

# Eye-CU: Sleep Pose Classification for Healthcare using Multimodal Multiview Data

Carlos Torres<sup>†</sup> Victor Fragoso<sup>‡</sup> Scott D. Hammond<sup>†</sup> Jeffrey C. Fried<sup>\*</sup> B.S. Manjunath<sup>†</sup>  
<sup>†</sup>Univ. of California Santa Barbara <sup>‡</sup>West Virginia University <sup>\*</sup>Santa Barbara Cottage Hospital  
{carlostorres@ece, shammond@tmrl, manj@ece}.ucsb.edu victor.fragoso@mail.wvu.edu jfried@sbch.org

## Abstract

*Manual analysis of body poses of bed-ridden patients requires staff to continuously track and record patient poses. Two limitations in the dissemination of pose-related therapies are scarce human resources and unreliable automated systems. This work addresses these issues by introducing a new method and a new system for robust automated classification of sleep poses in an Intensive Care Unit (ICU) environment. The new method, coupled-constrained Least-Squares (cc-LS), uses multimodal and multiview (MM) data and finds the set of modality trust values that minimizes the difference between expected and estimated labels. The new system, Eye-CU, is an affordable multi-sensor modular system for unobtrusive data collection and analysis in healthcare. Experimental results indicate that the performance of cc-LS matches the performance of existing methods in ideal scenarios. This method outperforms the latest techniques in challenging scenarios by 13% for those with poor illumination and by 70% for those with both poor illumination and occlusions. Results also show that a reduced Eye-CU configuration can classify poses without pressure information with only a slight drop in its performance.*

## 1. Introduction

New innovative methods for non-disruptive monitoring and analysis of patient-on-bed body configurations, such as those observed in sleep-pose patterns, add objective metrics for evaluating and predicting health status. Clinical scenarios where body poses of patients correlate to medical conditions include sleep apnea – where the obstructions of the airway are affected by supine positions [16]. Mothers-to-be are recommended to lay on their sides to improve fetal blood flow [11]. The findings of [2, 8, 19] correlate sleep positions with various effects on patient health. In these studies, the findings highlight the importance of automated analysis of patient sleep poses in natural scenarios. They substantiate the need of this work and its potential benefits. The benefits include improving patient quality of life and quality of care through continuously monitoring patient poses, correlating poses to medical diagnosis, and optimiz-

ing treatments by manipulating poses. The proposed Eye-CU system and cc-LS fusion method tackles the classification of sleep poses in a natural ICU environment with conditions that range from bright and clear to dark and occluded. The system collects sleep-pose data using an array of RGB-D cameras and a pressure mat. The method extracts features from each modality, estimates unimodal pose labels, fuses unimodal decisions based on trust (priors) values, and infers a multimodal pose label. The trusts are estimated via cc-LS optimization, which minimizes the distance between the oracle and multimodal matrices. In this context, the term multimodal refers to the various Eye-CU sensor measurements.

### 1.1. Related Work

Computer vision methods using RGB data to detect body configurations of patients on beds are discussed in [9, 10, 13] but are limited to scenes with constant illumination and/or without occlusions. The deformable parts model approach, commonly used in RGB images presented in [20] requires images with relatively uniform illumination and is limited to minor self-occlusions. The discriminative approach from [17] uses depth images and is robust to illumination changes. It requires clean depth segmentation and contrast and is susceptible to occlusions. A controlled method to classify human sleep poses using RGB images and a low-resolution pressure array is presented in [7]. It uses normalized geometric and load distribution features interdependently and requires a clear view of the patient.

The cc-LS work builds upon our previous work [18], where features from  $R$ ,  $D$ , and  $P$  sensors from a single view are combined to overcome challenging scene conditions. The trust method uses unimodal features to propose label candidates and infer a multimodal label. It improves unimodal decisions of Linear Discriminant Analysis (LDA) and Support Vector Classifier (SVC) via modality trust. Modality trust is defined as the mean classification accuracy of the unimodal pose classifiers (under the measured scene conditions). The trust system uses a high-resolution pressured mat and its performance relies heavily on a fixed camera over the patient's bed. A trust-adjustment method

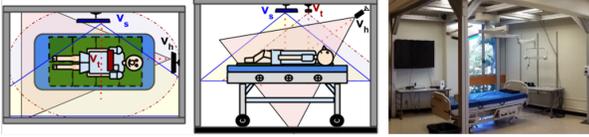


Figure 1: Diagram of the Eye-CU physical setup showing the pressure mat (left) in green and the camera views(center): top ( $v_t$ ) in red, side ( $v_s$ ) in blue, and head ( $v_h$ ) in black; and the mock-up ICU (right) where the system is tested.



Figure 2: Multimodal Eye-CU node with environmental sensors, RGB-D camera, aluminum enclosure, Panda Board, and battery pack. Four nodes are used to monitor a medical ICU room.

accounts for sensor failures; however, performance declines greatly without pressure data.

## 1.2. Proposed Work

The work presented in this paper differs from [18] by introducing a new probabilistic method to estimate trusts. We use cc-LS optimization (section 4) to estimate trusts, learn modality priors, and improve classification accuracy by up to 30%. Instead of using a multimodal system with a single camera view and a pressure mat, the Eye-CU system uses multimodal and multiview ( $MM$ ) data. Results suggest that combining reduced Eye-CU configurations with cc-LS robustly classifies sleep poses with incomplete views and without pressure information. Figure 1 shows two perspective views of the system in the mock-up ICU room.

