T}$ are used to represent non-directly observable poses ($\mathbf{Y} = Y_{1:T}$). The first objective of MASH is to find the sequence of poses ($\mathbf{Y} = Y_{1:T}$) that probabilistically can best represent the observations, as in: $\Pr\left(\mathbf{Y}, \mathbf{X}\right) = \Pr\left(Y_{1:T}, X_{1:T}\right)$. Temporal patterns caused by sleep-pose transitions are simulated and analyzed using Hidden Semi-Markov Modeling (HSMMs) technique, which is described in Section IV-B. The interactions between the modalities for accurate pose representation are encoded into the emission probabilities. Scene conditions are encoded into the set of states (the analysis of two scenes doubles the number of poses). Conventional Markov assumptions support MASH and ideally fit most of its analysis. However, HMMS are limited in their ability to distinguish between poses and pseudo-poses based on pose duration. This is because, by design, HMMs model the probability of staying in a given pose as a geometric distribution $\Pr_i(d) = (a_{ii})^{d-1}(1 - a_{ii})$, where d is the duration in pose i, and a_{ii} is the self-transition probability of pose i. More details are discussed in Sections IV-A and IV-B. Table I describes the variables used in MASH.

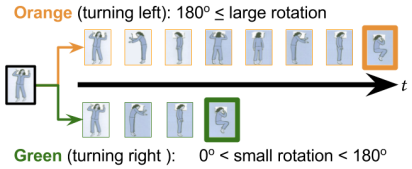


Fig. 6. Pose transitions require patients to reconfigure their body. Transitioning from the faller facing up (*falU*) position to the fetal laying on the left (*fetL*) position. The transition is achieved by either a long rotation (180^o; top row) or by a short rotation ($0 - 180^o$; bottom row).

A. Hidden Markov Models (HMMs)

HMM is a generative modeling approach that represents pose history and transitions as states. The hidden variable or state at time step k (i.e., $t = k$) is y_k (state$_k$ or pose$_k$) and the observable or measurable variables ($x^{(v)}_{k,m}$, the vector of image features corresponding to the k-th frame, using modality m, and view v) at time $t = k$: x_k such that $x_k = x^{(v)}_{k,m} = \{R_k, D_k, \ldots M_k\}$). The Markovian assumptions indicate that at t, the hidden variable y_t, depends only on the previous hidden variable y_{t-1}, and at t the observable variable x_t depends on the hidden variable y_t. These two assumptions are used to compute $\Pr(Y, X)$ given by:

$$\Pr\left(Y_{1:T}, X_{1:T}\right) = \Pr(y_1) \prod_{t=1}^{T} \Pr\left(x_t|y_t\right) \prod_{t=2}^{T} \Pr\left(y_t|y_{t-1}\right), \quad (1)$$

where $\Pr(y_1)$ is the initial state probability distribution ($\mathbf{\Pi}$). It represents the probability of a sequence starting at ($t = 1$) pose$_i$ (state$_i$). $\Pr\left(x_t|y_t\right)$ is the emission probability distribution ($\mathbf{B}$) and represents the probability that at time t, y_i (state$_i$) can generate the observable multimodal multiview vector x_t.

MASH VARIABLES

SYMBOL	DESCRIPTION
$\mathbf{A}$	Transition probability matrix $\mathbf{A} \in \mathbb{R}^{\lvert P\rvert \times \lvert P\rvert}$ and $\mathbf{A} = \{a_{ij}\}$
$a_{i,j}$	Probability of transition from pose i to j
$\mathbf{B}$	Emission probability matrix $\in \mathbb{R}^{\lvert P\rvert}$ and $\mathbf{B} = \{\mu_{in}\}$
b_u	Beginning of the u-th segment with $b_1 = 1$
D_k	k-th frame from the depth modality video
D	Face-Down patient pose
d	Segment duration
d_u	Segment duration for u-th segment
HMM	Abbreviation for Hidden Markov Model
HSMM	Abbreviation for Hidden Semi-Markov Model
K	Data set size, $K = \lvert\mathcal{X}\rvert$
k	Data point index, $1 \leq k \leq K$
KF	Set of sequential keyframes representing a transition between
L	Laying-Left patient pose
l, m, and n	Dummy variables
R_k	k-th frame from the rgb modality video
R	Laying-Right patient pose
μ_i	Probability that state i generates the observation x at time t
π	Initial state probability vector $\in \mathbb{R}^{\lvert}P\rvert$ and $\pi_i \in \pi$
k	The time step index (i.e., $k = t$)
P	Set of patient poses $P = \{p_i\}$
P_{mock}	Set of actor poses in the mock-up ICU room
P_{micu}	Set of patient poses in the real medical ICU (micu) room
$\Pr(Y, X)$	The joint PDF: sequence of states and observations
S	Set of time segments $S = \{s_u\}$ for $1 \leq u \leq U$
s	Segment element $s \in S$
t	Time tick with $1 \leq t \leq T$
τ_{td}	Stores the estimated duration ($1 \leq d \leq D$) at time (t)
θ	HMM model with probabilities $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{\Pi}$
U	Number of segments $U = \lvert S\rvert$
U	Face-Up patient pose
u	Segment index: $1 \leq u \leq U$
$\mathcal{V}$	View set $\mathcal{V} = \{$left, center, right$\}$
V	Number of views $V = \lvert\mathcal{V}\rvert$
v	View index, $1 \leq v \leq V$
y_k	k-th hidden state $y_k \in \mathbf{Y}$
$\mathbf{Y}$	Sequence of hidden states $\lvert\mathbf{Y}\rvert = T$
$\mathcal{X}$	Dataset indexed by k (i.e., $\mathcal{X}_k$)
$\mathcal{X}_k$	k-th datapoint with $\{f_{N_m}\}_k = \{f_R, f_D, f_P\}_k$
x_k	k-th observation feature vector
$\mathbf{x}_{km}^{(v)}$	The k-th observable variable from view v and modality m
δ	Kroenecker delta function
δ_t	The maximum probability duration
θ	Dummy variable used in inference
ζ	Stores the state label (for a pose) of the previous segment
ϕ	Stores the best duration
$\psi_t(i)$	Stores the label with the best duration for state i at t

