ChaLearn Looking at People Challenge 2014: Dataset and Results

http://gesture.chalearn.org/

Sergio Escalera, UB & CVC & ChaLearn
Xavier Baró, UOC & CVC
Jordi Gonzàlez, UAB & CVC
Miguel Ángel Bautista, UB & CVC
Meysam Madadi, UB & CVC
Miguel Reyes, UB & CVC
Víctor Ponce-López, UB & CVC & UOC
Hugo J. Escalante, INAOE & ChaLearn
Jamie Shotton, Microsoft Research
Isabelle Guyon, ChaLearn
Challenge on pose recovery, action/interaction, multi-modal gesture recognition

• **Track 1: Human Pose Recovery:** More than 8,000 frames of continuous RGB sequences are recorded and labeled with the objective of performing human pose recovery by means of recognizing more than 120,000 human limbs of different people.

• **Track 2: Action/Interaction Recognition:** 235 performances of 11 action/interaction categories are recorded and manually labeled in continuous RGB sequences of different people performing natural isolated and collaborative behaviors.

• **Track 3: Gesture Recognition:** The gestures are drawn from a vocabulary of Italian sign gesture categories. The emphasis of this third track is on multi-modal automatic learning of a set of 20 gestures performed by several different users, with the aim of performing user independent continuous gesture spotting.

http://gesture.chalearn.org/
• Track 1: Human Pose Recovery: More than 8,000 frames of continuous RGB sequences are recorded and labeled with the objective of performing human pose recovery by means of recognizing more than 120,000 human limbs of different people.

<table>
<thead>
<tr>
<th>Training frames</th>
<th>Validation frames</th>
<th>Test frames</th>
<th>Sequence duration</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>4,000</td>
<td>2,000</td>
<td>2,236</td>
<td>1-2 min</td>
<td>15</td>
</tr>
<tr>
<td>Modalities</td>
<td>Num. of users</td>
<td>Limbs per body</td>
<td>Labeled frames</td>
<td>Labeled limbs</td>
</tr>
<tr>
<td>RGB</td>
<td>14</td>
<td>14</td>
<td>8,234</td>
<td>124,761</td>
</tr>
</tbody>
</table>

Human pose recovery data characteristics.

• 9 videos (RGB sequences) and a total of 14 different actors. Stationary camera with the same static background.
• 15 fps rate, resolution 480x360 in BMP file format.
• For each actor 14 limbs (if not occluded) were manually tagged: Head, Torso, R-L Upper-arm, R-L Lower-arm, R-L Hand, R-L Upper-leg, R-L Lower-leg, and R-L Foot.
• Limbs are manually labeled using binary masks and the minimum bounding box containing each subject is defined.

• The actors appear in a wide range of different poses and performing different actions/gestures which vary the visual appearance of human limbs. So there is a large variability of human poses, self-occlusions and many variations in clothing and skin color.
**Track 1: Human Pose Recovery:** More than 8,000 frames of continuous RGB sequences are recorded and labeled with the objective of performing human pose recovery by means of recognizing more than 120,000 human limbs of different people.

**Overlap evaluation**

\[
J_{i,n} = \frac{A_{i,n} \cap B_{i,n}}{A_{i,n} \cup B_{i,n}},
\]

\[
H_{i,n} = \begin{cases} 
1 & \text{if } \frac{A_{i,n} \cap B_{i,n}}{A_{i,n} \cup B_{i,n}} \geq 0.5 \\
0 & \text{otherwise}
\end{cases}
\]

\[
J_{i,head} = \frac{A_{i,head} \cap B_{i,head}}{A_{i,head} \cup B_{i,head}} = 0.82
\]

\[J_{i,head} > 0.5 \rightarrow HR_{i,head} = 1\]
• **Track 2: Action/Interaction Recognition**: 235 performances of 11 action/interaction categories are recorded and manually labeled in continuous RGB sequences of different people performing natural isolated and collaborative behaviors.

