University of Burgundy

Master Thesis

Surgical Activity Recognition in the Operating Room using RGBD Cameras

Author:
Emre Ozan Alkan

Supervisors:
Dr. Nicolas Padoy
Andru Putra Twinanda

A thesis submitted in fulfilment of the requirements for the degree of Master of Science

in the

University of Burgundy

and carried at

Research Group CAMMA, ICube, University of Strasbourg

June 2015

Declaration of Authorship

I, Emre Ozan Alkan, declare that this thesis titled, 'Surgical Activity Recognition in the Operating Room using RGBD Cameras' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a master degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"When David Marr at MIT moved into computer vision, he generated a lot of excitement, but he hit up against the problem of knowledge representation; he had no good representations for knowledge in his vision systems."

Marvin Minsky

Abstract

Surgical Activity Recognition in the Operating Room using RGBD Cameras

by Emre Ozan Alkan

Operating room management and surgical procedure evaluation are time consuming tasks for operating room managers and clinicians. The demanding workload can be reduced thanks to context-aware system that use signals and data from operating rooms to automate processes such as surgical skill analysis, transcription of medical produces and evaluation of quality of procedures. Activity recognition is one of the key factors to enable the context-aware systems in the operating rooms. In this work, we address the problem of activity recognition in an operating room using a multi-view RGBD camera system. We adopt the activity recognition pipeline based on bag-of-words approach by [1] and extend it with a 4D spatio-temporal voting scheme. In the recognition pipeline, a data-driven non-rigid layout is learnt to divide the 4D spatio-temporal space of the features into 4D patches in order to recover the information loss caused by the bag-of-words approach. Since each patch from the learnt layout carries meaningful and semantic information of the 4D spatio-temporal space for an activity, the proposed voting scheme collects votes from the patches to determine the activity. We tested the proposed method on a multi-view RGBD dataset [1] using two classifiers, i.e., Support Vector Machines (SVM) and Random Forest (RF). We also compare the results of the proposed voting scheme approach to the non-voting approach [1]. The experiments show that the proposed voting scheme shows promising results with 83.1% accuracy.

Acknowledgements

I would first like to thank my supervisors Dr. Nicolas Padoy and Andru Putra Twinanda for all the support, valuable guidance and encouragement during this thesis. I would also like to thank my family for all their support...

Contents

List of Figures

List of Tables

Dedicated to my professors...

Chapter 1

Introduction

Activity recognition using cameras is a very active research topic since the past several decades in the domain of computer vision. It has been widely applied in many fields, such as video surveillance, human computer interaction, robotics and medicine. However, activity recognition is still a very challenging problem due to the large amount of data, amount of computational power required, and high intra-class variations of activities. Over the past decades, researchers have mainly focused on using image sequences from single-view RGB cameras for activity recognition which have many inherent limitations. Single-view camera approaches have limitations due to narrow field of view which makes it more vulnerable to occlusions. Moreover, RGB cameras are very sensitive to illumination changes.

Thanks to recent technological advances and the emergence of cost-effective depth sensors, depth cameras have become more popular and attracted many researchers to use them in activity recognition studies. Depth sensors have many advantages over RGB cameras. For example, they can provide structural information of the scenes and they can even work in total darkness since infrared structured light is used to reconstruct the scenes. All these advantages make it interesting to incorporate the RGBD cameras into a more challenging environments, e.g., operating rooms (OR).

Recent progresses in medicine transformed operating rooms into hybrid rooms equipped with new imaging devices, robotic arms, sensors and electronic devices. These changes require more management and tracking of information. Therefore, activity recognition in operating rooms has become an important research field in the domain of medicine. It

enables applications, such as OR time management for hospitals [2], surgical work-flow modeling and monitoring [3, 4], surgical skill analysis [5] and automatic transcription of medical procedures [6]. In particular, activity recognition in operating rooms equipped with x-ray imaging devices can be used to estimate the radiation exposure of the patients and the clinicians and correlate it to their activities [7].

In order to recognize activities, many data sources from operating rooms can be used, e.g, the vital signs of the patients [2] and tool usage signals [4]. However, they are only providing local information that is not enough to capture global activities and events in the operating room. Thus, in this thesis, we focus on performing activity recognition on videos. Specifically we are identifying what activities is happening in a video clip.

Twinanda et al. [1] have addressed the same problem. In their work, they proposed an action recognition pipeline with a novel feature encoding scheme that achieved 85.53% accuracy. The proposed novel feature encoding extends the bag-of-words (BoW) approach to learn a data-driven non-rigid layout to divide the 4D spatio-temporal (3D spatial + 1D temporal) space of the feature locations. Their work showed that encoding features with the learned non-rigid layout retains more spatio-temporal information compared to the rigid counterpart, i.e, SPM [8].

In this thesis, we propose to adopt the activity recognition pipeline from [1] and extend it with a voting scheme to solve activity recognition in the operating room. Firstly, spatio-temporal interest points (STIP) and depth spatio-temporal interest points (DSTIP) are detected from intensity and depth video clips respectively. Secondly, histogram of optical flows (HOF) from intensity and depth cuboid similarity features (DCSF [9]) from depth data are extracted around the detected interest points. Then we learn two separate dictionaries to encode our extracted features. A visual dictionary is learned to encode visual features: the HOF and the DCSF [9]. On the other hand, the spatio-temporal dictionary is learned to define a 4D non-rigid layout. The non-rigid layout divides the 4D spatio-temporal space into smaller 4D patches. From each patch, a histogram of feature points will be counted. Finally, histograms coming from the same video clip will be passed to the classifiers to vote for the action. We also introduce two voting approaches: one-model and multi-model. In the one-model approach, each patch from all video clips are used to train a classifier. On the other hand, the multi-model approach trains a separate classifier for each patch. In order to provide a comprehensive

comparison, we present the results for voting sheme and non-voting scheme approaches using non-linear SVM and Random Forest classifiers.

The rest of the thesis is organized as follows. Chapter 2 introduces the related work on activity recognition. Chapter 3 provides the details of the methodology used in this work. Chapter 4 describes and discusses the experimental setup and results. Finally, Chapter 5 presents the conclusions of the thesis.

Chapter 2

Related Work

The problem of activity recognition has been widely studied in the computer vision field. Recent works show excellent performances using local features such as local motions [10, 11], human detectors [12] and skeleton tracking [13–15]. In [11], Yang et al. presented an activity recognition framework for depth sequences. Clustered hypersurface normals are used to construct polynormals to get shape information from depth sequences. The method achieved good results on public Microsoft human action dataset MSR [16–19]. However, super normal vectors (SNV) encoded method is not reliable either on a highly cluttered and occluded environment or when the actions have very low depth variance.

More recently, great improvements were obtained by using skeleton tracking based activity recognition approaches. Vemulapalli et al. [14] proposed a novel skeleton representation that is modeling the 3D geometric relationships between various body parts. They outperformed most of the skeleton-based state-of-the-art methods with their representation. In other work, Lin et al. [13] argues that skeleon-based approaches are limited by the effective working distance of the RGBD cameras and practically they are not online all the time. They proposed using the combination of intensity, depth and skeleton structures. In [20], Want et al. stated that action recognition is view-dependent and proposed a novel multiview spatio-temporal graphical representation for cross-view action recognition. The proposed method leverages from 3D human skeleton data. Despite the satisfactory performance of skeleton tracking approaches, frontal camera views of the persons with low occlusion is needed which are not always

possible in environments such as the operating rooms. Kadkhodamohammadi et al. [21] addressed the problem of skeleton tracking of clinicians in the operating rooms and showed that off-the-shelf skeleton tracking methods fail most of the time.

In one of the most related work with a multi-view systems and voting scheme, Zhu et al. [22] addressed the problem of multi-view activity recognition using voting scheme and random forest classifier. They use IXMAS human action dataset [23, 24] which consist of 5 cameras, 13 daily-live actions with various data, e.g., silhouettes, reconstructed volumes. However, the dataset does not consist of real activities, i.e., the dataset is recorded in a laboratory environment. The method is dependent on the extracted 2D silhouettes from each camera. The segments in temporal domain consist of 2D binary silhouettes which are used to train the Random Forest classifier to get prediction histograms used in voting. The voting strategy collects the vote from each segment of video clip and apply the weights on the features and on the camera views. This voting strategy is different than our proposed voting scheme where we divide 4D space into cells to collect votes. Their results showed that combination of multi-view cameras, voting scheme and Random Forest classifier produce good results. On the other hand, in multi-view camera systems, actions may not be available from each camera view all the time. Furthermore, methods like silhouette extraction, human tracking are not feasible solutions to be used in high cluttered and occluded environment, e.g., the crowded places, operating rooms etc.

Similarly, in this thesis, we address the problem of activity recognition in the operating room using a multi-view camera system. It is not a trivial problem because of various challenges. Firstly, camera positioning in operating rooms for better field of view while accommodating the articulation of the surgical equipments is very challenging. In addition, operating rooms are very dense and dynamic with a cluttered background and reflective objects, high occlusions, illumination changes , upfront camera view and very tiny or slow movements. Hence, methods in [11, 13, 14, 20, 22] are not sufficient to deal with these problems. Due to these difficulties, it is required to have lower level features to represent the video clips.

