RAM: Role Representation and Identification from combined Appearance and Activity Maps

Carlos Torres[†] Archith J. Bency[†] Jeffrey C. Fried[‡] B. S. Manjunath[†]

[†]University of California Santa Barbara [‡]Santa Barbara Cottage Hospital
{carlostorres, archith, manju}@ece.ucsb.edu, jfried@sbch.org

ABSTRACT

This work introduces a multimodal multiview camera network for role identification and re-identification in an Intensive Care Unit (ICU) room, where identifying individuals is not permitted. The analysis challenges include imaging conditions such as medical isolation (where all visitors wear scrubs), poor and non-uniform illumination, or variable camera views. We propose a role representation, which combines static appearance features such as texture and color, together with a dynamic quantification of human locations and interactions that results in a semantic map. The proposed representation is easy to compute and robust to varying ICU conditions and network configurations, which make the methods suitable for low-power distributed sensor network deployment. Thorough evaluations and comparisons with competing methods are performed. The findings from this approach enable the compliant analysis of workflows in healthcare, while protecting the privacy of patients and medical staff.

CCS CONCEPTS

•**Computer systems organization** →**Sensor networks;** *External interfaces for robotics;*

KEYWORDS

Appearance, Healthcare, Role, Representation, Workflows, Activity, Multimodal, Semantic Map, HIPAA, Identification, Monitoring, ICU

ACM Reference format:
Carlos Torres[†] Archith J. Bency[†] Jeffrey C. Fried[‡] B. S. Manjunath[†]
[†]University of California Santa Barbara [‡]Santa Barbara Cottage Hospital
{carlostorres, archith, manju}@ece.ucsb.edu, jfried@sbch.org . 2016. RAM: Role Representation and Identification
from combined Appearance and Activity Maps. In *Proceedings of ICDSC, 2017, Session on Target Tracking and Person Re-Identification,* 7 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Person identification, re-identification, and tracking are essential in many healthcare settings. However, collecting and using such identifiable information is prohibited in most cases by the The Health Insurance Portability and Accountability Act (HIPAA) [7]. To address this limitation, this work introduces a novel framework for Role Representation and identification from Appearance and Activity Maps (RAM). We demonstrate its application in an Intensive Care Unit setting in a real hospital environment where RAM learns ICU-associated roles. RAM combines static appearance features (texture and color) with quantified dynamic human locations and interactions (semantic maps) to describe these roles.

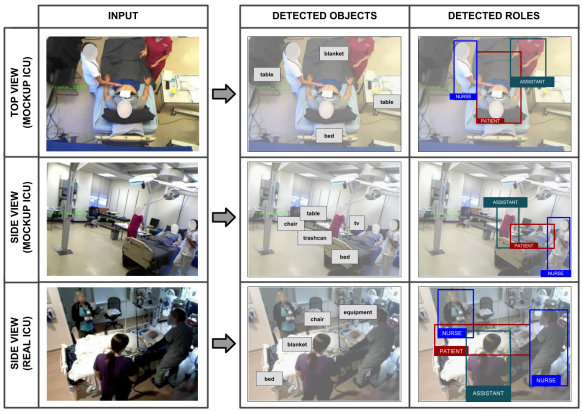


Figure 1: Views of the mock-up ICU room (top two rows) and the medical ICU room (bottom). The columns are the input frames (left), the detected objects (center), and the detected roles in the color bounding boxes (right).

The proposed representation is simple, easy to compute, and robust to natural conditions, which makes it suitable for low-power distributed network deployment. The semantic maps make RAM independent of network configuration. The performance of RAM is evaluated on 11 days of multimodal multiview data and compared with the latest methods. Thorough evaluation of RAM is performed to justify its components and to compare its performance with competing appearance-based and tracking-based methods. The findings from this approach will enable the privacy-compliant analysis of workflows in healthcare and other areas where identifying individuals is not permitted. RAM identifies roles (not individuals), protecting patient and staff privacy, while ensuring workflows remain unaffected by surveillance mechanisms. Figure 1 shows sample inputs, detected relevant objects, and estimated roles in a mock-up and a medical ICU room.

