Effective Performance Metrics for Evaluating Activity Recognition Methods

Tim van Kasteren, Hande Alemdar and Cem Ersoy

Boğaziçi University, Istanbul, Turkey

February 23, 2011
Outline

1. Introduction

2. Evaluation
   - Recognition Performance Metrics
   - Analyzing Errors

3. Experiments
   - Models compared
   - Datasets
   - Results

4. Conclusions
What is activity recognition?

- Extracting higher level information from low level sensor data.
- For example: recognizing human activities of daily living such as cooking, sleeping and bathing.
- Characteristics: Temporal data, fuzzy boundaries.
Introduction: Evaluation Methods

Problem: No clear methodology for evaluation

- Methods for activity recognition are often evaluated using standard pattern recognition metrics.
- Such metrics were initially developed for independently and identically distributed (I.I.D.) data.
- Those methods are generally inappropriate because:
  1. Activity recognition datasets often contain imbalanced classes.
  2. Temporal nature of activities results in different qualifications of errors.
Solution: New set of metrics needed

- We propose to use two types of performance metrics:
  1. Summarizing the overall performance of a recognition method.
  2. Analyzing the strengths and weaknesses of a recognition method.
Outline

1. Introduction

2. Evaluation
   • Recognition Performance Metrics
   • Analyzing Errors

3. Experiments
   • Models compared
   • Datasets
   • Results

4. Conclusions
<table>
<thead>
<tr>
<th>Ground truth</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>$TP_1$</td>
<td>$\epsilon_{12}$</td>
<td>$\epsilon_{13}$</td>
<td>$NG_1$</td>
</tr>
<tr>
<td>Class 2</td>
<td>$\epsilon_{21}$</td>
<td>$TP_2$</td>
<td>$\epsilon_{23}$</td>
<td>$NG_2$</td>
</tr>
<tr>
<td>Class 3</td>
<td>$\epsilon_{31}$</td>
<td>$\epsilon_{32}$</td>
<td>$TP_3$</td>
<td>$NG_3$</td>
</tr>
<tr>
<td></td>
<td>$NI_1$</td>
<td>$NI_2$</td>
<td>$NI_3$</td>
<td>Total</td>
</tr>
</tbody>
</table>

**Table:** Confusion Matrix showing true positives (TP), total number of ground truth labels (NG) and total number of inferred labels (NI) for each class.

**Evaluation: Confusion Matrices**

- We assume data is discretized into timeslices of constant length.
- The correct and incorrect classifications can be represented in a confusion matrix.
<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Inferred</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>$TP_1$</td>
</tr>
<tr>
<td>2</td>
<td>$\epsilon_{21}$</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$\epsilon_{31}$</td>
<td>$\epsilon_{32}$</td>
</tr>
</tbody>
</table>

Table: Confusion Matrix showing true positives (TP), total number of ground truth labels (NG) and total number of inferred labels (NI) for each class.

**Good:**
- Provides a lot of insight about the classification

**Bad:**
- Take up a lot of space
- Do not take temporal aspects of errors into account
<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Inferred</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$TP_1$</td>
<td>$\epsilon_{12}$</td>
<td>$\epsilon_{13}$</td>
<td>$NG_1$</td>
</tr>
<tr>
<td>2</td>
<td>$\epsilon_{21}$</td>
<td>$TP_2$</td>
<td>$\epsilon_{23}$</td>
<td>$NG_2$</td>
</tr>
<tr>
<td>3</td>
<td>$\epsilon_{31}$</td>
<td>$\epsilon_{32}$</td>
<td>$TP_3$</td>
<td>$NG_3$</td>
</tr>
<tr>
<td></td>
<td>$NI_1$</td>
<td>$NI_2$</td>
<td>$NI_3$</td>
<td>$Total$</td>
</tr>
</tbody>
</table>

**Evaluation: Accuracy**

- Results of can be compactly shown with the accuracy measure.

\[
\text{Accuracy} = \frac{\sum_{i=1}^{Q} TP_i}{Total}
\]

- However, this measure does not take class imbalance into account.
### Evaluation: Precision, Recall and F-measure

