Title
A multiview multimodal system for monitoring patient sleep

Permalink
https://escholarship.org/uc/item/9j73f2sr

Journal
IEEE Transactions on Multimedia, 20(11)

ISSN
1520-9210

Authors
Torres, C
Fried, JC
Rose, K
et al.

Publication Date
2018-11-01

DOI
10.1109/TMM.2018.2829162

Peer reviewed
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 subtly 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:
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.

Fig. 1. MASH framework blocks. The process starts with Block #1 and flows clockwise: data collection, motion threshold, deep feature extraction, mm3 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 [24] 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 [8], [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
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 (mmn) data collection camera network, a mmn keyframe extraction algorithm, and a mmn 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.
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 Prime-sense’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 of 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. 5 shows the observed counts of poses in number of minutes. Fig. 6 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.](image)

**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 process 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 (felL), respectively. The top sequence (orange) shows the left rotation and the bottom sequence shows the right (green) rotation.
small (≤ 180°) rotation or a large (> 180°) 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.

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, true pose, or pose). The observable or measurable variables \( x_k \) are the visible states. The vector of image features corresponding to the \( k \)-th frame, using modality \( m \), and view \( v \) at time \( t = k \) is \( x_k \). 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 \( \text{Pr}(Y|X) \) given by:

\[
\text{Pr}(Y_{1:T}, X_{1:T}) = \text{Pr}(y_1) \prod_{t=1}^{T} \text{Pr}(x_t|y_t) \prod_{t=2}^{T} \text{Pr}(y_t|y_{t-1}) \tag{1}
\]

where \( \text{Pr}(y_1) \) is the initial state probability distribution. It represents the probability of a sequence starting at \( (t = 1) \) pose. \( \text{Pr}(x_t|y_t) \) is the emission probability distribution and \( \text{Pr}(y_t|y_{t-1}) \) represents the probability that at time \( t \), \( y_t \) can generate the observable multimodal multiview vector \( x_t \).
Finally, $\Pr(y_i|y_{i-1})$ is the transition probability distribution (A) and represents the probability of going from pose i to pose o (state i to o). The HMM parameters are $A = \{a_{ij}\}$, $B = \{b_i\}$, and $\Pi = \{\pi_i\}$, 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: $Pr_t(d) = (a_{ii})^{d-1}(1-a_{ii})$, where $Pr_t(d)$ 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 HSM, hidden variables are segments, which have useful properties.