Medical Background. There is an increasing interest in role identification and analysis in healthcare [24], due to its potential benefits in improving and optimizing care. One major gain from role identification and analysis is in defining each person's responsibilities, ensuring appropriate implementation of each professional's role,

ICDSC, 2017

DOI: 10.1145/nnnnnnn.nnnnnnn

optimizing professional scopes of practice, and thereby ensuring efficient patient management [3]. Although clinicians agree that detailed understanding of workflows is essential to quality of care, healthcare restrictions prohibit the use of people's identifiable information. To circumvent these data restrictions, RAM introduces methods for role representation and role identification based on appearance features (for training and system initialization) and semantic activity maps (location and interactions).The idea behind semantic maps is that different roles interact with different objects, visit different locations, and maintain a certain distance from certain objects in the ICU. This is further enhanced by observing subjects over time. For example, a patient might walk past a ventilator, but a medical practitioner will spend more time in that area. Similarly, visitors spend more time close to the patient sitting in chairs, while staff mostly avoids chairs. Although semantic map features alone can be used to infer roles, their combination with appearance information achieves greater accuracy.

Related Technical Work. Studies analyzing healthcare environments include using a single RGB-D sensor, RFIDs, and proximity sensors to record activities in a neo-natal ICU as in [12] Workflows in an operating room are analyzed in [18] and the analysis tasks are very complex. One significant limitation is considering that any activity can be performed by any individual. This makes the action space relatively large, which decreases accuracy. One helpful concept in improving outcomes includes identifying roles who perform distinctive and common activities and using this information to identify roles (e.g., patient, doctor, or staff). The surveys from [25] and [11] describe the challenges and most popular techniques in person re-identification. Existing methods for identification and re-identification range from methods leveraging deformable parts [4] to feature representation and metric learning [13] to video ranking [26] (as an alternative to single-frame approaches). The work in [15] introduced a distributed network framework for node performance comparison and person re-identification that can be used to estimate optimal camera topology. The authors in [8] argue that most existing methods depend on person pose and orientation variations and introduce a technique to model such variations in the feature space. Also, there are several feature representations that have pushed the limits of performance to new levels. Appearance-based representations such as the ensemble of local features (ELF) [9] and symmetry-driven accumulation of local features (SDALF) [1] encode color properties. Similarly, salience matching and learning [28], [29] and mid-level filters [30] depend on relative patch contrast and distinctiveness. Although the previously cited research achieves impressive results, their appearance-based methods directly depend on proper imaging conditions, such as bright, and uniform illumination, and view angle between the individual and the camera. In addition, appearance-based role representation alone is not sufficient. For instance, medical isolation procedures to protect compromised patients require that all people entering the ICU room wear disposable isolation scrubs, so all roles appear identical. Another limitation of these representations in real-world applications, such as healthcare, is their inability to evolve over time (i.e., to consider temporal information) and to integrate interaction information. The proposed approach introduces a novel role representation; a semantic activity abstraction and extraction algorithm to identify; and a method for role identification based on the sequence of observed activities, visited locations, and detected interactions (cones for orientation and proximity). The proposed methods are capable of dealing with cases when role-based visual features are obfuscated by extreme scene and appearance changes.

The main contributions of RAM are:

- A simple, novel, and effective hybrid method for role representation, combining visual appearance (static) and semantic activity map (dynamic) features.
- A modular, non-disruptive, and inexpensive multimodal sensor network controlled by Raspberry Pi3 devices [23]. The network performance and ability to work with existing hospital infrastructure is being tested in a medical ICU.
- An experimentally validated decentralized method for learning role representations and inferring unknown roles.
- A score aggregation method, which combines the decentralized decisions to improve role identification accuracy.

2 DATA COLLECTION

The sensor locations and camera views are shown in Figure 2. The methods from [10] are used to calibrate the three cameras and estimate the floor plane of the ICU. A total of eleven days of video data (approximately a total of 264 hours, 15, 840 minutes, or 950, 400 seconds) are collected covering six nurse assistants, four caterers, five medical doctors, four facilities and janitorial personnel, ten nurses, five patients, twelve visitors, and two days of isolation.

Recall that when the room is in isolation, all visitors as well as hospital staff are required to wear disposable scrubs, causing all roles to appear identical. Multiple roles were observed manually and added to the counts for these two days.

The set of roles $\mathcal{R}$ and symbols representing each of the eight observed roles are: nurse (A)ssitant, (C)aterer, medical (D)octor, (F)acilities, (I)solation, (N)urse, (P)atient, (V)isitor. The role set is indexed by r, where R_r with $r = 1$ is used to indicate the role of nurse assistant. Figure 3 shows samples of the various roles. The total frequency count of observed/collected instances for each role (in number of minutes on the vertical axis) is show in Figure 4. The data also include three hours collected in a mock-up ICU room with actors playing four patients, one nurse, one visitor, and one nurse assistant. Note that about 30% of the data contains more than one role, patient-visitor being the most common. The scope of this work is focused on role representation and identification. A future paper will describe the analysis of activities using the complete anonymized (HIPAA compliant) and fully annotated dataset.