<table>
<thead>
<tr>
<th>Training actions</th>
<th>Validation actions</th>
<th>Test actions</th>
<th>Sequence duration</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>90</td>
<td>95</td>
<td>9× 1-2 min</td>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Modalities</th>
<th>Num. of users</th>
<th>Action categories</th>
<th>Interaction categories</th>
<th>Labeled sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGB</td>
<td>14</td>
<td>7</td>
<td>4</td>
<td>235</td>
</tr>
</tbody>
</table>

Action and interaction data characteristics.

• **235 action/interaction** samples performed by **14 actors**.
• Large **difference in length** about the performed actions and interactions.
• Several **distracter actions** out of the 11 categories are also present.
• **11 action categories, containing isolated and collaborative actions**: Wave, Point, Clap, Crouch, Jump, Walk, Run, Shake Hands, Hug, Kiss, Fight. There is a high intra-class variability among action samples.

Overlap evaluation

\[
J_{s,n} = \frac{A_{s,n} \cap B_{s,n}}{A_{s,n} \cup B_{s,n}},
\]

\[
J_{s,\text{fight}} = \frac{A_{s,\text{fight}} \cap B_{s,\text{fight}}}{A_{s,\text{fight}} \cup B_{s,\text{fight}}} = 0.46
\]
• **Track 2: Action/Interaction Recognition**: 235 performances of 11 action/interaction categories are recorded and manually labeled in continuous RGB sequences of different people performing natural isolated and collaborative behaviors.
Track 3: Gesture Recognition: The gestures are drawn from a vocabulary of Italian sign gesture categories. The emphasis of this third track is on multi-modal automatic learning of a set of 20 gestures performed by several different users, with the aim of performing user independent continuous gesture spotting.

<table>
<thead>
<tr>
<th>Training seq.</th>
<th>Validation seq.</th>
<th>Test seq.</th>
<th>Sequence duration</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>393 (7,754 gestures)</td>
<td>287 (3,362 gestures)</td>
<td>276 (2,742 gestures)</td>
<td>1-2 min</td>
<td>20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Modalities</th>
<th>Num. of users</th>
<th>Gesture categories</th>
<th>Labeled sequences</th>
<th>Labeled frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGB, Depth, User mask, Skeleton</td>
<td>27</td>
<td>20</td>
<td>13,858</td>
<td>1,720,800</td>
</tr>
</tbody>
</table>

Main characteristics of the Montalbano gesture dataset.
Track 3: Gesture Recognition: The gestures are drawn from a vocabulary of Italian sign gesture categories. The emphasis of this third track is on multi-modal automatic learning of a set of 20 gestures performed by several different users, with the aim of performing user independent continuous gesture spotting.

- **Largest dataset** in the literature with a large duration of each individual performance showing no resting poses and self-occlusions.
- There is **no information about the number of gestures to spot** within each sequence, and several distracter gestures (out of the vocabulary) are present.
- **High intra-class variability** of gesture samples and **low inter-class variability** for some gesture categories.

Overlap evaluation

\[
J_{s,n} = \frac{A_{s,n} \cap B_{s,n}}{A_{s,n} \cup B_{s,n}}
\]

\[
J_{s,fight} = \frac{A_{s,fight} \cap B_{s,fight}}{A_{s,fight} \cup B_{s,fight}} = 0.46
\]
Track 3: Gesture Recognition

1. Vattene
2. Viene qui
3. Perfetto
4. E un furbo
5. Che due palle
6. Che vuoi
7. d’accordo
8. Sei pazzo
9. Cos hai com- binato
10. Non mi me friega niente

