HHS Public Access

Author manuscript

Ubiquitous Comput Electron Mob Commun Conf (UEMCON) IEEE Annu. Author manuscript; available in PMC 2018 October 21.

Published in final edited form as:

Ubiquitous Comput Electron Mob Commun Conf (UEMCON) IEEE Annu. 2016 October ; 2016: .

Online Process Phase Detection Using Multimodal Deep Learning

Xinyu Li¹, Yanyi Zhang¹, Mengzhu Li¹, Shuhong Chen¹, Farneth R. Austin², Ivan Marsic¹, and Randall S. Burd²

¹Rutgers University, Piscataway, NJ, USA

²Children's National Medical Center, Washington, DC, USA

Abstract

We present a multimodal deep-learning structure that automatically predicts phases of the trauma resuscitation process in real-time. The system first pre-processes the audio and video streams captured by a Kinect's built-in microphone array and depth sensor. A multimodal deep learning structure then extracts video and audio features, which are later combined through a "slow fusion" model. The final decision is then made from the combined features through a modified softmax classification layer. The model was trained on 20 trauma resuscitation cases (>13 hours), and was tested on 5 other cases. Our results showed over 80% online detection accuracy with 0.7 F-Score, outperforming previous systems.

Keywords

Activity recognition; process phase recognition; deep learning; multimodal sensing; trauma resuscitation; Kinect

I. Introduction

We introduce a novel system capable of detecting phases of the trauma resuscitation process in real-time. We identified six consecutive phases of resuscitation: pre-arrival, patient arrival, primary survey, secondary survey, post-secondary survey, and patient departure from the trauma bay. Most work activities tend to occur during their associated process phase, so the phase provides context for activity recognition models. Also, because the phase indicates the overall progress of a work process, having real-time phase detection helps organizations optimize team placement and manage resources.

Real-time detection of process phases in medical settings is challenging for two reasons. First, process phase is a more abstract concept than activity—usually a phase consists of one or more activities that might be performed in various orders. Second, collecting spatial-temporal data in a medical setting is difficult: hospitals are crowded, dynamic, and privacy-sensitive, putting constraints on the sensing platform and the types of data collection allowed.

Related research on surgical phase detection used several strategies: manually coding input activity logs for phase detection [1], recording "machine signals" generated by medical

equipment [2], and placing wearable sensors on medical team members [3]. There are drawbacks for each of these methods: manual activity log coding requires employing an additional person and is too slow for real-time operation; the “machine signal” approach only works with certain medical equipment; and wearable sensors require regular maintenance and may interfere with work.

Our proposed system consists of a Microsoft Kinect V2 sensor and a mini PC. To protect the privacy of doctors and patients, we avoided capturing recognizable facial details by using the Kinect’s microphone array and depth sensor. In addition, the Kinect is comparatively cost effective (\$150) and can be easily deployed. There are already many Kinect-based activity recognition systems in the medical field. Most past research uses the Kinect’s body joint/skeleton recognition for patient fall detection or for monitoring daily living [4][5]. A key limitation of Kinect is that it can only track the joints of up to 6 people at once, which restricts its application in trauma or surgical rooms where 10 or more people may be working simultaneously.

To overcome the skeleton number limitation, we perform feature extraction directly on the Mel-frequency spectral coefficients and depth frames from the Kinect’s data stream. We extract the video and audio features independently, in the “video branch” and “audio branch.” Both branches use convolutional neural networks (ConvNets) [6] with the final softmax classification layer removed [7].

A key aspect of the video branch is our partitioning of each depth image into subzones. Each partitioned area is independently processed by a feature-extracting ConvNet, and then the features from all areas are merged in fusion layers.

Another important aspect to our approach is our implementation of “slow fusion,” a concept proposed in video classification research [8]. Instead of simply taking a vote between the audio and video classifiers, we integrate the features extracted from the different branches through fully-connected “fusion” layers [9].

Finally, we introduce the “constrained softmax layer” that makes the final classification decision. This constrained softmax layer still classifies based on the softmax function, but its decisions are constrained by additional rules obtained from domain-expert knowledge.

