ChaLearn Looking at People Challenge 2014: Dataset and Results

Sergio Escalera[1,2,3], Xavier Baró[1,4], Jordi Gonzàlez[1,5],
Miguel A. Bautista[1,2], Meysam Madadi[1,2], Miguel Reyes[1,2], Víctor Ponce[1,2,4],
Hugo J. Escalante[3,6], Jamie Shotton[7], Isabelle Guyon[3]

[1] Computer Vision Center, Campus UAB, Barcelona
[2] Dept. Mathematics, University of Barcelona
[3] ChaLearn, Berkeley, California
[4] EIMT at the Open University of Catalonia, Barcelona
[5] Dept. Computer Science, Univ. Autònoma de Barcelona
[6] INAOE, Puebla, Mexico
[7] Microsoft Research, Cambridge, UK

Abstract. This paper summarizes the ChaLearn Looking at People 2014 challenge. The competition was split into three independent tracks: human pose recovery from RGB data, action and interaction recognition from RGB data sequences, and multi-modal gesture recognition from RGB-Depth sequences. For all the tracks, the goal was to perform user-independent recognition in sequences of continuous images using the overlapping Jaccard index as the evaluation measure. In this edition of the ChaLearn challenge, two large novel datasets were made publicly available and the Microsoft Codalab platform were used to manage the competition. Results achieved an overlapping accuracy about 0.20 and 0.50 for pose recovery and action/interaction spotting, showing still much margin for improvement, meanwhile an overlapping about 0.85 was achieved for multi-modal gesture recognition, making it feasible to be applied in real applications.

Keywords: Human Pose Recovery, Behavior Analysis, Action and interactions, Multi-modal gestures, recognition.

1 Introduction

The automatic, computational analysis of the human body in image sequences, referred to as Looking at People (LAP) in [1], keeps making rapid progress with the constant improvement of (i) new published methods that constantly push the state-of-the-art, and (ii) the recent availability of inexpensive 3D video sensors such as Kinect. Applications are countless, like HCI, surveillance, communication, entertainment, safety, e-commerce and sports, thus having an important social impact in assisting technologies for the handicapped and the elderly, for example.

In 2011 and 2012, ChaLearn[1] organized a challenge on single user one-shot-learning gesture recognition with data recorded with Kinect. In 2013, 54 teams

[1] http://gesture.chalearn.org/

participated in the ChaLearn challenge which was devoted to Multimodal Gesture Recognition. In that edition, we proposed a user-independent gesture recognition task in visual data recorded with Kinect and containing a large set of continuously performed Italian gestures.

In the edition of 2014, we have organized a second round of the same gesture recognition task including a finer begin-end labeling of gestures with the objective of performing gesture recognition. Additionally, for the 2014 edition, we have organized two competitions for human pose recovery and action recognition in RGB data. One goal of the challenge, inspired by the previous 2005-2012 Pascal VOC image recognition challenges on Human Layout Analysis successfully organized by Everingham et al. [2], was also to automatically recognize human limbs from RGB data. Another goal was to run a competition for human action and interaction recognition on RGB data.

In this paper we detail how the ChaLearn LAP 2014 challenge was organized, the datasets, the results achieved by almost 200 participants that joined the competition, and the main characteristics of the winning methods.

2 Challenge tracks and schedule

The ChaLearn LAP 2014 challenge featured three quantitative evaluations: automatic human pose recovery on RGB data, action/interaction recognition on RGB data, and gesture recognition from a multi-modal dataset recorded with Kinect. The characteristics of each competition track are the following:

- Track 1: Human Pose Recovery: a novel dataset containing 120K+ manually annotated limbs for 8K+ frames showing actors performing natural motion was provided for automatic body limb detection.
- Track 2: Action/Interaction recognition: in total, 235 action samples performed by 17 actors were provided. The selected actions involved the motion of most of the limbs and included interactions among various actors.
- Track 3: Multi-modal gesture recognition: The RGBD data contains nearly 14K manually labeled (beginning and ending frame) gesture performances in continuous video sequences, with a vocabulary of 20 Italian gesture categories. This third track focused on multi-modal automatic learning of a set of gestures with the aim of performing user independent continuous gesture recognition.

2.1 Competition schedule

The challenge was managed using the Microsoft Codalab platform[2]. 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.

[2] https://www.codalab.org/competitions/

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.

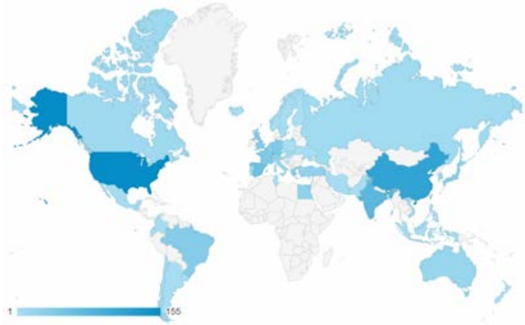
June 1, 2014: Deadline for submitting the fact sheets.

June 10, 2014: Publication of the competition results.

2.2 User Connection

The competition has been conducted using Codalab[3], a novel challenge open-source platform. 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, and 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, and 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 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. During the Challenge period, the download page have had a total of 2.895 visits from 920 different users of 59 countries. Details are shown in Fig. 2.2.



