Action recognition in robot-assisted minimally invasive surgery

Candidate: Laura Erica Pescatori
Co-Tutor: Hirenkumar Chandrakant Nakawala
Tutor: Elena De Momi
Project Objective

Surgeon’s action recognition during robot-assisted minimally invasive surgery (RAMIS) performed with da Vinci robot

APPLYCONTEXT
context-aware tool to aid in surgical training and in decision-making

• Automatic UNDERSTANDING and ANALYSIS of the workflow
• Real-time FEEDBACK on the surgeon’s action and gesture
  Ex: NEXT action to be performed

da Vinci Robot:
  • Console
  • Robot
Partial Nephrectomy

Renal cells carcinoma is the seventh most common cancer in men and ninth in woman with increasing cases every year (Rini et al. 2009)

Important steps of Partial Nephrectomy (PN)

1) Temporal occlusion of the renal vasculature
2) Resection of tumor with a rim of normal parenchyma
3) Reconstruction of the kidney
Procedure granularity levels

Each granularity level describes the surgical procedure as a sequential list of events. [Lalys et al. 2014]

- **PRO Procedure**
  - Major types of events occurring during surgery.
  - Phases are semantically meaningful group of steps occurring sequentially in a process.

- **Phases**
  - Sequence of activities used to achieve a surgical objective.

- **Steps**
  - Physical task: the fundamental element in the semantic interpretation of a scene. Each activity is composed of a list of motions.

- **Activities**
  - A motion is a surgical task involving only one hand trajectory.

Lower granularity levels

[Diagram showing the process of Tumor Resection with steps such as Clamp, Rejoin, Dissect]
VIDEO DATASET:
- 8 surgical procedures performed by 4 different surgeons
- Annotated with Anvil software by an expert urologist

<table>
<thead>
<tr>
<th>Phase</th>
<th>Step</th>
<th>Instrument-Right</th>
<th>Instrument-Left</th>
<th>Instrument1-Assistant</th>
<th>Instrument2-Assistant</th>
<th>Anatomy</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nephrectomy</td>
<td>00:00</td>
<td>bowelMobilization</td>
<td>mobilization</td>
<td>MonopolarCurvedScissors</td>
<td>fenestratedBipolar</td>
<td>peritoneum</td>
<td>Incise</td>
</tr>
<tr>
<td>1</td>
<td>00:01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>00:02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>00:03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>00:04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>00:05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>00:06</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>00:07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>00:08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>00:09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>00:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>00:11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>00:12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>00:13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>00:14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>00:15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>00:16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>00:17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>00:18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>00:19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>00:20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>00:21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>00:22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>00:23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>00:24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>00:25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>00:26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>00:27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>00:28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>00:29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>00:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Entities:
- Annotated elements
- Video seconds
Images features extraction

**HANDCRAFTED FEATURES**

- Designed to capture certain characteristics from data (e.g., corners).
- Robust to some common variations (translation, rotation, illumination, occlusion and scale variance).

**NEURAL NETWORKS**

- **CONVOLUTIONAL NEURAL NETWORKS** (CNN) allow the **AUTOMATICAL LEARNING** of discriminative **FEATURES** in the image.
- Classification provided in a one-step-solution.
- They outperform the results obtained with handcrafted features.

---

[LeCun et al. 2015]

Features cannot be automatically learnt

CLASSIFICATION algorithms

[Awas et al. 2016]
Representation-learning methods with multiple levels of representation.

They are composed of the so called hidden layers, each one composed by a certain number of neurons.

Neurons have trainable parameters called weights that are initialized and gradually adjusted during the training.

In CNNs layers parameters consist of a set of learnable filters convolved across the input volume. The network learns filters that activate when they see some type of visual feature.

Probability vector: one probability for each class of the dataset.
State of the art: Surgical field

**PHASE RECOGNITION**

**ENDONET:**
- CNN for features extraction
- Tool presence detection (7 tools, binary classification)
- Phase recognition

**[Twinanda et al. 2016]**

- CNN for features extraction
- Recurrent unit (LSTM) to analyze time-dependent data

**[Namazi et al. 2018]**
State of the art: Human action recognition

LRCN (Long-Term Recurrent Convolutional Networks): [Donahue et al. 2014]

Basic architecture:

Pretrained Alexnet → CNN → LSTM → Average → Y → RGB images

CNN → LSTM → Average → OPTICAL flow images
Optical flow estimation consists in the computation of the displacement vector field \( \mathbf{d}_t \) between two consecutive frames images at time \( t \) and \( t + 1 \).

\( \mathbf{d}_t(u, v) \) denote the displacement vector at the point \((u, v)\) in frame \( t \), which moves the point to the corresponding point in the following frame \( t + 1 \).

\[
\mathbf{p}_1 = (u, v) \quad \mathbf{d}_t(p_1)
\]

[Simonyan et al. 2014]
Proposed approaches

1) Action Recognition
1a. «SINGLE-STREAM»
1b. 5 classes action dataset
1c. CONFUSION MATRIX

2a. «SINGLE-STREAM» vs «DOUBLE-STREAM»
2b. 2 classes dataset obtained from dissect
2c. 10-fold cross validation

2) Increased Granularity
Action Recognition
1) Action Recognition

**«Single-stream» architecture**

- **Timestep t**: (224x224) image
  - Conv. Layer 32x3x3-s-1
  - Conv. Layer 48x3x3-s-1
  - Conv. Layer 64x3x3-s-1
  - Conv. Layer 128x3x3-s-1
  - Maxpool 2x2-s-2
  - LSTM 256
- **Timestep t+1**: (224x224) image
  - Conv. Layer 32x3x3-s-1
  - Conv. Layer 48x3x3-s-1
  - Conv. Layer 64x3x3-s-1
  - Conv. Layer 128x3x3-s-1
  - Maxpool 2x2-s-2
  - LSTM 256

- **Hidden states of the LSTMs**
  - At time t influence the hidden states of the LSTMs at time t+1.

