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

All rights reserved HuBPA©

http://gesture.chalearn.org/

Challenge on pose recovery, action/interaction, multi-modal gesture recognition

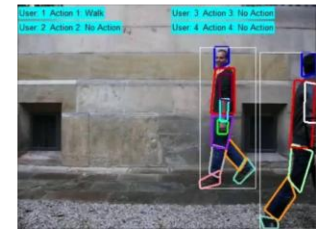
•**Track 1: Human Pose Recovery:** More than 8,000 frames of continuous RGB sequences are recorded and labeled with the objective of performing human pose recovery by means of recognizing more than 120,000 human limbs of different people.

•**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/

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



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

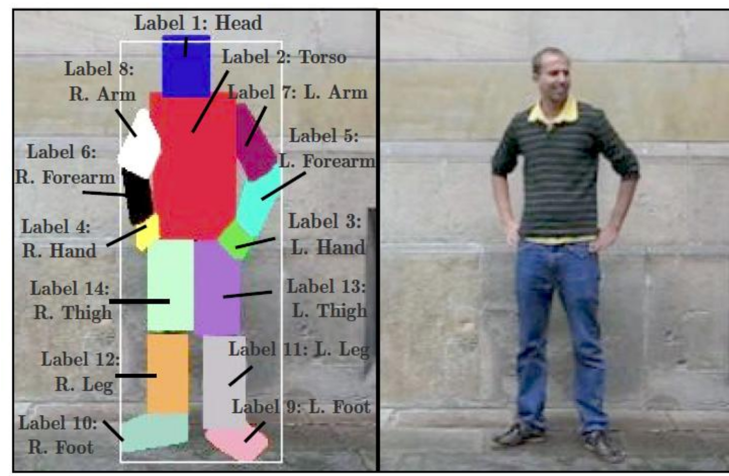
Human pose recovery data characteristics.

•**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.

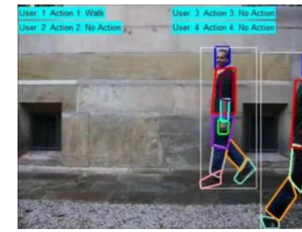
• **Limbs are manually labeled** using binary masks and the minimum bounding box containing each subject is defined.



• 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.

http://gesture.chalearn.org/

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

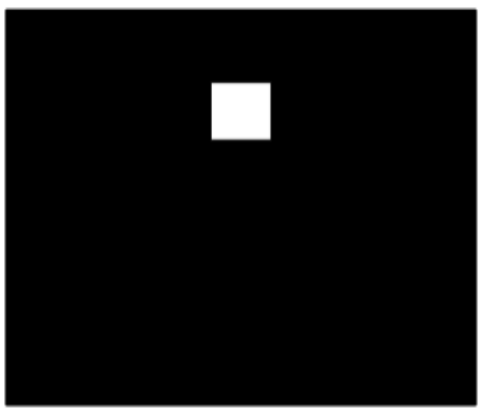


Overlap evaluation

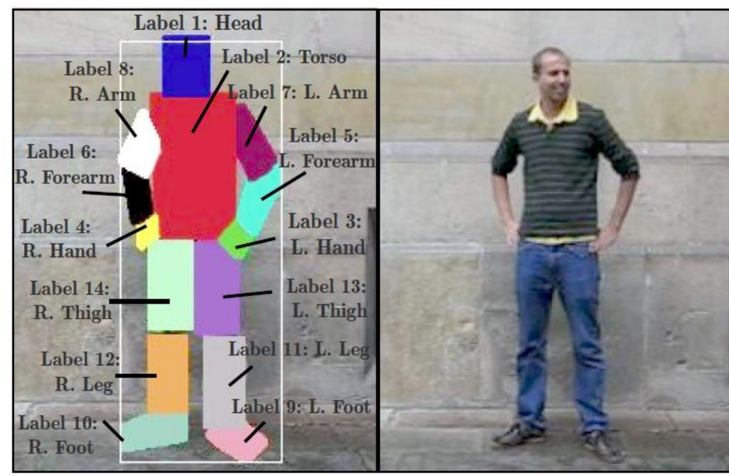
$$J_{i,n} = \frac{A_{i,n} \bigcap B_{i,n}}{A_{i,n} \bigcup B_{i,n}},$$

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

$A_{i,head}$



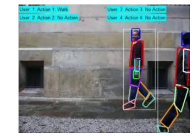
$B_{i,head}$



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

$$J_{i,head} > 0.5 \longrightarrow HR_{i,head} = 1$$

•**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.