We collected depth and audio data during 25 trauma resuscitations at a Level 1 trauma center. Our experiments showed an average of 80% accuracy for 6 resuscitation phases with 0.72 micro-precision and 0.76 micro-recall. The contributions of this paper are:

1. A deep-learning structure that classifies resuscitation phase by processing and merging different data from different sensory modalities.
2. An image segmentation method that optimizes the performance and efficiency of deep-learning systems dealing with complex images containing multiple targets.
3. A constrained-softmax strategy for sequential phase prediction that significantly improves performance.

The rest of this paper is organized as follows. Section 2 reviews related research. Section 3 describes our data collection method. Section 4 presents our multimodal deep-learning structure. Section 5 describes our model training and system implementation techniques and Section 6 shows the system's experimental results. Section 7 discusses the current system's limitations and potential improvements. Finally, Section 8 discusses possible future work.

II. Related Work

The best performance at medical process phase detection was achieved by previous research on surgical process phases that used a decision tree fed with manually generated activity logs [1]. Its performance was satisfactory with around a 0.74 F-Score on recognizing five surgical process phases, but the approach inconveniently required human input that could not be generated in real-time. Another research team predicted the surgical phases of laparoscopic cholecystectomy by using a Hidden Markov Model (HMM) fed with machine signals from medical equipment [2]. Dynamic time warping (DTW) and HMMs were also used for recognizing surgical process phases for endoscopic operations [10]. This research demonstrated the feasibility of modeling process phases with signal recordings, but the required specificity of medical equipment severely limited the system's implementation and deployment. Moreover, HMMs and DTW may not be robust enough for application to similar phases in different medical processes. More recently, a research team proposed phase detection in trauma resuscitation using passive RFIDs and wearable antennas in a simulated environment [3]. This approach was novel and did not raise privacy concerns, but the wearable antennas required active user participation and may have potentially interfered with work in an actual medical setting.

Video and audio information are commonly used for activity recognition in different applications beyond the medical field, and there have been many audio-visual detection frameworks using Kinect or other types of cameras. Finite state machines have been used to model events in video clips [4] and HMMs have been used for sports activity recognition [5]. However, these frameworks rely heavily on manually selected features and detectors (pedestrian detectors, skin detectors, etc), and are often limited to using one type of data. Addressing the unimodal limitation, another approach is to make final classification decision based on votes from independent classifiers fed with different types of features [11]. However, the shallow classifiers are generally too weak to predict complex activities, and such a voting strategy ignores connections between features of different data types.

Recent multimodal deep-learning research has better addressed this problem. Research on speech recognition that used video clips of mouth shapes, in addition to audio, showed the potential of multimodal deep belief networks on multimedia data understanding [12]. This multimodal structure has been extended to convolutional neural networks for RGB-D-based object recognition [9], where the data fusion was accomplished by establishing full connectivity between the layers of different ConvNets. The fusion strategy, as well as the efficiency and performance of several multimodal deep learning structures, was analyzed in large-scale video classification research [8].

To overcome the drawbacks of previous surgical phase classifiers [1][2][3][10], we designed our system to be privacy preserving, cost efficient, non-obtrusive, and easily deployed. We also adopted and extended ideas from pilot research on multimodal deep learning structures [12][9][8], such as the slow online merging of data from different sources. These factors contributed to our structure outperforming the existing systems for surgical process phase detection.

III. Data Collection

A. Hardware Configuration

Our proposed system uses a single Kinect V2 sensor. The hardware was deployed in an active trauma room of a large hospital in the US North-East region. It is fundamentally required that the hardware preserves the privacy of those observed and does not interfere with their work. To meet these requirements, we mounted the Kinect on the trauma room's side wall 2.5 meters above the ground and tilted it downwards at 20° to get a clear view of the room, as suggested by previous research [13]. Because the USB 3.0 cable required for fast Kinect V2 data transfer was not extendable, we also mounted the processing unit (mini PC) to the sidewall (Figure 1). Neither the Kinect nor the mini PC interferes with medical activities and, since we only use the Kinect's low-resolution depth frames, no facial details are captured. The system automatically starts based on the number of people in the room, detected by the Kinect sensor, and stops when no people are present in the room. This system was deemed adequately privacy-preserving and was approved by the hospital's Institutional Review Board (IRB).