One example of the lower level features for an image sequence is spatio-temporal interest point (STIP)[25]. STIPs are sparse interest points along spatial and temporal domain, which are detected by Harris corner detection in 3D spatio-temporal domain. In [26], Dollar et al. proposed a modified sparse and informative STIP and showed that it is

robust to pose, clutters and occlusions. The STIPs [26] are described with cuboids extracted from surrounding spatio-temporally windowed data. The proposed interest points are more robust to gradual gradient changes and periodic movements. However, the approach was designed to work mainly for intensity data and did work properly on depth data because of the nouse. Xia and Aggarwal [9] proposed filtering method to detect STIP in depth data, namely depth spatio-temporal interest points (DSTIP). They also introduced a novel feature (i.e., Depth Cuboid Similarity Feature) to describe local similarity of 3D depth cuboids around DSTIP.

In a recent work, Twinanda et al. [1] showed that intensity or depth data is not sufficient by itself to recognizing some actions in the operating rooms. Some actions are well recognized in depth videos due to illumination problems in intensity data, and some actions give better results in intensity data due to low depth variance of depth data. Hence, the combination of both intensity and depth data have more discriminating power than using only intensity or depth data. The problem of activity recognition in an operating room using multi-view RGBD camera systems was addressed. A dataset of 1734 video clips of 15 surgical activity is generated using 2 RGBD cameras from real surgical operations in a hybrid operating room. Then, STIP [26] and DSTIP [9] are detected from intensity and depth respectively. They extracted Histogram of Optical Flows (HOF) and Depth Cuboid Similarity Feature (DCSF [9]) features around the STIP and DSTIP respectively. These features are encoded by a novel feature encoding scheme that extends the bag-of-words (BoW) approach to learn a non-rigid layout that divides the 4D spatio-temporal space of the feature locations. Achieving 85.53% accuracy, the work showed that encoding features with the non-rigid layout retains more spatio-temporal information compared to Spatial Pyramid Matching (SPM) [8].

Chapter 3

Methodology

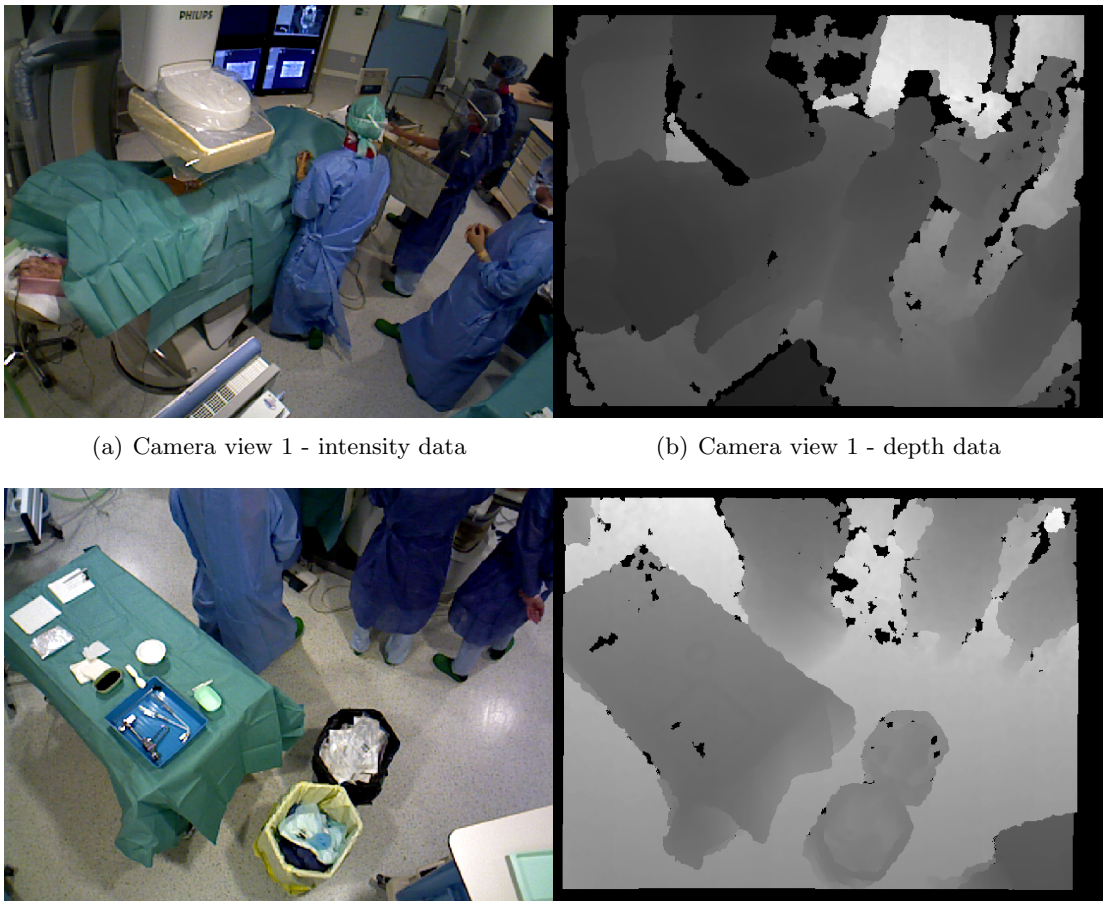
In this section, the adopted activity recognition pipeline from [1] and the proposed voting scheme are explained in detail with various classification strategies.

3.1 Adopted Classification Pipeline [1]

The adopted pipeline [1] is as follows: acquisition from a multi-view camera setup, interest point detection, feature extraction and building a data-driven non-rigid layout for feature encoding.

3.1.1 Multi-View RGBD Camera Setup

The proposed multi-view camera system contains two RGBD sensors, i.e. Asus Xtion-Pro Live, mounted on the ceiling of the operating room. The cameras record both intensity and depth data synchronously at 14 fps with 640 x 480 resolution. While one of the sensors is capturing the area around the operating room bed, the other sensor captures the area with equipment table, in which large area of the operating rooms is covered. Furthermore, thanks to minimal overlapping of the views of the sensors, interference of infrared patterns of depth sensors are very less. The intrinsic parameters for each camera are obtained by calibration with checkerboard pattern. The extrinsic parameters are obtained by using laser pointer method described in [27]. Both camera



(a) Camera view 1 - intensity data (b) Camera view 1 - depth data

(c) Camera view 2 - intensity data (d) Camera view 2 - depth data

FIGURE 3.1: Sample images from RGBD sensors with view 1 and view 2.

views are shown in 3.1. Panoramic view of the operating room that is showing the camera setups is provided in Appendix A.

3.1.2 Interest Point Detection

In the activity recognition pipeline, spatio-temporal interest points (STIP) by Dollar et al. [26] are used for intensity data. STIPs are computed by applying separable linear filters to get local maximum of the response. The response fuction R of an intensity image I is calculated as follows:

$$R_I = (I * g * h_{even})^2 + (I * g * h_{odd})^2 , \tag{3.1}$$

where $g(x, y; \sigma)$ is the 2D Gaussian smoothing kernel with scale σ, applied on the spatial (x, y) dimensions. Terms h_{even} and h_{odd} are the quadrature pair of 1D Gabor filters applied in temporal domain defined as: $h_{even}(t|\tau, \omega,) = \cos(2\pi\omega t) e^{\frac{-t^2}{2\tau^2}}$ and $h_{odd}(t|\tau, \omega) = \sin(2\pi\omega t) e^{\frac{-t^2}{2\tau^2}}$.

However, the STIP detector does not perform well and falsely detected on depth data because of the nature of the depth data where it may contain noise from sensors, pattern interference and fast foreground/background change of the border pixels. In order to cope with these problems, DSTIP detector by Xia and Aggarwal [9] is used to detect depth spatio-temporal interest points (DSTIP) on depth data which introduced filtering method to reduce the noise and the false detection. Additional noise suppression term is proposed for noise filtering to equation 3.1:

$$R_D = (D * g * h_{even} \circ \bar{s})^2 + (D * g * h_{odd} \circ \bar{s})^2, \tag{3.2}$$

where D is the depth image, and $\bar{s}$ is the noise suppression term.

3.1.3 Feature Extraction

We compute features by extracting 3D cuboid around each detected interest points in spatio-temporal location (x, y, t). Thus, the cuboids are the spatio-temporally windowed pixel values that contain most of the data contributed to response function 3.1 and 3.2 to detect interest points. The size of the cuboids $(\square_x \times \square_y \times \square_t)$ is proportionally assigned according to the spatial scale σ and the temporal scale τ.

Intensity cuboids are extracted from intensity data and histograms of optical flow (HOF) are calculated from the intensity cuboids. HOFs are used in intensity feature extraction process due to its distinguishable properties for direction of motion. These properties help to distinguish activities which have similar movements with different directions.

Moreover, 3D depth cuboids are extracted from depth data. Then, the depth cuboid similarity features (DCSF) [9] are computed from the cuboids. DCSFs are computed based on self-similarity to encode spatio-temporal structure of the 3D cuboid. DCSF divides cuboids into smaller voxels, compute histograms of the depth pixels from each

voxel, calculating relationship of each voxel by defining similarity between voxels using Bhattacharyya distance, then combination of all similarity scores from the voxels are concatenated to generate a DCSF feature.