http://gesture.chalearn.org/
Track 3: Gesture Recognition

(11) Ok

(12) Cosa ti farei

(13) Basta

(14) Le vuoi prendere

(15) Non ce ne piu

(16) Ho fame

(17) Tanto tempo

(18) Buonissimo

(19) Si sono messi d’accordo

(20) Sono stufo
Datasets of the three challenge tracks

- State of the art comparison

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Labeling at pixel precision</th>
<th>Number of limbs</th>
<th>Number of labeled limbs</th>
<th>Number of frames</th>
<th>Full body</th>
<th>Limb annotation</th>
<th>Gesture-action annotation</th>
<th>Number of gestures-actions</th>
<th>Number of gest-act. samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Montalbano[8]</td>
<td>No</td>
<td>16</td>
<td>27 532 800</td>
<td>1 720 800</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>20</td>
<td>13 858</td>
</tr>
<tr>
<td>HuPBA 8K+ [7]</td>
<td>Yes</td>
<td>14</td>
<td>124 761</td>
<td>8 234</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>11</td>
<td>235</td>
</tr>
<tr>
<td>LEEDS SPORTS[4]</td>
<td>No</td>
<td>14</td>
<td>28 000</td>
<td>2 000</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>UIUC people[10]</td>
<td>No</td>
<td>14</td>
<td>18 186</td>
<td>1 299</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pascal VOC[2]</td>
<td>Yes</td>
<td>5</td>
<td>8 500</td>
<td>1 218</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BUFFY[3]</td>
<td>No</td>
<td>6</td>
<td>4 488</td>
<td>748</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PARSE[11]</td>
<td>No</td>
<td>10</td>
<td>3 050</td>
<td>305</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MPII Pose[12]</td>
<td>Yes</td>
<td>14</td>
<td>40 522</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>20</td>
<td>491</td>
</tr>
<tr>
<td>FLIC[13]</td>
<td>No</td>
<td>29</td>
<td>-</td>
<td>5 003</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>H3D[14]</td>
<td>No</td>
<td>19</td>
<td>-</td>
<td>2 000</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Actions[15]</td>
<td>No</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>6</td>
<td>600</td>
</tr>
<tr>
<td>HW[5]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>8</td>
<td>430</td>
</tr>
</tbody>
</table>

Comparison of public dataset characteristics.

ChaLearn LAP data sets, public available at: [http://sunai.uoc.edu/chalearnLAP/](http://sunai.uoc.edu/chalearnLAP/)
Competition schedule

The challenge was managed using the Microsoft Codalab platform*. The schedule of the competition was as follows:

- **February 9, 2014:** Beginning of the quantitative competition, release of development and validation data.
- **April 24, 2014:** Beginning of the registration procedure for accessing to the final evaluation data.
- **May 1, 2014:** Release of the encrypted final evaluation data and validation labels. Participants started training their methods with the whole dataset.
- **May 20, 2014:** Release of the decryption key for the final evaluation data. Participants started predicting the results on the final evaluation labels. This date was the deadline for code submission as well.
- **May 28, 2014:** End of the quantitative competition. Deadline for submitting the predictions over the final evaluation data. The organizers started the code verification by running it on the final evaluation data.
- **June 1, 2014:** Deadline for submitting the fact sheets.
- **June 10, 2014:** Publication of the competition results.

* [https://www.codalab.org/competitions/](https://www.codalab.org/competitions/)
Participation

• We created a different competition for each track, having the specific information and leaderboard.
• A total of 278 users has been registered in the Codalab platform:
  – 70 for track1
  – 79 for track2
  – 129 for track3 (some users have been registered for more than one track)
• All these users were able to access the data for the Developing stage, and submit their predictions for this stage. For the final evaluation stage, a team registration was mandatory, and a total of 62 teams were successfully registered:
  – 9 for track1
  – 15 for track2
  – 39 for track3
• Only registered teams has access to the data for the last stage.
• The data was distributed in three mirrors to facilitate the data download, using a single web page for integrating all the links and information.
• Google Analytics was activated on this page in order to track the connection on this page, and have an idea of the user details.
Participation

- **Connectivity:** During the Challenge period, the download page had a total of 2.895 visits from 920 different users of 59 countries.