(a)

1	United States	155(16,85%)	13	South Korea	21(2,28%)
2	China	113(12,28%)	14	Taiwan	21(2,28%)
3	India	74(8,04%)	15	Italy	19(2,07%)
4	Spain	58(6,30%)	16	Netherlands	19(2,07%)
5	France	41(4,46%)	17	Singapore	19(2,07%)
6	Germany	40(4,35%)	18	Australia	18(1,96%)
7	Brazil	36(3,91%)	19	Vietnam	12(1,30%)
8	United Kingdom	34(3,70%)	20	Canada	11(1,20%)
9	Japan	31(3,37%)	21	Switzerland	11(1,20%)
10	Egypt	26(2,83%)	22	Belgium	9(0,98%)
11	Greece	26(2,83%)	23	Russia	9(0,98%)
12	Turkey	24(2,61%)	24	Hong Kong	7(0,76%)

(b)

Fig. 1. (a) Geographic distribution of users. (b) Distribution of the visitors of the 24 first countries in terms of users.

3 Competition data

In the next subsections we describe the datasets and their characteristics provided for the three challenge tracks[4].

[3] www.codalab.org

[4] Datasets are available at http://sunai.uoc.edu/chalearnLAP/

Training frames	Validation frames	Test frames	Sequence duration	FPS
4,000	2,000	2,236	1-2 min	15

Modalities	Num. of users	Limbs per body	Labeled frames	Labeled limbs
RGB	14	14	8,234	124,761

Table 1. Human pose recovery data characteristics.

3.1 Track 1: Human Pose Recovery dataset

Publicly available datasets for human pose recovery lack of refined labeling or contain a very reduced number of samples per limb (e.g. *Buffy Stickmen V*3.01, *Leeds Sports* and *Hollywood Human Actions* [3–5]). In addition, large datasets often use synthetic samples or capture human limbs with sensor technologies such as *MoCap* in very controled environments [6].

Being aware of this lack of public available datasets for multi-limb human pose detection, we presented a novel fully limb labeled dataset, the Human Pose Recovery and Behavior Analysis *HuPBA 8k+* dataset [7]. This dataset is formed by more than 8000 frames where 14 limbs are labeled at pixel precision, thus providing $124,761$ annotated human limbs. The characteristics of the dataset are:

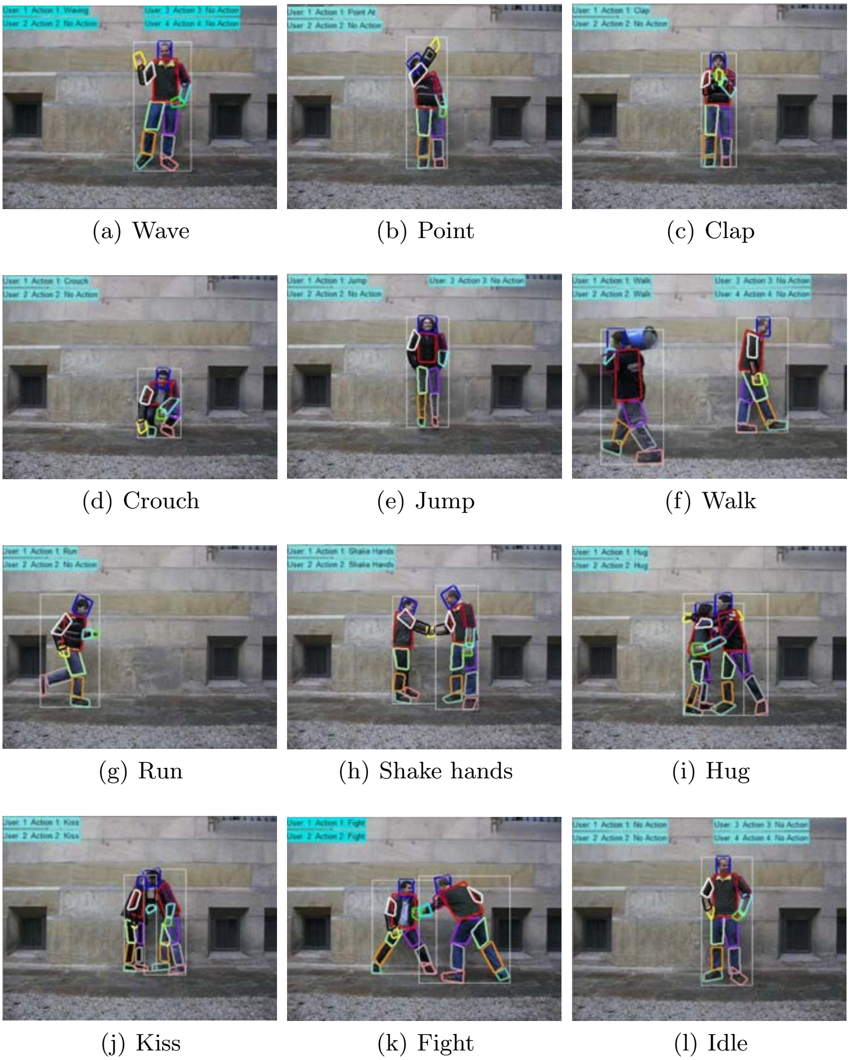
- The images are obtained from 9 videos (RGB sequences) and a total of 14 different actors appear in the sequences. The image sequences have been recorded using a stationary camera with the same static background.
- Each video (RGB sequence) was recorded at 15 fps rate, and each RGB image was stored with resolution 480 × 360 in BMP file format.
- For each actor present in an image 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.