3 DESCRIPTION OF THE PROBLEM

There are multiple problems and stages in role representation and role identification. At the high level, the objective can be described as assigning roles to people observed in an ICU room using multimodal and multiview videos. The challenges come at the lower levels of the analysis. For instance, what are good appearance role representations, how do roles evolve over time, and how can one infer roles when appearance features are not discernible. Therefore, given a set of training multimodal multiview videos, the first problem involves identifying a role using appearance features to create an appearance dictionary ($\mathcal{A}$). Nurses tend to wear unique uniforms in order to keep a sense of their individuality in the work

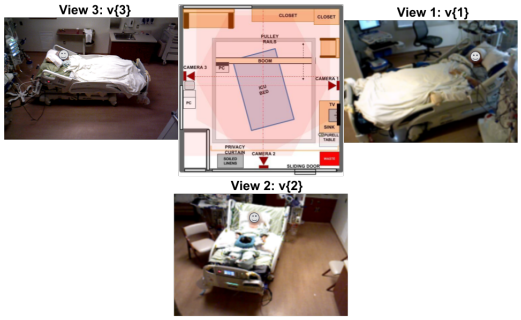


Figure 2: Location and views of the three nodes in the ICU room. Each node is composed of an RGB-D sensor, a Raspberry Pi3, and a battery, all inside an aluminum enclosure.

Figure 3: The eight roles associated with the ICU room.

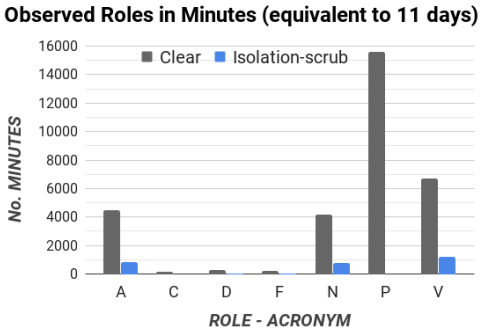


Figure 4: Number of minutes each of the roles is observed, over the 11 days of data. Gray bars indicate that differentiating roles based on appearance is possible, while blue bars indicate isolation scrubs are used (hence no appearance difference). Note: isolation scrubs were observed for a total of 2, 795.04 minutes or 1.91 days.

place [3]. This makes identification based on appearance alone unreliable. The solution is related to the second problem, which corresponds to learning the semantic motion dynamics associated with each of the observed roles and creating a dictionary of such representations and roles ($\mathcal{H}$). The last problem involves matching an unseen video with information extracted at training to find the best matching role R^* by maximizing the combined score $(S_r^A + S_r^M)$ from matching appearance features (f_A) to obtain the appearance score S^A and semantic activity features f_M to obtain the semantic score S^M) for all roles indexed by r, where $1 \leq r \leq R$ and R is the number of roles under consideration in the set of roles $\mathcal{R}$.

4 APPROACH AND METHODS

Consider the ICU rooms in Figure 1. Intuitively, hospital visitors look different from healthcare staff, often dressing differently. In addition, different roles perform different activities and visit different locations in the room. Some activities such as entering the ICU are performed by all roles, while some social or medical activities are performed only by specific roles. For example, nurses check ventilators and janitorial personnel clean rooms and empty trashcans, while visitors sit and interact with patients for longer time intervals and in closer proximity. The objective is to identify the set of locations (corresponding to the various activities) that each role visits along with the associated objects and interaction cone configurations. The interaction cones are used to quantify a person's relative distance and orientation to objects of interest. The variability of visit duration by a role and the set of observations (semantic features) is unbounded and can be very short or very large. A method to extract semantic features over time to deal with the variability in visit duration by certain roles is also proposed. The features are discriminative and informative and their nature allows them to be independent of their chronological order.

