ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/266590834

Video based Activity Recognition in a Trauma Center

Article · May 2012

CITATION
1

READS
61

2 authors:

Ishani Chakraborty
Rutgers, The State University of New Jersey
10 PUBLICATIONS **61** CITATIONS
SEE PROFILE

Ahmed Elgammal
Rutgers, The State University of New Jersey
203 PUBLICATIONS **7,586** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Subspace Learning View project

Zero-Shot Learning View project

All content following this page was uploaded by Ahmed Elgammal on 15 December 2014.

The user has requested enhancement of the downloaded file.

Video based Activity Recognition in a Trauma Center

Ishani Chakraborty, Ahmed Elgammal
Rutgers University
ishanic, elgammal@cs.rutgers.edu

Abstract

We present a feasibility study of automated, vision-based detection and recognition of trauma procedures in a medical emergency room. Given a ceiling-mounted camera view of the trauma room, our goal is to track and transcribe the activities performed during resuscitation of a patient, the time instances of their initiation and their temporal durations. We represent activities through complex spatio-temporal relationships between image features based on scene dynamics, patient localization, clinicians' hand motions and medical devices. We design an activity grammar based on trauma specific domain-knowledge and model the resulting logic as a Markov Logic Network. *Probabilistic inference of activity posterior is computed efficiently in the presence of observed features. To this end, we demonstrate our approach on videos of realistic trauma simulations in challenging, multi-agent. multi-task settings. This study primarily aims at exploring the overall problem of visual recognition of trauma procedures. The accuracy of the results we obtained from our recognition scheme confirms the suitability of our framework.*

1. Introduction

Trauma refers to "a body wound or shock produced by sudden physical injury, as from violence or accident"[1]. A patient suffering from trauma undergoes resuscitation in an emergency room. The resuscitation process involves multiple procedures including patient stabilization, monitoring vital signs, and determining the extent of injury. This process is performed by a team of specialist clinicians and documented in the medical record by a dedicated nurse recorder for future reference. The accuracy of the process as well as its transcription is limited due to the complexity of the performed tasks. The clinicians need to coordinate on time-critical procedures while constantly exchanging information about relevant statistics verbally. The task of the transcriber is to pick up relevant information in this noisy environment as well as keep an observation on multiple tasks being performed [22].

To alleviate the responsibilities of medical personnel and to aid procedure coordination, several researchers have proposed to automate activity recognition using sensory cues [4, 2, 1]. These works have primarily used embedded and RFID sensors that provide limited, object-centric signals. In contrast, a vision sensor is nonobtrusive and provides a scene-centric, rich description that is amenable to both computer and human interpretability. Consequently, a vision sensor has become an ubiquitous component in all types of activity surveillance systems [24, 10]. The focus of activity recognition on complex activities in specific environments e.g., sports, office etc. [28, 13, 20] has led to the inclusion of domain knowledge in system design.

In this paper, we consider a novel problem of recognition of trauma procedures in a hospital domain. The trauma room scenario presents a *complex, multi-agent and multi-task setting*, where procedures are simultaneous, interleaved and follow a flexible sequence. Moreover, since all activity is confined to the patient area, key observations are often occluded from camera view. In this context, we need to develop algorithms that specifically address these challenges while maintaining the generality of the solutions to be useful in other scenarios.

Activities are generally observed through spatio-temporal features extracted from the image sequence. Thus, the problem of activity recognition can be broadly divided into two steps; *feature extraction step*, in which we compute low-level, object and people related features using object detection, motion tracking and patient pose estimation, and *activity inference* step in which we represent activities through feature measurements and inter-feature relations. We encode the feature relations using logic based on domain knowledge and perform probabilistic inference using a Markov Logic Network(MLN) [18].

Multi-feature integration for activity recognition has been investigated in prior work. In [3], object detectors, motion tracks and people pose are combined in an MLN. Similar features and an innovative object-reaction cue are inputs to a graphical model in [9]. In [13], a specific scenario of

[1] www.thefreedictionary.com/trauma

Activity	*Patient body part*	*Image features*	*ATLS denotation*	*Duration, Repeated?*
Ventilation	Mouth	Oxygen Mask	Primary(Breathing)	Variable, Multiple
Intubation	Mouth	Laryngoscope	Primary(Airway)	Variable, Once
Listening to breath	Chest	Stethoscope	Primary(Breathing)	Fixed, Multiple
Chest compression	Chest	Hand position	Primary(Breathing)	Variable, Multiple
Check pulse	Wrist/feet/neck	Hand position	Primary(Circulation)	Fixed, Multiple
Cover/uncover	Body	Skin region	Primary(Exposure)	Variable, Once
Roll patient	Body	Scene dynamics	Secondary	Fixed, Once
Check blood pressure	Arms	BP cuff	Secondary	Fixed, Multiple
Check ECG	Chest	ECG leads,	Primary	Fixed, Once
Apply medication	Arms/feet	Syringe	Secondary	Variable, Multiple
Inspect head and eyes	Head	-	Secondary	Fixed, Once
Establish IV	Arms	Equipment approach	Secondary	Fixed, Once
Inspect body	Full body	-	Secondary	Fixed, Once

Table 1. Description of common trauma procedures. The top panel lists the *visually identifiable activities*.

basketball play is represented using player, ball tracks and field position cues. In our work, in addition to local, object and person based features we also include global scene dynamics to capture multi-person and agent-independent activities.