**Main Contributions of this work:** (1) **cc-LS** a simple and elegant method to estimate modality trusts, which improves pose classification accuracy; (2) **Eye-CU** a complete modular  $MM$  system that performs sleep pose classification with very high accuracy in healthcare. One node is shown in Figure 2 and the system is currently deployed in a medical ICU; and (3) a fully annotated  $MM$  dataset of 66,000 sleep-pose images<sup>1</sup>.

## 2. Eye-CU System Description

The various Eye-CU system configurations depend on the combination of modalities used: RGB ( $R$ ), depth ( $D$ ), and pressure ( $P$ ), and available camera views: head ( $h$ ), top

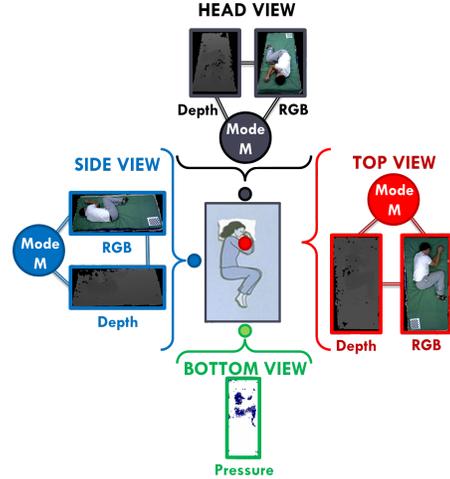


Figure 3: Multimodal and multiview representation of the fetal left-oriented pose observed by three RGB-D cameras and one pressure-mat collected using the Eye-CU system.

( $t$ ), and side ( $s$ ). The following configurations are explored:

- Multimodal and Multiview ( $MM$ ) uses  $R, D, P$  data and the  $h, s, t$  views. It is the most complex and has the best performance, but is difficult to deploy.
- Multimodal partial-Multiview ( $MpM$ ) uses  $RDP$  data and less than three views.  $MpM$  with a top view is equivalent to the one used in competing methods.
- Partial-Multimodal and Multiview ( $PMM$ ) uses  $R, D$  or  $RD$  data from three camera views ( $hst$ ). Its performance depends on having all views available.
- Partial-Multimodal partial-Multiview ( $PMPM$ ) is the simplest configuration. It uses  $RD$  data from two views ( $hs, ht, st$ ) and sets the lower bound in performance.

**Why Multimodal?** Suitability tests (section 5) of existing methods and available modalities indicate that neither a single modality nor the concatenation of modalities can be used to classify poses in a natural ICU environment.

**Why Multiview?** The ICU is a dynamic environment, where equipment is moved around continuously and can block sensors and view of the patients. A multiview system improves classification performance, increases the chances of observing the patients, and enables monitoring using simple and affordable sensor. Cameras do not have contact with patients and avoid the risk of infections by touch.

## 3. Data Collection

Sample  $MM$  data collected from one actor in various poses and scene conditions using all camera views and modalities is shown in Figure 4. The complete dataset is constructed with sleep poses collected from five actors in a mock-up ICU setting with a real ICU bed and equipment. The observations are the set of sleep poses  $\mathcal{Z} = \{\text{Background, Soldier U, Soldier D, Faller R, Faller L, Log R, Log L, Yearner R, Yearner L, Fetal R, Fetal L}\}$  of size

<sup>1</sup>will be available online at <http://vision.ece.ucsb>

$L (= |\mathcal{Z}|)$  and indexed by  $l$ . The letters U and D indicate the patient is up-facing and down-facing and letters L and R indicate lying-on-Left and lying-on-Right sides. The variable  $Z_l$  is used to identify one specific pose label (e.g.,  $Z_0 = \text{Background}$ ). The scene conditions are simulated using three illumination levels: bright (light sensor with 70-90% saturation), medium (50-70% saturation), and dark (below 50% saturation), as well as four occlusion types: clear (no occlusion), blanket (covering 90% of the actor's body), blanket and pillow, and pillow (between actor's head and upper back and the pressure mat). The illumination intensities are based on percent saturation values of an illumination sensor and the occlusions are detected using radio-frequency identification (RFID) and proximity sensors, all by .NET Gadgeteer. The combination of the illumination levels and occlusions types generates a 12-element scene set  $\mathcal{C} = \{(\text{bright, medium, dark}) \times (\text{clear, blanket, pillow, blanket+pillow})\}$ . The variable  $c \in \mathcal{C}$  is used to indicate a single illumination and occlusion combination (e.g.,  $c = 1$  indicates bright and clear scene). The dataset is created by allowing one scene to be the combination of one actor in one pose and under a single scene condition. Ten measurements are collected from one scene – three modalities ( $R, D$ , and synthetic binary masks) from each of the three camera views in the set  $\mathcal{V} = \{t, h, s\}$  and one pressure image ( $P$ ). The data collection process includes acquiring the background (empty bed), and asking the actors to rotate through the 10 poses (11 classes including the background) under each of the 12 scene conditions. The process is repeated 10 times for each of the five actors. In total, this process generates a dataset of 66,000 images (five actors  $\times$  10 sessions  $\times$  10 images  $\times$  11 classes  $\times$  12 scenes).

### 3.1. Feature Extraction

The sensors and camera views are calibrated using the standard methods from [5]. Homography transformations are computed relative to the top view and gradient and shape features are then extracted from the transformed images.

**Histogram of Oriented Gradients (HOG).** HOG features are extracted from RGB images to represent sleep-pose limb structures as demonstrated by [3, 20]. The HOG extraction parameters are: four orientations, 16-by-16 pixels per cell, and two-by-two cells per block, which yield a 5776-element vector per image.