The first step is detecting individuals in the scene via [5]. The set of semantic features representing the r-th role ($R_r \in \mathcal{R}$) at time t is represented as $f_{M,r,t} = \{g_x, g_y, C_{q,d,o}\}_{r,t}$, where g_x and g_y are the grid coordinates, $C_{q,d,o}$ is the interaction cone vector for quadrant (q), and interaction distance (d) in reference to objects in the set O = {bed, chair, computer, doorway, person, sink, table, trashcan, tray, and ventilator} indexed by o for $1 \leq o \leq 10$. Object detection for $o \in O$ is applied to the RGB modality. The semantic features $f_{M,r,t,v}$ are extracted for each role r seen by each view v ($1 \leq v \leq V$) at time t ($1 \leq t \leq T$) from a network with V views over a duration T. The $f_{M,r,t,v}$ vectors are used to build the dictionary of semantic activity maps across all roles $\mathcal{M} = \{M\}_{r,t,v} \in \mathbb{R}^{\text{RTV}}$. The RGB modality is used to perform the initial person detection, which is used to initialize a blob tracker in the Depth modality. The variables g_x and g_y are used to impose smoothness and serve to constrain the blob tracker.

4.1 Training

Training is split into two stages: static, where appearance features f_A are used to create a dictionary of roles $\mathcal{A}$; and dynamic, where a dictionary of semantic histograms $\mathcal{M}$ for all roles is learned from the semantic activity map features $f_{M,r,t,v}$ extracted for all roles, over time, and across all views.

4.1.1 Role Representation. The process of learning a role representation starts with identifying the set of semantic states corresponding to each role. Given a set of videos $\mathcal{K}$, each with $K = |\mathcal{K}|$ frames, the first step is to extract the appearance vectors f_{Ak} for $1 \leq k \leq K$. At time $t = 0$, appearance vectors are extracted from the detected person (given by the tracked blob form the depth modality). The vectors are composed of two elements: a 128-dimension GIST vector (one scale) for texture [17], and the 96-dimension (first and second order) color histogram vector [27]. These appearance vectors serve to identify distinct visitor clothing patterns and generic healthcare staff uniforms. The combination of appearance and semantic map features yields a reliable role representation that is easy to compute. This gives a great advantage for deployment in low-power distributed sensor networks.

Appearance Dictionary. GIST features [17] are used to represent texture patterns in clothing corresponding to the various roles.

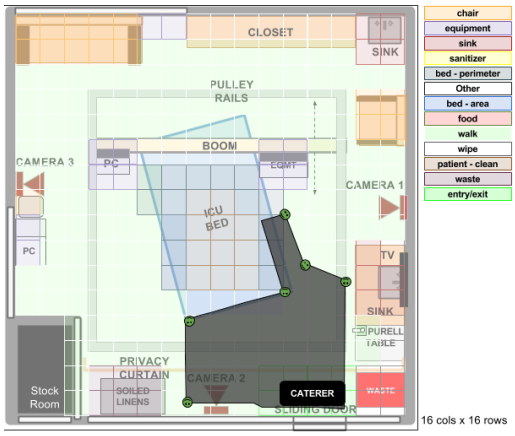


Figure 5: Semantic activity map for the caterer role in a 16x16 grid overlayed in black.

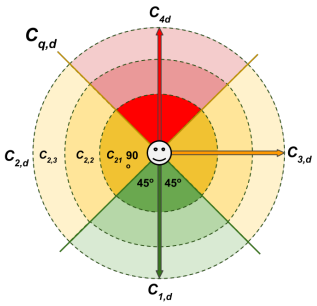


Figure 6: Role interaction cones $C_{q,d}$ for quadrant q and distance d showing regions of highest interaction (green), mid interaction (yellow), and lowest interaction (red). Darker disks indicate higher interaction probability (closer), while lighter disks indicate lower interaction (farther away). The color and color scales are used to indicate person orientation and proximity.

Color histograms are used to help characterize clothing styles such as uniforms. A color variance threshold (th) is used to indicate a person is wearing a uniform (e.g., janitorial or caterer uniforms). This information helps to distinguish between hospital staff and visitors. The 96-element color histogram is computed based on [27] by combining the first moment (mean) and second moment (standard deviation) on the 16-bin histograms extracted from each of the three channels in the HSV color space. The texture and color features are concatenated to form f_A. The f_A vectors are used to create the appearance dictionary of $\mathcal{A} \in \mathbb{R}^{AR}$, where A is the cardinality of the appearance feature vector ($A = 16$) and R is the number of roles ($R = 8$). The dictionary of appearance features ($\mathcal{A}$) is used to compute Linear Discriminant Analysis (LDA) [19] boundaries for each role in $\mathcal{R}$ and are represented by λ_r. The hyper-planes (decision boundaries) are used to score a new sample by computing the distance to all, but selecting the closes one i.e., S^A inversely proportional to the distance to the closest LDA-hyperplane.