B. Ground Truth Coding

Medical experts in our research team manually coded the ground truth resuscitation phases for each case by watching surveillance videos. To objectify the concept of resuscitation phase for this coding, we defined the resuscitation phases to be non-overlapping time intervals, initiated and terminated based on start or end points of key activities (Table 1).

Recognizing the key activities is not simple, even for medical experts, due to view occlusion caused by the team members surrounding the patient bed. When coding ground truth, our medical experts relied heavily on their domain knowledge, and often switched between several surveillance cameras for a better view of the room.

IV. Multimodal Deep Learning Structure

A. Overview

The key part of our system is a multimodal deep learning structure that makes predictions based on the features of different types of data from different sensory modalities. This structure takes as input visual-audio data from a 30-second sample window (the shortest resuscitation phase is ≈ 60 seconds, and so based on the Nyquist sampling theorem the sampling time should be 30 seconds). The input data is fed through the layers for feature extraction, feature fusion, and finally classification of the current resuscitation phase.

We implemented independent ConvNets for respectively extracting audio and depth features. ConvNets have shown proficiency in image feature extraction [14] and more recently in speech recognition [15]. We refer to the audio-processing ConvNet as the “audio branch” and the depth-video-processing ConvNet as the “video branch” (Figure 2, black and white boxes in feature extraction layers).

Unlike single-image classification, process and activity recognition detect events that occur over a time interval with multiple data frames. Therefore, the system should perform recognition on a multi-frame data sample over a time window [4][5]. There are three common ways of packaging the time window frames for deep learning structures: early fusion, late fusion, and slow fusion. The slow fusion structure, which slowly fuses information through many convolutional layers, has been successful at classifying activities from video [8]. However, unlike in previous research [8], we are working with both audio and video, and so cannot directly stack the layers of differing datatypes for shared convolution. Instead, we independently extract audio and video features, and then use the resulting ConvNet outputs to train fusion layers, as in other research [9]. The fusion layer itself is a fully-connected layer (Figure 2, gray boxes in fusion layers).

The softmax layer is commonly used for making the final classification decision in a deep learning structure [9][8][6][15]. However, we modified the softmax layer’s highest-scoring selection heuristic to better suit our classification task. Unlike subjects in sports videos or images, the placement and posture of medical subjects may be similar across several resuscitation phases (i.e. leaning over or standing around the patient bed is common in primary, secondary and post-secondary phases). Therefore, in some cases a prediction with the second-highest score may be the correct one. To account for this, we added the constraint layer after softmax to restrict the classification decisions using medical knowledge of resuscitation phase progression. The decision-making layer would then select the highest-scoring label out of those that meet the constraints. We call the layer that executes this procedure the “constrained softmax layer” (Figure 2, the patterned boxes at the top).

B. Video Branch

The convolutional neural network we used for extracting depth image features processes non-overlapping 3-second time intervals with fully-connected layers, roughly based on the ImageNet-winning structure [6].

Our image data’s complexity, however, called for modifications. To cope with handling multiple subjects, various equipment interactions, and rich backgrounds, we first split each image into three fixed work zones (head, body, and foot. See Figure 1, right). This segmentation adequately represents the image data in a more manageable format, allowing for independent contextual subzone feature extraction. Similar strategies, called “region proposals,” have been used to better locate and identify objects in images with rich backgrounds or multiple objects [16][17]. However, whereas these other methods propose new regions every frame, our technique segments the images based on patient bed location with dynamic room layout mapping strategies [13]. This more static image division is

possible because the sensor and most furniture (including patient bed) are fixed throughout each resuscitation. We perform this segmentation to:

1. **Improve system performance:** The segmentation takes advantage of the nature of resuscitation phases. Different resuscitation activities heavily tend to occur at their respective zones during their respective phases. This is why medical experts, when manually coding the ground truth, rely strongly on the zone in which the activity occurs. Thus, allowing the system to focus on separate areas facilitates more relevant preliminary feature extraction. The segmentation's advantages are shown by a 10% accuracy gain when switching from entire-image to segmented-image ConvNets (with the same structure).
2. **Make training more efficient:** The depth images we use are more “abstract” than a typical classification image. Our images contain several target objects that are difficult to distinguish from irrelevant backgrounds. Generally, learning more abstract concepts requires larger and deeper networks, which quickly become overwhelmingly expensive computationally. Instead of simply adding more layers onto the ConvNet, we should manipulate the data to make training more efficient [8]. Segmenting the input image allows the model's separate ConvNets to train in parallel on different machines, significantly decreasing the training time. With the same convolutional layers and kernel configuration, the proposed network with split-image input trained (on three computers) around 3.5 times faster than the network with whole-image input (training completed when the cross-validation error remained unchanged for 1 epoch).
3. **Satisfy hardware constraints:** Cutting edge computers with multiple GPUs may be able to handle an ultra-deep neural network, but it is not convenient nor practical in our scenario. To keep the trauma room organized and avoid potential interference with their work, we chose to use a mini PC (Intel NUC, i7, 4GB memory, no GPU) that can be mounted on the wall. Such a platform simply cannot handle simultaneous real-time data collection and a very deep online prediction model.
4. Each image in the 3-second window is fed to its own ConvNet, where it is first split into the three subzones (Figure 3). Each sub-image is convolved and pooled thrice. The resulting sub-image outputs are consolidated into a fusion layer representing their original single image. Then, three such layers (one for each second in the window) are finally fused into the video branch output layer.

C. Audio Branch

We did not perform explicit speech recognition or keyword detection in the audio branch—the trauma room is quite noisy and the microphone is too far to capture clear speech. Instead, after reviewing the data with medical experts, we found different sounds typical to specific phases. For example, in the patient arrival phase, there is usually an increasing sound from human conversation; in the primary phase, there are medical equipment sounds;

when the patient leaves, the room is silent. Taking advantage of these observations, we classified ambient sounds instead of recognizing speech.

Speech recognition and ambient sound classification is usually performed with time-frequency domains because direct analysis of only the time domain is inefficient [18][15]. As argued in previous research [15], the use of discrete cosine transformations on Mel-frequency cepstral coefficients (MFCCs) [19] compromises the energy's locality. Therefore, we instead extracted the sound's Mel-frequency spectral coefficients (MFSC), as described in previous speech recognition work [20]. By visual inspection of the MFCC feature maps, we observed certain patterns in the energy distributions of different environmental sounds. So, we decided to use a ConvNet, fed with time-frequency feature maps, similar to previous speech recognition research [15].

In our implementation (Figure 4), we used 40 frequency bands covering the range from 0 to 10,000 Hz. We made each audio frame 40ms long, with 50% overlap between adjacent frames, following the parameters from speech recognition research [15]. To make the system aware of energy distribution changes across different time windows, we did not generate a 3s MFSC feature map and perform 2D convolution. Instead, we generated a MFSC feature map for every second, and then stacked the three maps together as a three-layer feature matrix, ready for input into our ConvNet (Figure 4). We chose time window size $T=3s$ to fit both the feedforward multimodal structure and sensor data processing program into 4 GB RAM (1GB Windows OS, 2GB multimodal structure, and 1GB data handling). Although previous research suggests that using one-dimensional convolution with limited weight-sharing for each frequency band works better for speech recognition [15], our experimental results show similar performance compared to 2D ConvNets. This may be because ambient sounds are distributed over a wider range of frequency bands than human speech.