**Geometric Moments (gMOM).** Image gMOM features introduced in [6] and validated in [1, 15] are used to represent sleep-pose shapes. The in-house implementation uses the raw pixel values from tiled depth and pressure images, instead of standard binarized pixel values. The six-by-six tile dimensions are determined empirically, to balance accuracy and complexity. Finally, moments up to the third order are extracted from each block to generate a 10-element

vector per block. The vectors from each of the 36 blocks are concatenated to form a 360-element vector per image.

## 4. Multimodal-Multiview Formulation

Explanation of the method begins with the problem statement in section 4.1, followed by a description of the single-view multimodal formulation in section 4.2. This formulation is expanded to include multiview data in section 4.4. The multimodal classification framework for a single-view system is shown in Figure 5, which is applied to the set of pose labels  $Z$  of size  $L$  indexed by  $l$ . The multimodal dataset ( $\mathcal{X}$ ) of size  $K$  indexed by  $k$  is separated for each scene  $c \in \mathcal{C}$ . The dataset is composed of features extracted from a set of  $M$  modalities  $\mathcal{N} = \{R, D, P\}$  indexed by  $m$  (e.g.,  $f_{N_m}$  with  $m = 1$  gives  $f_R$ ). The  $k$ -th datapoint in the dataset has the form:

$$\begin{aligned} \mathcal{X}_k &= \{f_{N_m}\}_M = \{f_R, f_D, f_P\} \\ &= \{\text{HOG}(R), \text{gMOM}(D), \text{gMOM}(P)\}, \end{aligned} \quad (1)$$

where  $f_{N_m}$  is the feature vector extracted from the  $m$ -th modality. These features are used to train the ensemble of  $M$  unimodal SVM (and LDA) classifiers ( $\text{CLF}_m$ ). For a given input datapoint  $\mathcal{X}_k$ , each of the classifiers outputs a probability vector  $\text{CLF}_{k,m} = [s_{k,1,m}, \dots, s_{k,L,m}]^T$ , where the elements ( $s$ ) represent the probability of label  $l$  given modality feature  $m$ . The classifier label probabilities are computed using the implementations from [12] of Platt's method for SVC and Bayes' rule for LDA. The feature-classifier combinations are quantified at the trust estimation stage where the unimodal trust values  $w^c = [w_R^c, w_D^c, w_P^c]^T$  are computed for specific scene  $c$ . The multimodal trusted classifier is formed by fusing the candidate label decisions from the unimodal classifiers into one. The objective of this formulation is to find the pose label ( $Z_{\hat{l}}$ ) with the highest  $MM$  probability for a given input query  $\mathcal{X}_k$ , where  $\hat{l}$  is the estimated index label. The variables used through out this paper are listed in Table 1.

### 4.1. Problem Statement

The proposed fusion technique uses probabilistic concepts to compute the probability of a given class by marginalizing the joint probability over the modalities. The joint probability is calculated from the conditional probability of each class and the set of prior probabilities for each modality. The conditional probabilities are extracted from the classifiers in the ensemble of  $M$ -unimodal classifiers (i.e.,  $\mathbb{P}(Z = Z_l | \mathcal{X} = \mathcal{X}_k) = \mathbb{P}(Z_l | \mathcal{X}_k)$ ) and re-written as:

$$\mathbb{P}(Z_l | \mathcal{X}_k) = \sum_{m=1}^M \mathbb{P}(Z_l | \mathcal{X}_k, M = m) \mathbb{P}(M = m). \quad (2)$$

Methods such as Platt's [14] for SVMs enable the computation of conditional probabilities given by:

$$s_{k,l,m} = \mathbb{P}(Z_l | \mathcal{X}_k, M = m). \quad (3)$$

Symbol	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
POSE	Fetal L	Fetal R	Log L	Log R	Yearner L	Yearner R	Soldier D	Soldier U	Faller D	Faller U
RGB (R): Views: $t s h$										
Depth (D): Views: $t s h$										
Pressure (P)										
Light	Bright	Medium	Dark	Bright	Medium	Dark	Bright	Medium	Dark	Bright
Occlusion	Clear	Clear	Clear	Blanket	Blanket	Blanket	Pillow	Pillow	Pillow	Blanket Pillow

Figure 4: Multimodal and multiview dictionary of sleep poses for a single actor in various sleep configurations and scene conditions. It contains  $R, D$  (equalized for display) images from  $t, s$ , and  $h$  views and the pressure mat  $P$ . Images are transformed w.r.t the  $t$  view.

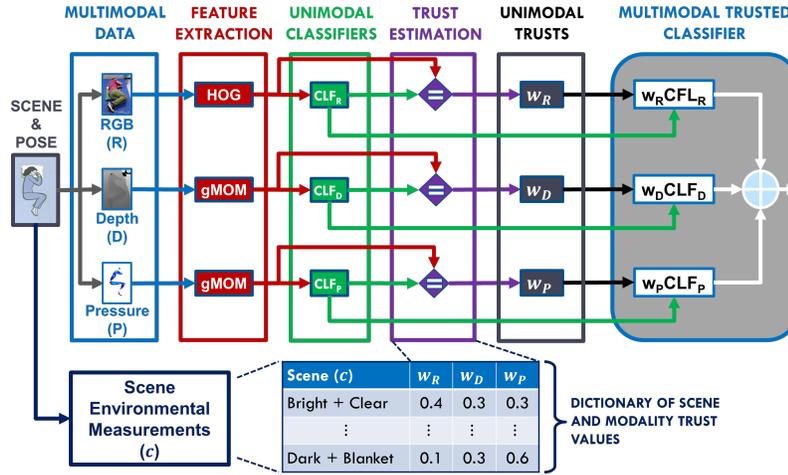


Figure 5: Diagram of the trusted multimodal classifier for the  $MpM$  configuration. Image features are extracted from the  $R, D, P$  camera and pressure data. Then the features are used to train unimodal classifiers ( $CLF_m$ ), which are in turn used to estimate the modality trust values. In the last stage of the  $MM$  classifier, the unimodal decisions are trusted and combined.