Semantic Activity Map. . The semantic activity map corresponds to the dynamic analysis of the roles. The floor plane is estimated from the depth modality. A set of maps is computed for each of the eight roles over time. A sample map for the caterer role is shown in Figure 5. A dictionary of semantic features ($\mathcal{M}$) is computed for all role instances in the training set. The semantic histogram features f_M are used to create a semantic action bag of visual words based on [16]. Various object detectors are tested based on their performance and complexity for offline and on-board processes. Evaluation of object detectors is beyond the scope of this work; however, the best performing detector for offline-ICU processes is YOLO [22], which uses convolutional neural network architecture. The best performing detector capable of running on the Raspberry Pi3 is [6], which detects objects from learned attributes.

Location and Interaction Quantification. The interaction cones in Figure 6 represent the 120-element vector $C_{q,d,o}$ for $1 \leq q \leq 3$ and $1 \leq d \leq 4$, with shape $[c_{q=1,d=1}, \ldots c_{q=1,d=3}, \ldots c_{q=4,d=3}]$. The feature vector is computed at each t with distance and orientation relative to each of the objects in O. The radius of each disk is computed based on distances greater than the average adult arm-length (outer disk:> 4ft), between forearm and full arm's length (middle disk:2 – 4ft), and within forearm length (inner disk: < 2ft). The poselets from [2], [14], and [21] are evaluated for usage in the ICU. Experimentally, the poselet estimator from [21] is used to compute the orientation cones, which is assigned to the closest quadrant. The poselets are given with respect to the camera that detected the person and mapped across the ICU floor plane. The distance (disk) is computed between detected objects and detected individual's blob centroids and assigned to the closest d.

The semantic map features M are composed of $[G, C]$, where $G = \{g\}_{x,y}$ and $C = \{c\}_{q,c,o}$. During implementation G is only used to track the various individuals and enforce smoothness and C is used to create semantic activity distance histograms (f_M for $r \in R$) each with a voting power of $\frac{1}{q+d}$. The histograms are combined to create the semantic dictionary ($\mathcal{M}$).

4.2 Testing

The objective at testing is to find the role R with the maximum score across all views over time ($0 \leq t \leq T$). The static scores ($t = 0$) are computed from appearance features alone as follows:

$$S_r^A = \frac{1}{V} \sum_{v=1}^{V} (S^A)_{r,v} \tag{1}$$

where S_r^A is the average score of role r for an individual seen through camera view v using appearance features f_A given by:

$$S^A = \min_r \mathrm{D}(f_A, \lambda_r) \text{ for } 1 \leq r \leq R, \tag{2}$$

where λ_r is the Linear Discriminant Analysis (LDA) boundary for role r. The method from [20] is modified to use scores as the distance ($\mathrm{D}(\cdot)$) to the LDA boundary (λ_r) for role r.

The dynamic scores ($1 \leq t \leq T$) are given by:

$$S_r^M = \sum_{v=1}^{V} \sum_{t=1}^{T} \min\ D(f_{M,t,v}, M_r) \text{ for } 1 \leq r \leq R. \tag{3}$$

The operator $D(\cdot)$ represents the computation of Earth Mover's Distance (EMD) between the observed instance $f_{M,t,v}$ and the role-histograms $M_r \in \mathcal{M}$ at time t from view v.

Finally, the role with the maximum score is found via:

$$R^* = \arg \max_{1 \leq r \leq R} (S_r^A + S_r^M) \tag{4}$$

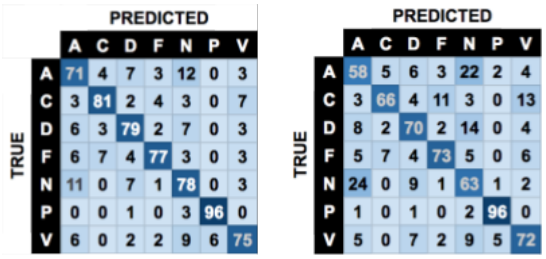
This approach has the additional advantage of ignoring the sequence of activities, which are not required to be sequential.

5 PERFORMANCE

The performance of RAM is evaluated under different camera views for accurate role identification using a stratified 10-fold cross-validation evaluation scheme. Average results across all folds are presented.