TABLE I. MASH VARIABLE SYMBOLS AND THEIR DESCRIPTIONS.

Finally, $\Pr\left(y_t|y_{t-1}\right)$ is the transition probability distribution $(\mathbf{A})$ and represents the probability of going from pose$_i$ to pose$_o$ (state i to o). The HMM parameters are $\mathbf{A} = \{a_{ij}\}$, $\mathbf{B} = \{\mu_{in}\}$, and $\mathbf{\Pi} = \{\pi_i\}$, which are standard to HMM.

Modeling Limitations of HMM. One critical limitation of HMM is its rigidity to model state duration. For instance, given an HMM in a state i (pose or transition), the probability that it stays there for d time slices is: $\Pr_i(d) = (a_{ii})^{d-1}(1 - a_{ii})$, where $\Pr_i(d)$ is the discrete probability density function (PDF) of duration d in pose i, and a_{ii} is the self-transition probability of pose i, given by a geometric distribution [26]. The inability to flexibly model pose and transition duration is observed when similar body positions can only be discerned by their distinctive duration (pose vs transitory pose). This limitation is tackled using HSMM and is described in section IV-B.

B. Hidden Semi-Markov Models (HSMMs)

HSMM serves to flexibly model state duration. It uses segments instead of time slices to sample observations. In HSMM, hidden variables are segments, which have useful properties. Fig. 1 Block #5 shows the HSMM trellis and indicates its main components. For instance, the sequence of states $y_{1:T}$ is represented by the segments (S). A segment is a sequence of unique, sequentially repeated poses (symbols), which serves to identify and track an observation's first instance and the observation's duration (based on the number of observed samples). From the original sequence, the elements of the j-th segment (S_j) are the indices at which the observation (b_j) is first detected; the number of sequential observations of the same symbol (d_j); and the state or pose symbol (y_j). For instance, the sequence $y_{1:9} = \{4, 4, 2, 2, 2, 3, 2, 1\}$ is represented by the set of segments $S_{1:U}$ with elements $S_{1:U} = \{S_1, S_2, S_3, S_4, S_5\} = \{(1, 2, 4),\ (3, 3, 2),\ (6, 1, 3),\ (7, 1, 2),\ (8, 1, 1)\}$, where U is the total number of segments (i.e., state changes). The elements of the segment $S_{j=1} = (b = 1, d = 2, y = 4)$ indicate that the segment started at the first observation, lasted for a duration of two time samples, and was observed to be the fourth state.

HSMM components. In conventional HMM, the hidden variables are y, but in HSMM, the hidden variables are now the segments $S_{1:U}$, while the observable features are the same in both methods $(X_{1:T})$. The joint probability of the segments $S_{1:U}$ and the observable variable $X_{1:T}$ is:

$$\begin{aligned}\Pr\left(S_{1:U}, X_{1:T}\right) &= \Pr\left(Y_{1:U}, b_{1:U}, d_{1:U}, X_{1:T}\right)\\ \Pr\left(S_{1:U}, X_{1:T}\right) &= \Pr(y_1)\Pr(b_1)\Pr(d_1|y_1)\prod_{t=b_1}^{b_1+d_1+1}\Pr(x_t|y_1)\times\\ &\quad\prod_{u=2}^{U}\Pr(y_u|y_{u-1})\Pr\left(b_u|b_{u-1}, d_{u-1}\right)\times\\ &\quad\Pr\left(d_u|y_u\right)\prod_{t=b_u}^{b_1+d_1+1}\Pr(x_t|y_u).\end{aligned} \tag{2}$$

Recall that U is the sequence of segments such that $S_{1:U} = \{S_1, S_2, ..., S_U\}$ for $S_u = \left(b_u, d_u, y_u\right)$, b_u as the start position (a bookkeeping variable to track the starting point of a segment), d_u is the duration, and y_u is the hidden state $(\in \{1, ..., Q\})$. The range of time slices starting at b_u and ending at $b_u + d_u$ have state label y_u. All segments have a positive duration, time-span $1 : T$ without overlap, and are constrained by $b_1 = 1$, $\sum_{u=1}^{U} = T$ and $b_{u+1} = b_u + d_u$.

