A RANDOM FOREST APPROACH TO SEGMENTING AND CLASSIFYING GESTURES

Ajjen Joshi, Camille Monnier, Margrit Betke, Stan Sclaroff
1. INTRODUCTION

• Gesture: “A movement of part of the body, especially a hand or the head, to express an idea or meaning.”
Examples of specific applications

- Aircraft communication gesture recognition

- Recognition of socio-cultural gestures

INTRODUCTION
Gesture spotting and recognition

- Gesture spotting

  INPUT
  Continuous input

  Spotting

  OUTPUT
  Start and End Points of Gestures

- Gesture recognition

  INPUT
  Segmented Gesture

  Recognition

  OUTPUT
  Class label
Problem Definition

- Problem:
  - Given a training set of multi-modal videos with multiple examples of all gestures in a gesture vocabulary, design a framework capable of automatically and accurately spotting and classifying gestures present in a set of test videos.
Related work

• Generative Graphical Models
  • Hidden Markov Model
    • Starner et al., 1997

• Discriminative Graphical Models
  • Hidden Conditional Random Fields
    • Song et al., 2011

• Other Discriminative Models
  • Support Vector Machines
    • Huang et al., 2009
  • Tree Ensembles
    • (ours)
2. RANDOM FORESTS

Training set

Bootstrap sample

Parent node

Leaf nodes

Random feature selection

Image reprinted from:
Information Gain

- Given a training set $S$ of data points and their labels, trees are built to optimize a certain function, e.g. Information gain ($I$)

\[
I_j = H(S_j) - \sum_{k \in \{L,R\}} \frac{|S_j^k|}{|S|} H(S_j^k)
\]

\[
H(S) = -\sum_{c \in C} p_c \log(p_c)
\]

$S$: set of training points
$p_c$: probability of a sample being in class $c$
Random forests have been used to good effect in:

- human pose recognition
  - Shotton et al., 2013

- object segmentation
  - Schroff et al., 2008

- image classification
  - Bosch et al., 2007
3. FRAMEWORK
3. FRAMEWORK
3. FRAMEWORK
3. FRAMEWORK
3. FRAMEWORK
Training: Input

Input

...
Training: Feature Extraction

Input

Feature Extraction

- Normalized positional coordinates of joints
- Rotation angles of joints
- Positional and angular velocity of joints

Joint-based features

- Histogram of Oriented Gradients computed on boxes centered around the left and right hands

Image-based features

Feature Fusion

- Feature are fused to create a combined feature descriptor
Training: Gesture Representation

1. Input feature matrix of a gesture, and smooth using a moving average filter

2. Divide into equal-sized temporal segments

3. Select median features and concatenate

Joint-based features

Image-based features

Feature Extraction

Gesture Representation
Training: Initial Random Forest Training

Input

Feature Extraction

Gesture Representation

1. Input feature matrix of a gesture

2. Divide into equal-sized temporal segments

3. Select median features and concatenate

Random Forest Training

Input to classifier:
Set of examples of all gestures in the training set, as well as randomly chosen examples of non-gestural intervals
Training: Test on continuous input of training set

- **Continuous Input**
  - Joint-based features
  - Image-based features

- **Feature Extraction**
  - Joint-based features
  - Image-based features

- **Multi-scale sliding window classification**
  - Frame: $F_t$
  - Frame: $F_{t+1}$
  - Scale: $S_0$
  - Scale: $S_n$

- **Trained Random Forest**

- **Output**
  - Gesture 18
  - New Buffer start
Training: Collection of hard negatives

1. Input feature matrix of a gesture

2. Divide into equal-sized temporal segments

3. Select median features and concatenate

Collect Hard Negatives
- Run classifier on continuous input of the training set
- Collect all misclassified instances and add them to the original training set and retrain the classifier

Random Forest Training

Joint-based features

Image-based features

Feature Extraction

Gesture Representation

Input feature frames

Random Forest Training

$P(y_i) = \sum P(y_i | f_i)$
Input

Feature Extraction

Joint-based features

Image-based features

Gesture Representation

1. Input feature matrix of a gesture after temporal smoothing

2. Divide into equal-sized temporal segments

3. Select median features and concatenate

Collect Hard Negatives

- Run classifier on continuous input of the training set
- Collect all misclassified instances and add them to the original training set and retrain the classifier

Random Forest Training

Until number of hard negatives fall below threshold

Training: Iterative Random Forest Training
Testing

Continuous Input

Feature Extraction
- Joint-based features
- Image-based features

Multi-scale sliding window classification
- Buffer start
- Frame: $F_t$
- Frame: $F_{t+1}$
- Scale: $S_0$
- Scale: $S_n$

Output
- New Buffer start
- Gesture 18
4. DATASETS

NATOPS

• Naval Air Training and Operating Procedures Standardization gestures
• 24 aircraft handling signals, performed by 20 subjects, 20 times
• Dataset includes RGB color images, depth maps, mask images and skeletal information
CHALEARN

• 20 unique Italian cultural and anthropological signs
• Development data: 7,754 labelled gestures
• Validation data: 3,363 gestures
• Test data: 2,742 gestures
• Dataset includes RGB color images, depth maps, mask images and skeletal information
5. EXPERIMENTAL RESULTS

• NATOPS
  • Classification only

• CHALEARN
  • Classification and Segmentation
NATOPS classification accuracy averaged over all subjects and gestures: 87.35%
Figure: Pairs of similar gestures
Confusion Matrix for pairs of similar gestures

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Average Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>*HMM</td>
<td>77.67%</td>
</tr>
<tr>
<td>*HCRF</td>
<td>87.0%</td>
</tr>
<tr>
<td>Random Forest (ours)</td>
<td>88.1%</td>
</tr>
</tbody>
</table>

* As shown in:
**EXPERIMENTAL RESULTS**

CHALEARN classification accuracy averaged over all subjects and gestures: **88.91%**
\[ J_{s,n} = \frac{A_{s,n} \cap B_{s,n}}{A_{s,n} \cup B_{s,n}} \]

Here, \( A_{s,n} \) is the vector describing the ground truth indices of gesture \( n \) at sequence \( s \), whereas \( B_{s,n} \) is the vector describing the predicted indices of gesture \( n \) at sequence \( s \).

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Jaccard score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Neural Network</td>
<td>0.84</td>
</tr>
<tr>
<td>Random Forest (our)</td>
<td>0.68</td>
</tr>
<tr>
<td>Competition baseline</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Table: CHALEARN 2014 Competition results
6. DISCUSSION

- We have presented a simple and efficient random forest framework.
- Reliable classifier that generalizes well
- The task of simultaneously detecting and classifying gestures is a more difficult challenge than solely classifying correctly segmented gestures.
THANK YOU!