A list of data attributes for this first track dataset is described in Table 1. Examples of images of the dataset are shown in Fig. 2.

3.2 Track 2: Action/interaction dataset

In addition to human-limb labelling, in the *HuPBA 8K+* dataset we also annotated the beginning and ending frames of actions and interactions. A key frame example for each gesture/action category is also shown in Fig. 2. The challenges the participants had to deal with for this new competition are:

- 235 action/interaction samples performed by 14 actors.
- Large difference in length about the performed actions and interactions. Several distractor 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.


(a) Wave (b) Point (c) Clap

(d) Crouch (e) Jump (f) Walk

(g) Run (h) Shake hands (i) Hug

(j) Kiss (k) Fight (l) Idle

Fig. 2. Key frames of the *HuPBA 8K+* dataset used in the tracks 1 and 2, showing actions ((a) to (g)), interactions ((h) to (k)) and the idle pose (l).

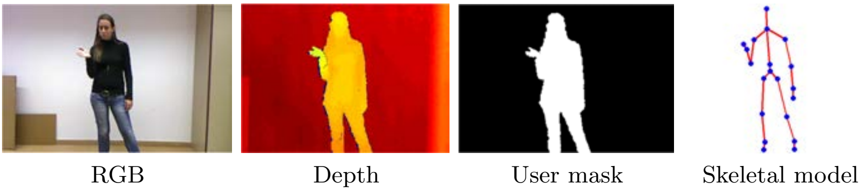

RGB Depth User mask Skeletal model

Fig. 3. Different modalities of the dataset used in track 3.

Table 2 summarizes the dataset attributes for this second track.

Training actions	Validation actions	Test actions	Sequence duration	FPS
150	90	95	9× 1-2 min	15

Modalities	Num. of users	Action categories	interaction categories	Labeled sequences
RGB	14	7	4	235

Table 2. Action and interaction data characteristics.

Training seq.	Validation seq.	Test seq.	Sequence duration	FPS
393 (7,754 gestures)	287 (3,362 gestures)	276 (2,742 gestures)	1-2 min	20

Modalities	Num. of users	Gesture categories	Labeled sequences	Labeled frames
RGB, Depth, User mask, Skeleton	27	20	13,858	1,720,800

Table 3. Main characteristics of the *Montalbano* gesture dataset.

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

Table 4. Comparison of public dataset characteristics.

3.3 Track 3: Multi-modal gesture dataset

This track is based on an Italian gesture dataset, called *Montalbano gesture dataset*, an enhanced version of the ChaLearn 2013 multi-modal gesture recognition challenge [8, 9] with more ground-truth annotations. In all the sequences, a single user is recorded in front of a Kinect, performing natural communicative gestures and speaking in fluent Italian. Examples of the different visual modalities are shown in Fig. 3. In ChaLearn LAP 2014 we have focused on the user-independent automatic recognition of a vocabulary of 20 Italian cultural/anthropological signs in image sequences, see Fig. 4.

The main characteristics of the database are:

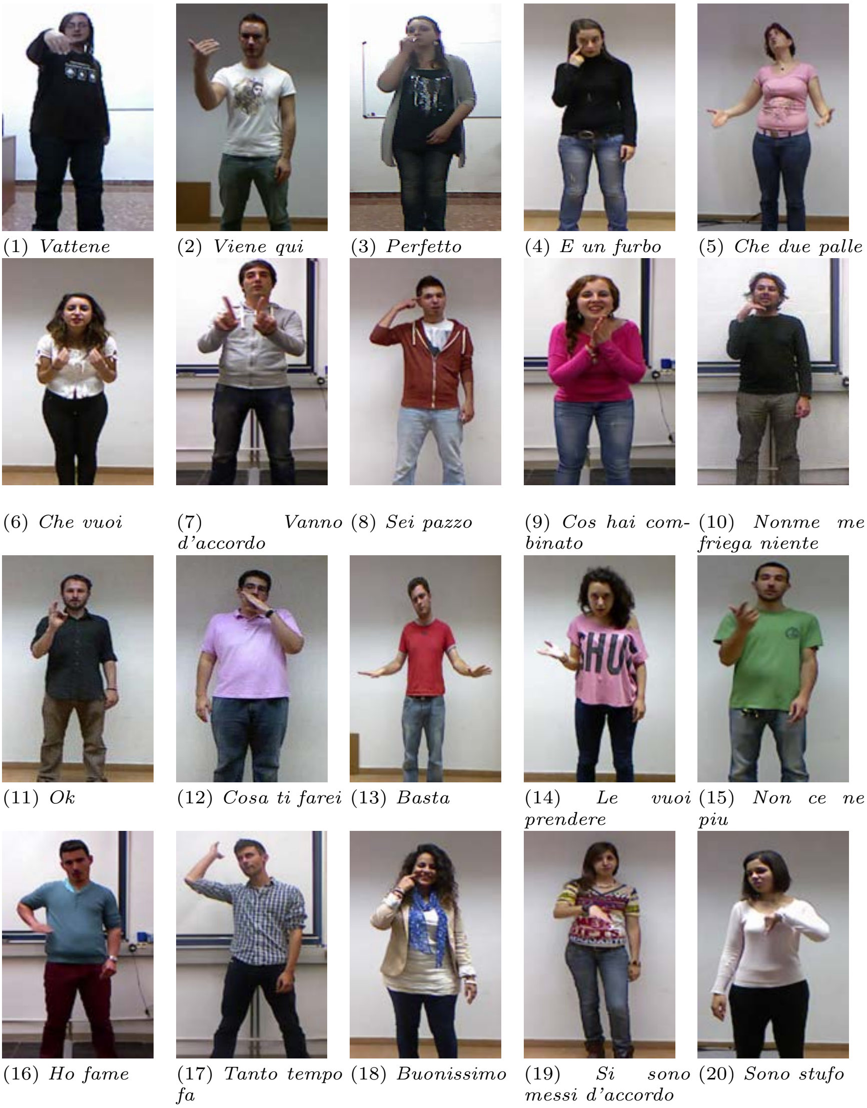
• Largest dataset in the literature, with a large duration of each individual performance showing no resting poses and self-occlusions.