However, the prior probability for each modality  $w_m = \mathbb{P}(M = m)$  remains unknown. The trust method finds the set of priors for each modality  $m$  in the ensemble of  $M$  modalities that approximates the following probability:

$$b_{k,l} = \mathbb{P}(Z = z_l | \mathcal{X} = \mathcal{X}_k, \text{Oracle}), \quad (4)$$

produced by an oracle-observed datapoint  $\mathcal{X} = \mathcal{X}_k$ . The estimation process is repeated for all  $c$ 's. However,  $c$  is omitted to simplify the notation (i.e.,  $\mathbf{w}^c$  becomes  $\mathbf{w}$ ).

The method uses the following coupled optimization problem to find the modality priors  $w_m$  for scene  $c$ :

$$\begin{aligned} & \underset{\mathbf{w}}{\text{minimize}} && \frac{1}{2} \sum_{k=1}^K \sum_{l=1}^L \left( \sum_{m=1}^M s_{k,l,m} w_m - b_{k,l} \right)^2 \\ & \text{subject to} && \mathbf{1}^T \mathbf{w} = 1 \\ & && 0 \leq w_m \leq 1, m = 1, \dots, M, \end{aligned} \quad (5)$$

The objective is to find the weights  $w_m$  that approximate the oracle  $b_{k,l}$  for every data point  $\mathcal{X}_k$ . Using the loss in Eq. 5, the problem becomes a cc-LS optimization problem.

This type of problem uses all points and poses labels from the training set to find the set of priors that approximates the values produced by the oracle for each point  $\mathcal{X}_k$  at once.

## 4.2. Multimodal Construction

The estimation method uses cc-LS optimization to minimize the difference between Oracle (b) and the multimodal matrix ( $\mathbf{A}$ ). It frames the trust estimation as a linear system of equations of the form  $\mathbf{A}\mathbf{w} - \mathbf{b} = 0$ , where the modality trust values are the elements of the vector  $\mathbf{w} = [w_R, w_D, w_P]^T$  that approximate  $\mathbf{A}\mathbf{w}$  to  $\mathbf{b}$ .

**Construction of the Multimodal Matrix ( $\mathbf{A}$ ).** The matrix  $\mathbf{A}$  contains label probabilities for each of the datapoints in the training set ( $K = |\mathcal{X}_{train}|$ ). This matrix has  $U$  rows ( $U = KL$ ) and  $M$  columns, where  $L$  is the total number of labels ( $L = |\mathcal{Z}|$ ), and  $M$  is the number of modalities ( $M = 3$ ) and has the following structure:

$$\mathbf{A} = [\mathbf{S}_{k=1}^T, \dots, \mathbf{S}_{k=K}^T]_{U \times M}^T \quad (6)$$

where  $\mathbf{S}_k(l, m) = s_{k,l,m}$ .

VARIABLES	
SYMBOL	DESCRIPTION
$\mathbf{A}$	Multimodal matrix $\in \mathbb{R}^{U \times M}$
$\mathbf{a}_m$	$m$ -th column vector of $\mathbf{A}$ with $U$ elements
$\mathbf{b}$	Oracle vector $\in \mathbb{R}^U$
$\mathbf{b}_m$	Oracle column vector for modality $m$
$\mathcal{C}$	Scenes set (light $\times$ occlusion) combination
$c$	Scene index, $1 \leq c \leq  \mathcal{C} $
$\text{CLF}_{k,m}$	Classifier for the $m$ -th modality
$\{f_{\mathcal{N}_m}\}_k$	Set of feature $M$ vectors for the $k$ -th datapoint
$D$	Depth modality
$h$	Head camera view
$K$	Dataset size, $K =  \mathcal{X} $
$k$	Datapoint index, $1 \leq k \leq K$
$L$	Size of the set of pose labels, $L =  \mathcal{Z} $
$l$	Index of the pose label ( $\mathcal{Z}_l$ ), $1 \leq l \leq L$
$\hat{l}$	Index of the estimated pose label, $1 \leq \hat{l} \leq L$
$l^*$	Index of the ground truth label ( $\mathcal{Z}_l$ ), $1 \leq l^* \leq L$
$MM$	Multimodal and Multiview
$MpM$	Multimodal and partial-Multiview
$M$	Size of modality set, $M =  N $
$m$	Modality index, $1 \leq m \leq M$
$\mathcal{N}$	Modality set $\mathcal{N} = \{R, D, P\}$ indexed by $m$
$P$	Pressure modality
$pMM$	Partial-Multimodal and Multiview
$PMpM$	Partial-Multimodal partial-Multiview
$R$	RGB modality
$s$	Side camera view
$s_{k,l,m}$	Probability of label $l$ from $\text{CLF}_{k,m}$
$t$	Top camera view
$U$	Multimodal dimension $U = KL$
$\mathcal{V}$	View set $\mathcal{V} = \{t, s, h\}$
$V$	Number of views $V =  \mathcal{V} $
$v$	View index, $1 \leq v \leq V$
$\mathbf{w}^c$	Trusts $\mathbf{w} = [w_R, w_D, w_P]^T$ for scene $c$
$w_{N_m}$	Modality trust value (e.g., $w_R$ for $m = 1$ )
$\mathcal{X}$	Dataset indexed by $k$ (i.e., $\mathcal{X}_k$ )
$\mathcal{X}_k$	$k$ -th datapoint with $\{f_{N_m}\}_k = \{f_R, f_D, f_P\}_k$
$Y$	$MM$ dimensions ( $= KLV$ )
$\mathcal{Z}$	Sleep-pose set

Table 1: Variables and their descriptions.