5.1 Non-Isolated and Isolated Environments

This experiment contains two parts. The first part uses appearance features and semantic maps for role identification in non-isolated environments. The confusion matrix in Figure 7(a) shows the qualitative performance of the proposed role representation. The second part of the experiments takes place in an isolated environment, where individuals have to wear blue disposable scrubs and appearance features are informative but non-discriminative. This means that appearance is used to detect the blue scrub but not to identify roles. In such case, the semantic features become the relevant input for classification and identification of roles. The confusion matrix for classification in isolated scenarios is shown in Figure 7(b).



(a) (b)

Figure 7: Role identification confusion matrices (a) confusion matrix of non-occluded identified roles. (b) Confusion matrix of the identified roles for individuals wearing isolation scrubs. All roles have the isolation pseudo-label, hence the symbol N/A on the row and column corresponding to the isolation role (I).

The role classification accuracy as a function of the number of semantic features observed over time ($t > 0$) is shown in Figure 8. The traces represent a one-vs-rest scheme for each role. The contribution of appearance features (f_A), semantic features (f_M), and their concatenated version ($f_A + f_M$) for identification of each of the roles is shown in Figure 9.

5.2 Decentralized Process: Camera Views

This experiment evaluates individual views and combinations of views by modifying equations 1 and 3. It serves to identify optimal views for accurate role identification. Obtaining a clear (unobstructed and direct) view of the activities and roles directly affects the role identification results shown in Figure 10. Camera locations and views are shown in Figure 2. There are two objectives behind this experiment: the first is to show that the decisions can be made at the individual nodes; and the second is to explore the best and worse case scenarios and to simulate the effects on identification performance due to sensor failures or sensor occlusions (i.e., in the ICU, views can be blocked by privacy curtains).

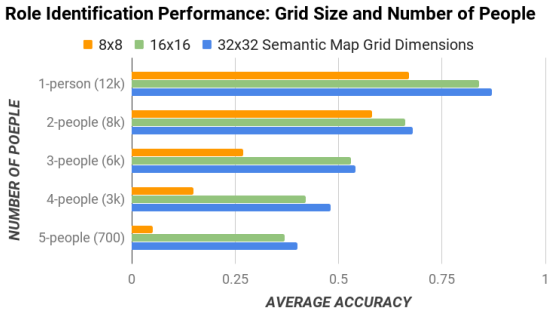


Figure 8: Mean role classification accuracy is based on the number of extracted semantic features. The duration of the observations helps to identify roles. Overall, roles observed for shorter periods of time are harder to identify. The vertical dashed line indicates that 74 observations is on average the best number for all detections. Although (I)solation is a scene condition, not a role, its identification accuracy is shown here as well.

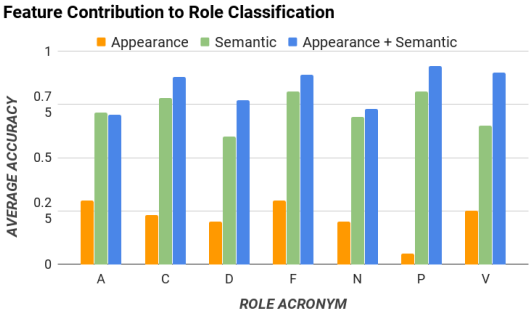


Figure 9: Mean role classification accuracy using appearance features (f_A), semantic features (f_M), and their combination ($f_A + f_M$).

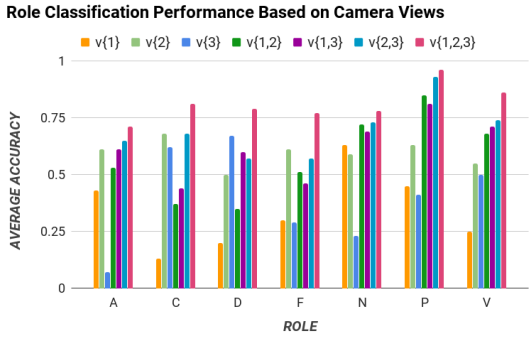


Figure 10: Average classification performance using all views and various reduced camera-view combinations. Sensor locations and views from right to left in clock-wise direction are: v{1}, v{2}, and v{3} shown in Figure 2. The bar plots indicate that better views of locations visited by specific roles help to better identify roles, while the best performing combination is the complete set of views.