3.1.4 Data Driven Feature Encoding

The bag-of-words (BoW) is a model that treats image features as words to construct vocabulary. Then, histogram of occurrences of the words from the vocabulary produces a sparse histogram vector. K-means is typically used to encode the features which provide a global representation with BoW where $\mathbf{x}$ is a feature, $\forall \mathbf{x}$, a value $w \in \{1, \cdots, K\}$, where K is number of K-means center, is assigned to describe the index of the K-mean center which is the closest to feature $\mathbf{x}$. Hence K-means clusters every feature vector $\mathbf{x} \in \mathbb{R}^n$ and express them in a sparse representation $\mathbf{s} \in \mathbb{R}^K$ which is the occurrences of the words from the learnt dictionary $\mathbf{D} \in \mathbb{R}^{n \times K}$

In [1], learning two separate dictionaries are proposed: a visual dictionary $\mathbf{D_v}$ with K_v centers and a spatio-temporal dictionary $\mathbf{D_{st}}$ with K_{st} centers. The visual dictionary $\mathbf{D_v}$ is learnt for encoding the visual features, i.e., the HOF and the DCSF for representing video clips. On the other hand, the spatio-temporal dictionary $\mathbf{D_{st}}$ is learnt by clustering the locations of the interest points (x, y, z, t), where $X = (x, y, z)$ and t are the 3D coordinates and the temporal location of the interest point, respectively. Since the video clips varies in length, t is normalized by length of its video clip. Multi-view camera systems require to define one camera's coordinate system as a reference coordinate system to transform the 3D coordinates into a common coordinate system. Hence, the final 3D coordinates are computed by:

$$\begin{aligned} X &= T_R \cdot Y + T_t \\ Y &= d \cdot C^{-1} \cdot \begin{pmatrix} u & v & 1 \end{pmatrix}^\top, \end{aligned} \tag{3.3}$$

where Y is the 3D point in the camera frame, (u, v) and d are respectively the image pixel coordinates and the corresponding depth value, C is the intrinsic camera matrix, T_R and T_t are respectively the rotation and translation from the camera frame to the reference frame. Then, learnt spatio-temporal dictionary $\mathbf{D_{st}}$ is used to place a non-rigid spatio-temporal grid to divide the spatio-temporal space into patches $\{P^1, ..., P^{K_{st}}\}$. These local patches from the non-rigid layout are high dimensional 4D patches. One

example of feature encoding in 2D space is shown in Figure 3.2 where non-rigid layout is learnt by clustering points indicated with colored patches and features are extracted from the patches.

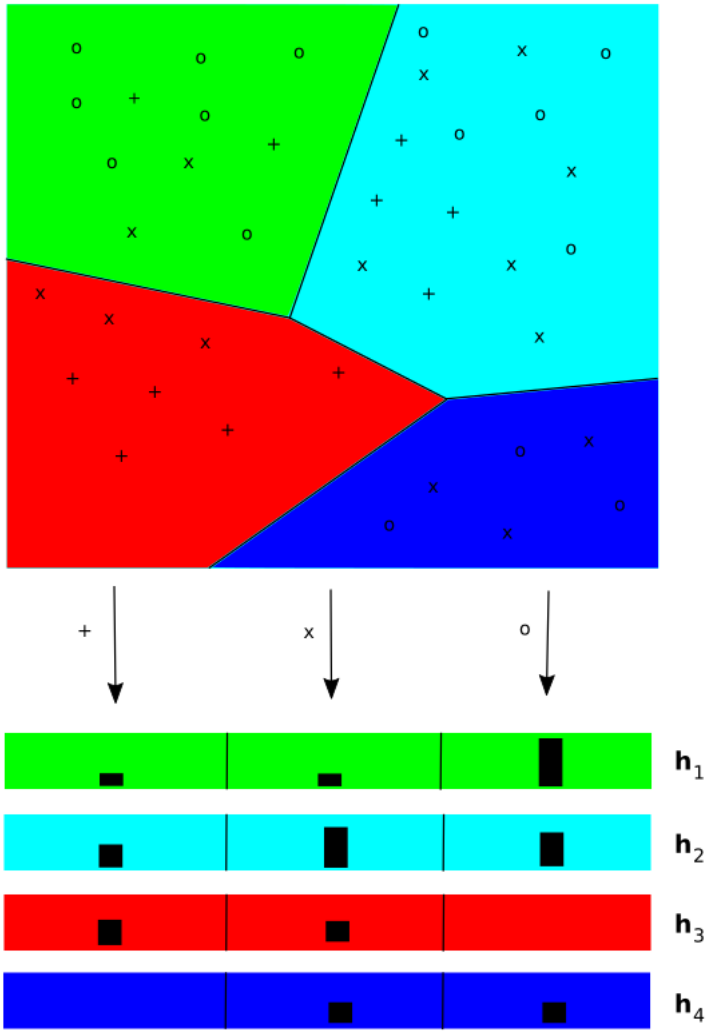


FIGURE 3.2: Example of data driven feature encoding in 2D space with $K_v = 3$ and $K_{st} = 4$

In order to get the final histogram representation of the patches for a video clip V_i, where i is video instance count, firstly, all features, $\forall \mathbf{x}$, belong to their corresponding patch, P^j, are found using the $\mathbf{D_{st}}$. Secondly, using the $\forall \mathbf{x} \in P^j$, the histogram of visual word occurrences from every patch P^j is computed using $\mathbf{D_v}$. Finally, we obtain K_{st} histograms, $\{{}_iH^1, ..., {}_iH^{K_{st}}\}$, representing the patches P^j from video clip V_i, where ${}_iH^j \in \mathbb{R}^{K_v}$.

3.2 Voting Scheme

We extended the pipeline aforementioned by proposing a voting scheme. The proposed voting scheme uses histograms H^j obtained from patches P^j of a data-driven non-rigid layout presented in Section 3.1.4. We classify every local patch P^j using H^j independently and collect their vote for calculating majority vote for global activity. Hence every patch P^j holds information about activity with semantic information that is contributing globally.

3.3 Activity Classification

In our multi-view camera system, every activity a_i is recorded from both of the views. Hence, it produces a video clip pair $\left(V_i^1, V_i^2\right)$ from the cameras. Firstly, interest point detection in Section 3.1.2 is applied on the video clip pair $\left(V_i^1, V_i^2\right)$ and features are extracted as described in Section 3.1.3. Then, two dictionaries are learnt for encoding as described in Section 3.1.4. The learnt spatio-temporal dictionary divides 4D space into smaller local 4D patches $\{P^1, ..., P^{K_{st}}\}$. Then, histogram representation H^j from each local patch P^j is calculated using the learnt visual dictionary for the video clip pair $\left(V_i^1, V_i^2\right)$. In order to get final histogram representation, $\{{}_iH^1, ..., {}_iH^{K_{st}}\}$, for video clip pair $\left(V_i^1, V_i^2\right)$, we merge information by computing ${}_iH^j = {}_i^1H^j + {}_i^2H^j$ where ${}_i^1H^j$ and ${}_i^2H^j$ are respectively histograms from patch P^j of V_i^1 and V_i^2. Then, final histograms, $\{H^1, ..., H^{K_{st}}\}$ from all video clip pairs are passed to the classifier for training and testing. In order to make performance comparison, we use the one-against all Support Vector Machine (SVM) with the nonlinear kernel, and we use Random Forest implementation to give comprehensive comparison of the voting scheme.

3.3.1 SVM

Support Vector Machine (SVM) is a well known supervised machine learning model for classification and regression. It uses hyper-planes in high dimensional space to divide training data with largest margin of the points. In this work, non-linear SVM with Chi-square $(\mathcal{X})^2$ and histogram intersection kernel is used. Since we have multi-class problem, we use one-against-all SVM to handle multi-class problem.

3.3.2 Random Forest

Random Forest is an algorithm of ensemble learning methods that used for classification and other tasks. Random Forest is collection of multiple decision trees that each produces a probabilistic response of the prediction of the classes with bagging and random selection of features. Then, these responses are combined to construct prediction which reduces over-fitting problem of the single decision tree models. Random Forest can be thought as a group of weak learners combined to get stronger learner.

3.3.3 Training Strategy

In this work, we propose to use a two-level classification strategy to carry on the voting scheme. In the first level, we learn a classification model to obtain the probability votes from the 4D patches. However, in the second level classification, we learn the weights for the probability votes under the assumption that each patch has different contribution. In addition to the proposed training strategy, we also propose to use two different approaches used in the first level of the proposed classification strategy, i.e., one-model and multi-model approaches. In the one-model approach, all histograms from each patch from each video is trained using one classifier. However, in the multi-model approach, separate classifiers are trained for each patch.

3.3.3.1 One-Model Approach

In one-model approach, we use one classifier model in each level, i.e., M_{level1} and M_{level2}. First, we train M_{level1} with the training set which consists of a collection of histogram vectors $\{H^1, ..., H^{K_{st}}\}$. Then, we pass the histogram vectors to M_{level1} to get the probability votes $\{W^1, ..., W^{K_{st}}\}$ from each video clip pair where $W^j \in \mathbb{R}^{N_c}$ and N_c is the number of activity classes. Then, we concatenate the probability vectors, $W = \left[(W^1)^T, ..., (W^{K_{st}})^T\right]^T$ where $W \in \mathbb{R}^{N_c \cdot K_{st}}$, of the same video clip. Finally, we train the second level classifier model M_{level2} with the concatenated probability vectors W. Testing is done with same order: (1) testing set is passed through the first level classifier model M_{level1} and the probability votes from the same video clips are concatenated, (2) concatenated probability votes are passed through the second

level model M_{level2} to to get the final classification probability. An illustration of the one-model approach is shown in Figure 3.3.