http://gesture.chalearn.org/
## Participation

- **Connectivity:** During the Challenge period, the download page had a total of 2,895 visits from 920 different users of 59 countries.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Visits</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>United States</td>
<td>155</td>
<td>16.85%</td>
</tr>
<tr>
<td>2</td>
<td>China</td>
<td>113</td>
<td>12.28%</td>
</tr>
<tr>
<td>3</td>
<td>India</td>
<td>74</td>
<td>8.04%</td>
</tr>
<tr>
<td>4</td>
<td>Spain</td>
<td>58</td>
<td>6.30%</td>
</tr>
<tr>
<td>5</td>
<td>France</td>
<td>41</td>
<td>4.46%</td>
</tr>
<tr>
<td>6</td>
<td>Germany</td>
<td>40</td>
<td>4.35%</td>
</tr>
<tr>
<td>7</td>
<td>Brazil</td>
<td>36</td>
<td>3.91%</td>
</tr>
<tr>
<td>8</td>
<td>United Kingdom</td>
<td>34</td>
<td>3.70%</td>
</tr>
<tr>
<td>9</td>
<td>Japan</td>
<td>31</td>
<td>3.37%</td>
</tr>
<tr>
<td>10</td>
<td>Egypt</td>
<td>26</td>
<td>2.83%</td>
</tr>
<tr>
<td>11</td>
<td>Greece</td>
<td>26</td>
<td>2.83%</td>
</tr>
<tr>
<td>12</td>
<td>Turkey</td>
<td>24</td>
<td>2.61%</td>
</tr>
<tr>
<td>13</td>
<td>South Korea</td>
<td>21</td>
<td>2.28%</td>
</tr>
<tr>
<td>14</td>
<td>Taiwan</td>
<td>21</td>
<td>2.28%</td>
</tr>
<tr>
<td>15</td>
<td>Italy</td>
<td>19</td>
<td>2.07%</td>
</tr>
<tr>
<td>16</td>
<td>Netherlands</td>
<td>19</td>
<td>2.07%</td>
</tr>
<tr>
<td>17</td>
<td>Singapore</td>
<td>19</td>
<td>2.07%</td>
</tr>
<tr>
<td>18</td>
<td>Australia</td>
<td>18</td>
<td>1.96%</td>
</tr>
<tr>
<td>19</td>
<td>Vietnam</td>
<td>12</td>
<td>1.30%</td>
</tr>
<tr>
<td>20</td>
<td>Canada</td>
<td>11</td>
<td>1.20%</td>
</tr>
<tr>
<td>21</td>
<td>Switzerland</td>
<td>11</td>
<td>1.20%</td>
</tr>
<tr>
<td>22</td>
<td>Belgium</td>
<td>9</td>
<td>0.98%</td>
</tr>
<tr>
<td>23</td>
<td>Russia</td>
<td>9</td>
<td>0.98%</td>
</tr>
<tr>
<td>24</td>
<td>Hong Kong</td>
<td>7</td>
<td>0.76%</td>
</tr>
</tbody>
</table>
Results

• Track1 results

<table>
<thead>
<tr>
<th>Team</th>
<th>Accuracy</th>
<th>Rank position</th>
<th>Features</th>
<th>Pose model</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZJU</td>
<td>0.194144</td>
<td>1</td>
<td>HOG</td>
<td>tree structure</td>
</tr>
<tr>
<td>Seawolf Vision</td>
<td>0.182097</td>
<td>2</td>
<td>HOG</td>
<td>tree structure</td>
</tr>
</tbody>
</table>

Both winner participants applied a similar approach based on [*].

• Mixture of templates for each part. This method incorporates the co-occurrence relations, appearance and deformation into a model represented by an objective function of pose configurations. Model is tree-structured, and optimization is conducted via dynamic programming.