D. Constrained Softmax Layer

Softmax (multinomial regression) is commonly used as the decision function after convolution [7]. In our experiments using softmax, we noticed that the system had difficulty distinguishing similar or adjacent phases. There are two main reasons for this problem. First, two phases may in reality have a short overlap when two team members concurrently perform activities of different phases, hindering the detection of phase transition. Second, some resuscitation phases are innately similar. For example, during both P and S, the number of team members standing near the patient bed is roughly equal, and their activities appear similar. Even medical experts may have difficulty determining the phase given only several low-resolution depth frames and audio feature maps. The experts produced many predictions by watching a longer video clip of what came before and using various contextual cues.

To simulate the expert's reliance on previous observations in making the final prediction, we apply another fusion technique on the softmax layer. We average each label's softmax scores within a 30-second time window (yielding 10 labels with averages, see Figure 2). These labels are sorted by descending average softmax score, effectively ordering the candidates by label confidence.

To include some of the medical experts' knowledge, we implemented two constraints that they applied during ground-truth coding. Interpreted generally, given the previous predicted phase P_p and currently predicted phase P_c , we have:

1. 1. The resuscitation phases have a progression that goes forward but never backwards ($P_c \geq P_p$).
2. 2. The resuscitation phases only occur in a specific sequential order, and will never jump to a non-adjacent phase. ($P_c - P_p \leq 1$)

In reality the strict sequential order might not always be maintained, and a jump to a non-adjacent phase may occur, although no such cases appeared in our data. It is also possible for resuscitation phases to loop back; in particular, the primary survey phase may be repeated after later phases, depending on a changing patient condition. Incorporating such rare situations into our system design and formulating the associated constraints will be part of our future work.

The final prediction is then the most confident label out of the candidates that satisfy these constraints.

V. Implementation

A. Model Training

We trained the model with TensorFlow [21] because of its strong ConvNet support. We used the rectified linear unit (ReLU) activation unit, following previous research [8][6]. We initiated the learning rate at 0.01 and dynamically updated it with the Adam optimizer [22].

Although the recording from 20 resuscitation cases produced around 14 hours of data, overfitting was still an issue. To avoid overfitting, we added two interventions. First, we adopted the widely-used "dropout" method [23]. Second, the medical ground-coding team avoided selecting too many cases with patients of similar ages and medical conditions, reducing overfitting to specific patient attributes.

Since Tensorflow does not currently support multimodal training, we could not directly train the entire model. Based on previous implementations [9], we first trained the video and audio branches separately. Then, when the cross-validation error remained constant for 1 epoch, we used each ConvNet's first fully-connected layer for fusion layer training. The fusion layers were also trained using Tensorflow.

B. Implementation of Online Prediction

TensorFlow trains well, but has caused some compatibility issues. The Kinect SDK is Windows-only and TensorFlow is Linux-only. To address this problem, we first trained the deep learning structure on Linux, then reloaded all the weights in C#. Since prediction requires only the feed-forward network, we didn't program back propagation with C#. We also programmed in C# our proposed constrained softmax layer.

VI. Experimental Results

We deployed our system in our hospital's trauma room. We observed 25 real resuscitations, 20 of which were used for training and 5 for testing. The ground truth was generated as described in Section III.B.

We first compare the results of three system structures: audio branch-only, video branch-only, and the proposed multimodal structure (all using ordinary softmax layers). Since each resuscitation phase had various durations in different cases, we use the micro-average (average over instances) for all averaging unless stated otherwise. The results show 41.57% average accuracy for the audio branch and 49.07% average accuracy for the video branch. The proposed multimodal structure significantly out-scored both structures with 67.51% average accuracy.

We then evaluated the effectiveness of the proposed constrained softmax layer, which achieved around 13% higher accuracy than the ordinary softmax's 67.51% micro-accuracy. The confusion matrices (Tables II) demonstrate that the proposed constrained softmax layer significantly improved performance, especially for the primary phase (Figure 5).