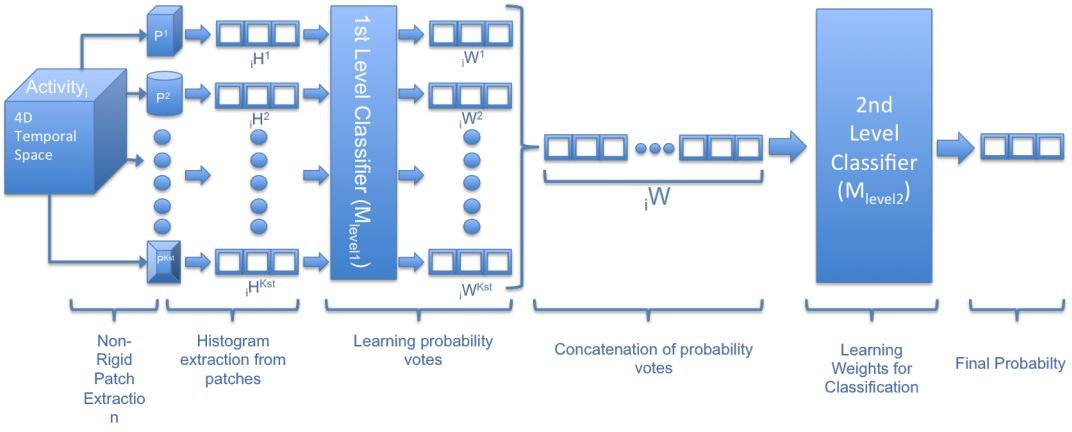


FIGURE 3.3: Illustration of one-model approach to classify a video clip

3.3.3.2 Multi-Model Approach

In contrast to the one-model approach, we propose to use multiple classifiers in the first level classification. Each patch P^j in the first level classification has its own classifier model, i.e., $\{M^1_{level1}, \ldots, M^{K_{st}}_{level1}\}$. In order to train the first level classifiers, $\{M^1_{level1}, \ldots, M^{K_{st}}_{level1}\}$, the training set is separated into K_{st} for each patch type. Each first level classifier, M^j_{level1}, is trained with its corresponding patch histograms, i.e., $\{_1H^j, ..., _{N_v} H^j\}$ where N_v is number of video clip pairs. Then, we use the same training set to test the models to get probability votes, $\{W^1, ..., W^{K_{st}}\}$ where $W^j \in \mathbb{R}^{N_c}$ and N_c is the number of activity classes. Then, we get concatenated probability vector W for each video clip as described in Section 3.3.3.1 and train the second level classifier. Testing is done as follows: (1) testing set is separated according to patch types, (2) each patch type is tested though the corresponding classifier in the first level to get probability votes, (3) histograms from the first level is concatenated and tested against the second level. An illustration of the multi-model approach is shown in Figure 3.4.

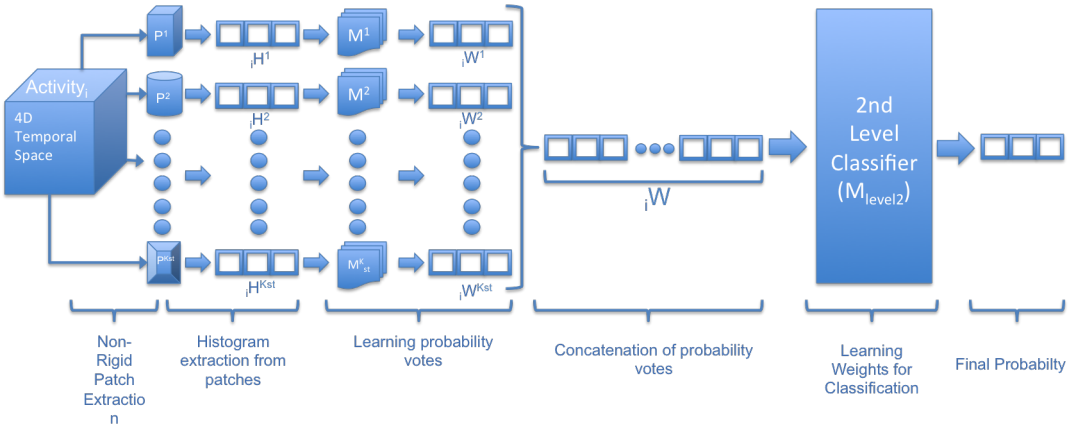


FIGURE 3.4: Illustration of multi-model approach to classify a video clip

Chapter 4

Experiments

In this chapter, experimental setup and results are described. First, the dataset is explained in detail with quantitative information. Then, feature extraction and encoding parameters and classification setups are shown and explained. Finally, results of evaluation of the proposed voting scheme are presented.

4.1 Experimental Setup

4.1.1 Dataset

In order to evaluate the performance of the proposed voting scheme with data-driven 4D spatio-temporal patches, the dataset collected by Twinanda et al. [1] is used. Composed of intensity and depth videos from 2 different views, the dataset has 11 days of recording of the operating room with real surgeries. The dataset is annotated with 15 different general and surgery-specific activities. The general activities are the common activities taking place in the operating room, e.g., bed entering, moving patient to OR bed, bed leaving, moving patient from OR bed, etc. The surgery-specific activities are picked from vertebroplasty procedures due to its frequency in the dataset. Vertebroplasty is a surgical procedure to stabilize the spinal fractures in which bone cement is injected through small punctures on the patient's skin to reduce pain. The dataset has 3 vertebroplasty specific activities, i.e., hammering, mixing cement and cement injection. It contains 1734 annotated video clips. Number of instances for each class are