Results

• Track2 results

<table>
<thead>
<tr>
<th>Team name</th>
<th>Accuracy</th>
<th>Rank</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUHK-SWJTU</td>
<td>0.507173</td>
<td>1</td>
<td>Improved trajectories [*]</td>
</tr>
<tr>
<td>ADSC</td>
<td>0.501164</td>
<td>2</td>
<td>Improved trajectories [*]</td>
</tr>
<tr>
<td>SBUVIS</td>
<td>0.441405</td>
<td>3</td>
<td>Improved trajectories [*]</td>
</tr>
<tr>
<td>DonkeyBurger</td>
<td>0.342192</td>
<td>4</td>
<td>MHI, STIP</td>
</tr>
<tr>
<td>UC-T2</td>
<td>0.121565</td>
<td>5</td>
<td>Improved trajectories [*]</td>
</tr>
<tr>
<td>MindLAB</td>
<td>0.008383</td>
<td>6</td>
<td>MBF</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Team name</th>
<th>Dimension reduction</th>
<th>Clustering</th>
<th>Classifier</th>
<th>Temporal coherence</th>
<th>Gesture representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUHK-SWJTU</td>
<td>PCA</td>
<td>-</td>
<td>SVM</td>
<td>Sliding windows</td>
<td>Fisher Vector</td>
</tr>
<tr>
<td>ADSC</td>
<td>-</td>
<td>-</td>
<td>SVM</td>
<td>Sliding windows</td>
<td>-</td>
</tr>
<tr>
<td>SBUVIS</td>
<td>-</td>
<td>-</td>
<td>SVM</td>
<td>Sliding windows</td>
<td>-</td>
</tr>
<tr>
<td>DonkeyBurger</td>
<td>-</td>
<td>Kmeans</td>
<td>Sparse code</td>
<td>Sliding windows</td>
<td>-</td>
</tr>
<tr>
<td>UC-T2</td>
<td>PCA</td>
<td>-</td>
<td>Kmeans</td>
<td>Sliding windows</td>
<td>Fisher Vector</td>
</tr>
<tr>
<td>MindLAB</td>
<td>-</td>
<td>Kmeans</td>
<td>RF</td>
<td>Sliding windows</td>
<td>BoW</td>
</tr>
</tbody>
</table>

### Results

**Track3 results**

<table>
<thead>
<tr>
<th>Team</th>
<th>Accuracy</th>
<th>Rank</th>
<th>Modalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIRIS</td>
<td>0.849987</td>
<td>1</td>
<td>SK, Depth, RGB</td>
</tr>
<tr>
<td>CraSPN</td>
<td>0.833904</td>
<td>2</td>
<td>SK, Depth, RGB</td>
</tr>
<tr>
<td>JY</td>
<td>0.826799</td>
<td>3</td>
<td>SK, RGB</td>
</tr>
<tr>
<td>CUHK-SWJTU</td>
<td>0.791933</td>
<td>4</td>
<td>RGB</td>
</tr>
<tr>
<td>Lpigou</td>
<td>0.788804</td>
<td>5</td>
<td>Depth, RGB</td>
</tr>
<tr>
<td>stevenwudi</td>
<td>0.787310</td>
<td>6</td>
<td>SK, depth</td>
</tr>
<tr>
<td>Ismar</td>
<td>0.746632</td>
<td>7</td>
<td>SK</td>
</tr>
<tr>
<td>Quads</td>
<td>0.745449</td>
<td>8</td>
<td>SK</td>
</tr>
<tr>
<td>Telepoints</td>
<td>0.688778</td>
<td>9</td>
<td>SK, Depth, RGB</td>
</tr>
<tr>
<td>TUM-fortiss</td>
<td>0.648979</td>
<td>10</td>
<td>SK, Depth, RGB</td>
</tr>
<tr>
<td>CSU-SCM</td>
<td>0.597177</td>
<td>11</td>
<td>Skeleton, Depth, mask</td>
</tr>
<tr>
<td>iva.mn</td>
<td>0.556251</td>
<td>12</td>
<td>Skeleton, RGB, depth</td>
</tr>
<tr>
<td>Terrier</td>
<td>0.539025</td>
<td>13</td>
<td>Skeleton</td>
</tr>
<tr>
<td>Team Netherlands</td>
<td>0.430709</td>
<td>14</td>
<td>Skeleton, Depth, RGB</td>
</tr>
<tr>
<td>VecsRel</td>
<td>0.408012</td>
<td>15</td>
<td>Skeleton, Depth, RGB</td>
</tr>
<tr>
<td>Samgest</td>
<td>0.391613</td>
<td>16</td>
<td>Skeleton, Depth, RGB, mask</td>
</tr>
<tr>
<td>YNL</td>
<td>0.270600</td>
<td>17</td>
<td>Skeleton</td>
</tr>
</tbody>
</table>