High level activity models can be broadly classified into probabilistic (e.g., Dynamic Baysian Networks and its variants [28, 8]) and plan and grammar based systems [11, 19]. Recently however, researchers have realized the benefits of fusing the two approaches through hybrid models [27]. Markov Logic Network is one such hybrid model that has been effectively applied on a variety of problems. For example, in [17], MLN is used to represent group activities for video matching in football datasets. A method for parking lot monitoring is proposed in [25] where multi-person, multi-vehicle interactions are expressed as first order logic formulas. Unlike aforementioned works which handle fixed time instances, Morariu et al [13], describe basketball rules using Interval Algebra to handle time durations. MLN depends on logic formulas that are usually hand-crafted for each environment. A set of general logic formulas are proposed in [3] for multi-cue activity recognition. However they are limited to classifying simple, single-person actions. We propose an activity grammar which comprises of generic rules applicable to many scenarios and trauma specific rules. Additionally, we also propose a generalized temporal duration model.

Activity recognition in a hospital domain is an active field of research. In [2], the phase of a surgical operation is inferred from RFID based people tracks and object use. Locations of people and their interactions with PDAs are mapped to activities using HMM in [21]. A method for automatic transcription of operating room events was proposed in [1], where the authors use a logic based approach to detect an activity sequence. To the best of our knowledge, ours is the first work that uses visual analysis to detect activities in a hospital domain. It is important to note here that vision is not a perfect technology and many activities cannot be explained by using visual analysis alone. The inclusion of complementary technologies such as RFID have been seen to improve the accuracy of activity recognition systems [16]. The aim of this work is to present the applicability of visual analysis in trauma procedures and more importantly, to point out the limitations which can be further improved by inclusion of other algorithms and technologies.

2. Overall approach

In this project, we perform a comprehensive overview of trauma workflow analysis using automated vision. We determine *visually identifiable* trauma activities and propose solutions based on state-of-the-art algorithms to represent activities using spatio-temporal, local and global features. For activity detection, we propose an activity grammar that is modeled by a temporal Markov Logic Network. We demonstrate the efficacy of our approach by applying our model on realistic simulations in a real trauma room.

Table 1 lists a set of common trauma activities. Based on empirical evaluation, we shortlisted a set of trauma procedures that are consistently well-represented using image-based features. These *visually identifiable* activities are listed in the top panel. The rest of the activities are either better detected based on non-visual signals (e.g., ECG can be detected from machine response) or require a more descriptive representation, e.g., through stereo or zoom camera. Evaluations with improved representations is part of future work.

Given a video of trauma resuscitation, our goal is to track and transcribe the procedures performed during the process. The overall approach is illustrated in **Figure 1**. First, patient arrival in the trauma room is detected and her body pose is localized in the image. Next, we extract features corresponding to scene dynamics, hand motions and object presence from the image sequence (Section 3). Each of these features are mapped to a one dimensional probability trajectory along the time axis. Based on trauma-domain specific knowledge, activities are expressed as com-

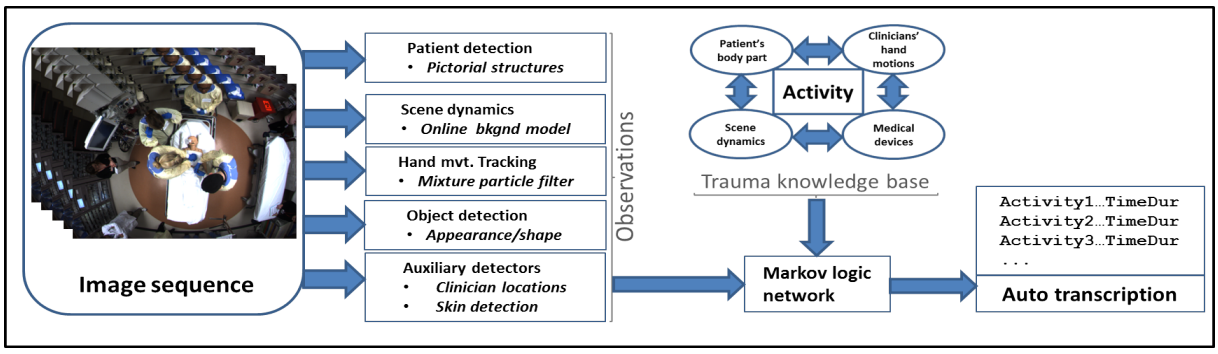

Figure 1. Overview of the proposed system.

plex spatio-temporal feature interactions using weighted logic rules (Section 4). Finally, we model the logic rules as a Markov Logic Network to perform efficient inference in activity space(Section 5). The output of our algorithm is a probability distribution of each activity over time.