The transition probability $\Pr(y_u|y_{u-1})$, is the probability of going from one segment to the next via:

$$\mathbf{A} : \Pr\left(y_u = j|y_{u-1} = i\right) \equiv a_{ij} \tag{3}$$

The first segment (b_u) starts at 1 $(u = 1)$ and consecutive points are calculated from the previous point via:

$$\Pr\left(b_u = m|b_{u-1} = n, d_{u-1} = l\right) = \delta\big(m - n - l\big) \tag{4}$$

where $\delta(i - j)$ is the Kroenecker function with 1 for $i = j$; 0 else (i.e., $m = n + l$). The duration probability is now given by $\Pr(d_u = l|y_u = i) = \Pr_i(l)$ with $\Pr_i(l)$ as a free parameter. This allows MASH to sample a distribution of the form $\Pr_i(l) = \mathcal{N}(\mu, \sigma)$ in the implementation.

A normal distribution allows computing the duration probability of the i-th state and distinguishing between slow

and fast pose duration/transitions. The estimation of MASH parameters, Viterbi, and inference are described as follows.

MASH Parameter Estimation. HSMM estimation of parameters is based on maximum likelihood (MLE). The training sequence of keyframes is fully annotated, including the start and end index frames for each segment $X_{1:T}, Y_{1:T}$. To find the parameters that maximize $\Pr\big(Y_{1:T}, X_{1:T}|\theta\big)$, the likelihood parameters of each of the factors in the joint probability must be maximized. In particular, the observation probability, $\Pr\big(x^n|y=i\big)$, is a Bernoulli distribution whose maximum likelihood is computed as follows:

$$\mu_{n,i} = \frac{\sum_{n=1}^{T} x_n^i \delta\big(y_n, i\big)}{\sum_{n=1}^{T} \delta\big(y_n, i\big)}, \tag{5}$$

where T is the number of time-series data points, $\delta(i,j)$ is the Kroenecker delta function, and $\Pr\big(y_t = j|y_{t-1} = i\big)$ is the multinomial distribution of the form:

$$a_{ij} = \frac{\sum_{n=2}^{N} \delta\big(y_n, j\big)\delta\big(y_{n-1}, i\big)}{\sum_{n=2}^{N} \delta\big(y_{n-1}, j\big)} \tag{6}$$

Viterbi for MASH. The segment notation is used to represent state sequences in HSMM modeling. The objective behind the inference is to find the state sequence that maximizes $P\big(S_{1:U}, X_{1:T}|\theta\big)$, for a new sequence of observations with unknown duration. The sequence corresponding to the duration with the highest probability is determined at each time step by iterating over all possible duration values from 1 to a predetermined duration D. This data is stored in:

$$\tau_{t,d,i} = \max_{s_1,\ldots,s_{k-1}} \Pr\Big(X_{1:t}, s_{1:k} = \big(t-d+1, d, i\big)|\theta\Big), \tag{7}$$

which represents the highest probability of a sequence of K segments, where the final segment starts at $t-d+1$ and has duration d and label i.

Note: in conventional HMMs, to compute the maximum probability of ending up in state s_k, it is sufficient to only keep track of the maximum probability of ending in state s_{k-1}.

The label for a pose or state of the previous segment is stored in $\zeta_t(d,i)$. The max probability duration (δ) is:

$$\delta_t(i) = \max_{s_1,\ldots,s_{k-1}} \Pr\Big(x_{1:t}, s_{1:k} = \big(t-d^*+1, d^*, i\big)|\theta\Big), \tag{8}$$

where d^* is the duration with the highest probability at time t for state i. The best duration is stored in $\phi_t(i)$ and the label of the previous segment is stored in $\psi_t(i)$.

Inference for MASH. Four steps for finding the best sequence:

1) *Initialization.* The label probability of the first segment is given by the initial state distribution π and computed via $\tau_{t,d} = \pi_i \Pr_i(d) \prod_{t=1}^{T} \Pr\big(x_t|y_t\big)$ and $\zeta_d(d,i) = 0$.
2) *Recursion.* Iterate over all possible duration values in $\tau_{t,d} = \max_{1\le i\le Q}\big[\delta_{t-d}(i)a_{ij}\big] \Pr_j(d) \prod_{m=m_1}^{t} \Pr\big(\vec{x}_m|y_m=j\big)$, with $m_1 = t-d+1$ and $\zeta_d(d,i) = \arg\max_{1\le i\le Q}\big[\delta_{t-d}(i)a_{ij}\big]$.
   The duration with max probability is $\delta_t(i) = \max_{1\le d\le D}\big[\delta_{t-d}(i)a_{ij}\big]$, which represents the best segment. The best duration for state i at time t is given by $\phi_t(i) = \arg\max_{1\le d\le D} \tau_{d,t}(i)$. Finally, $\psi_t(i) = \zeta_t\big(\phi_t(i), i\big)$ represents the label of the best duration at time t for state i.
3) *Termination.* Estimate the state with the highest probability in the last timeslice from $\Pr^* = \max_{1\le i\le Q}[\delta_T(i)]$, where $y_T^* = \arg\max[\delta_T(i)]$, $t=T$, and $u=0$.
4) *Backtracking.* From the termination, look up the duration and previous states stored in variables ϕ and ψ given by $d_t^* = \phi_t\big(y_t^*\big)$ and $s_u^* = \big(t-d_t^*+1, d_t^*, y_t^*\big)$, with $t = t-d_t^*$, $u = u-1$, and $y_t^* = \phi_{t+d}(y_{t+d}^*)$.

Note: negative indexing is used for the segments because the number of segments is not known in advance. This is corrected after inference by adding $|S^*|$ to all indices.

