Recognition of Transitional Action for Short-Term Action Prediction using Discriminative Temporal CNN Feature

Hirokatsu Kataoka1, Yudai Miyashita2, Masaki Hayashi3,4, Kenji Iwata2, Yutaka Satoh1
1AIST, Japan  2Tokyo Denki Univ., Japan  3Liquid Inc., Japan  4Keio Univ., Japan

Motivation

- Goal
  - Accurate “short-term action prediction”
- Problems in action analysis
  - Recognition is NOT predictable
  - Prediction is NOT reliable
- Applications
  - Active safety, autonomous driving
  - Robots

Framework

- IDEA
  - Action-class while an action is transitive (see below)
- Contributions
  - Definition of transitional action for short prediction
  - Subtle Motion Descriptor (SMD) to classify TA1 and NA2

<table>
<thead>
<tr>
<th>Framework</th>
<th>Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action Recognition</td>
<td>$f(F_{1:t}) \rightarrow A_t$</td>
</tr>
<tr>
<td>Early Action Recognition</td>
<td>$f(F_{1:t-L}) \rightarrow A_{t-L}$</td>
</tr>
<tr>
<td>Action Prediction</td>
<td>$f(F_{1:t}^A) \rightarrow A_{t+L}$</td>
</tr>
<tr>
<td>Transitional Action Recognition</td>
<td>$f(F_{1:t}^{TA}) \rightarrow A_{t+L}$</td>
</tr>
</tbody>
</table>

Transitional Action?

- Transitional action is defined as the transition class between actions (see below)
  - TA: “walking – put down item” between NA: “walking” and NA: “put down item”
  - The TA classes and NA classes are partially overlapped each other
    - But no more than 5 frames overlap

Subtle Motion Descriptor (SMD)

- SMD is used to identify the sensitive differences between TA and NA
  - The SMD is based on pooled time series (PoT) [Ryoo+ CVPR15]
  - Temporal pooling with zero-around elements ($x_{V0-}^{\text{SMD}}, x_{V0+}^{\text{SMD}}$)

$$
\begin{align*}
    & x_{V0+}^{\text{SMD}} = \sum_{i=t-L}^{t} h_{i}^{+} (t), \quad x_{V0-}^{\text{SMD}} = \sum_{i=t-L}^{t} h_{i}^{-} (t) \\
    & x_{V0+}^{\text{SMD}} = \sum_{i=t-L}^{t} h_{i}^{0} (t), \quad x_{V0-}^{\text{SMD}} = \sum_{i=t-L}^{t} h_{i}^{-} (t)
\end{align*}
$$

$$
\begin{align*}
    & h_{i}^{+} (t) = |\Delta V_{i} | \quad (\Delta V_{i} > TH) \\
    & h_{i}^{-} (t) = |\Delta V_{i} | \quad (0 < \Delta V_{i} < TH) \\
    & h_{i}^{0} (t) = |\Delta V_{i} | \quad (-TH < \Delta V_{i} < 0) \\
    & h_{i}^{-} (t) = |\Delta V_{i} | \quad (\Delta V_{i} < -TH)
\end{align*}
$$

Experiments

- NTSEL (Traffic; TA1, NA3), UTKinect-Action (Indoor; TA8, NA10), Watch-n-Patch (Indoor; TA10, NA10)
  - Threshold? – good for 0.03 ~ 0.05
  - Frame accumulation? – 10-frame for state-of-the-art, 3-frame for faster prediction
  - FC layer – Layer 6 is better

<table>
<thead>
<tr>
<th>Framework</th>
<th>% on NTSEL 10 frm / 3 frm</th>
<th>% on UT 10 frm / 3 frm</th>
<th>% on WnP 10 frm / 3 frm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>99.18 / 92.00</td>
<td>85.78 / 77.15</td>
<td>99.19 / 92.00</td>
</tr>
<tr>
<td>PoT, CVPR15</td>
<td>97.00 / 92.00</td>
<td>77.15 / 65.46</td>
<td>97.00 / 92.00</td>
</tr>
</tbody>
</table>

Conclusion

- Definition of transitional action (TA) for short-term prediction
  - The TA allows us to produce earlier and accurate prediction
- Proposal of subtle motion descriptor (SMD)
  - Outstanding results with 3-frame feature accumulation

Reference


http://hirokatsukataoka.net/