### Construction of the Multimodal Oracle Vector ( $\mathbf{b}$ ).

The vector  $\mathbf{b}$  is generated by the oracle and quantifies the classification ability of the combined modalities. It is used to corroborate estimation correctness when compared to the ground truth. The  $\mathbf{b}_m$  column vectors have  $U$  rows:

$$\mathbf{b}_m = [\mathbf{b}_{k=1}^T, \dots, \mathbf{b}_{k=K}^T]^T, \quad (7)$$

where  $\mathbf{b}_k = [b_{k,l=1}, \dots, b_{k,l=L}]^T$ . The values of the  $b_{k,l}$  elements are set using the following condition:

$$b_{k,l} = \begin{cases} 1, & \text{if } \hat{l} = l^* \text{ for } \mathcal{X}_k \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

where  $\hat{l} = \arg \max s_{k,l,m}$  is the index of the estimated label and  $l^*$  is the index of the ground truth label for  $\mathcal{X}_k$ .

The construction of the oracle  $\mathbf{b}$  depends on how the columns  $\mathbf{b}_m$  (i.e., unimodal oracles) are combined. The system is tested with a uniform construction and the results are reported in section 5. In the uniform construction, each modality has a  $\frac{1}{M}$  voting power and can add up to one via:

$$\mathbf{b} = \frac{\sum_{\forall m} \mathbf{b}_m}{M}. \quad (9)$$

### 4.3. Coupled Constrained Least-Squares (cc-LS)

Finally, the weight vector  $\mathbf{w} = [w_R, w_D, w_P]^T$  is computed by substituting  $\mathbf{A}$  and  $\mathbf{b}$  into Eqn. (5) and solving the cc-LS optimization problem:

$$\begin{aligned} & \underset{\mathbf{w}}{\text{minimize}} && \frac{1}{2} \|\mathbf{A}\mathbf{w} - \mathbf{b}\|_2^2 \\ & \text{subject to} && \mathbf{1}^T \mathbf{w} = 1 \\ & && 0 \leq w_m \leq 1, m = 1, \dots, M \end{aligned} \quad (10)$$

Intuitively, the cc-LS problem finds the modality priors that allow the method to fuse information from different modalities to approximate the oracle probabilities.

### 4.4. Multiview Formulation

The bounded multimodal formulation is expanded to include multiview data using  $V$  views indexed by  $v$ . The values that  $v$  can take indicate which camera view is used (e.g.,  $v = 1$  for the top view,  $v = 2$  for the side view, and  $v = 3$  for the head view.) The multimodal and multiview matrix  $\mathbf{A}$  has the following form:

$$\mathbf{A} = [[\mathbf{A}^{(v=1)}], \dots, [\mathbf{A}^{(v=V)}]]_{Y \times M}^T, \quad (11)$$

where  $Y = LKV$  for a system with  $V$  views and  $M$  modalities. The  $\mathbf{b}_m$  multimodal and multiview oracle vector is constructed by concatenating data from all the views in the set ( $\mathcal{V}$ ) via:

$$\mathbf{b}_m = \left[ [\mathbf{b}_{k=1}^{(v=1)}]^T, \dots, [\mathbf{b}_{k=1}^{(v=V)}]^T \right]_Y^T, \quad (12)$$

and the  $\mathbf{b}$  column vector is generated using (9).

### 4.5. Testing

The test process is shown in Figure 6. The room sensors in combination with  $\mathcal{N} = \{R, D, P\}$  measurements are collected from the ICU scene. Features ( $\{f_{N_m}\}_k$ ) are extracted from the modalities in  $\mathcal{N}$  and are used as inputs to the trusted multimodal classifier. The classifier outputs a set of label candidates from which the label with the largest probability for datapoint  $\mathcal{X}_k = \{f_{N_m}\}_k$  is selected via:

$$\hat{l}_k = \arg \max_{l \in L} (w_{N_m} \text{CLF}_m \{f_{N_m}\}_k), \forall m. \quad (13)$$

**Missing Modalities** Performance of the Eye-CU system with missing sensor information is evaluated using trust adjustment method from [18], in which trust value for a missing modality is set to zero and the originally estimated value is proportionally distributed to the remaining modalities.

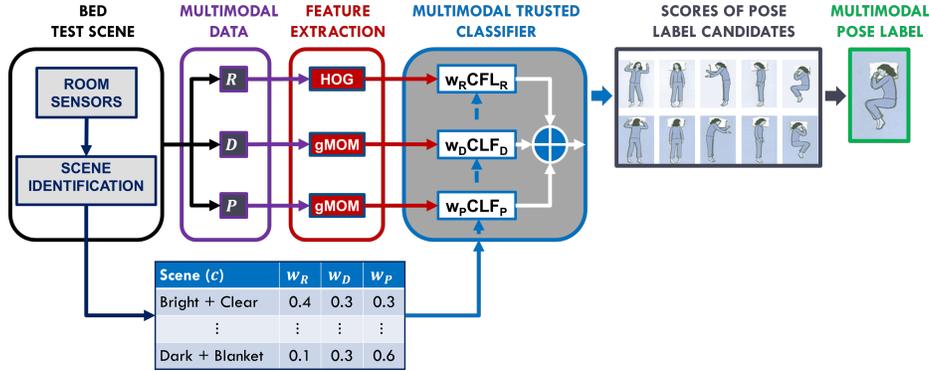


Figure 6: Block diagram for testing of a single view multimodal trusted classifier. Observations ( $R, D, P$ ) are collected from the scene. Features are extracted from the observations and sent to the unimodal classifiers to provide a set of score-ranked pose candidate labels. The set of candidates is trusted and combined into one multimodal set from which one with the highest score is selected.