• There is no information about the number of gestures to spot within each sequence, and several distractor gestures (out of the vocabulary) are present.

• High intra-class variability of gesture samples and low inter-class variability for some gesture categories.

A list of data attributes for dataset used in track 3 is described in Table 3.

In Table 4 we compare the *HuPBA 8K+* and *Montalbano* datasets used in the ChaLearn LAP 2014 with other publicly available datasets. These datasets are chosen taking into account the variability of limbs and gestures/actions. Considering limb labelling, the *HuPBA 8K+* dataset contains the highest number of annotated limbs at pixel precision. When compared with other action datasets, the number of action instances are similar. On the other hand, the *Montalbano* database contains many more samples and much more variety of gestures than any proposed dataset up to this date.



(1) *Vattene*
(2) *Viene qui*
(3) *Perfetto*
(4) *E un furbo*
(5) *Che due palle*
(6) *Che vuoi*
(7) *Vanno d'accordo*
(8) *Sei pazzo*
(9) *Cos hai combinato*
(10) *Nonme me friega niente*
(11) *Ok*
(12) *Cosa ti farei*
(13) *Basta*
(14) *Le vuoi prendere*
(15) *Non ce ne piu*
(16) *Ho fame*
(17) *Tanto tempo fa*
(18) *Buonissimo*
(19) *Si sono messi d'accordo*
(20) *Sono stufo*

Fig. 4. The *Montalbano* gesture dataset.

4 Protocol and evaluation

The evaluation metrics used to evaluate the participants for the three tracks, based on the Jaccard Index, are detailed in the following subsections.

4.1 Evaluation procedure for track 1

For all the $n \leq 14$ limbs labeled for each subject at each frame, the Jaccard Index is defined as:

$$J_{i,n} = \frac{A_{i,n} \bigcap B_{i,n}}{A_{i,n} \bigcup B_{i,n}}, \tag{1}$$

where $A_{i,n}$ is the ground truth of limb n, and $B_{i,n}$ is the prediction for the same limb at image i. For the *HuPBA 8K+* dataset used in this track, both $A_{i,n}$ and $B_{i,n}$ are binary images where pixels with value 1 denote the region in which the n-th limb is predicted, 0 otherwise. Particularly, since $A_{i,n}$ (ground truth) is a binary image and 1-pixels indicate the region of the $n-$th limb, this positive region does not necessarily need to be square. However, in all cases the positive region is a polyhedron defined by four points. Thus, the numerator in Eq. (1) is the number of 1-pixels that intersects in both images $A_{i,n}$ and $B_{i,n}$, and the denominator is the number of union 1-pixels after applying the logical OR operator.

The participants' methods were evaluated based on Hit Rate ($H_{i,n}$) accuracy for for each limb n at each image i. In essence, a hit is computed if $J_{i,n} \geq 0.5$. Then, the mean hit rate among all limbs for all images was computed (where all limb detections had the same weight) and the participant with the highest mean hit rate won the challenge.

$$H_{i,n} = \begin{cases} 1 & \text{if } \frac{A_n \bigcap B_n}{A_n \bigcup B_n} \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

In the case of false positives (e.g. predicting a limb that is not on the ground truth because of being occluded), the prediction did not affect the mean Hit Rate calculation. In that case where $n < 14$, participants do not need to provide any prediction for that particular limb. In other words, n is computed as the intersection of the limb categories in the ground truth and the predictions.

An example of the mean hit rate calculation for an example of $n = 3$ limbs and $i = 1$ image is show in Fig. 5(a). In the top part of the image the Jaccard Index for the head limb is computed. As it is greater than 0.5 then it is counted as a hit for image i and the head limb. Similarly, for the torso limb the Jaccard Index obtained is 0.72 (center part of the image) which also computes as a hit for torso limb. In addition, in the bottom of the image the Jaccard Index obtained for the left thigh limb is shown, which does not count as a hit since $0.04 < 0.5$. Finally, the mean hit rate is shown for the three limbs.