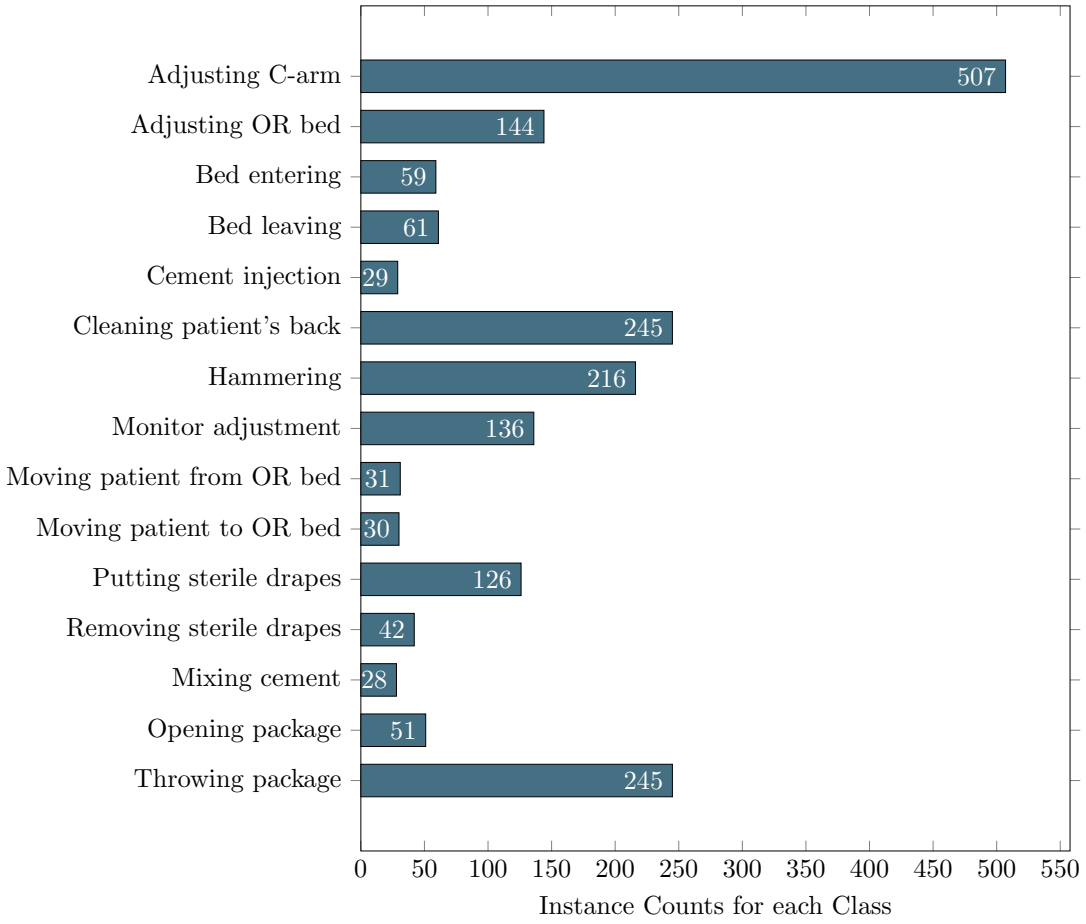


FIGURE 4.1: Dataset Information

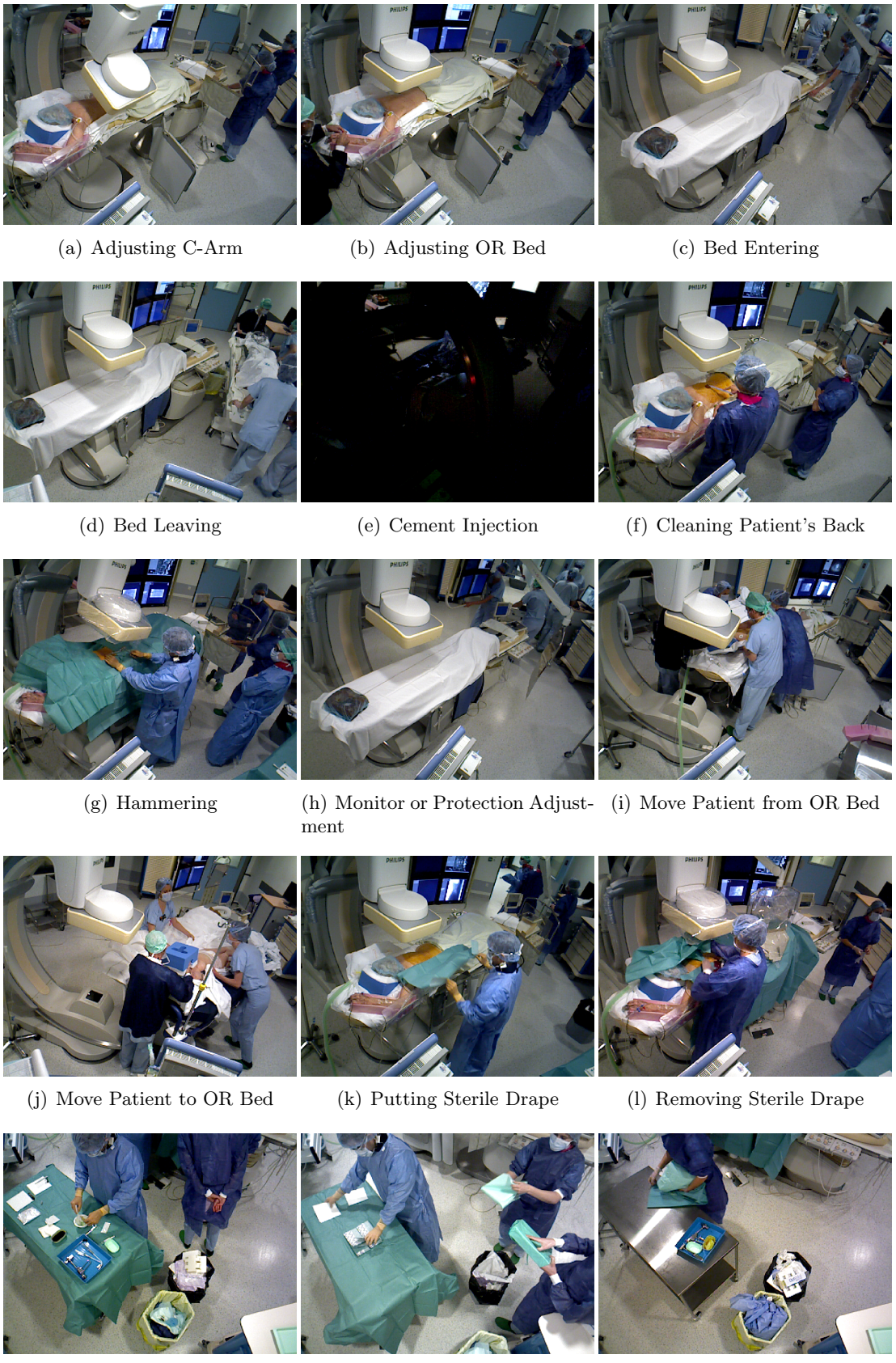
shown in Figure 4.1. Since the dataset is unbalanced as shown in Figure 4.1, bagging approach is used to balance instances from each class.

4.1.1.1 Activities in Dataset

The dataset has 15 different activities occurs in an hybrid operating room that are shown in Figure 4.2.

- **Adjusting C-Arm:** Most frequent activity in the dataset. Robotic C-Arm is moved by clinicians for taking x-ray images.

- **Adjusting OR Bed:** The operating room bed is moved by clinicians.
- **Bed Entering:** The bed enters the operating rooms empty or with a patient.
- **Bed Leaving:** The bed leaves the operating room empty or with a patient.
- **Cement Injection:** Cement is injected to patient and during the operation operating room is dark which cause high detection in intensity data. One of the imbalance instance in dataset.
- **Cleaning Patient's Back:** Clinicians sterilize patient by cleaning the operation area of the patient's skin.
- **Hammering:** Small needle is hammered through patient's fractured vertebra with continuous x-ray captures.
- **Monitor or Protection Adjustment:** Clinicians adjust display monitors according to surgeons mostly before the operations. Protection equipment for x-ray beams are moved during the x-ray acquisition.
- **Move Patient from OR Bed:** Patient is moved from operating room bed to portable bed after operation ends.
- **Move Patient to OR Bed:** After patient enters the operating room, anesthesia is applied and then the patient moved to operating room bed.
- **Putting Sterile Drape:** After cleaning patient's skin, sterile drapes are put to cover patient except the operating place.
- **Removing Sterile Drape:** After operation is ended, clinicians remove the sterile drapes and throw them to thrash.
- **Mixing Cement:** Cement is mixed by surgeon before cement injection procedure. Most Vertoblasty operations have one mixing cement activity. Mixing cement is the most imbalanced instance in our dataset.
- **Opening Package:** Packages that contain operation tools get opened on operating room tool table.
- **Throwing Package:** After operation ends, clinicians throw packages on the tool table to trash.



(a) Adjusting C-Arm (b) Adjusting OR Bed (c) Bed Entering

(d) Bed Leaving (e) Cement Injection (f) Cleaning Patient's Back

(g) Hammering (h) Monitor or Protection Adjustment (i) Move Patient from OR Bed

(j) Move Patient to OR Bed (k) Putting Sterile Drape (l) Removing Sterile Drape

(m) Mixing Cement (n) OpeningPackage (o) Throwing Package

FIGURE 4.2: Activity types in the dataset - samples of intensity images

4.1.2 Visual feature extraction and encoding

In interest point detection, we detect 1000 STIPs from intensity and 1000 DSTIPs from depth data from each video clip. Then, we extract 480-dimensional HOF from intensity and 4005-dimensional DCSF from depth data around each interest point. We use Principal Component Analysis (PCA) to reduce dimensions of HOF and DSCF to 250 and 800 respectively for visual dictionary construction. All these parameters have proven to produce good results in Twinanda et al. [1]. In [1], performance comparison of size of visual dictionaries are studied and it is shown that visual dictionary size 1500 has the highest accuracy and increasing the number of words in dictionary size has not much improve the performance. Then second dictionary, spatio-temporal dictionary, is built to divide 4D spatio-temporal space into non-rigid cells with 24 patches for intensity and depth data. When we use combination of intensity and depth data, the number of 4D patches becomes 48 (24 from both intensity and depth).

4.1.3 Classification Setup

In this work, we use Support Vector Machine (SVM) and Random Forest (RF) classifiers. In the experiments, two-level classification strategy is used and explained in detail in Section 3.3.3. We also use two different approaches: one-model and multi-model. In one-model approach, a single classifier model is trained on each level. In contrast, in the multi-model approach, multiple classifier models are trained on the first level. For intensity and depth data, 24 classifier models are trained. However, for the combination, 48 classifier models are trained. These numbers are used to match the number of 4D patches.

4.1.3.1 SVM Setup

SVM models in the experiments are trained using VLFeat toolbox [28]. SVM is used with non-linear kernels, i.e., Chi-square $(\mathcal{X})^2$ and histogram intersection kernels. Since there is 15 different classes for classification, one-against-all SVM is used to handle multi-class problem. K-Fold cross validation is used to evaluate performance by dividing the dataset into 10 folds.

4.1.3.2 Random Forest Setup

Random Forest models in the experiments are trained using our own implementation of Random Forest. Random Forests depend on many parameters, e.g., number of trees, maximum depth of the trees, minimum information gain and minimum number of samples in each tree node as stopping criteria to stop building a tree. Random Forests are trained with 100 trees, with maximum depth 15. The minimum information gain in the node is set to 0.001 to avoid trees with low information gain. One of the effective parameter in tree building is the minimum number of samples reach to each node before stopping. We use 32 samples as minimum number of samples in node as stopping criteria. Each tree in the random forest have seen 80% of the training data as randomly for bagging across trees. Finally K-Fold cross validation is used to evaluate the performance by dividing the dataset into 10 folds.

4.2 Experimental Results

In this section, we report results from the voting scheme as well as the comparison with non-voting scheme. We also compare the one-model and multi-model approaches using the proposed two-level classification strategy with the voting scheme, and their results compared on intensity, depth and the combination of the intensity and depth.

4.2.1 One Model Approach

In the one-model approach, we learn a single SVM and Random Forest model in each level with all patches from each activity. In Table 4.1, results for one-model approach is presented with intensity, depth and combination of the data with both of the classifiers. It is shown that non-linear SVM performs slightly better than the Random Forest. It is also shown that combination of the intensity and depth data has more discriminating power than using only intensity or depth. The combination of the depth and intensity profits the complementary information that intensity and depth carries. Because some activities are well classified in intensity due to illumination and intensity changes, e.g., cement injection, however some actions showed better performance in depth due to high variance in depth, e.g., Adjusting C-Arm. These effects are shown in confusion matrices in Figure 4.3 and Figure 4.4. The combination of the intensity and depth data

gets the highest accuracy with 83.1% accuracy using SVM and the confusion matrix is shown in Table 4.5.