## 5. Experiments

Validation of modalities and views for sleep-pose classification substantiates the need for a multiview and multimodal system. The cc-LS method is tested on the  $MpM$ ,  $MM$ ,  $PMM$  and  $PMpM$  Eye-CU configurations and data collected from scenes with various illumination levels and occlusion types. The labels are estimated using multi-class linear SVC ( $C=0.5$ ) and LDA classifiers from [12]. A validation set is used to tune the SVC’s  $C$  parameter and the Ada parameters. Classification accuracies are computed using five-fold cross validation using in-house implementations of competing methods and reported as percent accuracy values inside color-scaled cells.

### 5.1. Modality and View Assessment

Classification results obtained using unimodal and multimodal data without modality trust are shown in Figure 7. The cell values indicate classification percent accuracy for each individual modality and modality combinations with three common classification methods. The labels of the column blocks at the top of the figure indicates modalities used. The labels at the bottom of the figure show which classifier is used. The labels on the left and right indicate scene illumination level and type of occlusion. The figure only shows classification results for the top camera view because variation across views tested did not have statistical significance.

### 5.2. Performance of Reduced Eye-CUs

The complete  $MM$  configuration achieves the best classification performance, followed closely by the performances of the  $MpM$ ,  $PMM$ , and  $PMpM$  configurations, which is summarized in figure 8. The values inside the cells represent classification percent accuracy of the cc-LS method combined with various Eye-CU system configurations. The top row indicates the configuration. The second row indicates the views. The labels on the bottom of the figure identify the modalities. The labels on the left and right indicate illumination level and occlusion type. The

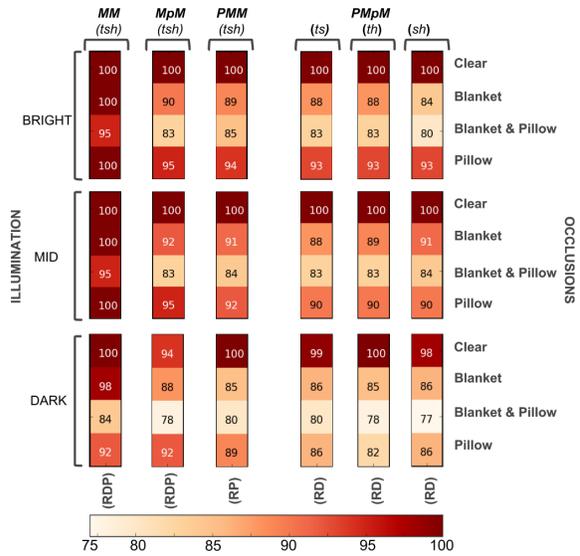


Figure 8: Classification performance in red scale (dark: best, light: worst) of the various Eye-CU configurations using LDA. The  $PMpM$  has the lowest performance of 76.7% using  $sh$  views of a dark and occluded scene. The method from [18] performs below 50% and the method from [7] is not suited for such conditions. The top row identifies the configuration. The second row indicates views used. The bottom labels indicate modalities used (in parenthesis). The labels on the left and right indicate scene illumination and occlusion type. Similar pattern is observed with SVC.

red scale ranges from light red (worst) to dark red (best). The figure shows that the complete  $MM$  system in combination with the cc-LS method performs the best across all scenes. However, it requires information from a pressure mat. The  $PMM$  and  $PMpM$  configurations do not require the pressure mat and are still capable of performing reliably and with only a slight drop in their performance. For example, in dark and occluded scenes the  $PMM$  and  $PMpM$  configurations reach 77% and 80% classification rates respectively (see row: DARK; Blanket & Pillow).

