Spontaneous Facial Expression Analysis Based on Temperature Changes and Head Motions

Peng Liu and Lijun Yin
State University of New York-at Binghamton
Motivation

Fear
disgust
Happy
Motivation

Fear
disgust
Happy
Motivation

Fear

disgust

Happy
Prior works

Video clips,
Multiple instance learning [K. Sikka et.al. FG 2013]

facial expression interaction of infants with their mothers, specific Action Units [N. Zaker et.al. FG 2013]

head nod, head shake and hand wag, Hidden Conditional Random Field [K. Bousmalis et.al. FG 2011]
Prior works

Face recognition [P. Buddharaju et.al. CVPR 2009]

Detection of Deceit [P. Tsiamyrtzis et.al. IJCV 2007]

Eye localization [S. Wang et.al. Pattern Recognition 2013]
Overview of our approach
Overview of our approach

Training video
Overview of our approach

Training video
Overview of our approach
Overview of our approach

Training video

Learning thermal word
Overview of our approach

Training video

Learning thermal word
Overview of our approach

Training video

Learning thermal word
Overview of our approach

Training video

Learning thermal word

Learning motion word
Overview of our approach

Training video → Learning thermal word → Codebook

Training video → Learning motion word → Codebook
Overview of our approach
Overview of our approach

Training video

Learning thermal word

Learning motion word

Test video

Codebook
Overview of our approach

Training video

Learning thermal word

Learning motion word

Test video

Codebook
Face region alignment and warping
Face region alignment and warping
Face region alignment and warping

\[ E(w) = \sum_p \min(||f(T_Q(p)) - f(T_R(p' + w(p)))||_1, t) + \]
\[ \sum_p \eta(|u(p)| + |v(p)|) + \]
\[ \sum_{(p,q) \in N} \min(\alpha|u(p) - u(q)|, d) + \min(\alpha|v(p) - v(q)|, d) \]

[\text{C. Liu et.al. Trans. PAMI 2011}]
Max pooling the most distinguished cubic

\[ y_j = \max\{|x_{ij}|, |x_{2j}|, \ldots, |x_{Nj}|\} \]
Visualization of SIFT flow and motion video cubic

\[ \hat{y}_j = \text{mean}\{\hat{x}_{1j}, \hat{x}_{2j}, \ldots, \hat{x}_{Nj}\} \]
Thermal video descriptor

Based on the codebook, the test video is represented by the thermal video word and SIFT flow motion video word.

\[
H^T = \left( \frac{N_{\Delta t_1}}{N_T}, \frac{N_{\Delta t_2}}{N_T}, \ldots, \frac{N_{\Delta t_n}}{N_T} \right)
\]

\[
H^M = \left( \frac{N_{\Delta v_1}}{N_M}, \frac{N_{\Delta v_2}}{N_M}, \ldots, \frac{N_{\Delta v_n}}{N_M} \right)
\]

\[
H = \{H^T, H^M\}
\]

\(H^T\) is the histogram of thermal video words

\(H^M\) is the histogram of motion video words

\(N_T\) is the number of thermal video cubic extracted from thermal video clips.

\(N_M\) is the number of motion video cubic extracted from motion video clips.
Experiments and Evaluation

Neutral thermal
Neutral texture
Expression thermal
Expression texture

Embarrassment  Upset  Disgust  Fear  Pain  Sadness  Surprise  Happiness
Evaluation of the thermal and motion descriptor

(a) The confusion matrix of utilizing both thermal and motion video words.

(b) The confusion matrix of just utilizing thermal video words.

(c) The confusion matrix of just utilizing motion video words.
Evaluation of the max pooling method and comparison

<table>
<thead>
<tr>
<th>Class</th>
<th>Max pooling</th>
<th>Forehead</th>
<th>Left cheek</th>
<th>Right cheek</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embarrassment</td>
<td>0.96</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Surprise</td>
<td>1</td>
<td>0.14</td>
<td>0.33</td>
<td>0.15</td>
</tr>
<tr>
<td>Happiness</td>
<td>0.88</td>
<td>0.88</td>
<td>0.84</td>
<td>0.65</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.95</td>
<td>0.9</td>
<td>0.78</td>
<td>0.73</td>
</tr>
<tr>
<td>Disgust</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Anger or upset</td>
<td>0.91</td>
<td>0.82</td>
<td>0.82</td>
<td>0.65</td>
</tr>
<tr>
<td>Pain</td>
<td>0.78</td>
<td>0.63</td>
<td>0.5</td>
<td>0.42</td>
</tr>
<tr>
<td>Fear or nervous</td>
<td>0.79</td>
<td>0.86</td>
<td>0.89</td>
<td>0.9</td>
</tr>
<tr>
<td><strong>Weighted average</strong></td>
<td><strong>0.91</strong></td>
<td><strong>0.77</strong></td>
<td><strong>0.76</strong></td>
<td><strong>0.68</strong></td>
</tr>
</tbody>
</table>
Comparison to traditional descriptor on thermal video

<table>
<thead>
<tr>
<th>Class</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0.91</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>HOG[Laptev, CVPR’08]</td>
<td>0.82</td>
<td>0.75</td>
<td>0.82</td>
<td>0.76</td>
</tr>
<tr>
<td>HOF[Laptev, CVPR’08]</td>
<td>0.80</td>
<td>0.83</td>
<td>0.78</td>
<td>0.76</td>
</tr>
<tr>
<td>HOG+HOF[Laptev, CVPR’08]</td>
<td>0.88</td>
<td>0.88</td>
<td>0.85</td>
<td>0.83</td>
</tr>
<tr>
<td>Cuboids[Doll´ar, PETS’05]</td>
<td>0.32</td>
<td>0.31</td>
<td>0.29</td>
<td>0.30</td>
</tr>
</tbody>
</table>
Comparison to prior method

Comparison on USTC-NVIE database

Comparison on our new database
### Comparison on USTC-NVIE database with two modalities

<table>
<thead>
<tr>
<th>Approach</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours on the thermal</td>
<td>0.72</td>
<td>0.70</td>
<td>0.71</td>
<td>0.68</td>
</tr>
<tr>
<td>Ours on the texture</td>
<td>0.63</td>
<td>0.61</td>
<td>0.60</td>
<td>0.59</td>
</tr>
<tr>
<td>Method in [Wang et. al. <em>IEEE Trans. on AC ’13</em>] on the thermal</td>
<td>0.64</td>
<td>0.61</td>
<td>0.60</td>
<td>0.59</td>
</tr>
</tbody>
</table>

### Comparison on our new database with two modalities

<table>
<thead>
<tr>
<th>Approach</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours on the thermal</td>
<td>0.91</td>
<td>0.88</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>Ours on the texture</td>
<td>0.73</td>
<td>0.72</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>GW-based on the texture [Zhang et. al. <em>FG’13</em>]</td>
<td>0.66</td>
<td>0.64</td>
<td>0.62</td>
<td>0.60</td>
</tr>
</tbody>
</table>
Conclusion

• We presented a new infrared thermal video descriptor which can compactly describe a spatio-temporal-temperature information.
• We demonstrated through many experiments that the new descriptor can be a very useful tool to spontaneous facial expression classification.
Future work

• We will utilize the spatial and temporal structural information for improving the classification accuracy.

• We will combine thermal data, texture data and physiology data to further improve the classification performance.
Acknowledgment

• This material is based on the work supported in part by the NSF under grant CNS-1205664 and the SUNY IITG.

• We thank Dr. Jeffrey Cohn and Dr. Qiang Ji and their groups for collaboration on data collection and annotation.
Thanks!