Keyframe (KF) Selection. Datasets collected from pose transition are very large and often repetitive, since the motion is relatively slow and subtle. The pre-processing stage incorporates a keyframe estimation step that integrates multimodal and multiview data. The algorithm used to select a set (KF) of K-transitory frames is shown in Fig. 7 and detailed in Algorithm 1. The size of the keyframe set is determined experimentally ($K=5$) on the feature space using Inception vectors.

Input: $\mathcal{X}$, set of mm features and dissimilarity threshold th;
Result: $KF = \{\text{Keyframes}\}_K$, $K \ge 1$
Initialize: $KF = \{\text{empty}\}_K$, $K \ge 1$ and $count = 0$;
Stage 1: Modality (m) and View (v) Selection;
for $1 < v < V$ *and* $1 < m < M$ **do**
  $D_m^{(v)} = \text{euclid}(x_{mn_i}^{(v)}, x_{mn_o}^{(v)})$, $n_i = 1, n_o = N$;
end
$\hat{v}, \hat{m} =$ max $D_m^{(v)} > th$;
$\{x_{\hat{m}n_1}^{(\hat{v})}, x_{\hat{m}n_N}^{(\hat{v})}\} \rightarrow FK$;
Stage 2: Find Complementary Frames to KF ;
for $1 < v < V$ *and* $1 < m < M$ *and* $1 < n < N$ **do**
  $D_1 = D_{m,n_1}^{(v)} = \text{euclid}(x_{mn_1}^{(v)}, x_{mn}^{(v)})$;
  $D_2 = D_{m,n_N}^{(v)} = \text{euclid}(x_{mn_N}^{(v)}, x_{mn}^{(v)})$;
end
Sort $D_1 = \{d_1 > d_2 > ... > d_{N-2}\}$ descending;
Sort $D_2 = \{d_1 > d_2 > ... > d_{N-2}\}$ descending;
$d_i \rightarrow KF$ if $\frac{d_i}{d_j} > th$, for $1 < i, j < N-2$;
Stage 3: Find Center Frame (i.e., Motion Peak);
for KF_2 *and* KF_{K-1} **do**
  Use Stage 2 to compute D_3 and D_4;
  if *max*$(D_3, D_4) > 0$) **then**
    max $(D_3, D_4) \rightarrow KF$;
  end
end

Algorithm 1: Multimodal multiview keyframe selection using euclidean dissimilarity measure. The algorithm is applied at training with labeled frames to estimate the number and indices of keyframes across views and modalities.

Let $\mathcal{X} = \{x_{m,n}^{(v)}\}_f$ be the set of training features extracted from V views and M modalities over N frames and let P_i and P_o represent the initial and final poses. The transition frames are indexed by n, $1 \le n \le |N|$; views are indexed by v, $1 \le v \le |V|$ and modalities are indexed by m, $1 \le m \le$

$|\mathcal{M}|$. Algorithm 1 uses this information to identify keyframes. Experimental evaluation of $|KF|$ is shown in Fig. 9.

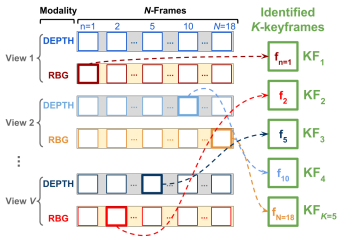


Fig. 7. Keyframe extraction for pose transition representation. The keyframe selection is based on Algorithm 1. This figure shows MASH's keyframe extraction process from three views and two modalities. The first two keyframes are extracted from the first camera's RGB modality (Views 1 and 2). Subsequent keyframes are selected from the View 2's depth, and from View V's RGB.

V. MASH Results and Analysis

MASH is evaluated using a five-fold cross-validation approach. The results indicate that deep features increase MASH's classification accuracy over engineered features by 7% in DO scenes (from 86.7% to 93.6%), while matching the performance of engineered features in BC scenes. The overall time tracing and summarization error rate between HMM and the proposed MASH approach increased from 46.4% to 83.2% in the mock-up ICU and from 35.8% to 80.1% in the medical ICU. In addition, the proposed keyframe transition representation achieves a classification of 78%.

Static Pose Analysis - Feature Validation. Static sleep-pose classification analysis is used to compare the MASH method to previous studies. Couple-Constrained Least-Squares (cc-LS)[36] and MASH are tested on the dataset from [36]. Combining the cc-LS method with deep features extracted from two common network architectures improved classification performance over the HOG and gMOM features in DO scenes by an average of eight percent with Inception and four percent with VGG. Deep features matched the performance of cc-LS (with HOG and gMOM) for a BC scenario. Results for both scenes are shown in Table **??**. Similarly, the contribution of each of the multimodal and multiview sources is analyzed and evaluated. The plot in Fig. 8 shows the contribution of each MASH sensor modality and view to the mean classification accuracy of static poses using cc-LS from [36].

Keyframe Performance. The effect of $|KF|$ $(= 5)$ and keyframe dissimilarity threshold th $(\geq .8)$ on pose transition classification accuracy is shown in Fig. 9. The traces indicate the portion of transitions correctly identified by MASH.