	SVM	**RF**
Intensity	**74.45%**	70.06%
Depth	**72.72%**	67.87%
Combination	**83.1%**	79.18%

TABLE 4.1: Classification results using one-model approach

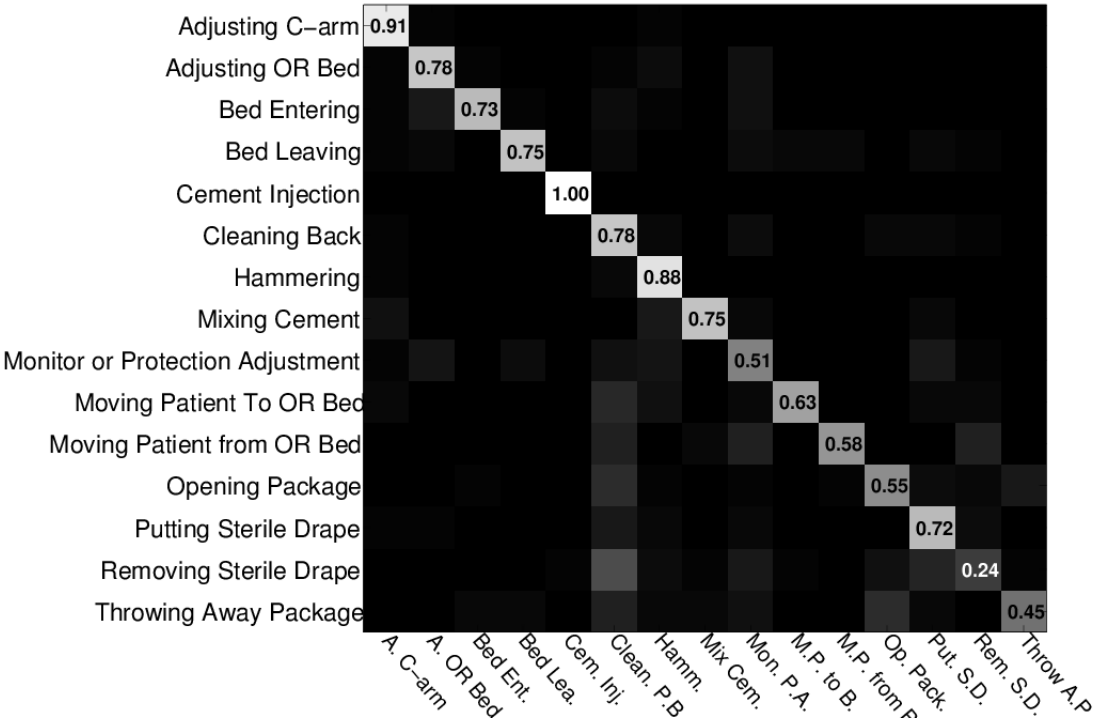


FIGURE 4.3: Confusion matrix of one-model approach using intensity features

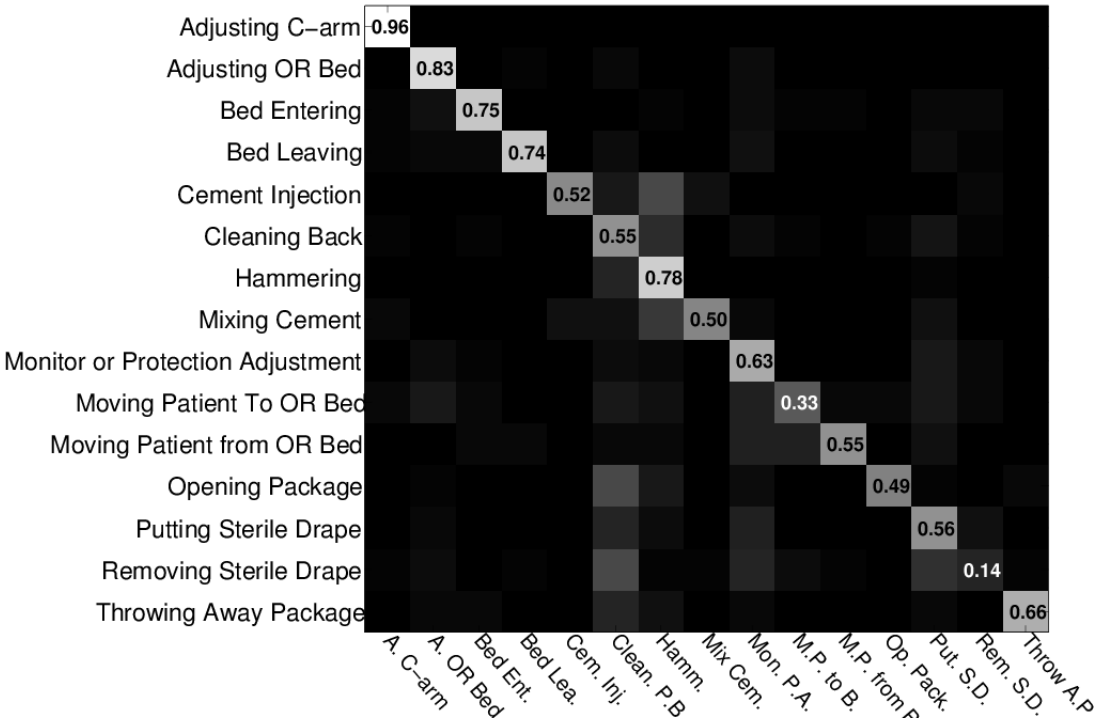


FIGURE 4.4: Confusion matrix of one-model approach using depth features

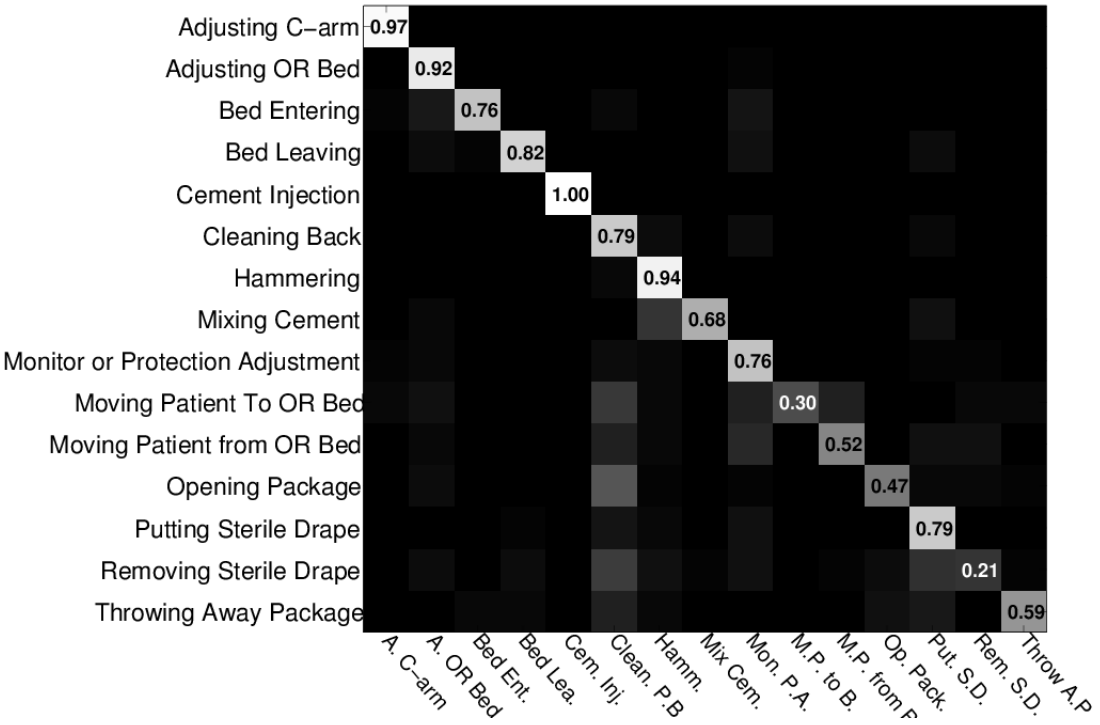


FIGURE 4.5: Confusion matrix of one-model approach using combination of features

4.2.2 Multi-Model

In the multi-model approach, we learn multiple models in the first level for each patch type P^j. However, the one-model approach in Section 4.2.1 trained a single model for the first level using all the patches from all the video clips. Hence, one-model approach profits from seeing more training data whereas in the multi-model approach, the first level classifiers sees less training data. Thus, multi-model has slight disadvantage over one-model approach due to training the each patch model. Thus, results presented in Table 4.2 have slighly lesser accuracy than one model results in Table 4.1. On the other hand, the combination of intensity and depth is still performing better than using only intensity or depth. We achieved 81.71% accuracy with combination of the data with the SVM classifier. The detailed results are presented with confusion matrices in Figure 4.6, 4.7 and 4.8.

	SVM	**RF**
Intensity	**77.28%**	69.20%
Depth	**71.91%**	64.45%
Combination	**81.71%**	75.54%