3. Feature detection

In this section, we provide details about the various feature extraction algorithms used in our framework. First, we describe a method to compute patient's body pose in Section 3.1. Scene-level dynamics are calculated in terms of Online background model (Section 3.2). Clinicians' actions are tracked using hand motion trajectories in Section 3.3 and finally algorithms to detect medical devices is detailed in Section 3.4.

3.1. Patient Detection

Since most trauma procedures involve examination of a specific body part of the patient, feature search related to such procedures can be spatially constrained around the relevant location. In particular, we detect the head, chest and the upper/lower limb locations of the patient by adapting the human pose estimation algorithm as proposed in Ferrari et al. [6]. In this method, a human body is modeled as a part based Pictorial Structure model, in which pairwise spatial priors between body parts enforce kinematic constraints on the human body. Local image evidence about the body parts is prefiltered using a simple human body detector. This method for pose estimation does not assume face or skin detection, although these may be used as additional cues. For upper body detection, we replace Histogram of Orientation Gradients(HOG) features with motion-based cues during patient transfer onto bed. Based on the observation that the patient is usually in a passivie state during pose estimation, we include an additional symmetry constraint on limb positions. This prevents abnormal pose discovery and makes the algorithm robust to partial limb occlusions (**Figure 2**).

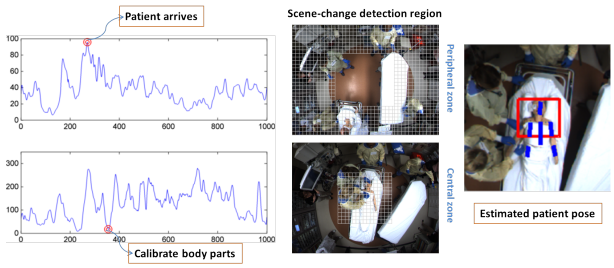

Figure 2. Left and center: ROI detection in peripheral zone signals patient arrival and momentary inactivity in central zone signals that patient is at rest and pose can be reliably estimated. Right: Symmetry constraints on pose estimation circumvents left hand occlusion.

3.2. Scene dynamics

The onset of an event usually manifests as rapid pixel intensity changes at the event's location in the image sequence. This cue can be computed through consecutive image differencing to identify movement [3]. A more robust approach is proposed in Grimson et al [23] where instead of merely calculating intensity changes, a memory-based online background model keeps track of the intensity changes to separate pixels into foreground and background classes. We define foreground as *Region of Interest*(ROI) and employ the latter approach to compute an ROI map of the scene. In particular, the ROI is an inactive region that undergoes large appearance change e.g., due to addition or removal of objects, pose changes etc. (**Figure 3**). In this approach, pixel intensities are represented as Mixture of Gaussians and each pixel in the new image is either accommodated into one of the background clusters or labeled as foreground with its own Gaussian distribution. Based on the persistence and variance of this Gaussian, a pixel either remains in foreground or is added to background. The ROI map is effective in detecting agent-independent, non-local procedures such as *rolling patient*, *cover/uncover* and *equipment approach* as well as to detect *patient arrival* (**Figures 2 and 3**).

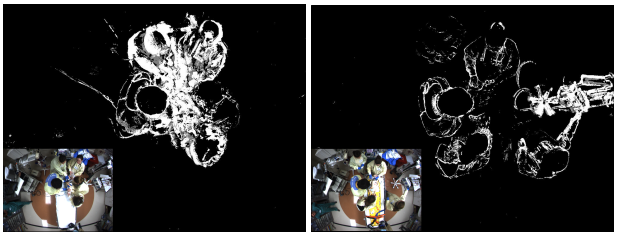

Figure 3. ROI map based on online background model. Left, patient area is ROI when rolling patient sideways and right, equipment approach. Inset: Corresponding image from video.

3.3. Hand motion tracking

To detect hand locations, we model the color of gloves that are mandatorily worn by all the clinicians. We use a simple two dimensional Gaussian model over the chrominance components U and V in the YUV color space to obtain a probability map of hand positions. The glove color model can be easily replaced to identify other colors or textures, e.g., skin.

To compute hand motion trajectories, we perform *tracking-by-detection*. We ignore identity association and jointly track hand positions as independent targets. We adopt Mixture Particle Filter(MPF) [26] for our multi-target tracking problem, where each target is modeled with an individual particle filter that forms part of a mixture. The state transition of hand positions is based on two kinds of dynamics - random Gaussian noise and constant velocity autoregressive model. As in [15], when observation likelihoods are present, a new detection is either included into existing targets or creates a new target. To avoid incorrect particle-cluster pairings, resampling is performed by weighting transition priors with observation likelihoods to increase particle accuracy (**Figure 4**).

