Variable-state Latent Conditional Random Fields for Facial Expression Recognition and Action Unit Detection

Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic and Maja Pantic

May 6, 2015
Motivation

- Human-computer interaction
- Patients in intensive care units
- Pain or stress detection
- Psychological studies
- Security application
- Interactive gaming
Facial Actions: Level of description

Figure 1: Action Units (AU) defined by Facial Action Coding System (FACS)

Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic and Maja Pantic
VSL-CRF for Emotion Recognition and Action Unit Detection
7 Basic Emotions: Anger, Contempt, Fear, Disgust, Happiness, Sadness and Surprise

Figure 2: AU 4, 5, 7, 23

Figure 3: AU 1, 2, 4, 5, 7, 20, 26

Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic and Maja Pantic
VSL-CRF for Emotion Recognition and Action Unit Detection
Our Goal

- Sequence classification of
  1. Action units
  2. Facial expressions
- Modeling of dynamics (temporal information)
- Modeling of ordinal constraints
Existing Work

- **Frame Based**
  - [BLFM03]
  - SVM or AdaBoost [VPP05]
  - CMIB [SGM05]
  - STM [CTC13]
  - ...
  - ...
  - ...

- **Sequence Based**
  - HMM (single HMM for each emotion) [SC09]
  - CRF (single linear chain crf for each emotion) [JHA11, CLL09, SLS+07]
  - HCRF [WQM+06]
  - HCORF [KP10a]
Hidden Conditional Random Fields

Score function:
\[ s(x, h_r, h_s) = \sum_{r \in V} v^T \Psi_r^{(V)}(x, h_r) + \sum_{e=(r,s) \in E} u^T \Psi_e^{(E)}(x, h_r, h_s) \]
Hidden Conditional Random Fields

Score function:
\[
s(x, h_r, h_s) = \sum_{r \in V} v^T \Psi_r^{(V)}(x, h_r) + \sum_{e = (r, s) \in E} u^T \Psi_e^{(E)}(x, h_r, h_s)
\]

(a) Logistic regression: \(\Psi_r^{(V)}(x, h_r) \rightarrow \log(\frac{\exp(f_n(x, c))}{\sum_{l=1}^C \exp(f_n(x, l))})\), where: \(f_n(x, c) = \beta_c^T \cdot [1, x]\)
Hidden Conditional Random Fields

Score function:
\[ s(x, h_r, h_s) = \sum_{r \in V} v^T \Psi_r^{(V)}(x, h_r) + \sum_{e = (r, s) \in E} u^T \Psi_e^{(E)}(x, h_r, h_s) \]

(a) Logistic regression:
\[ \Psi_r^{(V)}(x, h_r) \rightarrow \log\left(\frac{\exp(f^n(x, c))}{\sum_{l=1}^{C} \exp(f^n(x, l))}\right), \quad \text{where: } f^n(x, c) = \beta_c^T \cdot [1, x] \]

(b) Ordinal regression:
\[ \Psi_r^{(V)}(x, h_r) \rightarrow \log\left(\Phi\left(\frac{b_c - f^o(x)}{\sigma}\right) - \Phi\left(\frac{b_{c-1} - f^o(x)}{\sigma}\right)\right) \] \[ [KP10b, KP10a] \]
Ordinal vs Logistic Regression

\[ f(x) = Ax \]

\[ f_c(x) = V_c x , \ c = 1, ..., R \]

\[ h = c \text{ iff } f(x) \in [b_{c-1}, b_c), \ b_{c-1} < b_c \]

\[ h = \arg\max_c f_c(x) \]

Figure 4: Ordinal regression: (1) fewer parameters, (2) smaller error due to closest neighboring intervals
Hidden Conditional Random Fields

Inputs: (observed features per frame)
\[ x_i = \{ \vec{x}_{i1}, \vec{x}_{i2}, \vec{x}_{i3}, \ldots, \vec{x}_{it_i} \} \]

Outputs: (Emotion or AU label)
\[ y_i \in \mathcal{Y} \]

Marginal conditional probability of HCRFs
\[
P(y|x) = \sum_h P(y, h|x) = \frac{\sum_h \exp(s(y, x, h))}{Z(x)}
\]
where: \[ Z(x) = \sum_{k \in \mathcal{Y}} \sum_{h \in \mathcal{H}} \exp(s(k, x, h)) \]
Proposed Variable-State-Latent-CRF (VSL-CRF)

Generalization of H-CRF/H-CORF models that allow their latent states to be modeled using either nominal or ordinal potentials within each sequence.