TABLE 4.2: Classification results using multi-model approach

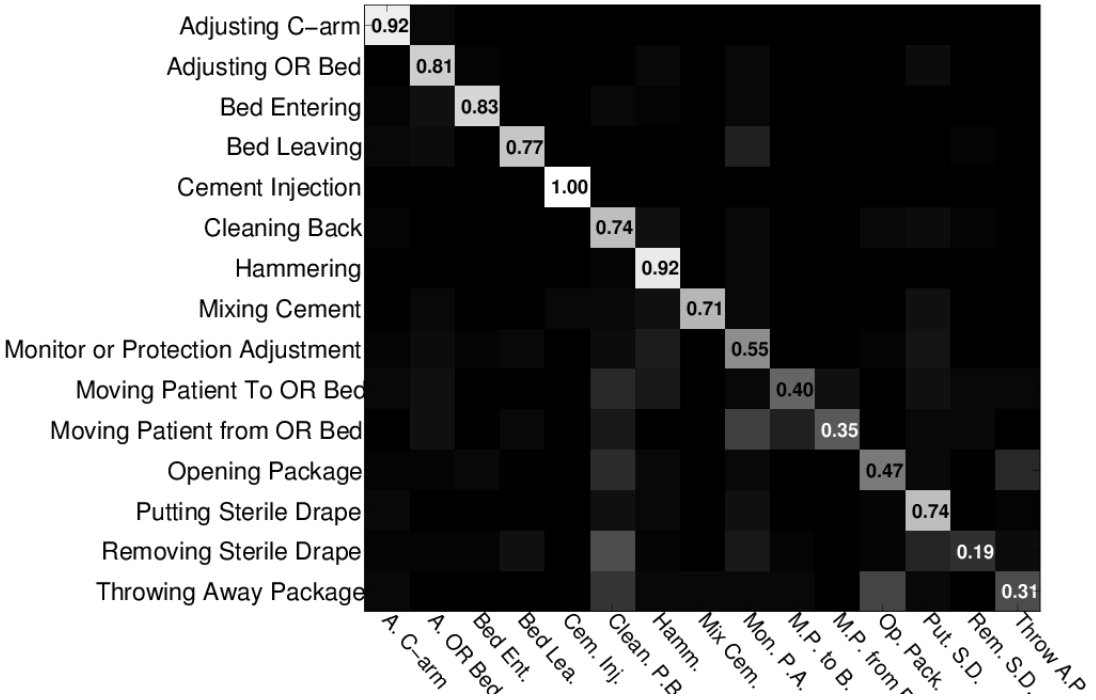


FIGURE 4.6: Confusion matrix of multi-model approach using intensity features

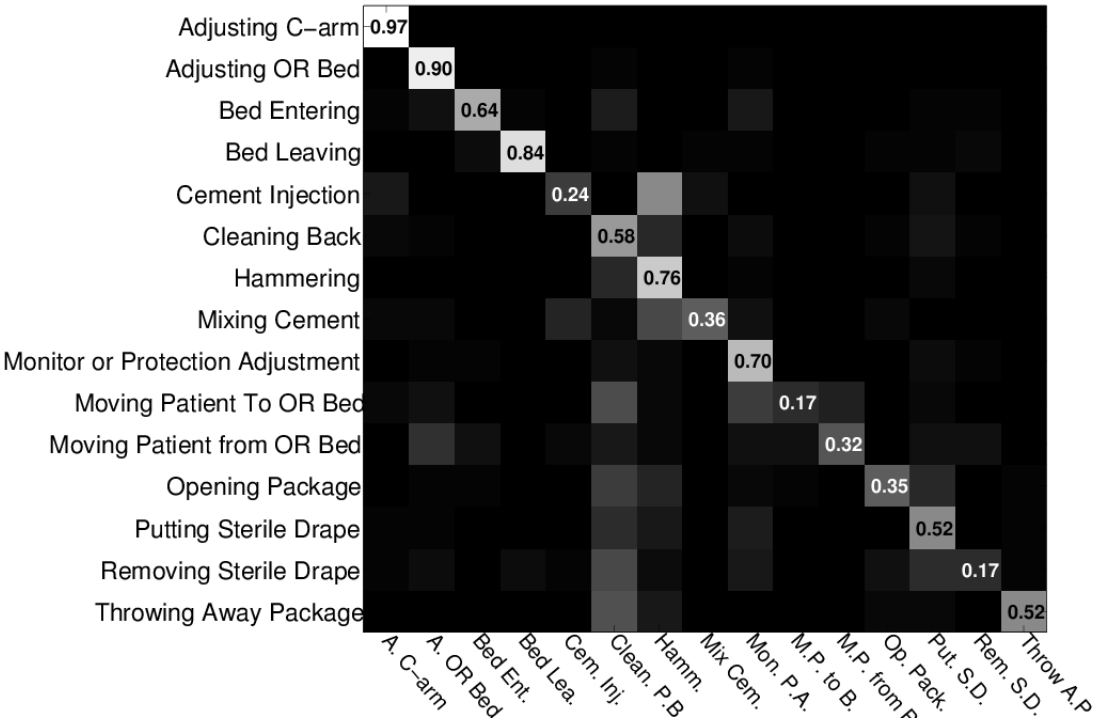


FIGURE 4.7: Confusion matrix of multi-model approach using depth features

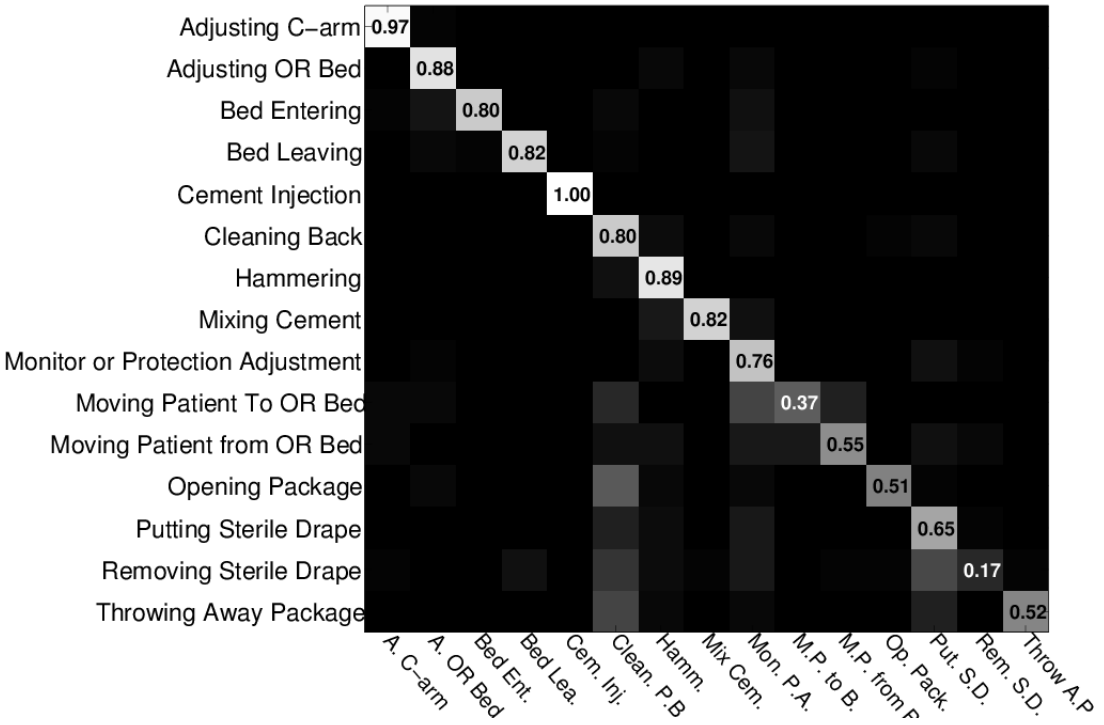


FIGURE 4.8: Confusion matrix of multi-model approach using combination of features

4.2.3 Non-Voting Approach

The proposed voting scheme approach also compared to non-voting scheme approach in [1]. In [1], a data-driven non-rigid layout is used by concatenating the histogram vectors from patches to construct single histogram vector to represent video clip. The histogram vector also concatenated with a histogram vector extracted from video clip without using any layout. Hence, the final concatenated histogram carries more information than using histograms from each patch separately as in Section 4.2.1 and Section 4.2.2. Using the patch histograms separately for training and collecting votes increased the sparsity problem in classification in the first level. The non-voting method reached 85.53% accuracy compared to our maximum accuracy 83.1%. The detailed comparison with intensity, depth data and their combination is shown in Table 4.3.

	Non-Voting [1]	**Voting**
Intensity	**80.40%**	74.45%
Depth	**78.37%**	72.72%
Combination	**85.53%**	83.1%

TABLE 4.3: Classification accuracy comparison of non-voting [1] and voting scheme

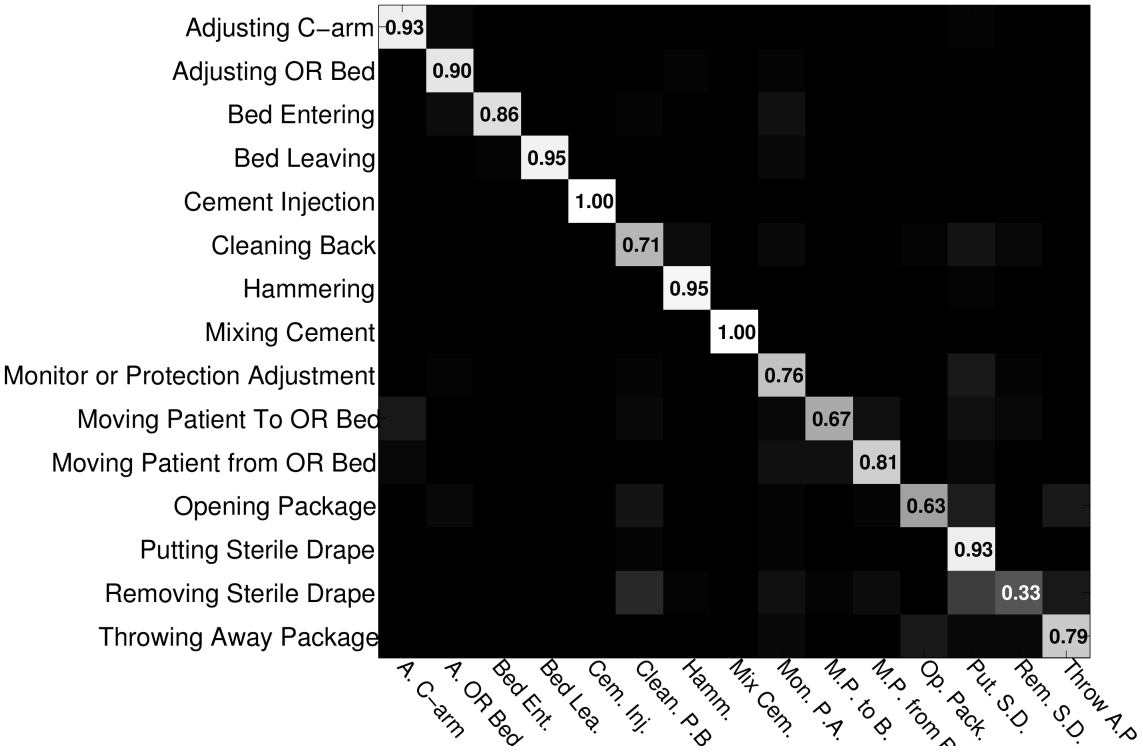


FIGURE 4.9: Confusion matrix of non-voting scheme using combination of features[1]