Instead of the common approach of slicing up the entire dataset and randomly choosing 20% of segments for testing, we chose 5 out of 25 whole cases to ensure that testing data had no overlap with training data. This approach is appropriate because no trauma resuscitation has the exact same room layout, environmental sounds, and personnel present. In addition, not having independent training and testing cases would introduce a form of look-ahead bias. In fact, our experiments show that the system trained on a randomly-selected set achieved over 90% testing accuracy, which is significantly higher than when trained on whole cases, strongly indicating that the results have lost authenticity.

To further evaluate system performance, we treated the recognition of each resuscitation phase as a one-vs-rest problem. To this end, we implemented the F-measure. The results (Figure 6) show 0.72 micro-average precision and 0.76 micro-average recall. The PA and P phases had noticeably lower precision. These two phases have short durations (~1 minute), so an error of a few seconds leads to a significant precision drop. We also noticed that P and S had relatively low recall. This is due to their less distinguishable teamwork appearance and ambient sounds. A possible future solution might be to install multiple Kinects for a more complete view of the trauma room.

Finally, we compared our system to existing ones [1][2][3][10]. We provide a full comparison between system data type, obtrusiveness, and performance (Table III). Note that the term "phase" has different meanings in different research contexts. For example, some research defines a "phase" as a set of activities in a surgical operation that are potentially detectable by the usage of certain signal-emitting objects [2][10]. In other cases [1][3], "phase" refers to a certain stage of an operation, which may contain multiple activities in various orders. The latter "phase" is much more abstract, and poses a greater challenge to detection. We follow the latter definition in the context of trauma resuscitation.

Our system is the least obtrusive and can operate in virtually any trauma or surgical room. Unlike the manually-generated activity log system [1], our implementation requires no human input. Unlike the system based on machine-signals[2][10], our structure does not rely on specific equipment, other than a Kinect sensor and mini PC. Finally, unlike the wearable sensor approach [3], our system neither interferes with work nor violates privacy, and so has been approved by our hospital's Institutional Review Board for use during actual resuscitations.

Our proposed system also meets or exceeds the performance of other existing systems (Table III). Although some evaluation scores may be similar, our system is more robust and less demanding than the others [1][2][3][10]. Most previous systems based on probabilistic models or shallow classifiers were designed for specific medical procedures. For example, one study focused solely on lumbar disk herniation operations [1], and another was only for cholecystectomies [10]. These approaches are narrow, and will ultimately face scalability issues, since each operation would require an independently trained model. Also, having separate models implies that the specific model for the medical process at large must be selected and loaded in advance, necessitating human input. Unlike existing systems, our implementation is practical and generalizable enough to work for different cases with various patients and medical teams.

VII. Discussion

A. Current Limitations

Although our system has performed well, it still faces several limitations. The most basic is the hardware's memory capacity; the current installation limits time windows to 3s, but a larger window may increase performance at the cost of more expensive hardware. Analyzing system performance on better hardware platforms will be our future work.

In addition, using a single fixed Kinect sensor makes it difficult to accurately distinguish some phases. We found that with a fixed camera view, even medical experts sometimes fail to predict the resuscitation phase correctly. When the medical experts code the ground truth, they often switch between three surveillance cameras, indicating that having more views may increase performance. Incorporating additional sensor data would occur naturally in our multimodal structure, since all the data is processed with independent ConvNets.

Another issue is that the Kinect depth sensor's limited resolution may fail to sense small objects critical to a key activity. However, RFID systems show promise in medical activity recognition [1], and may serve as an applicable workaround. We expect that having RFID data in addition to audio and video will further enhance system performance.

B. Potential Extensions

Considering that the system does not compromise user privacy and works online, it can be implemented in other scenarios. For example, the system may be deployed 24/7 in an elder patient's home, trained to detect patient emergencies or unusual behaviors. The system can also collect data for teamwork recognition, as the trauma resuscitation is a typical example

of teamwork – multiple people working together in the same room to accomplish the same task.

The proposed image segmentation idea can be further extended to discover critical “zone features” that may help reduce irrelevant input noise for future ConvNets. Such a method can also be used in many applications with fixed cameras recording less dynamic backgrounds.