In each output trajectory the interval relevant to activity detection lies in the vicinity of the patient. Hence, we divide the trajectories into short distance segments denoted by (1) *approach*, (2) *recede* and (3) *pause* relative to a body part of the patient. By considering short segments, we are also able to circumvent the problem of identity exchange that arises due to multiple-hand tracking.

3.4. Object detection

We consider two types of object detection methods, (a) color/texture based detection for oxygen mask and laryngoscope. and (b) shape based detection, specifically the tubular structure, which is a common shape of many devices such as stethoscope tube, ECG leads, IV access tubes etc. The methods are described below.

Texture based detection: First, an image is segmented based on color and texture to isolate small devices in separate regions. We use the algorithm in [5] since it can detect small, non-convex regions. Next, color SIFT features [12] are extracted at each region and matched to trained features. Since devices are hand-held, segmented regions close to hand positions are assigned high prior probability of object presence. The method is illustrated in **Figure 5**.

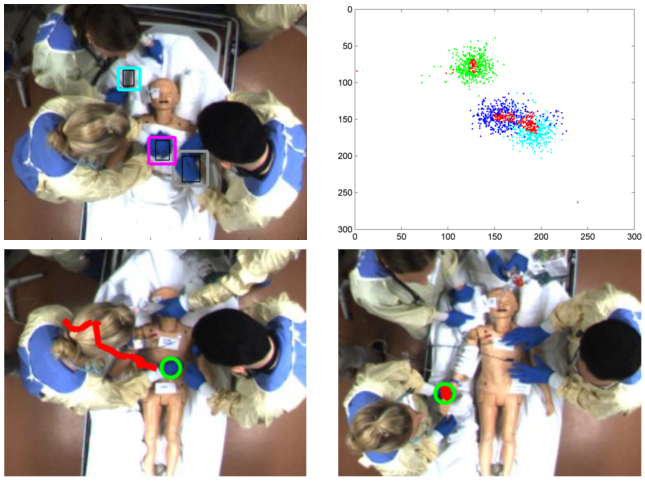

Figure 4. Top left: Tracked hand locations. Top right: The corresponding particles (green, blue, cyan) and the resampled particles based on observations(red) which shows more compact clusters. Bottom left: Tracked segment showing approach towards chest to attach ECG leads. Bottom right: Pause segment at left wrist to check pulse.

Tube detection: We use the Frangi vesselness filter [7] on an intensity image. The vesselness filter assumes pixel intensities at a tubular stucture to be distributed as a Gaussian and computes the strength of tubularness as a ratio of the principle axes after eigen decomposition. We further add a heuristic to separate the ridge (true tubes) and step edges (false detections). At each candidate edge the average intensity is computed over the normalized profile of its cross-section. If the edge is a step response, the average intensity is close to zero, otherwise it is a ridge.

4. Activity modeling

The goal of activity modeling is to combine low-level feature trajectories to accurately infer activities. Probabilistic graphical models such as Dynamic Bayesian Network and its variants [8] can effectively handle uncertainties in activity inference. However, due to a large state space, complex inter-feature relations and flexible within and between activity transitions, structure or parameter learning is an ill-posed problem for trauma procedures. On the other hand, trauma activities are well-defined in terms of domain-specific knowledge about medical devices, reference to patient's body parts and clinicians' actions. Hence, for efficient activity inference, a model should be able to compensate for uncertainty in feature observations within the framework of a knowledge-based activity description. To this end, we represent our problem as a Markov Logic Net-

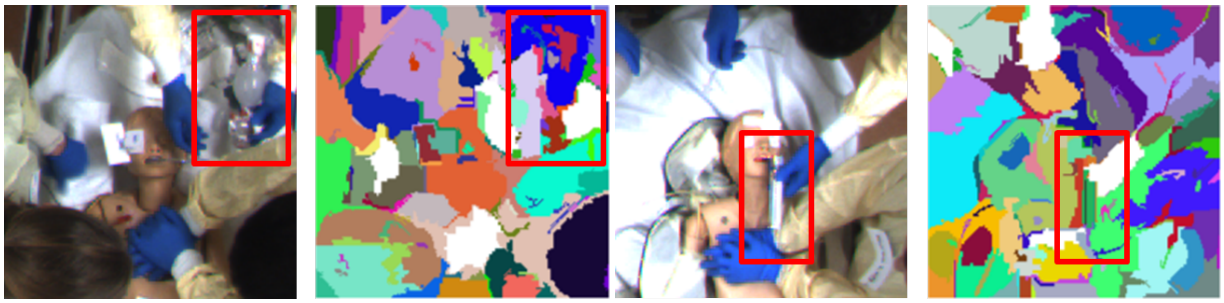

Figure 5. Object detection examples. Oxygen mask and Laryngoscope are detected by segmenting the image region around the patient's head and scanning segments features matches within the grasping radius around the hands.

work [18], a graphical model that fulfils the above requirements by allowing probabilistic inference on knowledge based rules.