In comparison of confusion matrices with voting and non-voting scheme from Figure 4.5 and Figure 4.9, it is shown that the voting scheme has better detection on activities, i.e., Adjusting C-Arm, Adjusting OR Bed and Cleaning Back, where in these activities, detected interest points are more separable and informative than other activities. For example, in Mixing Cement activity, 32% accuracy decreased in voting scheme approach. In this activity, interest points are detected in very narrow space, and they detected as cluster around the activity, Thus, detected interest points are not uniform

in 4D space but distributed around one point. The decrease in the accuracy can be explained as the detected interest points in 4D space is not separated informatively with non-rigid layout for the Mixing Cement activity.

Chapter 5

Conclusions

In this thesis, a voting scheme approach is proposed to address the problem of activity recognition in an operating room using a multi-view RGBD camera system. The 4D spatio-temporal space is divided into smaller local patches using a data-driven non-rigid layout by [1] which also helps with sparse interest points. A two-level classification strategy is introduced, i.e., learning the probability votes and learning the weights. Furthermore, two different classification approaches are compared using two-level classification: one-model and multi-model. The one-model approach trained a single classifier model with all patches from the all video clips in each level, however, the multi-model approach trained separate classifier models for each patch type in a video clip in the fist level. Finally, the proposed voting strategy collects votes from the each local patch from a video clip and the weights are used to recognize the activity. The voting scheme is evaluated on a new dataset from [1] which consists of annotated 1734 real surgical video clips with 15 different surgical activity types. In order to make comprehensive comparison, we compared the proposed voting scheme with non-voting scheme approach from [1]. We have shown that the proposed voting scheme gives promising results.

The current spatio-temporal interest point detection is based on movements occur in the video clips where the detection is in the borders of the action. Hence, the detected interest points are sparse and ignore the information from stationary parts of the videos. Additionally, the bag-of-words approach is another limit which encodes the features into sparse representation. It would be interesting to incorporate extraction of dense features and different feature encoding approaches. Since the voting scheme uses local

parts of 4D spatio-temporal space, it would be interesting to use the voting scheme to recognize concurrent activities.

Appendix A

The operating room

FIGURE A.1: The operating room - panoramic view

Bibliography

[1] Andru P Twinanda, Emre O Alkan, Afshin Gangi, Michel de Mathelin, and Nicolas Padoy. Data-driven spatio-temporal rgbd feature encoding for action recognition in operating rooms. *International journal of computer assisted radiology and surgery*, pages 1–11, 2015.

[2] Beenish Bhatia, Tim Oates, Yan Xiao, and Peter Hu. Real-time identification of operating room state from video. In *Proceedings of the 19th National Conference on Innovative Applications of Artificial Intelligence - Volume 2*, IAAI'07, pages 1761–1766. AAAI Press, 2007. ISBN 978-1-57735-323-2. URL http://dl.acm.org/citation.cfm?id=1620113.1620126.

[3] Florent Lalys, Laurent Riffaud, David Bouget, and Pierre Jannin. A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. *IEEE Trans. Biomed. Engineering*, 59(4):966–976, 2012. doi: 10.1109/TBME.2011.2181168. URL http://dx.doi.org/10.1109/TBME.2011.2181168.

[4] Nicolas Padoy, Tobias Blum, Seyed-Ahmad Ahmadi, Hubertus Feussner, Marie-Odile Berger, and Nassir Navab. Statistical modeling and recognition of surgical workflow. *Medical Image Analysis*, 16(3):632 – 641, 2012. ISSN 1361-8415. doi: http://dx.doi.org/10.1016/j.media.2010.10.001. URL http://www.sciencedirect.com/science/article/pii/S1361841510001131. Computer Assisted Interventions.

[5] Luca Zappella, Benjamín Béjar, Gregory Hager, and René Vidal. Surgical gesture classification from video and kinematic data. *Medical Image Analysis*, 17(7):732 – 745, 2013. ISSN 1361-8415. doi: http://dx.doi.org/10.1016/j.media.2013.04.007. URL http://www.sciencedirect.com/science/article/

pii/S1361841513000522. Special Issue on the 2012 Conference on Medical Image Computing and Computer Assisted Intervention.

[6] I. Chakraborty, A. Elgammal, and R.S. Burd. Video based activity recognition in trauma resuscitation. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8, April 2013. doi: 10.1109/FG.2013.6553758.

[7] Nicolas Loy Rodas and Nicolas Padoy. Seeing is believing: increasing intraoperative awareness to scattered radiation in interventional procedures by combining augmented reality, monte carlo simulations and wireless dosimeters. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–11, 2015. ISSN 1861-6410. doi: 10.1007/s11548-015-1161-x. URL http://dx.doi.org/10.1007/s11548-015-1161-x.

[8] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178, 2006. doi: 10.1109/CVPR.2006.68.

[9] Lu Xia and J.K. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2834–2841, June 2013. doi: 10.1109/CVPR.2013.365.

[10] Heng Wang, A. Klaser, C. Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176, June 2011. doi: 10.1109/CVPR.2011.5995407.

[11] Xiaodong Yang and YingLi Tian. Super normal vector for activity recognition using depth sequences. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 804–811, June 2014. doi: 10.1109/CVPR.2014.108.

[12] Alexander Klaser, Marcin Marszalek, Cordelia Schmid, and Andrew Zisserman. Human focused action localization in video. In *Proceedings of the 11th European Conference on Trends and Topics in Computer Vision - Volume Part I*,

ECCV'10, pages 219–233, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-35748-0. doi: 10.1007/978-3-642-35749-7_17. URL http://dx.doi.org/10.1007/978-3-642-35749-7_17.

[13] Yen-Yu Lin, Ju-Hsuan Hua, N.C. Tang, Min-Hung Chen, and H.-Y.M. Liao. Depth and skeleton associated action recognition without online accessible rgb-d cameras. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2617–2624, June 2014. doi: 10.1109/CVPR.2014.335.

[14] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 588–595, June 2014. doi: 10.1109/CVPR.2014.82.

[15] Di Wu and Ling Shao. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 724–731, June 2014. doi: 10.1109/CVPR.2014.98.

[16] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 9–14, June 2010. doi: 10.1109/CVPRW.2010.5543273.

[17] O. Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 716–723, June 2013. doi: 10.1109/CVPR.2013.98.

[18] Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu. Robust 3d action recognition with random occupancy patterns. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part II*, ECCV'12, pages 872–885, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-33708-6. doi: 10.1007/978-3-642-33709-3_62. URL http://dx.doi.org/10.1007/978-3-642-33709-3_62.

[19] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern*

Recognition (CVPR), 2012 IEEE Conference on, pages 1290–1297, June 2012. doi: 10.1109/CVPR.2012.6247813.

[20] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning, and recognition. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2649–2656, June 2014. doi: 10.1109/CVPR.2014.339.

[21] Abdolrahim Kadkhodamohammadi, Afshin Gangi, Michel de Mathelin, and Nicolas Padoy. Temporally consistent 3d pose estimation in the interventional room using discrete mrf optimization over rgbd sequences. In Danail Stoyanov, D.Louis Collins, Ichiro Sakuma, Purang Abolmaesumi, and Pierre Jannin, editors, *Information Processing in Computer-Assisted Interventions*, volume 8498 of *Lecture Notes in Computer Science*, pages 168–177. Springer International Publishing, 2014. ISBN 978-3-319-07520-4. doi: 10.1007/978-3-319-07521-1_18. URL http://dx.doi.org/10.1007/978-3-319-07521-1_18.

[22] Fan Zhu, Ling Shao, and Mingxiu Lin. Multi-view action recognition using local similarity random forests and sensor fusion. *Pattern Recognition Letters*, 34(1):20 – 24, 2013. ISSN 0167-8655. doi: http://dx.doi.org/10.1016/j.patrec.2012.04.016. URL http://www.sciencedirect.com/science/article/pii/S0167865512001407. Extracting Semantics from Multi-Spectrum Video.

[23] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2-3):249–257, 2006.

[24] Daniel Weinland, Mustafa Özuysal, and Pascal Fua. Making action recognition robust to occlusions and viewpoint changes. In *European Conference on Computer Vision*, 2010.

[25] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[26] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72, Oct 2005. doi: 10.1109/VSPETS.2005.1570899.

[27] F. Barrera and N. Padoy. Piecewise planar decomposition of 3d point clouds obtained from multiple static rgb-d cameras. In *3D Vision (3DV), 2014 2nd International Conference on*, volume 1, pages 194–201, Dec 2014. doi: 10.1109/3DV.2014.57.

[28] Andrea Vedaldi and Brian Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 1469–1472, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-933-6. doi: 10.1145/1873951.1874249. URL http://doi.acm.org/10.1145/1873951.1874249.