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Inferred</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$TP_1$</td>
<td>$\epsilon_{12}$</td>
<td>$\epsilon_{13}$</td>
<td>$NG_1$</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$\epsilon_{21}$</td>
<td>$TP_2$</td>
<td>$\epsilon_{23}$</td>
<td>$NG_2$</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$\epsilon_{31}$</td>
<td>$\epsilon_{32}$</td>
<td>$TP_3$</td>
<td>$NG_3$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$NI_1$</td>
<td>$NI_2$</td>
<td>$NI_3$</td>
<td>Total</td>
<td></td>
</tr>
</tbody>
</table>

**Precision**

$$\text{Precision} = \frac{1}{Q} \sum_{i=1}^{Q} \frac{TP_i}{NI_i} \quad (1)$$

**Recall**

$$\text{Recall} = \frac{1}{Q} \sum_{i=1}^{Q} \frac{TP_i}{NG_i} \quad (2)$$

**F-Measure**

$$\text{F-Measure} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$
Outline

1 Introduction

2 Evaluation
   - Recognition Performance Metrics
   - Analyzing Errors

3 Experiments
   - Models compared
   - Datasets
   - Results

4 Conclusions
We define four error types:

- Substitution errors
- Occurrence errors
- Timing errors
- Segmentation errors
Figure: Examples of substitution and occurrence error types, circles denote inferred (Inf.), squares denote ground truth (Gr.).

Substitution and occurrence errors:
- Substitution errors (S)
- Occurrence errors: Insertion (I) and Deletion(D)
Figure: Examples of the timing error type, circles denote inferred (Inf.), squares denote ground truth (Gr.).

Timing errors

- Overfilled error (O)
- Underfilled error (U)
Figure: Examples of the segmentation error type, circles denote inferred (Inf.), squares denote ground truth (Gr.).

Segmentation errors

- Merging errors (M)
- Fragmentation errors (F)
Error Metrics

### Substitution Errors:

\[
S_G^* = \frac{1}{Q} \sum_{i=1}^{Q} \frac{\sum_{j=1}^{Q} S(i, j)}{NG_i} \quad (4)
\]

\[
S_I^* = \frac{1}{Q} \sum_{j=1}^{Q} \frac{\sum_{i=1}^{Q} S(i, j)}{NI_i} \quad (5)
\]

### Occurrence Errors:

\[
D^* = \frac{1}{Q} \sum_{i=1}^{Q} \frac{D_i}{NG_i} \quad (6)
\]

\[
I^* = \frac{1}{Q} \sum_{i=1}^{Q} \frac{I_i}{NI_i} \quad (7)
\]

### Timing Errors:

\[
U^* = \frac{1}{Q} \sum_{i=1}^{Q} \frac{U_i}{NG_i} \quad (8)
\]

\[
O^* = \frac{1}{Q} \sum_{i=1}^{Q} \frac{O_i}{NI_i} \quad (9)
\]

### Segmentation Errors:

\[
F^* = \frac{1}{Q} \sum_{i=1}^{Q} \frac{F_i}{NG_i} \quad (10)
\]

\[
M^* = \frac{1}{Q} \sum_{i=1}^{Q} \frac{M_i}{NI_i} \quad (11)
\]
<table>
<thead>
<tr>
<th>Errors</th>
<th>Correct</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substitution</td>
<td>$S^*_G$</td>
<td>$S^*_i$</td>
<td></td>
<td>Sub*</td>
</tr>
<tr>
<td>Occurrence</td>
<td>$D^*$</td>
<td>$I^*$</td>
<td></td>
<td>Occ*</td>
</tr>
<tr>
<td>Timing</td>
<td>$U^*$</td>
<td>$O^*$</td>
<td></td>
<td>Tim*</td>
</tr>
<tr>
<td>Segmentation</td>
<td>$F^*$</td>
<td>$M^*$</td>
<td></td>
<td>Seg*</td>
</tr>
</tbody>
</table>

**Table**: Compact result matrix showing all the error metrics.
Outline

1. Introduction

2. Evaluation
   - Recognition Performance Metrics
   - Analyzing Errors

3. Experiments
   - Models compared
   - Datasets
   - Results

4. Conclusions
Two probabilistic models are compared

- Hidden Markov model (HMM) and Conditional Random Fields (CRF) are compared.
- Both temporal models with a similar dependency structure.
- HMM are generative models, CRFs are discriminative models.
Outline