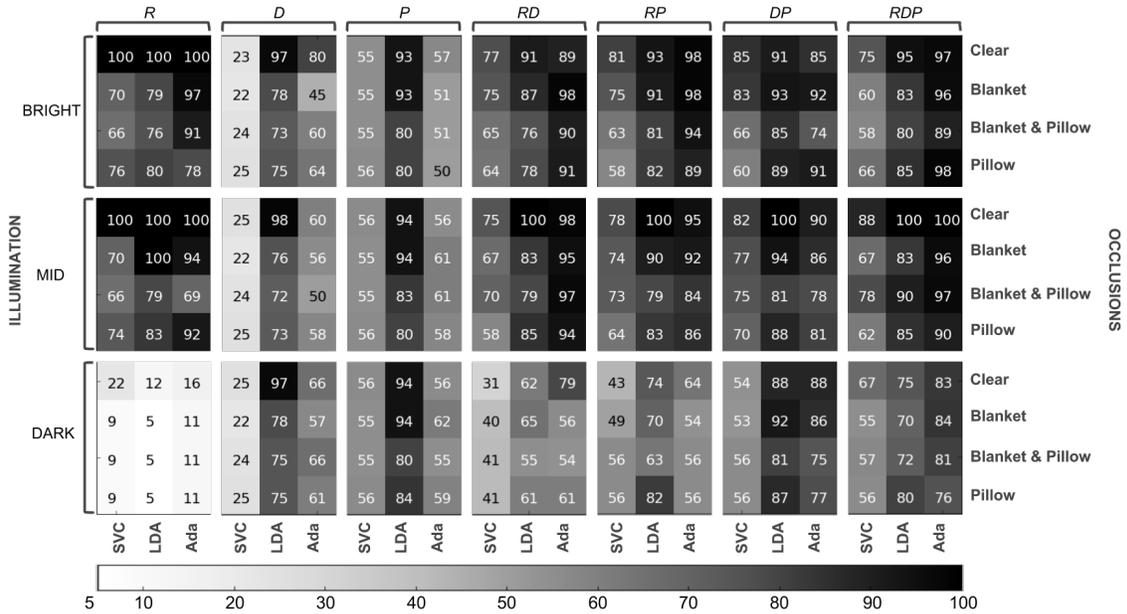


Figure 7: Performance evaluation of modalities and modality combinations using SVC, LDA, and Ada-Boosted SVC (Ada) based on their classification percent accuracy (cell values). The evaluation is performed over all the scene conditions considered in this study. The results indicate that no single modality ( $R, D, P$ ) or combination of concatenated modalities ( $RD, RP, DP, RDP$ ) in combination with one of three classification techniques cannot be directly used to recognize poses in all scenes. The top row indicates which modality or combination of modalities is used. The labels on the bottom indicate which classifier is used. The labels to the left and right indicate the scene’s illumination level and occlusion types. The gray-scaled boxes range from worst (white) to best (black) performance.

### 5.3. Comparison with Existing Methods

Performance of the cc-LS and the in-house implementations of the competing methods from [7] and [18] and Ada [4] are shown in Figure 9. The figure shows results using the  $MpM$  configuration, which more closely resembles those used in the competing methods. All the methods use a multimodal system with a top camera view and a pressure mat. The values inside the cells are the classification percent accuracy. The green scale goes from light green (worst) to dark green (best). The top row divides the methods into competing and proposed. The second row cites the methods. The bottom row indicates which classifier and, in parentheses, modalities are used. The labels on the left and right indicate illumination level and occlusion type. The results are obtained using the four methods with  $MM$  dataset.

**Confusion Matrices:** The confusion matrices in Figure 10 show how the indexes of estimated labels  $\hat{l}$  match the actual labels  $l^*$ . The top three matrices are from a scene with bright and clear ICU conditions (Figure 10a). The bottom three matrices illustrate the performance of the methods in a dim and occluded ICU scenario (Figure 10b). A dark blue diagonal in the confusion matrices indicates perfect classification. In the selected scenes, all methods achieved a 100% classification for the bright and clear scene. However, their performance varies greatly in dim and occluded scenes. The matrix generated using [7] achieves 7% classification accuracy (bottom left), matrix generated using [18] achieves

a 55% accuracy (bottom center), and the matrix generated with the cc-LS method achieves a 86.7% accuracy (bottom right). The  $MpM$  configuration with the cc-LS method outperforms the competing methods by an approximate 30%.

**Performance of Ada-Boost** The system is tested using Ada-Boost (Ada) algorithm [4] to improve the decision of weak unimodal SVCs. The results from Figure 7 show a slight SVC improvement. The comparison in Figure 9 shows that the Ada’s improvement is small. It barely outperforms the reduced  $MpM$  configuration with cc-LS method in some scenes (see row: MID-Blanket). Overall, Ada is outperformed by the combination of cc-LS and  $MpM$ .

## 6. Discussion

The results in Figure 9 show performance disparities between the results obtained with the in-house implementation and those reported in [7]. The data and code from [7] were not released, so the findings and implementation details reported in this paper cannot be compared at the finest level. Nevertheless, the accuracy variations observed are most likely due to differences in data resolutions, sensor capacities, scene properties, and tuning parameters.

The performance of the  $MM$  and  $MpM$  configurations, which use a pressure mat, is slightly improved. However, the deployment and maintenance of such systems in the real world can be very difficult and perhaps logistically impossible. The cc-LS method in combination with the  $PMM$  or

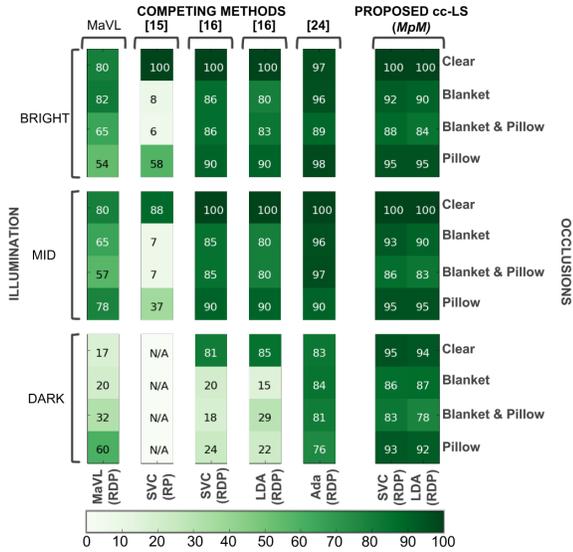


Figure 9: Mean classification performance in green scale (dark: best, light: worst) of MaVL, [7], [18], [4] and the cc-LS method using SVC and LDA. The combination of cc-LS and *MpM* matches the performance of competing methods in bright and clear scenes. Classification is improved with cc-LS by 70% with SVC and by 30% with LDA in dark and occluded scenes. The top row distinguishes between competing and proposed methods; the second row cites them. The bottom row indicates classifier and modalities (in parenthesis) used. The labels on the left and right indicate scene illumination and occlusion type. N/A indicates not suitable.