<table>
<thead>
<tr>
<th>SYMBOL</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Transition probability matrix $A \in \mathbb{R}^{(X \times T)}$ and $A = {a_{ij}}$</td>
</tr>
<tr>
<td>B</td>
<td>Probability of transition from pose i to j</td>
</tr>
<tr>
<td>$b_i$</td>
<td>Emission probability matrix $b_i \in [0,1]$ and $B = {b_i}$</td>
</tr>
<tr>
<td>$D_k$</td>
<td>k-th frame from the depth modality video</td>
</tr>
<tr>
<td>D</td>
<td>Face-Down patient pose</td>
</tr>
<tr>
<td>d</td>
<td>Segment duration</td>
</tr>
<tr>
<td>$D_u$</td>
<td>Segment duration for u-th segment</td>
</tr>
<tr>
<td>HMM</td>
<td>Abbreviation for Hidden Markov Model</td>
</tr>
<tr>
<td>HSMM</td>
<td>Abbreviation for Hidden Semi-Markov Model</td>
</tr>
<tr>
<td>$K$</td>
<td>Data set size, $K =</td>
</tr>
<tr>
<td>$k$</td>
<td>Data point index, $1 \leq k \leq K$</td>
</tr>
<tr>
<td>$K_F$</td>
<td>Set of sequential keyframes representing a transition between</td>
</tr>
<tr>
<td>L</td>
<td>Laying-Left patient pose</td>
</tr>
<tr>
<td>$l, m, n$</td>
<td>Dummy variables</td>
</tr>
<tr>
<td>$R_k$</td>
<td>k-th frame from the rgb modality video</td>
</tr>
<tr>
<td>R</td>
<td>Laying-Right patient pose</td>
</tr>
<tr>
<td>$\mu_i$</td>
<td>Probability that state i generates the observation $x$ at time t</td>
</tr>
<tr>
<td>$\pi$</td>
<td>Initial state probability vector $\pi \in \mathbb{R}^K$ and $\pi_i \in \pi$</td>
</tr>
<tr>
<td>$P$</td>
<td>Set of patient poses $P = {p_i}$</td>
</tr>
<tr>
<td>$P_{mock}$</td>
<td>Set of actor poses in the mock-up ICU room</td>
</tr>
<tr>
<td>$P_{real}$</td>
<td>Set of patient poses in the real medical ICU (mics) room</td>
</tr>
<tr>
<td>$Pr(Y, X)$</td>
<td>The joint PDF: sequence of states and observations</td>
</tr>
<tr>
<td>$S$</td>
<td>Set of time segments $S = {s_u}$ for $1 \leq u \leq U$</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>Segment element $s \in S$</td>
</tr>
<tr>
<td>$t$</td>
<td>Time tick with $1 \leq t \leq T$</td>
</tr>
<tr>
<td>$\tau_{id}$</td>
<td>Stores the estimated duration $(1 \leq d \leq D)$ at time (t)</td>
</tr>
<tr>
<td>$\theta$</td>
<td>HMM model with probabilities A, B, and $\Pi$</td>
</tr>
<tr>
<td>$U$</td>
<td>Number of segments $U =</td>
</tr>
<tr>
<td>U</td>
<td>Face-Up patient pose</td>
</tr>
<tr>
<td>u</td>
<td>Segment index: $1 \leq u \leq U$</td>
</tr>
<tr>
<td>$\mathcal{V}$</td>
<td>View set $\mathcal{V} = {\text{left, center, right}}$</td>
</tr>
<tr>
<td>v</td>
<td>Number of views $V =</td>
</tr>
<tr>
<td>$v_m$</td>
<td>View index, $1 \leq v \leq V$</td>
</tr>
<tr>
<td>$y_k$</td>
<td>k-th hidden state $y_k \in Y$</td>
</tr>
<tr>
<td>$\mathcal{Y}$</td>
<td>Sequence of hidden states $\mathcal{Y} = Y$</td>
</tr>
<tr>
<td>$X$</td>
<td>Dataset indexed by k (i.e., $A_k$)</td>
</tr>
<tr>
<td>$X_k$</td>
<td>k-th datapoint with ${x_{kn}} k = (f_r, f_d, f_P)_k$</td>
</tr>
<tr>
<td>$x_k$</td>
<td>k-th observation feature vector</td>
</tr>
<tr>
<td>$x_{k, m}$</td>
<td>The k-th observable variable from view v and modality m</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Kroenecker delta function</td>
</tr>
<tr>
<td>$\delta_i$</td>
<td>The maximum probability duration</td>
</tr>
<tr>
<td>$\vartheta$</td>
<td>Dummy variable used in inference</td>
</tr>
<tr>
<td>$\zeta$</td>
<td>Stores the state label (for a pose) of the previous segment</td>
</tr>
<tr>
<td>$\phi$</td>
<td>Stores the best duration</td>
</tr>
<tr>
<td>$\psi(i,t)$</td>
<td>Stores the label with the best duration for state i at time t</td>
</tr>
</tbody>
</table>

**TABLE 1. MASH VARIABLES AND THEIR DESCRIPTIONS.**

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 $(b_j)$; and the state or pose symbol $(y_j)$. For instance, the sequence $y_{1:5} = \{4, 4, 2, 2, 3, 2, 1\}$ is represented by the set of segments $S_{1:T}$ with elements $S_{1:T} = \{S_1, S_2, S_3, S_4, S_5\} = \{(1, 2, 4), (3, 3, 2), (6, 1, 3), (7, 1, 2), (8, 1, 1)\}$, where $U$ is the total number of segments (i.e., state changes). The elements of the segment $S_{1:1} = \{b = 1, d = 2, y = 4\}$ 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 $S_{1:T}$, while the observable features are the same in both methods ($X_{1:T}$). The joint probability of the segments $S_{1:T}$ and the observable variable $X_{1:T}$ is:

$$
Pr(S_{1:T}, X_{1:T}) = \prod_{t=1}^{T} Pr(y_t|y_{t-1}) \prod_{t=1}^{T} Pr(b_t|x_t|y_t) \prod_{t=1}^{T} Pr(x_t|y_t)
$$

Recall that $U$ is the sequence of segments such that $S_{1:T} = \{S_1, S_2, ..., S_U\}$ for $S_u = \{b_u, d_u, y_u\}$, $b_u$ 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−1 < 1 : T without overlap, and are constrained by $b_1 = 1$, $\sum_{u=1}^{U} T = T$ 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:

$$
A : Pr(y_u = j|y_{u-1} = i) = a_{ij}
$$

The first segment $(b_1)$ starts at 1 ($u = 1$) and consecutive points are calculated from the previous point via:

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

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 $Pr(d_u = l|y_u = i) = Pr_t(l)$ with $Pr_t(l)$ as a free parameter. This allows MASH to sample a distribution of the form $Pr_t(l) = N(\mu, \sigma)$ 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 \(\Pr (Y_{1:T}, X_{1:T}|\theta)\), the likelihood parameters of each of the factors in the joint probability must be maximized. In particular, the observation probability, \(\Pr (x^n|y = i)\), is a Bernoulli distribution whose maximum likelihood is computed as follows:

\[
\mu_{n,i} = \frac{\sum_{t=1}^{T} x^n_t \delta(y_n, i)}{\sum_{t=1}^{T} \delta(y_n, i)},
\]

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

\[
a_{ij} = \frac{\sum_{n=2}^{N} \delta(y_n, j) \delta(y_{n-1}, i)}{\sum_{n=2}^{N} \delta(y_{n-1}, j)}
\]

**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 \(P(S_{t:d}, X_{t:T}|\theta)\), 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:

\[
\tau_{t,d,i} = \max_{s_1,...,s_k \rightarrow 1} \Pr (X_{t:t}, s_{1:k} = (t - d + 1, d, i)|\theta),
\]

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 \(s_{k-1}\), 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 \(\zeta(d, i)\). The max probability duration (\(\delta\)) is:

\[
\delta_t(i) = \frac{\max_{s_1,...,s_k \rightarrow 1} \Pr (x_{t:t}, s_{1:k} = (t - d^* + 1, d^*, i)|\theta)}{\delta_t(i)},
\]

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 \(\pi\) and computed via \(\tau_0, \Pr (c_d|\pi) \prod_{t=1}^{T} \Pr (x_t|y_t)\) and \(\zeta(0, d) = 0\).

2) **Recursion.** Iterate over all possible duration values in \(\tau_t, \Pr (d|\tau_{t-1}) \prod_{m=1}^{D} \Pr (x_m|y_m = j)\), with \(m_t = t - d + 1\) and \(\zeta(d, i) = \arg \max_{1 \leq d \leq D} \delta_t(d, i)\).

The duration with max probability is \(\delta_t(i) = \max_{1 \leq d \leq D} \delta_t(d, i)\), which represents the best segment.

The best duration for state \(i\) at time \(t\) is given by

\[
\phi_t(i) = \arg \max_{1 \leq d \leq D} \tau_t, d(i).
\]

Finally, \(\psi_t(i) = \zeta(\phi_t(i), i)\) 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 \(\Pr^* = \max_{1 \leq i \leq Q} \delta_T(i)\), where \(y^*_t = \arg \max_{1 \leq i \leq Q} \delta_T(i), t = T\), and \(u = 0\).

4) **Backtracking.** From the termination, look up the duration and previous states stored in variables \(\phi\) and \(\psi\) given by \(d^*_t = \phi_t(y^*_t)\) and \(s^*_0 = (t - d^*_t + 1, d^*_t, y^*_t)\), 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:** \(X\), set of mm features and dissimilarity threshold \(th\);

**Result:** \(KF\) = \{Keyframes\} \(K \geq 1\)

**Initialize:** \(KF = \{empty\} \_K \geq 1\) and \(count = 0\);

**Stage 1:** Modality \((m)\) and View \((v)\) Selection;

for \(1 < v < V\) and \(1 < m < M\)

\[
\forall_{m} \forall_{v} \forall_{n} \forall_{o} \forall_{count}
\]

**Stage 2:** Find Complementary Frames to \(KF\);

for \(1 < v < V\) and \(1 < m < M\) and \(1 < n < N\)

\[
D_m(v) = \text{euclid}(x_{v,m,n}, x_{v,m,n}), n_i = 1, o_n = N;
\]

**Stage 3:** Find Center Frame (i.e., Motion Peak);

for \(KF2\) and \(KF_{K-1}\)

Use Stage 2 to compute \(D_1\) and \(D_2\);

if \(\max(D_1, D_2) > 0\) then

\[
\max(D_1, D_2) \rightarrow KF;
\]

**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 \(X = \{x_{m,n}^{(v)}\}_f\) be the set of training features extracted from \(V\) views and \(M\) modalities over \(N\) frames and let \(P_f\) and \(P_o\) represent the initial and final poses. The transition frames are indexed by \(n\), \(1 \leq n \leq |N|\); views are indexed by \(v\), \(1 \leq v \leq |V|\) and modalities are indexed by \(m\), \(1 \leq m \leq |M|\).
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 II. 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 II contains the numerical symbols of the various poses and the names used in the summarization traces.

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

<table>
<thead>
<tr>
<th>Scene</th>
<th>Average Detection Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC</td>
<td>85</td>
</tr>
<tr>
<td>DO</td>
<td>76</td>
</tr>
</tbody>
</table>

**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 4 hr span.

Fig. 11. History summarization traces of HMM, C3D, and MASH in (a) the mock-up ICU under BC scene conditions 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.

**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 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.

**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
The 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 default 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