Summarization Performance. Pose history summarization is important to decubitus ulceration prevention and analysis. An example of the objective behind history summarization is shown in Fig. 10, where the sequence of poses is identified as A or B. History summarization is the coarser time resolution. The mock-up ICU enables staging the motion and scene condition variations necessary to carry out this experiment. In particular, it avoids disturbing real patients in the medical ICU. Table V contains the numerical symbols of the various poses and the names used in the summarization traces.

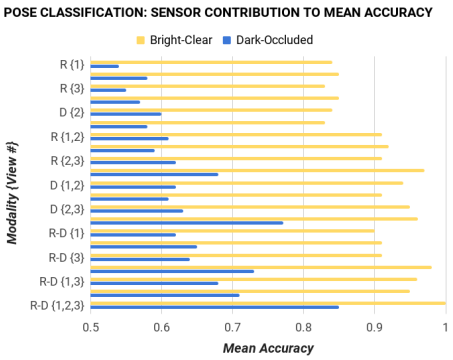


Fig. 8. Sensor contribution to mean classification of sleep poses. MASH performance is evaluated in BC (yellow) and DO (blue) scenes. RGB and Depth are R and D, while {view #} is the camera view.

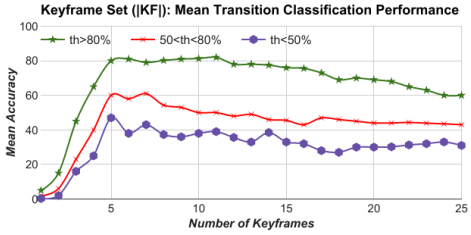


Fig. 9. Motion summarization performance for pose transition classification as a function of the number of keyframes used to represent transitions and rotations between poses. The best set uses $th = 0.80$.

MASH: Pose History Summarization

Symbol	Pose Name
0	Aspiration
+1 / -1	Soldier (+Up / -Down)
+2 / -2	Yearner (+R / -L)
+3 / -3	Log (+R, -L)
+4 / -4	Faller (+Up / -Down)
+5 / -5	Other / Background
+6 / -6	Fetal (+R / -L)

TABLE II. Pose symbols and descriptions used for ICU pose history summarization in the mock-up and the real environment.

Pose History Summarization in the ICU. Summarization history results are shown in Fig. 11 for the mock-up ICU room in (a) and for the medical ICU room in (b). The accuracy is computed as the percent overlap between the trace representing the true poses and the traces representing MASH and HMM in orange and gray, respectively. The pose history summarization experiments are staged using a sampling rate of one second and an pose duration of 10 seconds, with a minimum average detection of 80 percent. A pose is assigned a label if it is consistently detected (i.e., 80% of the time), including the label "other". Poses that are not consistently detected are ignored. The system is tested in the mock-up setting using a randomly selected sequence of ten poses starting with a randomly selected scene condition. The duration of the poses is also selected at random with one scene transition (from BC to DO or from DO to BC). The history summarization performance is shown in Table III.

Pose Transition Dynamics: Motion Direction. The detection and quantization of transitions and directions of rotations is

MASH: Pose History Summarization	
Scene	Average Detection Rate
BC	85
DO	76

TABLE III. POSE HISTORY SUMMARIZATION PERFORMANCE (PERCENT ACCURACY) OF THE MASH FRAMEWORK IN BRIGHT AND CLEAR (BC) AND DARK AND OCCLUDED (DO) SCENES IN THE MOCK-UP ICU. THE SEQUENCES ARE COMPOSED OF 10 POSES WITH DURATION THAT RANGES FROM 10 SECONDS TO 1 MINUTE. THE SAMPLING RATE IS ONE SECOND.

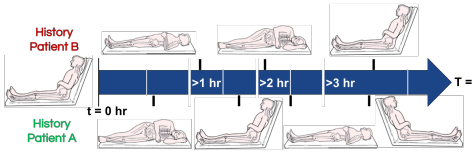


Fig. 10. Sample pose history summarization log of patient motion in medical ICUs over a 4hr span.

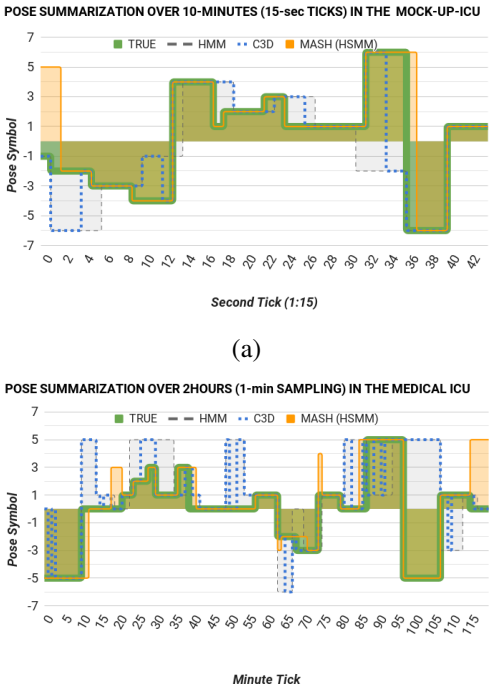


Fig. 11. History summarization traces of HMM, C3D, and MASH in (a) the mock-up ICU under BC conditions on a 10-minute video and (b) in medical ICU under random scene conditions on a two-hour video with a reduced set of poses due to patient immobility. The medical summary is based on a two-hour medical round standard. The green solid traces represent the ground-truth.

important to physical therapy and recovery rate analysis.