An MLN is represented by a set of weighted, first-order logic formulas. A weight serves as a measure of confidence in its formula; if a world violates the formula, it only becomes less probable, not impossible. A formula consists of objects, predicates and weights. In our model, objects represent features, e.g., medical devices, patient body parts, time etc. whereas predicates represent inter-object relations or object attributes, e.g., the relation between device use and patient body part. The weights can be learnt, or as in our case, specified by the modeler. Each object variable corresponds to a finite set of constants,e.g., a time variable corresponds to 1...T, the time intervals in a trauma process. Based on these components, an MLN is an abstract description of a Markov Random Field(MRF). In the presence of observations an MLN is transformed into an MRF by *grounding*, i.e., replacing the objects by observation constants. The MRF hence formed, models the joint distribution over a set of weighted, binary formulas of the *grounded* MLN. All the methods that are applicable for estimation and inference of MRF are then applicable on this grounded MLN. We model activity as a dependent variable and compute its posterior probability. To perform inference, we use ALCHEMY, an MLN software which uses MCSAT, a combination of Markov Chain Monte Carlo (MCMC) and WalkSAT, a Constraint Satisfaction sampling technique.

5. Description of Logical Rules

In the aforementioned section we have explained the formal description of MLN. But the core of an MLN is the set of logic rules that encode its graphical structure. The design of rules should be driven by *modularity* (amenable to addition of new rules), *complexity* (minimal set of variables) and *consistency* (agreeability among rules). In this section, we describe the underlying assumptions of our activity grammar to design the MLN graph.

We infer activities through spatio-temporal combination of feature trajectories. An activity duration is defined using three states - {**Begin**}, {**During**} and {**End**}. Each state is independently modeled through feature observations and other states. **Figure 6** describes the states of *Listen to breathing* procedure. This activity's *Begin* state (blue curve) is based on a clinician attaching the stethoscope earpiece (Frame 1) and approaching the patient's chest (Frame 2). The *End* state (green curve) is based on a recede motion away from the patient's chest and detachment of earpiece (Frame 4). The *During* state (green curve) is supported by two clauses, an *evidential support* from object presence (stethoscope tube, Frame 3) and the *anchor support* provided by detection of *Begin* and *End* states (Rule 4). By modeling duration through independent individual states, we achieve linear time complexity. Moreoever, in comparison to Morariu et al's approach [13], where a relevant *duration* is directly based on consecutive start and end times of observations, our model is based on high-level encoding of *Begin* and *End* of states, which allows interleaved and simultaneous activities to be detected.

Activity-specific and Activity-independent Modeling: Rules can be activity-independent e.g., Rule 1 in which an approach towards a patient's body part increases the probability of all procedures specific to that part, or activity-specific, e.g., Rule 11 which exclusively encodes the *CheckPulse* procedure. The overall probability of an activity is based on the combined inference on both types of rules. The aim of activity modeling should be to maximize activity-independent grammar in order to minimize overfitting.

Static, Dynamic and Causal Activity Modeling: Activities are **static** if they can be detected based on spatial features only. The *During* phase of such activities attach high weights on feature observations, e.g., the *Intubation* procedure in Rule 3. **Dynamic** activities on the other hand, are based on temporally and spatially constrained interactions among features, e.g., *Listen to breathing* procedure as described above. A **causal** activity is marked by an absence of any salient, individual action or object. However, the effect of the activity is measurable based on scene-level changes, e.g., either through ROI(Section x) or as in *Cover/Uncover Patient* based on skin appearance around the patient area