**Percentage of methods using each independent modality**

![Bar chart showing the percentage of methods using each independent modality](http://gesture.chalearn.org/)
## Results

### Track3 results

<table>
<thead>
<tr>
<th>Team</th>
<th>Gesture representation</th>
<th>Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIRIS</td>
<td>-</td>
<td>DNN</td>
</tr>
<tr>
<td>CraSPN</td>
<td>BoW</td>
<td>Adaboost</td>
</tr>
<tr>
<td>JY</td>
<td>-</td>
<td>MRF, KNN</td>
</tr>
<tr>
<td>CUHK-SWJTU</td>
<td>Fisher Vector, VLAD</td>
<td>SVM</td>
</tr>
<tr>
<td>Lpigou</td>
<td>-</td>
<td>CNN</td>
</tr>
<tr>
<td>stevenwudi</td>
<td>-</td>
<td>HMM, DNN</td>
</tr>
<tr>
<td>Ismar</td>
<td>-</td>
<td>RF</td>
</tr>
<tr>
<td>Quads</td>
<td>Fisher Vector</td>
<td>SVM</td>
</tr>
<tr>
<td>Telepoints</td>
<td>-</td>
<td>SVM</td>
</tr>
<tr>
<td>TUM-fortiss</td>
<td>-</td>
<td>RF, SVM</td>
</tr>
<tr>
<td>CSU-SCM</td>
<td>2DMTM</td>
<td>SVM, HMM</td>
</tr>
<tr>
<td>iva.mm</td>
<td>BoW</td>
<td>SVM, HMM</td>
</tr>
<tr>
<td>Terrier</td>
<td>-</td>
<td>RF</td>
</tr>
<tr>
<td>Team Netherlands</td>
<td>-</td>
<td>SVM, RT</td>
</tr>
<tr>
<td>VecsRel</td>
<td>-</td>
<td>DNN</td>
</tr>
<tr>
<td>Samgest</td>
<td>-</td>
<td>HMM</td>
</tr>
<tr>
<td>YNL</td>
<td>Fisher Vector</td>
<td>HMM, SVM</td>
</tr>
</tbody>
</table>

### Percentage of methods using each gesture classification strategy

![Graph showing percentages of methods using different classification strategies]
### Results

- **Track3 results**

<table>
<thead>
<tr>
<th>Team</th>
<th>Features</th>
<th>Fusion</th>
<th>Temp. segmentation</th>
<th>Dimension reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIRIS</td>
<td>RAW, SK joints</td>
<td>Early</td>
<td>Joints motion</td>
<td>-</td>
</tr>
<tr>
<td>CraSPN</td>
<td>HOG, SK</td>
<td>Early</td>
<td>Sliding windows</td>
<td>-</td>
</tr>
<tr>
<td>JY</td>
<td>SK, HOG</td>
<td>Late</td>
<td>MRF</td>
<td>PCA</td>
</tr>
<tr>
<td>CUHK-SWJTU</td>
<td>Improved trajectories</td>
<td>-</td>
<td>Joints motion</td>
<td>PCA</td>
</tr>
<tr>
<td>Lpigou</td>
<td>RAW, SK joints</td>
<td>Early</td>
<td>Sliding windows</td>
<td>Max-pooling CNN</td>
</tr>
<tr>
<td>stevenwudi</td>
<td>RAW</td>
<td>Late</td>
<td>Sliding windows</td>
<td>-</td>
</tr>
<tr>
<td>Ismar</td>
<td>SK</td>
<td>-</td>
<td>Sliding windows</td>
<td>-</td>
</tr>
<tr>
<td>Quads</td>
<td>SK quads</td>
<td>-</td>
<td>Sliding windows</td>
<td>-</td>
</tr>
<tr>
<td>Telepoints</td>
<td>STIPS, SK</td>
<td>Late</td>
<td>Joints motion</td>
<td>-</td>
</tr>
<tr>
<td>TUM-fortiss</td>
<td>STIPS</td>
<td>Late</td>
<td>Joints motion</td>
<td>-</td>
</tr>
<tr>
<td>CSU-SCM</td>
<td>HOG, Skeleton</td>
<td>Late</td>
<td>Sliding windows</td>
<td>-</td>
</tr>
<tr>
<td>iva.mm</td>
<td>Skeleton, HOG</td>
<td>Late</td>
<td>Sliding windows</td>
<td>-</td>
</tr>
<tr>
<td>Terrier</td>
<td>Skeleton</td>
<td>-</td>
<td>Sliding windows</td>
<td>-</td>
</tr>
<tr>
<td>Team Netherlands</td>
<td>MHI</td>
<td>Early</td>
<td>DTW</td>
<td>Preserving projections</td>
</tr>
<tr>
<td>VecsRel</td>
<td>RAW, skeleton joints</td>
<td>Late</td>
<td>DTW</td>
<td>-</td>
</tr>
<tr>
<td>Samgest</td>
<td>Skeleton, blobs, moments</td>
<td>Late</td>
<td>Sliding windows</td>
<td>-</td>
</tr>
<tr>
<td>YNL</td>
<td>Skeleton</td>
<td>-</td>
<td>Sliding windows</td>
<td>-</td>
</tr>
</tbody>
</table>
Conclusion (1/2)