Transition Summarization in the Mock-Up ICU. The performance of MASH summarizing fine motion to describe transitions between poses is shown in Figs. 12 and 13 for (a) singleview and (b) multiview system configurations, while (c) shows the scale and font legend.

Peak performance is attributed to the combination of multiple views and modalities. The contributions of each sensor and view are shown in Fig. 14.

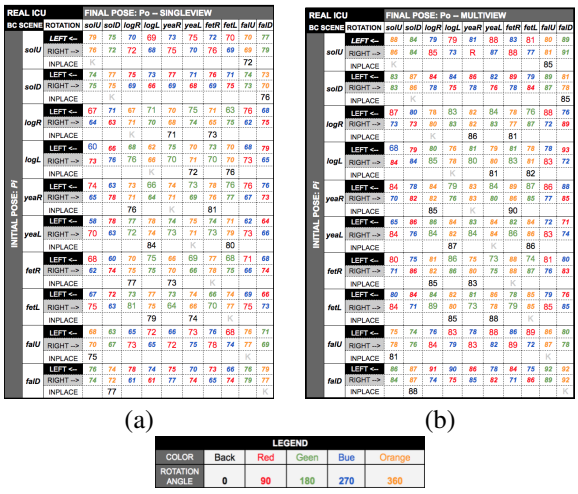


Fig. 12. MASH pose transition detection and classification mean accuracy in the mock-up ICU under BC scene conditions. The detection scores are shown for the singleview (a) and multiview (b) system configurations with the legend in (c).

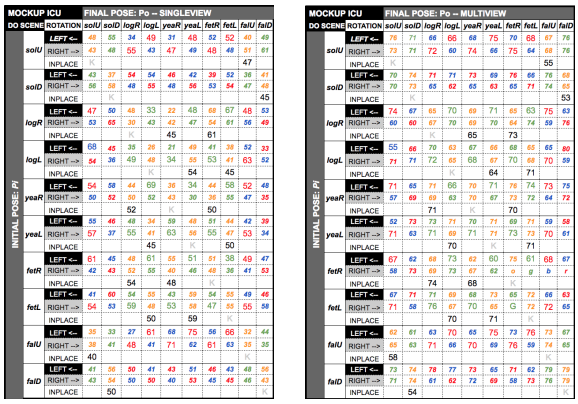


Fig. 13. MASH pose transition detection and classification mean accuracy in the mock-up ICU under DO scene conditions. The detection scores are shown for the singleview (a) and multiview (b) system configurations.

Transition Summarization in the Medical ICU. Note that it is logistically impossible to control ICU work flows and to account for unpredictable patient motion in a medical ICU. ICU patients do not have the same rotation range as the patients/actors in the mock-up ICU. This mobility constraint reduces the set of poses and pose transitions (unavailable transitions are marked N/A). The timeline in Fig. 10 shows the overall clinical objective behind the pose history summarization. Once in production, clinicians will be able to label the pose history summaries correlate pose patterns with patient health status (i.e., replacing the labels sequence A and B with actual medically validated labels).

Views of the medical ICU room are shown in Fig. 4 and the traced detections are shown in Fig. 11 (b). The green trace represents the true transition labels and the red trace indicates the predicted labels. Table V shows the pose descriptions used in the summarization plots. MASH's summarization results for fast motion of four patients are shown in Fig. 15(a) using

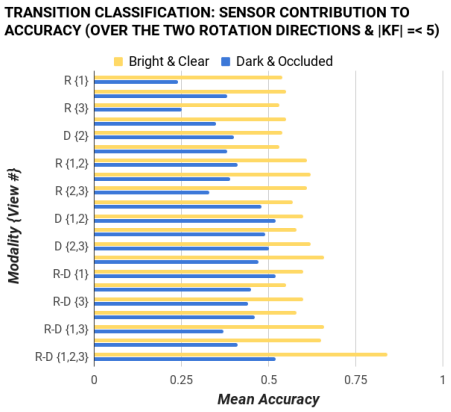
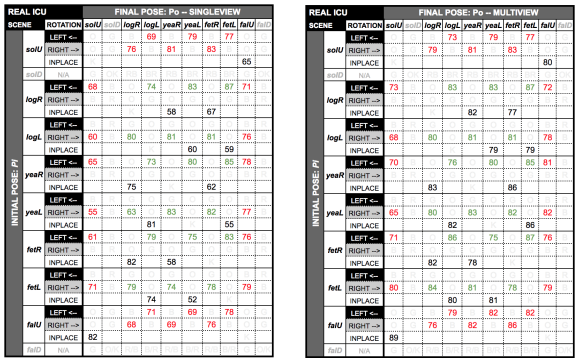


Fig. 14. Sensor contribution to the mean precision classification of pose transitions. The MASH sensors and views are tested in BC (blue) and DO (yellow) scenes. The RGB and Depth modalities are represented by R and D, respectively. The views are marked {view number} shown in 4.



(a) (b)

Fig. 15. Pose transition detection and classification in the medical ICU. Mean accuracy scores are shown for the singleview (a) and multiview (b) configurations. The set of poses is reduced due to patient's constrained mobility. Scenes are not controlled since the priority is patient care.

a singleview and (b) using a multiview configuration. Note that overall, the performance of the MASH solution using multimodal singleview in (a) is matched or outperformed by the multmimodal multiview solution in (b).