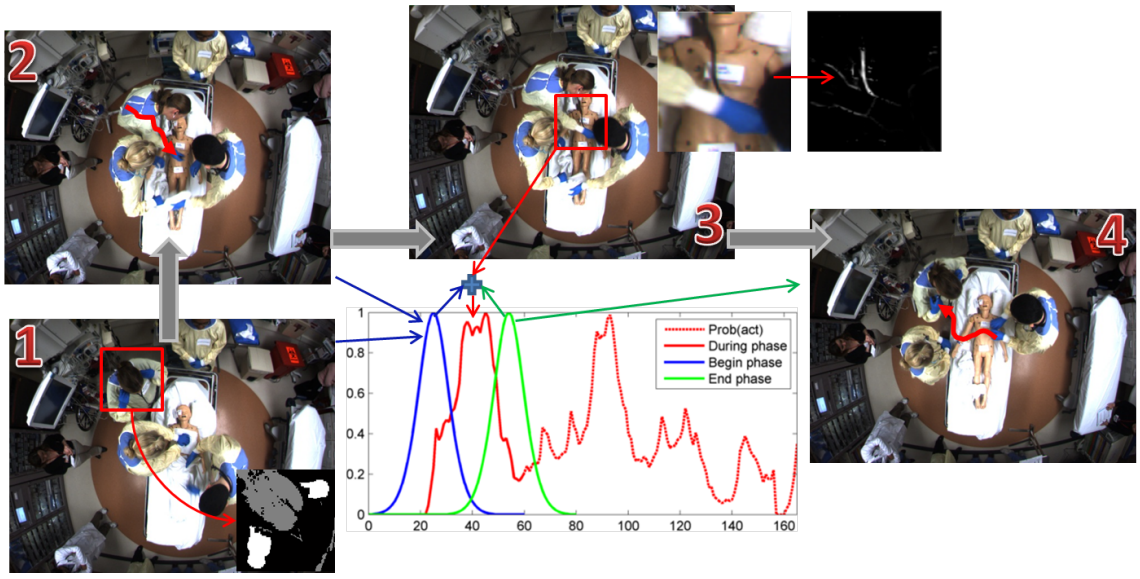

Figure 6. *Listen to breathing* procedure described in terms of spatio-temporal features. See section 5 for details.

(Rule 16).

Temporal Sequence modeling: The sequence in which the trauma procedures should be performed is defined in *Advanced Trauma Life Support*(ATLS)[2], a worldwide accepted protocol for performing resuscitation. According to ATLS, the procedures are broadly divided into primary, secondary and tertiary surveys based on their importance in identifying and solving life threatening conditions. While ATLS provides a simple and standardized approach to resuscitation, its effect on patient outcomes is debatable. Hence, most doctors often deviate from the code based on their own assessment of the patient's medical condition. To model this high-level procedure sequencing, we broadly divide the trauma activities into three phases based on their prior probability of occurrence phase. For example, in Rules 18 and 19, *Uncover* and *Rolling patient* to assigned to Phases 1 and 3 respectively. Furthermore, activities can occur simultaneously only if different body parts of the patient are examined. This exclusion principle is described in Rule 9.

6. An Illustrative Example

Our research is a work-in-progress in collaboration with a teaching hospital in the US Northeast region. The trauma resuscitation simulations are conducted in a Level-1 pediatric trauma center on the hospital premises by professional clinicians on a patient simulator manikin. The process is captured using a ceiling fitted Bumblebee2 camera.

To illustrate the overall approach of activity inference, we show results of our algorithm on a short test video of approx. 4 mins. duration. All the visually identifiable procedures, except *Roll Patient* are performed in the simulation. The video with 3270 image frames is segmented into twenty-frame intervals. Thus, the process duration is 167 intervals. Patient arrival is detected in the 19th interval. After running pose detection, features are tracked during the entire duration which are combined post-process in the MLN. The logic formulas and corresponding weights are defined in **Table 2**. MLN software for learning and inference is available online at *(http://alchemy.cs.washington.edu/)*. The trauma activities performed in this video and the grammar used to infer them are listed below. The inferred activity trajectories are mapped to the corresponding groundtruth in **Figure 7**.

- *Chest compression* is directly defined in terms of paused hand-motion tracks around patient chest area. The begin and end states of this Phase 1 procedure are described using activity-specific Rules 9 and 10 (Panel 3, **Figure 7**).
- *Listen to breathing* procedure (described in Section 5) is a dynamic, phase-independent activity described by a combination of activity-specific and independent rules. (Panel 1, **Figure 7**)
- *Intubation* is the process of inserting a tube through the patient's mouth into the trachea. The standard way of insertion is with the aid of a laryngoscope, which can be accurately detected based on appearance cues (Table 6). However, in this video it is inserted by hand. In the absence of the object cue, the activity inference is based only on hand motion tracks which generates several false detections.
- *Ventilation* is the act of pumping air into the lungs of the patient. Using a bag valve mask, the air can be forced either by pumping air through the mouth or post-intubation, through a tube inserted in the trachea. It is a static, phase