4.2 Evaluation procedure for tracks 2 and 3

To evaluate the accuracy of action/interaction recognition, we use the Jaccard Index as in track 1, the higher the better. Thus, for the n action, interaction, and gesture categories labeled for a RGB/RGBD sequence s, the Jaccard Index is defined as:

$$J_{s,n} = \frac{A_{s,n} \bigcap B_{s,n}}{A_{s,n} \bigcup B_{s,n}}, \tag{3}$$

where $A_{s,n}$ is the ground truth of action/interaction/gesture n at sequence s, and $B_{s,n}$ is the prediction for such an action at sequence s. $A_{s,n}$ and $B_{s,n}$ are binary vectors where 1-values correspond to frames in which the $n-$th action is being performed. The participants were evaluated based on the mean Jaccard Index among all categories for all sequences, where motion categories are independent

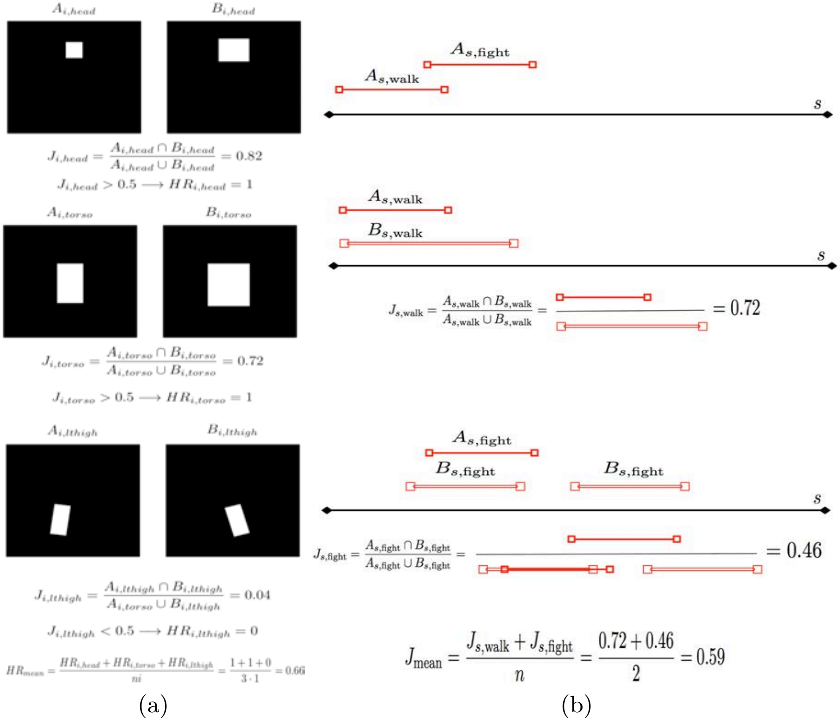

$A_{i,head}$
$B_{i,head}$
$J_{i,head} = \frac{A_{i,head} \cap B_{i,head}}{A_{i,head} \cup B_{i,head}} = 0.82$
$J_{i,head} > 0.5 \longrightarrow HR_{i,head} = 1$
$A_{i,torso}$
$B_{i,torso}$
$J_{i,torso} = \frac{A_{i,torso} \cap B_{i,torso}}{A_{i,torso} \cup B_{i,torso}} = 0.72$
$J_{i,torso} > 0.5 \longrightarrow HR_{i,torso} = 1$
$A_{i,lthigh}$
$B_{i,lthigh}$
$J_{i,lthigh} = \frac{A_{i,lthigh} \cap B_{i,lthigh}}{A_{i,torso} \cup B_{i,lthigh}} = 0.04$
$J_{i,lthigh} < 0.5 \longrightarrow HR_{i,lthigh} = 0$
$HR_{mean} = \frac{HR_{i,head} + HR_{i,torso} + HR_{i,lthigh}}{ni} = \frac{1+1+0}{3 \cdot 1} = 0.66$
(a)
$A_{s,\text{fight}}$
$A_{s,\text{walk}}$
s
$A_{s,\text{walk}}$
$B_{s,\text{walk}}$
s
$J_{s,\text{walk}} = \frac{A_{s,\text{walk}} \cap B_{s,\text{walk}}}{A_{s,\text{walk}} \cup B_{s,\text{walk}}} =$
$= 0.72$
$A_{s,\text{fight}}$
$B_{s,\text{fight}}$
$B_{s,\text{fight}}$
s
$J_{s,\text{fight}} = \frac{A_{s,\text{fight}} \cap B_{s,\text{fight}}}{A_{s,\text{fight}} \cup B_{s,\text{fight}}} =$
$= 0.46$
$J_{\text{mean}} = \frac{J_{s,\text{walk}} + J_{s,\text{fight}}}{n} = \frac{0.72 + 0.46}{2} = 0.59$
(b)

Fig. 5. (a) Example of Mean hit rate calculation for track 1. (b) Example of mean Jaccard Index calculation for tracks 2 and 3.

but not mutually exclusive (in a certain frame more than one action, interaction, gesture class can be active).

In the case of false positives (e.g. inferring an action, interaction or gesture not labeled in the ground truth), the Jaccard Index is 0 for that particular prediction, and it will not count in the mean Jaccard Index computation. In other words n is equal to the intersection of action/interaction/gesture categories appearing in the ground truth and in the predictions.