Comparison with Popular Methods. The performance of MASH is compared with C3D [39] and the summarization and detection performance is shown in Fig. 11. The sequence overlaps achieved by each method in the mock-up ICU and the medical ICU are: 46.4% and 35.8% for conventional HMM, 70.5% and 63.3% for C3D, and 83.2% and 80.1% for MASH, respectively. Using a combined average, MASH outperforms HMM by 20% and C3D by 15% overlap.

Online and Offline Processing Speeds. The **online** performance of MASH includes collecting data at 12 fps using the RPi3 devices. Real-ICU data collection is very critical and the main objective of the deployed system. Each device controls two modalities and synchronizes the data collection over the three views. RPi3s are incapable of extracting Inception features; therefore, feature extraction is defaulted to gMOM and HOG vectors as described in [36]. The performance is then extrapolated from the data collection to optic-flow computation and feature extraction. The average running speed is approximately 6 fps: 12 fps for data collection (with pre-buffering), a drop of $2-3$ frames for optic-flow and a similar drop for feature extraction, with a under one frame for inference. The **offline** performance is extrapolated using desktop computers with GPUs to process the data frames and extract capable of extracting Inception features. The data collection in the mock-up ICU achieves 30 fps (with pre-buffering). Offline runtime performance is approximately 23 fps: 30fps video with a drop of 3 frames for optic-flow computation and 3 frames for Inception features and 2 fps for inference resulting in an average performance 22fps, which is just under four times faster than the simulated online approach using RPi3s. The results in Fig. 11 are obtained using the offline approach.

VI. Conclusion and Future Work

Current computational abilities can help address the challenges of today's healthcare system. The application have the potential to improve patient care, develop new techniques, and objective evaluate and validate medical treatments. The MASH framework is such an example. It can analyze patient poses in healthcare environments. Thorough evaluation highlights the feasibility of the detection and quantification of patient poses and motion dynamics for healthcare applications. The mm sensor network is robust to variations in illumination, view, orientation, and partial occlusions. MASH is non-obtrusive and non-intrusive, but not without a cost, as the patient-motion monitoring performance of MASH in dark and occluded scenes is far from perfect; however, most medical applications that analyze motion, such as physical therapy sessions, are carried out under less severe conditions. Although the deployed version of the system suffers slightly from under-powered devices, the findings reported in this paper open up the door to new studies and optimization opportunities.

Future studies will focus on system optimization. Also, future studies will investigate the recognition and analysis of activities and events in the ICU, such as hand sanitation. The continuous growth of the MASH dataset will soon enable deep learning analysis. An important future study will incorporate additional modalities, such as thermography, to validate findings and close the learning loop. Finally, effective medical applications require generating semantically meaningful logs. MASH will explore natural language understanding to create such logs and narrate ICU activities and events.

Acknowledgments

This research is sponsored in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

The authors thank Dr. Richard Beswick, Paula Gallucci, Mark Mullenary, and Dr. Leilani Price from Santa Barbara Cottage Hospital for their support. Special thanks to Professor Victor Fragoso and Archith John Bency for their feedback.

REFERENCES

[1] B. B. Amor, J. Su, and A. Srivastava. Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[2] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *Springer Int'l Workshop on Human Behavior Understanding*, 2011.

[3] S. Bihari, R. D. McEvoy, E. Matheson, S. Kim, R. J. Woodman, and A. D. Bersten. Factors affecting sleep quality of patients in intensive care unit. *Journal of Clinical Sleep Medicine*, 2012.

[4] G. Chéron, I. Laptev, and C. Schmid. P-cnn: Pose-based cnn features for action recognition. In *IEEE Int'l Conf. on Computer Vision*, 2015.

[5] C. de Leo and B. Manjunath. Multicamera video summarization and anomaly detection from activity motifs. *ACM Transactions on Sensor Networks*, 2014.

[6] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 2013.

[7] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. *Springer Image analysis*, 2003.

[8] T. Giraud, J.-f. Dhainaut, J.-f. Vaxelaire, T. Joseph, D. Journois, G. Bleichner, J.-p. Sollet, S. Chevret, and J.-f. Monsallier. Iatrogenic complications in adult intensive care units: a prospective two-center study. *Critical care medicine*, 1993.

[9] T. Grimm, M. Martinez, A. Benz, and R. Stiefelhagen. Sleep position classification from a depth camera using bed aligned maps. In *IEEE Int'l Conf. on Pattern Recognition (ICPR)*, 2016.

[10] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge Univ. Press, 2nd edition, 2004.

[11] E. Hoque and J. Stankovic. Aalo: Activity recognition in smart homes using active learning in the presence of overlapped activities. In *IEEE Int'l Conf. on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, 2012.

[12] W. Huang, A. A. P. Wai, S. F. Foo, J. Biswas, C.-C. Hsia, and K. Liou. Multimodal sleeping posture classification. In *IEEE Int'l Conf. on Pattern Recognition (ICPR)*, 2010.

[13] C. Idzikowski. Sleep position gives personality clue. *BBC News (September 16)*, 2003.

[14] Itseez. Open source computer vision library. https://github.com/itseez/opencv, 2015.

[15] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[16] A.-A. Liu, N. Xu, W.-Z. Nie, Y.-T. Su, Y. Wong, and M. Kankanhalli. Benchmarking a multimodal and multiview and interactive dataset for human action recognition. *IEEE Transactions on Cybernetics*, 2017.