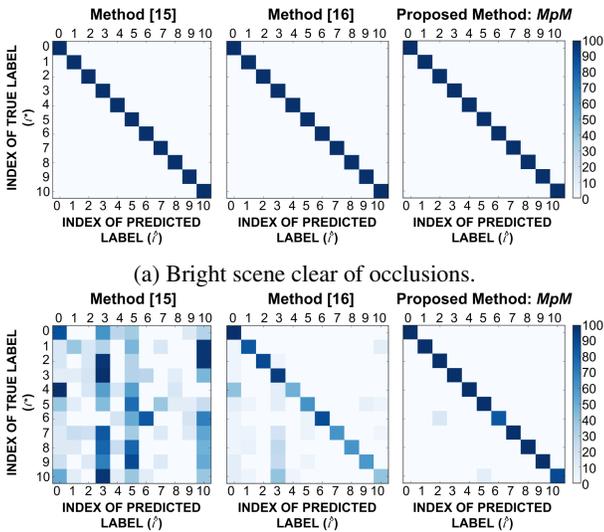


Figure 10: Confusion matrices generated in blue scale (dark: best, light: worst) using a top camera view and applying the methods from [7], [18], and cc-LS. The top matrices show all methods have perfect classification in ideal scenes (i.e., main diagonal). The bottom matrices are [7] with 7%, [18] with 55%, and cc-LS with 86.7% for dark and occluded scenes. The matrices show the matches between estimated ( $\hat{l}$ ) and ground truth ( $l^*$ ) indexes.

*PMpM* configurations, which do not use a pressure mat, match and outperform the competing techniques in ideal and challenging scenarios (see Figure 9).

## 7. Conclusion and Future Work

This work introduced a new modality trust estimation method based on cc-LS optimization. The trust values approximate the difference between the multimodal candidate labels **A** and the expected oracle **b** labels. The Eye-CU system uses the trust to weight label propositions of available modalities and views. The cc-LS method with the *MM* Eye-CU system outperforms three competing methods. Two reduced Eye-CU variations reliably classify sleep poses without pressure data. The *MM* properties allow the system to handle occlusions and avoid problems associated with a pressure mat (e.g., sanitation and sensor integrity).

Reliable pose classification methods and systems enable clinical researchers to design, enhance, and evaluate pose-related healthcare protocols and therapies. Given that the Eye-CU system is capable of reliably classifying human sleep poses in an ICU environment, expansion of the system and methods is under investigation to include temporal information. Future analysis will seek to quantify and typify pose sequences (i.e., duration and transition). Future work will investigate removing the constraints that clearly define the set of sleep poses and explore tools from novelty detection to identify other (e.g., helpful and harmful) patient poses that occur in an ICU. Recent studies indicate that deep features might improve the classification performance of the Eye-CU system in the most challenging healthcare scenarios. Hence, future work will investigate the performance and integration of deep features into the cc-LS method and the Eye-CU system.

**Acknowledgements** This project was supported in part by the Institute for Collaborative Biotechnologies (ICB) through grant W911NF-09-0001 from the U.S. Army Research Office; and by the U.S. Office of Naval Research (ONR) through grant N00014-12-1-0503. The content of the information does not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred.

## References

- [1] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa. Motion history image: its variants and applications. *Machine Vision and Applications*, 2012.
- [2] 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.
- [3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.
- [4] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Jrn'l of computer and sys. sci.*, 1997.
- [5] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2004.
- [6] M.-K. Hu. Visual pattern recognition by moment invariants. *IEEE IRE Trans. on Info. Theory*, 1962.
- [7] W. Huang, A. A. P. Wai, S. F. Foo, J. Biswas, C.-C. Hsia, and K. Liou. Multimodal sleeping posture classification. In *Proc. of the IEEE Int'l Conf. on Pattern Recognition (ICPR)*, 2010.
- [8] C. Idzikowski. Sleep position gives personality clue. *BBC News. September 16*, 2003.
- [9] C.-H. Kuo, F.-C. Yang, M.-Y. Tsai, and L. Ming-Yih. Artificial neural networks based sleep motion recognition using night vision cameras. *Biomedical Engineering: Applications, Basis and Communications*, 2004.
- [10] W.-H. Liao and C.-M. Yang. Video-based activity and movement pattern analysis in overnight sleep studies. In *Proc. of the IEEE Int'l Conf. on Pattern Recognition (ICPR)*, 2008.
- [11] S. Morong, B. Hermsen, and N. de Vries. Sleep position and pregnancy. In *Positional Therapy in Obstructive Sleep Apnea*. Springer, 2015.
- [12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.
- [13] T. Penzel and R. Conradt. Computer based sleep recording and analysis. *Sleep medicine reviews*, 2000.
- [14] J. Platt et al. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods: support vector learning*, 1999.
- [15] S. Ramagiri, R. Kavi, and V. Kulathumani. Real-time multi-view human action recognition using a wireless camera network. In *Proc. of the ACM/IEEE Int'l Conf. on Distributed Smart Cameras (ICDSC)*, 2011.
- [16] C. Sahlin, K. A. Franklin, H. Stenlund, and E. Lindberg. Sleep in women: normal values for sleep stages and position and the effect of age, obesity, sleep apnea, smoking, alcohol and hypertension. *Sleep medicine*, 2009.
- [17] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, and A. Kipman. Efficient human pose estimation from single depth images. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2013.
- [18] 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 *Proc. of Springer Int'l Conf. on Computer Vision Sys. (ICVS)*, 2015.
- [19] G. L. Weinhouse and R. J. Schwab. Sleep in the critically ill patient. *Sleep-New York Then Westchester*, 2006.
- [20] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2013.