An example of the calculation for two actions is shown in Fig. 5(b). Note that in the case of recognition, the ground truth annotations of different categories can overlap (appear at the same time within the sequence). Also, although different actors appear within the sequence at the same time, actions/interactions/gestures are labeled in the corresponding periods of time (that may overlap), there is no need to identify the actors in the scene.

The example in Fig. 5(b) shows the mean Jaccard Index calculation for different instances of actions categories in a sequence (single red lines denote ground truth annotations and double red lines denote predictions). In the top part of the image one can see the ground truth annotations for actions walk and fight at sequence s. In the center part of the image a prediction is evaluated obtaining a Jaccard Index of 0.72. In the bottom part of the image the same procedure is performed with the action fight and the obtained Jaccard Index is 0.46. Finally, the mean Jaccard Index is computed obtaining a value of 0.59.

5 Challenge results and methods

In this section we summarize the methods proposed by the participants and the winning methods. For the three tracks, 2 (track 1), 6 (track 2) and 17 (track 3) teams submitted their code and predictions for the test sets. Tables 5, 6 and 7 summarize the approaches of the participants who uploaded their models.

5.1 Track 1: RGB Pose recovery results

For the first track, as shown in Table 5, both winner participants applied a similar approach based on [17]. Basically, both methods estimate human pose based on static images employing a 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. When co-occurrence and spatial relations are tree-structured, optimization can be efficiently conducted via dynamic programming. Inference is conducted via maximizing the objective function with respect to the most probable configuration.

5.2 Track 2: RGB action/interaction recognition results

Table 6 summarizes the methods of the six participants that participated on the test set of track 2. One can see that most methods are based on similar approaches. In particular, alternative representations to classical BoW were considered, as Fisher Vector and VLAD [18]. Most methods perform sliding windows and SVM classification. In addition, to refine the tracking of interest points, 4 participants used improved trajectories [16]. Next, we describe the main characteristics of the three winning methods.

First place: The method was composed of two parts: video representation and temporal segmentation. For the representation of video clip, the authors first extracted improved dense trajectories with HOG, HOF, MBHx, and MBHy descriptors. Then, for each kind of descriptor, the participants trained a GMM and used Fisher vector to transform these descriptors into a high dimensional super vector space. Finally, sum pooling was used to aggregate these codes in the whole video clip and normalize them with power L2 norm. For the temporal recognition, the authors resorted to a temporal sliding method along the time dimension. To speed up the processing of detection, the authors designed a temporal integration histogram of Fisher Vector, with which the pooled Fisher Vector was efficiently evaluated at any temporal window. For each sliding window, the authors used the pooled Fisher Vector as representation and fed it into the SVM classifier for action recognition.

Second place: a human action detection framework called ”mixture of heterogeneous attribute analyzer” was proposed. This framework integrated heterogeneous attributes learned from various types of video features including static and dynamic, local and global features, to boost the action detection accuracy.

Team	Accuracy	Rank position	Features	Pose model
ZJU	**0.194144**	1	HOG	tree structure
Seawolf Vision	**0.182097**	2	HOG	tree structure

Table 5. Track 1 Pose Recovery results.

Team name	Accuracy	Rank	Features	Dimension reduction	Clustering	Classifier	Temporal coherence	Gesture representation
CUHK-SWJTU	**0.507173**	1	Improved trajectories [16]	PCA	-	SVM	Sliding windows	Fisher Vector
ADSC	**0.501164**	2	Improved trajectories [16]	-	-	SVM	Sliding windows	-
SBUVIS	**0.441405**	3	Improved trajectories [16]	-	-	SVM	Sliding windows	-
DonkeyBurger	0.342192	4	MHI, STIP	-	Kmeans	Sparse code	Sliding windows	-
UC-T2	0.121565	5	Improved trajectories [16]	PCA	-	Kmeans	Sliding windows	Fisher Vector
MindLAB	0.008383	6	MBF	-	Kmeans	RF	Sliding windows	BoW

Table 6. Track 2 action/interaction recognition results. MHI: Motion History Image; STIP: Spatio-Temporal interest points; MBF: Multiscale Blob Features; BoW: Bag of Visual Words; RF: Random Forest.