[17] L. Liu, L. Shao, X. Li, and K. Lu. Learning spatio-temporal representations for action recognition: A genetic programming approach. *IEEE Transactions on Cybernetics*, 2016.

[18] P. Liu, W. Liu, and H. Ma. Weighted sequence loss based spatial-temporal deep learning framework for human body orientation estimation. In *IEEE Intn'l Conf. on Multimedia and Expo (ICME)*, 2017.

[19] W. Liu, Y. Zhang, S. Tang, J. Tang, R. Hong, and J. Li. Accurate estimation of human body orientation from rgb-d sensors. *IEEE Transactions on Cybernetics*, 2013.

[20] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. 1981.

[21] P. E. Morris. Moving our critically ill patients: mobility barriers and benefits. *Critical care clinics*, 2007.

[22] S. Obdržálek, G. Kurillo, J. Han, T. Abresch, R. Bajcsy, et al. Real-time human pose detection and tracking for tele-rehabilitation in virtual reality. *Studies in health technology and informatics*, 2012.

[23] H. S. of Medicine. Finding Top-Line Opportunities in a Bottom-Line Healthcare Market. Technical report, Harvard School of Med., 2016.

[24] R. Panda and A. R. Chowdhury. Multi-view surveillance video summarization via joint embedding and sparse optimization. *IEEE Transactions on Multimedia*, 2017.

[25] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Elsevier Computer Vision and Image Understanding*, 2016.

[26] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989.

[27] S. Ramagiri, R. Kavi, and V. Kulathumani. Real-time multi-view human action recognition using a wireless camera network. In *ACM/IEEE Int'l Conf. on Distributed Smart Cameras (ICDSC)*, 2011.

[28] Y. Shi, Y. Tian, Y. Wang, and T. Huang. Sequential deep trajectory descriptor for action recognition with three-stream cnn. *IEEE Transactions on Multimedia*, 2017.

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[30] J. Song, H. Jegou, C. Snoek, Q. Tian, and N. Sebe. Guest editorial: Large-scale multimedia data retrieval, classification, and understanding. *IEEE Transactions on Multimedia*, 2017.

[31] Y. Song, Y.-T. Zheng, S. Tang, X. Zhou, Y. Zhang, S. Lin, and T.-S. Chua. Localized multiple kernel learning for realistic human action recognition in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2011.

[32] B. Soran, A. Farhadi, and L. Shapiro. Generating notifications for missing actions: Don't forget to turn the lights off! In *IEEE Int'l Conf. on Computer Vision (ICCV)*, 2015.

[33] S. Sunderrajan and B. S. Manjunath. Context-aware hypergraph modeling for re-identification and summarization. *IEEE Transactions on Multimedia*, 2016.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[35] The Raspberry Pi Foundation. *Raspberry Pi 3 Model B*, 2017 (accessed July 17th, 2017). https://www.raspberrypi.org/products/raspberry-pi-3-model-b/.

[36] C. Torres, V. Fragoso, S. D. Hammond, J. C. Fried, and B. S. Manjunath. Eye-cu: Sleep pose classification for healthcare using multimodal multiview data. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2016.

[37] C. Torres, J. C. Fried, K. Rose, and B. Manjunath. Deep eye-cu (decu): Summarization of patient motion in the icu. In *European Conf. on Computer Vision, Workshop on Assistive Computer Vision and Robotics (ACVR)*, page 178194. Springer, 2016.

[38] C. Torres, S. D. Hammond, J. C. Fried, and B. S. Manjunath. Multimodal pose recognition in an icu using multimodal data and environmental feedback. In *Springer Int'l Conf. on Computer Vision Systems (ICVS)*, 2015.

[39] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE Int'l Conf. on Computer Vision (ICCV)*, 2015.

[40] O. Ulutan, B. Riggan, N. Nasrabadi, and B. S. Manjunath. An order preserving bilinear model for person detection in multi-modal data. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2018.

[41] V. Veeriah, N. Zhuang, and G.-J. Qi. Differential recurrent neural networks for action recognition. In *IEEE Int'l Conf. on Computer Vision (ICCV)*, 2015.

[42] G. L. Weinhouse and R. J. Schwab. Sleep in the critically ill patient. *Sleep-New York Then Westchester*, 2006.

[43] C. Wu, A. H. Khalili, and H. Aghajan. Multiview activity recognition in smart homes with spatio-temporal features. In *ACM/IEEE Int'l Conf. on Distributed Smart Cameras (ICDSC)*, 2010.

[44] J. Wu, Y. Zhang, and W. Lin. Good practices for learning to recognize actions using fv and vlad. *IEEE Transactions on Cybernetics*, 2016.

[45] J. Xu, V. Jagadeesh, Z. Ni, S. Sunderrajan, and B. Manjunath. Graph-based topic-focused retrieval in a distributed camera network. *IEEE Transaction on Multimedia*, 2013.

[46] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, and X. Gao. Discriminative multi-instance multitask learning for 3d action recognition. *IEEE Transactions on Multimedia*, 2017.

[47] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall. A survey on human motion analysis from depth data. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. Springer, 2013.

[48] X. Zhen, F. Zheng, L. Shao, X. Cao, and D. Xu. Supervised local descriptor learning for human action recognition. *IEEE Transactions on Multimedia*, 2017.