VSL-CRF score function:

\[
s(y, x, h, \nu; \Omega) = \begin{cases} 
  \sum_{k=1}^{K} I(k = y) \cdot s(x, h; \theta^n_y), & \text{if } \nu_y = 0 \quad \text{(nominal)} \\
  \sum_{k=1}^{K} I(k = y) \cdot s(x, h; \theta^o_y), & \text{if } \nu_y = 1 \quad \text{(ordinal)}
\end{cases}
\]

Marginal conditional probability of VSL-CRFs

\[
P(y|x, \Omega) = \frac{\max(\sum_{h} \exp(s(y, x, h, \nu, \Omega)))}{Z(x)}
\]

\[
Z(x) = \sum_{k} Z_k(x) = \sum_{k} \max(\sum_{h} \exp(s(k, x, h, \nu))) \quad \text{and } \Omega = \{\theta^n_k, \theta^o_k\}_{k=1}^{K}
\]

Prediction: \( y^* = \arg \max_{y} P(y|x^*) \)
Learning and Inference

\[
\arg\min_{\Omega} \sum_i - \log P(y_i|x_i; \Omega) + \lambda_n^{(o)} \| \theta^{n^{(o)}}_{k=1..K} \|^2
\]

\[
P(y_i|x_i; \Omega) = \frac{\max( \sum_{h} \exp(s(y_i, x_i, h, \nu, \Omega)))}{Z(x)}
\]

\[
\Omega = \{ \theta^{n}_{k}, \theta^{o}_{k} \}_{k=1}^{K}
\]

\[
\theta^{o} = \{ b_1, \ldots, b_{r-1}, u, \beta^0 \}
\]

\[
\theta^{n} = \{ \beta^1, \ldots, \beta^{r-1} u \}
\]

\[\longrightarrow\] Quasi Newton Limited-memory BFGS subgradients for optimization

Prediction: \[ y^* = \arg\max_y P(y|x^*) \]
Experimental Procedure

Facial points extracted from target images

Align the facial points to the mean faces

Reduced feature size, retaining 97% of energy (~20 D)

Generating balanced training sets with positive and negative samples

SVM, HCRF, HCRF, VSL-CRF

Figure 5 : Pre-processing and training
## Experiments on Emotion Classification

### Method (7 emotions on AFEW dataset)

<table>
<thead>
<tr>
<th>Method</th>
<th>F-1 [%]</th>
<th>Accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>EmotiW [DGJ^{+13}]</td>
<td>—</td>
<td>27.3</td>
</tr>
<tr>
<td>SVM-SB (RBF)</td>
<td>26.3</td>
<td>31.2</td>
</tr>
<tr>
<td>H-CRF</td>
<td>19.7</td>
<td>22.6</td>
</tr>
<tr>
<td>H-CORF</td>
<td>22.4</td>
<td>27.4</td>
</tr>
<tr>
<td><strong>VSL-CRF</strong></td>
<td><strong>28.1</strong></td>
<td><strong>32.2</strong></td>
</tr>
</tbody>
</table>

### Method (6 basic emotions on CK+ dataset)

<table>
<thead>
<tr>
<th>Method</th>
<th>F-1 [%]</th>
<th>Accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bartlett et al. [BLF^{+05}]</td>
<td>—</td>
<td>87.5</td>
</tr>
<tr>
<td>PO-HCRF9 [CLL09]</td>
<td>—</td>
<td>92.9</td>
</tr>
<tr>
<td>TMS [JHA11]</td>
<td>—</td>
<td>91.2</td>
</tr>
<tr>
<td>SVM-SB (RBF)</td>
<td>89.8</td>
<td>91.3</td>
</tr>
<tr>
<td>H-CRF</td>
<td>91.2</td>
<td>93.2</td>
</tr>
<tr>
<td>H-CORF</td>
<td>90.4</td>
<td>92.3</td>
</tr>
<tr>
<td><strong>VSL-CRF</strong></td>
<td><strong>94.5</strong></td>
<td><strong>96.7</strong></td>
</tr>
</tbody>
</table>

### Method (7 emotions on CK+ dataset)

<table>
<thead>
<tr>
<th>Method</th>
<th>F-1 [%]</th>
<th>Accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITBN [WWJ13]</td>
<td>—</td>
<td>86.3</td>
</tr>
<tr>
<td>STM-ExpLet [LSWC13]</td>
<td>—</td>
<td>94.2</td>
</tr>
<tr>
<td>SVM-SB (RBF)</td>
<td>89.5</td>
<td>91.1</td>
</tr>
<tr>
<td>H-CRF</td>
<td>85.0</td>
<td>89.1</td>
</tr>
<tr>
<td>H-CORF</td>
<td>91.7</td>
<td>93.5</td>
</tr>
<tr>
<td><strong>VSL-CRF</strong></td>
<td><strong>93.9</strong></td>
<td><strong>95.8</strong></td>
</tr>
</tbody>
</table>

Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic and Maja Pantic

VSL-CRF for Emotion Recognition and Action Unit Detection
Experiments on AU Sequence Classification

- 5 upper face AUs

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68.1</td>
<td>69.4</td>
<td>70.0</td>
<td>70.0</td>
<td>72.9</td>
<td>85.6</td>
<td>81.8</td>
<td>87.4</td>
<td>87.5</td>
<td>91.7</td>
</tr>
<tr>
<td>2</td>
<td>65.5</td>
<td>69.7</td>
<td>70.5</td>
<td>73.5</td>
<td>85.9</td>
<td>79.4</td>
<td>79.6</td>
<td>87.4</td>
<td>78.5</td>
<td>87.5</td>
</tr>
<tr>
<td>4</td>
<td>43.3</td>
<td>58.8</td>
<td>61.2</td>
<td>61.3</td>
<td>70.0</td>
<td>81.2</td>
<td>58.8</td>
<td>57.0</td>
<td>63.8</td>
<td>69.5</td>
</tr>
<tr>
<td>6</td>
<td>71.6</td>
<td>76.9</td>
<td>63.7</td>
<td>70.1</td>
<td>79.1</td>
<td>87.2</td>
<td>63.9</td>
<td>67.0</td>
<td>63.3</td>
<td>78.7</td>
</tr>
<tr>
<td>7</td>
<td>66.2</td>
<td>60.3</td>
<td>55.8</td>
<td>58.1</td>
<td>74.3</td>
<td>80.9</td>
<td>77.9</td>
<td>66.1</td>
<td>86.7</td>
<td>87.7</td>
</tr>
<tr>
<td>AVG</td>
<td>62.9</td>
<td>67.0</td>
<td>64.2</td>
<td>66.6</td>
<td>76.4</td>
<td>82.6</td>
<td>72.4</td>
<td>73.0</td>
<td>76.0</td>
<td>83.0</td>
</tr>
</tbody>
</table>

Table 1: AU detection from the GEMEP-FERA and MMI datasets. The numbers show the F-1 scores in % for each method.
Figure 6: Number of training sequences from AU7 - active (upper) and AU7 - not active (lower) being assigned nominal/ordinal
Conclusion

- We proposed a novel Latent Conditional Random Field model for dynamic facial expression recognition and AU detection.
- We showed that the proposed model better discriminates between different facial expressions than the existing models that restrict their latent states to have the same and pre-defined structure for all classes (nominal or ordinal).
- We showed on four facial expression datasets that the proposed model outperforms other traditional classification models.
Thank You! / Questions?
Recognizing facial expression: machine learning and application to spontaneous behavior.
In *CVPRW*, 2005.

M. S. Bartlett, G. Littlewort, I. Fasel, and J. Movellan.
Real time face detection and facial expression recognition: Development and applications to human computer interaction.

K. Chang, T. Liu, and S. Lai.
Learning partially-observed hidden conditional random fields for facial expression recognition.

W. Chu, F. Torre, and J. F Cohn.
Selective transfer machine for personalized facial action unit detection.


S. Jain, Changbo Hu, and J.K. Aggarwal.
Facial expression recognition with temporal modeling of shapes.

B. Jiang, M. Valstar, and M. Pantic.
Action unit detection using sparse appearance descriptors in space-time video volumes.
In *Automatic Face & Gesture Recognition and Workshops (FG 2011)*, 2011.
M. Kim and V. Pavlovic.
Hidden conditional ordinal random fields for sequence classification.

M. Kim and V. Pavlovic.
Structured output ordinal regression for dynamic facial emotion intensity prediction.

M. Liu, S. Shan, R. Wang, and X. Chen.
Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition.

L. Shang and K. Chan.
Nonparametric discriminant hmm and application to facial expression recognition.

C. Shan, S. Gong, and P. W McOwan.
Conditional mutual information based boosting for facial expression recognition.
In *BMVC*, 2005.

Authentic facial expression analysis.

M. F. Valstar, I. Patras, and M. Pantic.
Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data.
In *CVPRW*, 2005.
Hidden conditional random fields for gesture recognition.

Z. Wang, S. Wang, and Q. Ji.
Capturing complex spatio-temporal relations among facial muscles for facial expression recognition.