1. Introduction

2. Evaluation
   - Recognition Performance Metrics
   - Analyzing Errors

3. Experiments
   - Models compared
   - Datasets
   - Results

4. Conclusions
<table>
<thead>
<tr>
<th>House</th>
<th>House A</th>
<th>House B</th>
<th>House C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>26</td>
<td>28</td>
<td>57</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>Male</td>
<td>Male</td>
</tr>
<tr>
<td>Setting</td>
<td>Apartment</td>
<td>Apartment</td>
<td>House</td>
</tr>
<tr>
<td>Rooms</td>
<td>3</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Duration</td>
<td>25 days</td>
<td>13 days</td>
<td>18 days</td>
</tr>
<tr>
<td>Sensors</td>
<td>14</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td>Activities</td>
<td>10</td>
<td>14</td>
<td>16</td>
</tr>
</tbody>
</table>

Table: Information about the datasets recorded in three different houses.

Datasets:

- Publicly available from [http://sites.google.com/site/tim0306/](http://sites.google.com/site/tim0306/).
Figure: House A

Figure: House B

Figure: House C, first floor

Figure: House C, second floor
Outline

1. Introduction

2. Evaluation
   - Recognition Performance Metrics
   - Analyzing Errors

3. Experiments
   - Models compared
   - Datasets
   - Results

4. Conclusions
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>HMM</td>
<td>66</td>
<td>68</td>
<td>67</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>57</td>
<td>68</td>
<td>62</td>
<td>91</td>
</tr>
<tr>
<td>B</td>
<td>HMM</td>
<td>55</td>
<td>42</td>
<td>48</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>46</td>
<td>54</td>
<td>49</td>
<td>92</td>
</tr>
<tr>
<td>C</td>
<td>HMM</td>
<td>40</td>
<td>37</td>
<td>39</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>30</td>
<td>36</td>
<td>33</td>
<td>78</td>
</tr>
</tbody>
</table>

**Table**: Recall, Precision (Prec.), F-measure (F-Meas.) and Accuracy (Acc.) results for the HMM and the CRF on all three houses.
**Table:** Result matrices for House A, values are percentages.

<table>
<thead>
<tr>
<th></th>
<th>Ground</th>
<th>Infer.</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HMM</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td>66</td>
<td>68</td>
<td>67</td>
</tr>
<tr>
<td>Substitution</td>
<td>21</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td>Occurrence</td>
<td>5</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>Timing</td>
<td>6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Segmentation</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>CRF</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td>57</td>
<td>68</td>
<td>62</td>
</tr>
<tr>
<td>Substitution</td>
<td>30</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>Occurrence</td>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Timing</td>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Segmentation</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Errors</td>
<td>HMM</td>
<td>CRF</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------</td>
<td>-------------------</td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td>55</td>
<td>42</td>
<td>48</td>
</tr>
<tr>
<td>Substitution</td>
<td>36</td>
<td>41</td>
<td>38</td>
</tr>
<tr>
<td>Occurrence</td>
<td>6</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>Timing</td>
<td>3</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Segmentation</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table:** Result matrices for House B, values are percentages.
<table>
<thead>
<tr>
<th>Errors</th>
<th>HMM</th>
<th>CRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>40</td>
<td>37</td>
</tr>
<tr>
<td>Substitution</td>
<td>47</td>
<td>42</td>
</tr>
<tr>
<td>Occurrence</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>Timing</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Segmentation</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table:** Result matrices for House C, values are percentages.
Outline

1. Introduction

2. Evaluation
   - Recognition Performance Metrics
   - Analyzing Errors

3. Experiments
   - Models compared
   - Datasets
   - Results

4. Conclusions
Conclusions

- We showed how Precision, Recall and F-measure can be used in a multi-class problem taking class imbalance into account.
- Presented several error metrics for compactly presenting the strengths and weaknesses of a recognition method.
- But I don’t believe this work is final, so please bring on the criticism.
Conclusions

- We showed how Precision, Recall and F-measure can be used in a multi-class problem taking class imbalance into account.
- Presented several error metrics for compactly presenting the strengths and weaknesses of a recognition method.
- But I don’t believe this work is final, so please bring on the criticism.
Thank you for your time.