[2]http://www.facs.org/trauma/atls/index.html

#	*Formula*	*Weight*
1	$Approach(part, t)\ AND\ Located(part, act) => Activity(act, t, Begin)$	$W/2$
2	$Recede(part, t)\ AND\ Located(part, act) => Activity(act, t, End)$	$W/4$
3	$observeObject(o, t)\ AND\ Use(o, act) => Activity(act, t, During)$	W
4	$Activity(act, t_1, Begin)\ AND\ Activity(act, t_2, End)\ AND\ (t_1 < t_2)\ AND\ (t_2 > t > t_1) => Activity(act, t, During)$	$W/2$
5	$Activity(act, t_1, Begin)\ AND\ Activity(act, t_2, End)\ AND\ (t_1 > t_2)\ AND\ (t_1 > t > t_2) =>!Activity(act, t, During)$	W
6	$Activity(act, t_1, During) => Activity(act, succ(t_1), During)$	$W/4$
7	$FGchange(part, t_1)\ AND\ Located(part, act)\ AND\ Follows(t_1, t_2) => Activity(act, t_2, Begin)\ OR\ Activity(act, t_2, End)$	$W/4$
8	$Activity(act_1, t, During)\ AND\ Located(part, act_1) => (a_1 = a_2)\ OR\ !(Activity(act_2, t, During)\ AND\ Located(part, act_2))$	W
9	$PauseBegin(Chest, t) => Activity(Compression, t, Begin)$	$W/2$
10	$PauseEnd(Chest, t) => Activity(Compression, t, End)$	$W/2$
11	$PauseBegin(part, t)\ AND\ !(part = Chest) => Activity(CheckPulse, t, Begin)$	$W/2$
12	$PauseEnd(part, t)\ AND\ !(part = Chest) => Activity(CheckPulse, t, End)$	$W/2$
13	$FGchange(Head, t)\ AND\ FGchange(Chest, t) => Activity(RollPatient, t, Begin)\ OR\ Activity(RollPatient, t, End)$	w
14	$AttachEarpiece(t_1)\ AND\ Approach(Chest, t_2)\ AND\ Follows(t_1, t_2) => Activity(Stethoscope, t_2, Begin)$	w
15	$Skindetect(t_1)\ AND\ ^{!}Skindetect(t_2)\ AND\ Follows(t_1, t_2) => Activity(Cover, t_1, During)$	W
16	$exist\ t\ Activity(RollPatient, t, During) => Phase(t, 3)$	W
17	$exist\ t\ Activity(Uncover, t, During) => Phase(t, 1)$	W
18	$Follows(t, t)$	∞
19	$Follows(t, succ(t))$	∞

Table 2. Activity rules with weights

	Approach	Recede	Pause
Head	0.06	0.11	0.03
Chest	0.19	0.29	0.27
Limbs	-	-	0.37

Table 3. False alarm rate of hand motion detection averaged over 10 video sequences.

	OxyMask	L.scope	Ste. Tube
Per frame(%)	88.7	76.0	62.7
Per usage	23 / 24	10 / 13	39 / 39

Table 4. Object detection average from 10 video sequences.

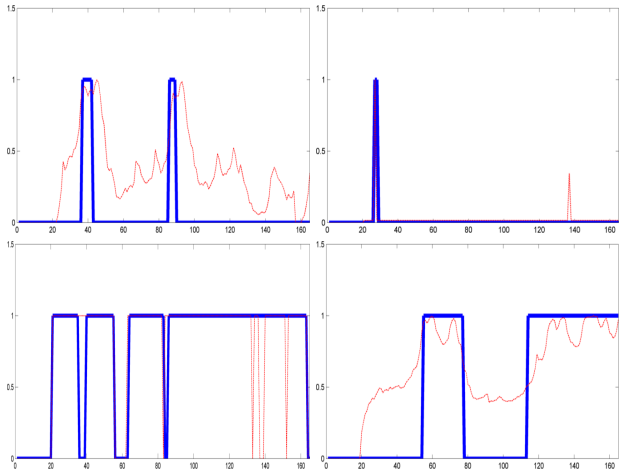

Figure 7. Activity detection in test video. Blue shows the groundtruth and red shows the intervals detected by our algorithm.

independent activity based on activity-specific and independent rules. (Panel 4, **Figure 7**

- *Check pulse* begin and end states of this phase independent procedure are described using activity-specific paused hand tracks around patient's wrist, feet and neck (Rules 11, 12). Other methods employed to measure pulse is by using a stethoscope or a pulse oximeter.
- *Cover/Uncover*: This causal activity is detected by scene-level skin detection around the patient's body area. Uncover is a Phase 1 (Panel 2, **Figure 7**) whereas cover is a Phase 3 procedure (Rule 15).
- *Rolling patient* sideways is a phase 3 procedure to identify any side injuries. This is a causal activity detected by measuring scene changes around patient's body (Rules 13, **Figure 3**). This procedure is not performed in this simulation.

Quantitative evaluation of hand tracking and object detection algorithms are evaluated on images from 10 video simulations, as shown in **Tables 3 and 4**. In the object detection experiment, per frame accuracy is based on 1000 randomly chosen images containing the object. Per usage accuracy is the average detection over the entire video segment in which the procedure is performed. For example, stethoscope tube detection is unreliable based on individual images but is accurate over an image sequence. In the hand tracking experiment, we concern ourselves with false detections since the overall accuracy is 98%. Hand tracks related to patient's head area are more accurately detected than chest because of less occlusion and sparser activities. Approach movements are better detected since receding tracks often switch identities in the presence of multiple targets close to the patient's body. Paused tracks around limbs get confused with idle hands leading to higher false alarms than chest or head area.

7. Discussion and Future Work

In this paper, we have proposed an overview of a complete recognition system for trauma procedures. We have identified the key activities and described them through *visually salient* actions and objects. We believe that our first steps towards understanding video activity recognition in a hospital domain will have far-reaching effects on automated transcription, content storage and retrieval, and training based on trauma recordings. The accuracy of the results, albeit on a small dataset, signifies the suitability of our framework. At present, we are in the process of annotating a large dataset of simulations for a complete evaluation. We conclude by listing a few challenges that we seek to address in our future work.