5.3 Multiple-Target Role Identification

This experiment uses the combined RAM elements to represent, track, and identify roles in the ICU. The experiment is performed

on video instances with one or more people present in the scene. This experiment also validates the dimensionality of the semantic map based on accuracy and complexity as shown in Figure 11.

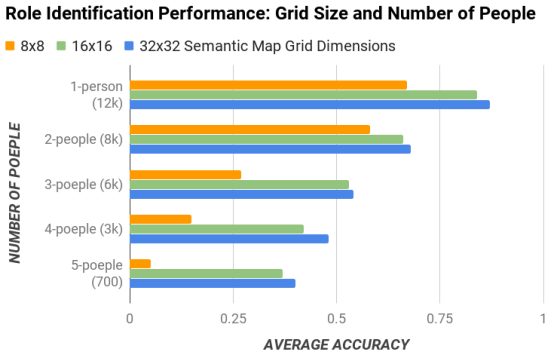


Figure 11: Average role classification accuracy as a function of semantic map grid dimensions and number of people in the scene. The worst performing grid size is 8×8 due to the artifact in which multiple people can occupy the same grid. The best performing grid size is 32×32; however, it is also the more complex.

5.4 Performance Comparison

The performance of RAM is compared with competing state-of-the-art methods as shown in Figure 12. The contrast methods are You-Only-Look-Once (YOLO) [22] and the method based on deformable part models and appearance from [13]. However, the competing methods only apply to the non-isolated environments, where appearance can be used to identify roles. The methods can detect isolation scrubs but cannot identify the occluded role.

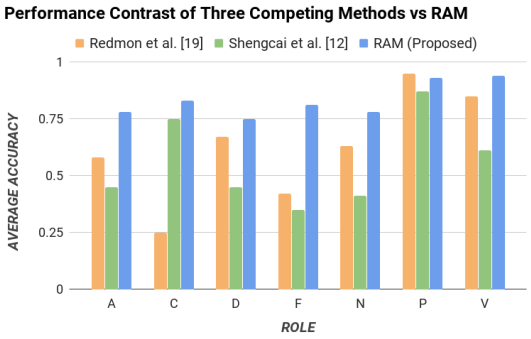


Figure 12: Performance contrast of four competing methods and RAM for ICU role identification in clear and isolation (I) conditions.

6 DISCUSSION AND CONCLUSION

The classification results indicate that roles can be identified by using a combined appearance and semantic activity approach. In some cases, individuals can be identified at the moment of first detection ($t = 0$) based on appearance features only. Although decentralized decisions are possible at the node level, the best individual decisions depend on having the optimal view of the activities in the room. The best role identification performance is achieved when appearance and semantic data from all nodes are combined.

The grid dimensions are evaluated experimentally. The best compromise between complexity and performance is met with the 16×16 grid size, as shown in Figure 11. In this case, each grid covers an approximate area of one square foot, which coincidentally is also the average area covered by a standing person (scene's top-view).

Recall that the objective of RAM is to identify various specific roles in the ICU. To reduce the study's search space, re-ranking was explored, but it was omitted due to its minimal impact. The intra-class similarities (e.g., among nurses, nurse assistants, and doctors) are small, compared with the large inter-class differences (e.g., between medical and non-medical staff).

7 FUTURE WORK

Future work will explore the evolution of roles in healthcare. For instance, due to the scarcity in the healthcare workforce, regular ICU visitors are often trained on basic healthcare tasks, alleviating some of the task load on the staff. This evolution can obfuscate RAM's re-identification analysis and can have a negative impact in its performance. In addition, future studies will investigate the identification of new roles based on anomaly detection.

Experiments involving object detectors and human poselet estimators indicate that good object detectors and poselet estimators directly affect the performance of RAM. Although multimodal-based detectors are still in their infancy, they are shown to outperform unimodal (e.g., RGB-only) object detectors. One improvement in the system and and future area of research would be the incorporation of object detectors and poselet estimators that combine RGB-D data. RAM uses a predefined set of objects; however, these directly depend on the specific environment and application. Future studies will explore identifying relevant objects and estimating an object's importance in role representation and identification. It is important to note that not all roles are explored in this study due to the limited number of observed instances. The continuous expansion of this study will allow the integration and analysis of additional roles.

High-level semantic activities are used in this study. However, finer analysis can be be used to better infer the roles. Potential studies will explore a finer level of dynamic motion information.