- For the case of **pose recovery**, tree-structure models were mainly applied.
- The winner achieved almost **0.2 of accuracy**.
- In the case of **action/interaction** RGB data sequences, methods for refining the tracking process of visual landmarks while considering alternatives to the classical BoW feature representation have been used.
- So the general trend was to compute a quantification of visual words present in the image and performing **sliding windows classification using discriminative classifiers** (note that limbs from track1 were not available!).
- Most top ranked participants used **SVMs**, although **random forests** were also considered.
- The winner achieved an **accuracy of over 0.5**.
- In the case of **multi-modal gesture recognition**, and following current trends in the computer vision literature, a **deep learning architecture** achieved the first position, with an **accuracy score of almost 0.85**.
- Most approaches were based on **skeleton joint information and several state-of-the-art descriptors were jointly used** by the participants without showing a generic common trend.
- **Temporal segmentation** was usually considered by **sliding windows or skeleton motion information**.
Conclusion (2/2)

- **SVM, RF, HMM, and DTW** algorithms were widely considered.
- Interestingly, it is the first time that some participants used deep learning architectures such as Convolutional Neural Networks.
- The winner of the competition used all the modalities and information of the human joints to segment gesture candidates.
- The code of the participants using deep learning took a lot more time for training than the rest of approaches.
- There are still much ways for improvement in the two RGB domains considered, namely human pose recovery and action/interaction recognition from RGB data.
- **Future trends in Looking at People** may include group interactions and cultural event classification, where context also places an important role, while including the analysis of social signals (also maybe considering multi-modal input data), affective computing, and face analysis as relevant information cues.
Organizers

Sergio Escalera     Jordi Gonzàlez     Xavier Báró     Miguel Reyes     Víctor Ponce

Meysam Madadi     M. Ángel Bautista     Isabelle Guyon     Hugo J. Escalante     Jamie Shotton
Next events of ChaLearn LAP

ChaLearn LAP 2015: Age recognition on RGB data, be prepared for the challenge and workshop!!!

Call for Papers IEEE Transactions on Pattern Analysis and Machine Intelligence

Special Issue on Multimodal Human Pose Recovery and Behavior Analysis – M2HuPBA

Important Dates

Submission Deadline: December 1, 2014
First round of Reviews: March 15, 2015
First revisions of Submissions: April 2015
Final Decisions/Manuscript: August 2015
Estimated Online Publication: End of 2015

Guest Editors

Dr. Sergio Escalera, University of Barcelona & Computer Vision Center
Dr. Jordi González, Universitat Autònoma de Barcelona & Computer Vision Center
Dr. Xavier Baró, Universitat Oberta de Catalunya & Computer Vision Center
Prof. Jamie Shotton, Microsoft Research
Contact email: sergio.escalera.guerrero@gmail.com

We wait for your contributions!
Thank you and hope to see you in our next event!