**Convolutional Layer**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>4</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Maxpooling Layer**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

3x3 filter stride1

2x2 filter stride 2

* «Single-stream» architecture

**1a. SINGLE-STREAM**

1b. Action Dataset | 1c. CONFUSION MATRIX | 2a. SINGLE-STREAM vs DOUBLE-STREAM | 2b. Action motion Dataset | 2c. 10-fold cross validation
Action Dataset

Data augmentation

-10° Rotation
+10° Rotation
Original
Cropping
Contrast
Brightness

<table>
<thead>
<tr>
<th>Action</th>
<th>N. videos</th>
<th>N. images</th>
</tr>
</thead>
<tbody>
<tr>
<td>cut</td>
<td>262</td>
<td>8342</td>
</tr>
<tr>
<td>dissect</td>
<td>444</td>
<td>14821</td>
</tr>
<tr>
<td>mark</td>
<td>86</td>
<td>3621</td>
</tr>
<tr>
<td>resect</td>
<td>179</td>
<td>6268</td>
</tr>
<tr>
<td>suture</td>
<td>98</td>
<td>2730</td>
</tr>
</tbody>
</table>

Train: 72%
Validation: 18%
Test: 10%

1a. SINGLE - STREAM
1b. Action Dataset
1c. CONFUSION MATRIX
2a. SINGLE-STREAM vs DOUBLE-STREAM
2b. Action motion Dataset
2c. 10-fold cross validation

POLITECNICO MILANO 1863
The confusion matrix visually represents the capacity of the algorithm to predict the correct class. It shows if the system is confusing different classes.

Elements along the diagonal are those correctly predicted.
2) Increased Granularity Action Recognition

«Single-stream» architecture

«Double-stream» architecture

Same layers characteristics as before
Optical flow images

Optical flow estimated with Brox and Bruhn computation [Brox et al. 2004]

PROBLEM:
Organs motion presence

Optical flow image extraction

Subsequent images

Instrument detection: YOLOv2 [Redmond et al. 2016]

Instruments flow extraction

1a. SINGLE-STREAM
1b. Action Dataset
1c. CONFUSION MATRIX
2a. SINGLE-STREAM vs DOUBLE-STREAM
2b. Action motion Dataset
2c. 10-fold cross validation
The original action **Dissect** was divided into two classes characterized by a lower granularity.

Each class is characterized by a specific **MOTION**.

<table>
<thead>
<tr>
<th>Subclass</th>
<th>N. videos</th>
<th>N. RGB images</th>
<th>N. Flow images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dissect-Tissue_cut</td>
<td>60</td>
<td>1682</td>
<td>1622</td>
</tr>
<tr>
<td>Dissect-Vertical_movement</td>
<td>68</td>
<td>1607</td>
<td>1539</td>
</tr>
</tbody>
</table>
10-fold Cross Validation

The 60 videos are separated into 10 clusters each one containing 6 videos

10 experiments are performed: at each experiment the test cluster changes

ADVANTAGE: the result does not depend on the chosen train and test set

RESULTS

AVERAGE ACCURACY over the 10 experiments

«Single-stream» | «Double-stream»
--- | ---
73,3% | 86,7%
Conclusion and future development

1) Action recognition
   - «Single-stream» could be a valid tool to be used to recognize surgeon’s actions.
   - Variation in the number of filters could be explored
   - Cross validation

2) Increased granularity action recognition
   - «Double-stream»: optical flow can provide more discriminative information that the only RGB images
   - Wider dataset characterized by a proper redefinition

FUTURE DEVELOPMENTS
   - Instrument SEGMENTATION
   - Introduction of a TRACKING algorithm for the instrument
THANK YOU FOR YOUR ATTENTION
Medical errors related to surgical skills

- Lower **technical skills** score associated to: higher complication rates (14.5% vs. 5.2%, $P<0.001$) higher mortality (0.26% vs. 0.05%, $P=0.01$) longer operations (137 minutes vs. 98 minutes, $P<0.001$) higher rates of reoperation (3.4% vs. 1.6%, $P=0.01$) readmission (6.3% vs. 2.7%) ($P<0.001$). [Birkmeyer et al. 2013]

- Factors contributing to errors were: inexperience/lack of competence in a surgical task (53% of incidents), communication breakdowns among personnel (43%), fatigue or excessive workload (33%). (technical skills) (non-technical skills) [Gawande et al 2003]

- Poor **non-technical skills** were identified as a cause of death in 46.6% of the cases (41.2%) of problems with situation awareness (23.5%) with team working (8.8%) with decision making [Uramatzu et al 2017]