ACKNOWLEDGMENTS

Research was partially sponsored by the Army Research Laboratory (ARL) and was accomplished 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 Drs. R. Beswick (Director of Research), and Leilani Price (Compliance Officer), Mark Mullenary (BioMedical Department), and (ICU Nurse Manager) at Santa Barbara Cottage Hospital for their help deploying the system and their help consenting patients for the study.

REFERENCES

[1] Loris Bazzani, Marco Cristani, and Vittorio Murino. 2013. Symmetry-driven accumulation of local features for human characterization and re-identification. In *Elsevier Computer Vision and Image Understanding (CVIU)*.

[2] Lubomir Bourdev and Jitendra Malik. 2009. Poselets: Body part detectors trained using 3d human pose annotations. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*. IEEE.

[3] Isabelle Brault, Kelley Kilpatrick, Danielle DfiAmour, Damien Contandriopoulos, Véronique Chouinard, Carl-Ardy Dubois, Mélanie Perroux, and Marie-Dominique Beaulieu. 2014. Role clarification processes for better integration of nurse practitioners into primary healthcare teams: a multiple-case study. In *Nursing research and practice*. Hindawi Publishing Corporation.

[4] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. 2011. Custom Pictorial Structures for Re-identification. In *British Machine Vision Conf. (BMVC)*.

[5] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[6] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[7] Centers for Medicare & Medicaid Services and others. 1996. The health insurance portability and accountability act of 1996 (HIPAA). *Online at http://www. cms. hhs. gov/hipaa* (1996).

[8] Jorge García, Niki Martinel, Alfredo Gardel, Ignacio Bravo, Gian Luca Foresti, and Christian Micheloni. 2016. Modeling feature distances by orientation driven classifiers for person re-identification. *Elsevier Journal of Visual Communication and Image Representation* (2016).

[9] Douglas Gray and Hai Tao. 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Springer Proc. of European Conf. on Computer Vision (ECCV)*.

[10] R. I. Hartley and A. Zisserman. 2004. Multiple View Geometry in Computer Vision. Cambridge University Press.

[11] Srikrishna Karanam, Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, and Richard J Radke. 2016. A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets. *arXiv preprint arXiv:1605.09653* (2016).

[12] Colin Lea, James Facker, Gregory Hager, Russell Taylor, and Suchi Saria. 2013. 3d sensing algorithms towards building an intelligent intensive care unit. In *AMIA summits on translational science proceedings*. American Medical Informatics Association.

[13] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[14] Subhransu Maji, Lubomir Bourdev, and Jitendra Malik. 2011. Action recognition from a distributed representation of pose and appearance. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[15] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. 2016. Person re-identification in a distributed camera network framework. *IEEE Transactions on Cybernetics* (2016).

[16] David Nister and Henrik Stewenius. 2006. Scalable recognition with a vocabulary tree. In *IEEE Proc. on Computer Vision and Pattern Recognition (CVPR)*.

[17] Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. In *Springer Int'l Journal of Computer Vision (IJCV)*.

[18] Nicolas Padoy, Diana Mateus, Daniel Weinland, Marie-Odile Berger, and Nassir Navab. 2009. Workflow monitoring based on 3d motion features. In *IEEE Proc. of Int'l Conf. on Computer Vision Workshops (ICCV Workshops)*.

[19] 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. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* (2011).

[20] Sricharan Ramagiri, Rahul Kavi, and Vinod Kulathumani. 2011. Real-time multi-view human action recognition using a wireless camera network. In *ACM/IEEE Proc. of Int'l Conf. on Distributed Smart Cameras (ICDSC)*.

[21] Michalis Raptis and Leonid Sigal. 2013. Poselet key-framing: A model for human activity recognition. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[22] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*.

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

[24] Paul R Torrens. 2010. The health care team members: Who are they and what do they do. (2010).

[25] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. 2013. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)* (2013).

[26] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. 2014. Person re-identification by video ranking. In *Proc. of European Conf. on Computer Vision (ECCV)*.

[27] Hui Yu, Mingjing Li, Hong-Jiang Zhang, and Jufu Feng. 2002. Color texture moments for content-based image retrieval. In *IEEE Proc. of Int'l Conf. on Image Processing (ICIP)*.

[28] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2013. Person re-identification by salience matching. In *IEEE Proc. of Int'l Conf. on Computer Vision (ICCV)*.

[29] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2013. Unsupervised salience learning for person re-identification. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[30] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2014. Learning mid-level filters for person re-identification. In *IEEE Proc. of Conf. on Computer Vision and Pattern Recognition (CVPR)*.