Finally, the proposed multimodal structure can be used for fine-level activity recognition, since activities are naturally recognized based on visual and acoustic data. We did not perform medical activity recognition with the proposed model because of the limited data available to us now. Unlike resuscitation phase, specific activities may be very short (a few seconds), and may require data with higher sampling frequencies. In addition, many important activities are performed only once per resuscitation, severely limiting the amount of training data available. It may take months to gather enough resuscitation cases, and even longer to generate their ground truths for model training.

VIII. Conclusion And Future Work

We presented an online trauma resuscitation detection system that uses only a Kinect and a multimodal deep learning structure. The system achieved satisfying performance in an actual medical setting. The proposed framework can be transferred to other activity recognition applications with more sensors. Our future work will investigate how including other sensors, such as passive RFIDs, may help in recognizing resuscitation/surgical phases and activities.

Acknowledgements

This research was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM011834.

References

- [1]. Forestier G, Riffaud L and Jannin P, "Automatic phase prediction from low-level surgical activities", *International Journal of Computer Assisted Radiology and Surgery*, vol. 10, no. 6, pp. 833–841, 2015. [PubMed: 25900340]
- [2]. Blum T, Padoy N, Feußner H and Navab N, "Modeling and Online Recognition of Surgical Phases Using Hidden Markov Models", *Springer Berlin Heidelberg*, pp. 627–635, 2008.
- [3]. Bardram J, Doryab A, Jensen R, Lange P, Nielsen K and Petersen S, "Phase recognition during surgical procedures using embedded and body-worn sensors", *Pervasive Computing and Communications (PerCom)*, 2011 IEEE International Conference on, pp. 45–53, 2011.
- [4]. Bai L, Lao S, Zhang W, Jones G and Smeaton A, "A Semantic Event Detection Approach for Soccer Video based on Perception Concepts and Finite State Machines", *Image Analysis for Multimedia Interactive Services*, 2007. *WIAMIS'07. Eighth International Workshop on*, pp. 30–30, 2007.
- [5]. Brand M and Kettner V, "Discovery and segmentation of activities in video", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 844–851, 2000.
- [6]. Krizhevsky A, Sutskever I and Hinton G, "ImageNet Classification with Deep Convolutional Neural Networks", *Papers.nips.cc*, 2012.

- [7]. Dunne RA, Campbell NA, "On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function", InProc. 8th Aust. Conf. on the Neural Networks, Melbourne, 181 1997 (Vol. 185).
- [8]. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R and Fei-Fei L, "Large-scale Video Classification with Convolutional Neural Networks", [Research.google.com](https://research.google.com), 2014.
- [9]. Eitel A, Springenberg J, Spinello L, Riedmiller M and Burgard W, "Multimodal deep learning for robust RGB-D object recognition", Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on, pp. 681–687, 2015.
- [10]. Padoy N, Blum T, Ahmadi SA, Feussner H, Berger MO and Navab N, "Statistical modeling and recognition of surgical workflow", Medical image analysis, 16(3), pp.632–641. [PubMed: 21195015]
- [11]. Brezmes T, Gorricho J and Cotrina J, "Activity Recognition from Accelerometer Data on a Mobile Phone", Springer Berlin Heidelberg, pp. 796–799, 2009.
- [12]. Ngiam J, Khosla A, Kim M, Nam J, Lee H and Ng AY, "Multimodal deep learning", In Proceedings of the 28th international conference on machine learning (ICML-11) (pp. 689–696).
- [13]. Li X, Zhang Y, Marsic I and Burd R, "Privacy Preserving Dynamic Room Layout Mapping", Springer International Publishing, pp. 61–70, 2016.
- [14]. Yosinski J, Clune J, Nguyen A, Fuchs T and Lipson H, "Understanding Neural Networks Through Deep Visualization", [Arxiv.org](https://arxiv.org), 2015.
- [15]. Abdel-Hamid O, Mohamed A, Jiang H, Deng L, Penn G and Yu D, "Convolutional Neural Networks for Speech Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, 2014 (pp. 1533–1545).
- [16]. Girshick R, Donahue J, Darrell T and Malik J, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", Proceedings of the IEEE conference on computer vision and pattern recognition 2014 (pp. 580–587).
- [17]. Wang X, Yang M, Zhu S and Lin Y, "Regionlets for Generic Object Detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 10, pp. 2071–2084, 2015. [PubMed: 26353185]
- [18]. Lane ND, Georgiev P and Qendro L, "DeepEar: robust smartphone audio sensing in unconstrained acoustic environments using deep learning", In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 283–294).
- [19]. Sahidullah M and Saha G, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition", Speech Communication, 54(4), pp.543–565, 2012.
- [20]. Mohamed A, "Deep Neural Network Acoustic Models for ASR", [Tspace.library.utoronto.ca](https://tspace.library.utoronto.ca), 2014..
- [21]. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems", [Arxiv.org](https://arxiv.org), 2016.
- [22]. Kingma D and Ba J, "Adam: A Method for Stochastic Optimization", [Arxiv.org](https://arxiv.org), 2014.
- [23]. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I and Salakhutdinov R, "Dropout: a simple way to prevent neural networks from overfitting", Journal of Machine Learning Research, pp. 1929–1958.
- [24]. Li X, Yao D, Pan X, Johannaman J, Yang J, Webman R, et al., "Activity recognition for medical teamwork based on passive RFID", 2016 IEEE International Conference on RFID (RFID), pp. 1–9, 2016.