Team	Accuracy	Rank	Modalities	Features	Fusion	Temp. segmentation	Dimension reduction	Gesture representation	Classifier
LIRIS	**0.849987**	1	SK, Depth, RGB	RAW, SK joints	Early	Joints motion	-	-	DNN
CraSPN	**0.833904**	2	SK, Depth, RGB	HOG, SK	Early	Sliding windows	-	BoW	Adaboost
JY	**0.826799**	3	SK, RGB	SK, HOG	Late	MRF	PCA	-	MRF, KNN
CUHK-SWJTU	0.791933	4	RGB	Improved trajectories [16]	-	Joints motion	PCA	Fisher Vector, VLAD	SVM
Lpigou	0.788804	5	Depth, RGB	RAW, SK joints	Early	Sliding windows	Max-pooling CNN	-	CNN
stevenwudi	0.787310	6	SK, depth	RAW	Late	Sliding windows	-	-	HMM, DNN
Ismar	0.746632	7	SK	SK	-	Sliding windows	-	-	RF
Quads	0.745449	8	SK	SK quads	-	Sliding windows	-	Fisher Vector	SVM
Telepoints	0.688778	9	SK, Depth, RGB	STIPS, SK	Late	Joints motion	-	-	SVM
TUM-fortiss	0.648979	10	SK, Depth, RGB	STIPS	Late	Joints motion	-	-	RF, SVM
CSU-SCM	0.597177	11	Skeleton, Depth, mask	HOG, Skeleton	Late	Sliding windows	-	2DMTM	SVM, HMM
iva.mm	0.556251	12	Skeleton, RGB, depth	Skeleton, HOG	Late	Sliding windows	-	BoW	SVM, HMM
Terrier	0.539025	13	Skeleton	Skeleton	-	Sliding windows	-	-	RF
Team Netherlands	0.430709	14	Skeleton, Depth, RGB	MHI	Early	DTW	Preserving projections	-	SVM, RT
VecsRel	0.408012	15	Skeleton, Depth, RGB	RAW, skeleton joints	Late	DTW	-	-	DNN
Samgest	0.391613	16	Skeleton, Depth, RGB, mask	Skeleton, blobs, moments	Late	Sliding windows	-	-	HMM
YNL	0.270600	17	Skeleton	Skeleton	-	Sliding windows	-	Fisher Vector	HMM, SVM

Table 7. Track 3 Multi-modal gesture recognition results. SK: Skeleton; DNN: Deep Neural Network; RF: Ranfom Forest; 2DMTM: 2D motion trail model; RT: Regression Tree.

The authors first detected a human from the input video by SVM-HOG detector and performed forward-backward tracking. Multiple local human tracks are linked into long trajectories by spatial-temporal graph based matching. Human key poses and local dense motion trajectories were then extracted within the tracked human bounding box sequences. Second, the authors proposed a mining method that learned discriminative attributes from three feature modalities: human trajectory, key pose and local motion trajectory features. The mining framework was based on the exemplar-SVM discriminative middle level feature detection approach. The learned discriminative attributes from the three types of visual features were then mixed in a max-margin learning algorithm which also explores the combined discriminative capability of heterogeneous feature modalities. The learned mixed analyzer was then applied to the input video sequence for action detection.

Third place: The framework for detecting actions in video is based on improved dense trajectories applied on a sliding windows fashion. Authors independently trained 11 one-versus-all kernel SVMs on the labeled training set for 11 different actions. The feature and feature descriptions used are improved dense trajectories, HOG, HOF, MBHx and MBHy. During training, for each action, a temporal sliding window is applied without overlapping. For every action, a segment was labeled 0 (negative) for a certain action only if there is no frame in this segment labeled 1. The feature coding method was bag-of-features. For a certain action, the features associated with those frames which are labeled 0 (negative) are not counted when we code the features of the action for the positive segments with bag-of-features. On the basis of the labeled segments and their features, a kernel SVM was trained for each action. During testing, non-overlap sliding window was applied for feature coding of the video. Every frame in a segment was consistently labeled as the output of SVM for each action. The kernel type, sliding window size and penalty of SVMs were selected during validation. When building the bag-of-features, the clustering method was K-means and the vocabulary size is 4000. For one trajectory feature in one frame, all the descriptors were connected to form one description vector. The bag-of-features were built upon this vector.

5.3 Track 3: Multi-modal gesture recognition recognition results

Table 7 summarizes the methods of the participants that contributed to the test set of track 3. Although DTW and HMM were mainly applied in the last edition of the ChaLearn competition [8, 9], random forest has been widely applied in this edition. Also, 3 participants used deep learning architectures. Next, we describe the main characteristics of the winner methods.

First place: The proposed method was based on a deep learning architecture that iteratively learned and integrated discriminative data representations from individual channels, modeling cross-modality correlations and short- and long-term temporal dependencies. This framework combined three data modalities: depth information, grayscale video and skeleton stream (”articulated pose”). Articulated pose served as an efficient representation of large-scale body motion of the upper body and arms, while depth and video streams contained complementary information about more subtle hand articulation. The articulated pose was formulated as a set of joint angles and normalized distances between upper-body joints, augmented with additional information reflecting speed and acceleration of each joint. For the depth and video streams, the authors did not rely on hand-crafted descriptors, but on discriminatively learning joint depth-intensity data representations with a set of convolutional neural layers. Iterative fusion of data channels was performed at output layers of the neural architecture. The idea of learning at multiple scales was also applied to the temporal dimension, such that a gesture was considered as an ordered set of characteristic motion impulses, or dynamic poses. Additional skeleton-based binary classifier was applied for accurate gesture localization. Fusing multiple modalities at several spatial and temporal scales led to a significant increase in recognition rates, allowing the model to compensate for errors of the individual classifiers as well as noise in the separate channels.

Second place: The approach combined a sliding-window gesture detector with multi-modal features drawn from skeleton, color, and depth data produced by Kinect sensor. The gesture detector consisted of a set of boosted classifiers, each tuned to a specific gesture or gesture mode. Each classifier was trained independently on labeled training data, employing bootstrapping to collect hard examples. At run-time, the gesture classifiers were evaluated in a one-vs-all manner across a sliding window. Features were extracted at multiple temporal scales to enable recognition of variable-length gestures. Extracted features included descriptive statistics of normalized skeleton joint positions, rotations, and velocities, as well as HOG descriptors of the hands. The full set of gesture detectors was trained in under two hours on a single machine, and was extremely efficient at runtime, operating at 1700 fps using skeletal data.

