Weakly Supervised Pain Localization using Multiple Instance Learning

Karan Sikka¹, Abhinav Dhall², and Marian S. Bartlett¹

¹Machine Perception Lab
University of California, San Diego

²Australian National University

Pain/No-pain
(where)

FG 2013
Motivation

• Pain monitoring critical for clinical applications.

• Spontaneous expression.
  – Classification difficult compared to posed expressions (CK+ dataset).

• Pain has high variability (expression, perception, location and duration)
  – Efficient prediction algorithms.
Problem Definition

• Subjects undergoing shoulder pain in videos.
  – UNBC MC-Master Pain Dataset*.
  – Ranging from 60-600 frames.

• Classifying and localizing pain in videos.
  – Sequence level ground-truth labels.

Lucey et. al., PAINFUL DATA: The UNBC-McMaster Shoulder Pain Expression Archive Database, FG’11
Challenges

• Ambiguity introduced by sequence level labels.
  – Time points and duration of pain unknown apriori.

• Incorporating dynamics/temporal information.

• Temporal segmentation is hard in itself.
Previous Approaches

‘Classical’ Fixed Length Features

\[ f \sim \text{Features for each frame} \]

1. BoW using Local Features.
2. Active Appearance (AAM) based Texture and Shape features.
3. Gabor
4. LBP

Pooling

- Avg or max
- Fixed Length Features

*Laptev et. al., Learning Realistic Human Actions From Movies, CVPR’08*
Previous Approaches
‘Classical’ Fixed Length Features

• Most common approach.

• Works well when action spans whole videos
  – Facial expression classification (CK+ dataset).
  – Action classification (KTH dataset).

• Pooling features will not work well for long videos.
  – Kills the signal of interest.
  – Localized instead of global approaches required.
Previous Approaches
‘Classical’ Fixed Length Features

• Most intuitive approach.

• Works well when action spans whole videos
  – Facial expression classification (CK+ dataset).
  – Action classification (KTH and hollywood dataset).

• Pooling features doesn’t work well in all cases.
  – Kills the signal of interest.
  – Localized instead of global approaches required.
Previous Approaches
Frame Level Features

Label = 1

- AAM \( \Rightarrow \) Clustering \( \Rightarrow \) SVM
- Assign labels of sequence to each frames.
- Test
  - \( \text{Score(} \text{video} \text{)} = \text{Avg(} \text{Output(} \text{frames} \text{)}) \).
Previous Approaches
Frame Level Features

AAM ➔ DCT ➔ SVM

Assign labels of sequence to each frames.

Test
• \( \text{Score( video )} = \text{Avg( Output( frames )).} \)

Label = 1

\[ \begin{align*}
&f \quad f \\
&f \\
\end{align*} \]

Lucey et al., ICASP’08.
1. Assigning sequence label to each frame.
   - Label Ambiguity.
   - ML methods like SVM not robust to outliers.

• **Solution:** Multiple Instance Learning (MIL).
  - Efficiently handle weakly labeled data.

![MIL](image)
Previous Approaches

Limitations

2. Treated videos as individual frames.
   - Lack of temporal information.
   - Vital for pain classification.

• **Solution:** Represent sequences as sets of frames: “Multiple segments”
Multiple Segment Representation

- Extracting at multiple scales and can overlap (no-restriction).
- Allow multiple hypothesis.
Multiple Segments based Multiple Instance Learning (MS-MIL)

**Train**

1. Video → Feature Extraction → Multiple Segment Representation → Train Using MIL

**Test**

1. Test Video → Classification with MIL model → Soft-Localization
Feature Extraction - Frames

BoW model

2 pixel SIFT (5 scales) LLC* Max-pooling within spatial pyramids

*LLC- Locality constrained Linear Encoding

Sikka et al., ECCV’12.
Multiple Segment Representation

Pooling

Max pooling summarizes sparse signals better than average pooling.
• Training videos as Bags

MS-MIL

Train using MilBoost (Viola’06)
• MIL has a joint optimization framework.
MS-MIL Localization

- Hamming window
MS-MIL Localization

- \( P(y=1|\text{frame}_i) \)
- Max over weighted segment scores
Experiments

• Leave one subject out protocol.

• 147 videos from 23 subjects.

• Observer Pain Intensity as ground-truth labels.
  – Binarized.

• Faces aligned using provided AAM features.

• Total classification rate at Equal Error Rate.
## Classification Performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>#Subjects</th>
<th>#Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS-MIL</td>
<td>83.7</td>
<td>23</td>
<td>147</td>
</tr>
<tr>
<td>Lucey et.al</td>
<td>80.99</td>
<td>20</td>
<td>142</td>
</tr>
<tr>
<td>Ashraf et.al</td>
<td>68.31</td>
<td>20</td>
<td>142</td>
</tr>
<tr>
<td>(as shown in Lucey et.al)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML-SVM&lt;sub&gt;max&lt;/sub&gt;</td>
<td>70.75</td>
<td>23</td>
<td>147</td>
</tr>
<tr>
<td>ML-SVM&lt;sub&gt;avg&lt;/sub&gt;</td>
<td>76.19</td>
<td>23</td>
<td>147</td>
</tr>
</tbody>
</table>

- Shows gains over previous methods.
# Classification Performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>#Subjects</th>
<th>#Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS-MIL</td>
<td>83.7</td>
<td>23</td>
<td>147</td>
</tr>
<tr>
<td>Lucey et.al</td>
<td>80.99</td>
<td>20</td>
<td>142</td>
</tr>
<tr>
<td>Ashraf et.al</td>
<td>68.31</td>
<td>20</td>
<td>142</td>
</tr>
<tr>
<td>(as shown in Lucey et.al)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS-SVM_{avg}</td>
<td>70.75</td>
<td>23</td>
<td>147</td>
</tr>
<tr>
<td>MS-SVM_{max}</td>
<td>76.19</td>
<td>23</td>
<td>147</td>
</tr>
</tbody>
</table>

- **MS-SVM**
  - Each segment assigned the label of the video
  - SVM + score combining rule (max and avg).
## Classification Performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>#Subjects</th>
<th>#Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS-MIL</td>
<td>83.7</td>
<td>23</td>
<td>147</td>
</tr>
<tr>
<td>Lucey et.al</td>
<td>80.99</td>
<td>20</td>
<td>142</td>
</tr>
<tr>
<td>Ashraf et.al (as shown in Lucey et.al)</td>
<td>68.31</td>
<td>20</td>
<td>142</td>
</tr>
<tr>
<td>MS-SVM(_{avg})</td>
<td>70.75</td>
<td>23</td>
<td>147</td>
</tr>
<tr>
<td>MS-SVM(_{max})</td>
<td>76.19</td>
<td>23</td>
<td>147</td>
</tr>
</tbody>
</table>

- MS-MIL performs better than its traditional ML counterparts.
Localization Performance

• Compared per-frame probabilities predicted by MS-MIL with human expert pain labels.

• PSPI computes pain intensity based on FACS.
  – PSPI sums intensities of 4 Action Units.
  – Prkachin & Solomon’08.

• Normalized PSPI to 0-1.
Localization Performance
Conclusion

• Proposed Novel approach to problem of classifying and localizing pain.

• Highlighted limitations of previous approaches and motivations for current algorithm.

• Compared MS-MIL with
  – Previous Approaches
  – Traditional ML counterparts.

• Localization compared with ground-truth index (PSPI).
Questions?

Karan Sikka
Abhinav Dhall
Dr. Marian S. Bartlett

Machine Perception Lab, UCSD

Thanks