Fig. 1.

Our hardware configuration in the trauma room. Left: The Kinect V2 sensor mounted on the trauma room's side wall. Middle: a more detailed view of the Kinect and Mini PC. Right: The depth image captured by Kinect.

Fig. 2. Our multimodal deep learning structure for process phase detection.

Fig. 3. The video branch structure covers three zones of the patient bed (head, body, foot). FC stands for fully-connected layer.

Fig. 4.
The network structure for the audio branch of our system.

Fig. 5.
The accuracy of different structures.

Fig. 6. The accuracy, precision, recall, and F-score of our system at detecting different resuscitation phases.

TABLE I

THE DEFINITION OF RESUSCITATION PHASES

Phase	Start Signal (and end of previous)
Pre-Arrival (PA)	Entrance of first trauma team member
Patient Arrival (A)	Patient head appears in video
Primary (P)	First “primary survey activity” starts
Secondary (S)	First “secondary survey activity” starts
Post-Secondary (PS)	First “post-secondary activity” starts
Patient Leave (PL)	Patient disappears from camera

TABLE II

THE CONFUSION MATRIX USING ORDINARY SOFTMAX LAYER (LEFT) AND CONSTRAINED SOFTMAX LAYER. NUMBER IS IN PERCENTAGE.

	PA	A	P	S	PS	PL		PA	A	P	S	PS	PL
PA	72	1	2	1	20	3		98	2	0	0	0	0
A	3	61	7	6	21	2		13	87	0	0	0	0
P	1	1	11	26	60	0		3	36	48	1	2	0
S	0	0	10	38	52	0		0	0	13	38	47	0
PS	8	0	0	1	89	0		0	0	4	1	93	1
PL	0	0	0	0	0	1		0	0	0	0	6	94

TABLE III
COMPARISON OF OUR SYSTEM TO EXISTING SURGICAL PHASE RECOGNITION APPROACHES.

System	Data used	Obrusive	Acc.	Prec.	Rec.	F-S
Automatic phase detection from low-level surgical activities [1]	Activity log	No	n/a	0.75	0.74	0.74
Modeling and online recognition of surgical phases using hidden Markov models [2]	Machine signal	No	83%	n/a	n/a	n/a
Phase recognition for surgical procedures using embedded and body-worn sensors [3]	Wearable sensor data	Yes	77%	n/a	n/a	n/a
Statistical modeling and recognition of surgical workflow [10]	Machine signal	No	84%	0.85	0.84	n/a
Our Method	Depth + Audio	No	80%	0.72	0.76	0.70