Third place: The proposed method was based on four features: skeletal joint position feature, skeletal joint distance feature, and histogram of oriented gradients (HOG) features corresponding to left and right hands. Under the naïve Bayes assumption, likelihood functions were independently defined for every feature. Such likelihood functions were non-parametrically constructed from the training data by using kernel density estimation (KDE). For computational efficiency, k-nearest neighbor (kNN) approximation to the exact density estimator was proposed. Constructed likelihood functions were combined to the multimodal likelihood and this serves as a unary term for our pairwise Markov random field (MRF) model. For enhancing temporal coherence, a pairwise term was additionally incorporated to the MRF model. Final gesture labels were obtained via 1D MRF inference efficiently achieved by dynamic programming.

6 Discussion

This paper has described the main characteristics of the ChaLearn Looking at People 2014 Challenge which included competitions on (i) RGB human pose recovery, (ii) RGB action/interaction recognition, and (iii) multi-modal gesture recognition. Two large datasets ($HuPBA8K$ and $Montalbano$ datasets) were designed, manually-labeled, and made publicly available to the participants for a fair comparison in the performance results. Analyzing the methods used by the 25 teams that finally participated in the test set and uploaded their models, several conclusions can be drawn.

For the case of pose recovery, tree-structure models were mainly applied. Both participants used pictorial structures for inferring best configuration of body parts. 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. Most top ranked participants used SVMs, although random forests were also considered. It has been proven that removing incoherent visual words based on a background motion estimation before performing vector quantification was useful to improve the final recognition score. 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. As in our previous ChaLearn gesture recognition challenges, SVM, RF, HMM, and DTW algorithms were widely considered.

Interestingly, as said before, it is the first time that participants used deep learning architectures such as Convolutional Neural Networks, which exhibited high recognition rates. In particular, the winner of the competition used all the modalities and information of the human joints to segment gesture candidates. As expected, the code of the participants took a lot more time for training than the rest of approaches.

As a conclusion, there are still much ways for improvement in the two RGB domains considered, namely human pose recovery and action/interaction recognition from RGB data. On the other hand, for multi-modal gesture recognition, although the achieved performance make it feasible to be applied in real applications, there is still room for improvement in the precise begin-end frame level segmentation of gestures, a challenging task to perform even by humans. 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, affective computing, and face analysis as relevant information cues.

Acknowledgements

We thank the teams who participated in ChaLearn LAP 2014. Special thanks to Pau Rodríguez for annotating part of the multi-modal gestures. We thank Microsoft Codalab submission website and researchers who joined the PC and reviewed for the workshop. This work was supported by Spanish projects TIN2009-14501-C02-02, TIN2012-39051, TIN2012-38187-C03-02, and TIN2013-43478-P.

References

1. Moeslund, T., Hilton, A., Krueger, V., Sigal, L., (Eds): Visual Analysis of Humans: Looking at People. Springer-Verlag, Netherlands (2011)
2. Everingham, M., Gool, L.V., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. IJCV **88**(2) (2010) 303–338
3. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR. (2008)
4. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC. (2010) doi:10.5244/C.24.12.
5. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008) 1–8
6. De la Torre, F., Hodgins, J.K., Montano, J., Valcarcel, S.: Detailed human data acquisition of kitchen activities: the CMU-multimodal activity database (CMU-MMAC). Technical report, RI-TR-08-22h, CMU (2008)
7. Sánchez, D., Bautista, M.A., Escalera, S.: HuPBA 8k+: Dataset and ECOC-graphcut based segmentation of human limbs. Neurocomputing (2014)
8. Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Lopes, O., Guyon, I., Athitsos, V., Escalante, H.J.: Multi-modal gesture recognition challenge 2013: Dataset and results. ChaLearn Multi-modal Gesture Recognition Grand Challenge and Workshop (ICMI) (2013) 445–452
9. Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Guyon, I., Athitsos, V., Escalante, H.J., Sigal, L., Argyros, A., Sminchisescu, C., Bowden, R., Sclaroff, S.: Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. 15th ACM International Conference on Multimodal Interaction (2013) 365–368
10. Tran, D., Forsyth, D.: Improved human parsing with a full relational model. In: ECCV. IEEE (2010) 227–240
11. Ramanan, D.: Learning to parse images of articulated bodies. In: NIPS. (2006) 1129–1136
12. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: Human pose estimation: New benchmark and state of the art analysis. In: CCVPR, IEEE (2014)
13. Sapp, B., Taskar, B.: Modec: Multimodal decomposable models for human pose estimation. In: CVPR, IEEE (2013)
14. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV, IEEE (2009) 1365–1372
15. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: ICPR. Volume 3. (2004) 32–36
16. Wang, H., Schmid, C.: Action recognition with improved trajectories. ICCV (2013)
17. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE TPAMI (2013)
18. Jegou, H., Perronnin, F., Douze, M., Sanchez, J., Perez, P., Schmid, C.: Aggregating local image descriptors into compact codes. IEEE TPAMI **34**(9) (2012) 1704–1716