- *Multi-modality*. Team activities are multi-modal, hence inclusion of additional sensors such as RFID and audio would improve detection. In the vision context, we hope to improve coverage and reduce occlusion through multi-camera network.
- *Activity structure.* Instead of describing activities solely through logic, we are looking at combining logic rules and bayesian learning to define activity structure, as recently proposed in [27]. Such a model would also allow semi-supervised learning of activities based on weakly labeled sequences and reduce annotation effort.
- *Device search.* Recent work has explored *functional objects* [9, 14], in which objects are defined more by the action and scene context than their appearance. Several medical devices, such as syringes and blood pressure cuffs fall into this category. By correlating object detection with other cues, we hope to improve detection accuracy.

References

[1] S. Agarwal, A. Joshi, T. Finin, Y. Yesha, and T. Ganous. A pervasive computing system for the operating room of the future. *Mobile Networks and Applications*, 12:215–228, March 2007.

[2] J. E. Bardram, A. Doryab, R. M. Jensen, P. M. Lange, K. L. G. Nielsen, and S. T. Petersen. Phase recognition during surgical procedures using embedded and body-worn sensors. In *IEEE International Conference on Pervasive Computing and Communications*, pages 45–53, 2011.

[3] R. Biswas, S. Thrun, and K. Fujimura. Recognizing activities with multiple cues. In *Proc. HUMO Workshop (ICCV)*, 2007.

[4] H. B. Christensen. Using logic programming to detect activities in pervasive healthcare. In *International Conference on Logic Programming*, pages 421–436, 2002.

[5] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision(IJCV)*, 59(2), September 2004.

[6] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[7] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever. Multiscale vessel enhancement filtering. In *Medical Image Computing and Computer-Assisted Intervention(MICCAI)*, 1998.

[8] S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. *IEEE international conference on computer vision (ICCV)*, 2003.

[9] A. Gupta, A. Kembhavi, and L. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)*, 31(10):1775–1789, October 2009.

[10] J. Hoey, P. Poupart, A. v. Bertoldi, T. Craig, C. Boutilier, and A. Mihailidis. Automated handwashing assistance for persons with dementia using video and a partially observable markov decision process. *Comput. Vis. Image Underst.*, 114:503–519, May 2010.

[11] H. A. Kautz and J. F. Allen. Generalized plan recognition. In *National Conference on Artificial Intelligence*, pages 32–37, 1986.

[12] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision(IJCV)*, 60, 2004.

[13] V. Morariu and L. Davis. Multi-agent event recognition in structured scenarios. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2011.

[14] S. Oh, A. Hoogs, M. Turek, and R. Collins. Content-based retrieval of functional objects in video using scene context. In *European Conference on Computer Vision(ECCV)*, 2010.

[15] K. Okuma, A. Taleghani, N. de Freitas, J. Little, and D. Lowe. A boosted particle filter: Multitarget detection and tracking. In *European Conference on Computer Vision(ECCV)*, 2004.

[16] S. Park and H. Kautz. Hierarchical recognition of activities of daily living using multi-scale, multi-perspective vision and rfid. In *4th International Conference on Intelligent Environments (IET)*, pages 1–4, July 2008.

[17] Q. Qiu and R. Chellappa. A unified approach for modeling and recognition of individual actions and group activities. *Technical Report*.

[18] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62:107–136, 2006.

[19] P. C. Roy, S. Giroux, B. Bouchard, A. Bouzouane, C. Phua, A. Tolstikov, and J. Biswas. A possibilistic approach for activity recognition in smart homes for cognitive assistance to alzheimer's patients. *Activity Recognition in Pervasive Intelligent Environments*, 4:1–22, 2010.

[20] M. Ryoo and J. Aggarwal. Stochastic representation and recognition of high-level group activities. *International Journal of Computer Vision(IJCV)*, 93(2):183–200, June 2011.

[21] D. Sánchez, M. Tentori, and J. Favela. Activity recognition for the smart hospital. *IEEE Intelligent Systems*, 23:50–57, March 2008.

[22] A. Sarcevic and R. Burd. Whats the story? information needs of trauma teams. In *American Medical Informatics Association Annual Symposium (AMIA)*, 2008.

[23] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 1999.

[24] R. Stephen. Digital video surveillance: enhancing physical security with analytic capabilities. *IBM Global Services*.

[25] S. D. Tran and L. S. Davis. Event modeling and recognition using markov logic networks. In *European Conference on Computer Vision(ECCV)*, pages 610–623, 2008.

[26] J. Vermaak, A. Doucet, and P. Perez. Maintaining multi-modality through mixture tracking. In *International Conference on Computer Vision(ICCV)*, 2003.

[27] Z. Zeng and Q. Ji. Knowledge based activity recognition with dynamic bayesian network. In *European Conference on Computer Vision(ECCV)*, 2010.

[28] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Modeling individual and group actions in meetings with layered hmms. *IEEE Transactions on Multimedia*, 8:509–520, 2006.

View publication stats