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

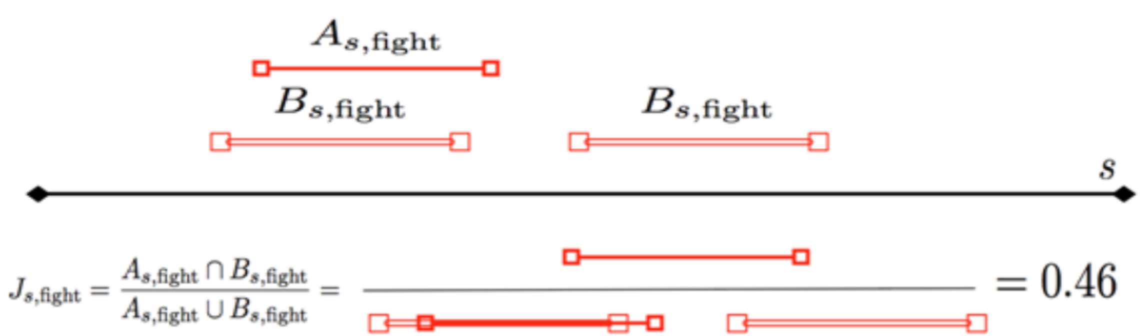
Modalities	Num. of users	Action categories	interaction categories	Labeled sequences
RGB	14	7	4	235

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.

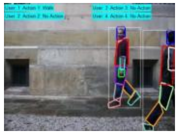
Overlap evaluation

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



•**<u>Track 2: Action/Interaction Recognition</u>:** 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.

http://gesture.chalearn.org/



Wave

Point

Clap

Crouch

Jump

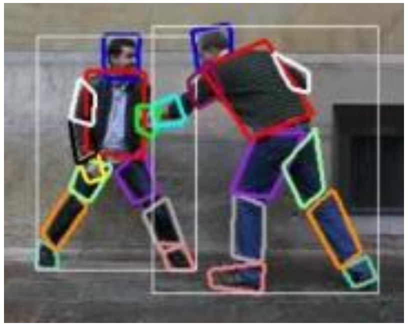
Walk

Run

Shake Hands

Hug

Kiss



Fight

http://gesture.chalearn.org/

<u>Track 3: Gesture Recognition</u>: 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 <u>user independent continuous gesture spotting</u>.

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

Main characteristics of the *Montalbano* gesture dataset.

RGB Depth User mask Skeletal model

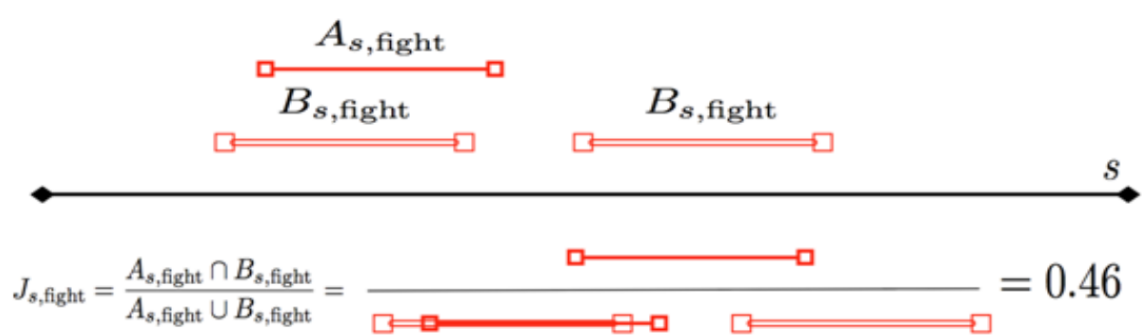
Track 3: Gesture Recognition: The gestures are drawn from a vocabulary of Italian sign gesture categories. The emphasis of this third track is on multi-modal automatic learning of a set of 20 gestures performed by several different users, with the aim of performing user independent continuous gesture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

http://gesture.chalearn.org/

- **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

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



Track 3: Gesture Recognition

http://gesture.chalearn.org/

(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*

Human Pose Recovery and Behavior Analysis Group

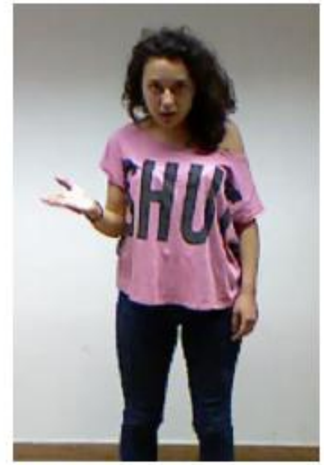
Track 3: Gesture Recognition

http://gesture.chalearn.org/

(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*

Datasets of the three challenge tracks

- State of the art comparison

	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 8K+ [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

Comparison of public dataset characteristics.

ChaLearn LAP data sets, public available at:
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/ CodaLab ALPHA

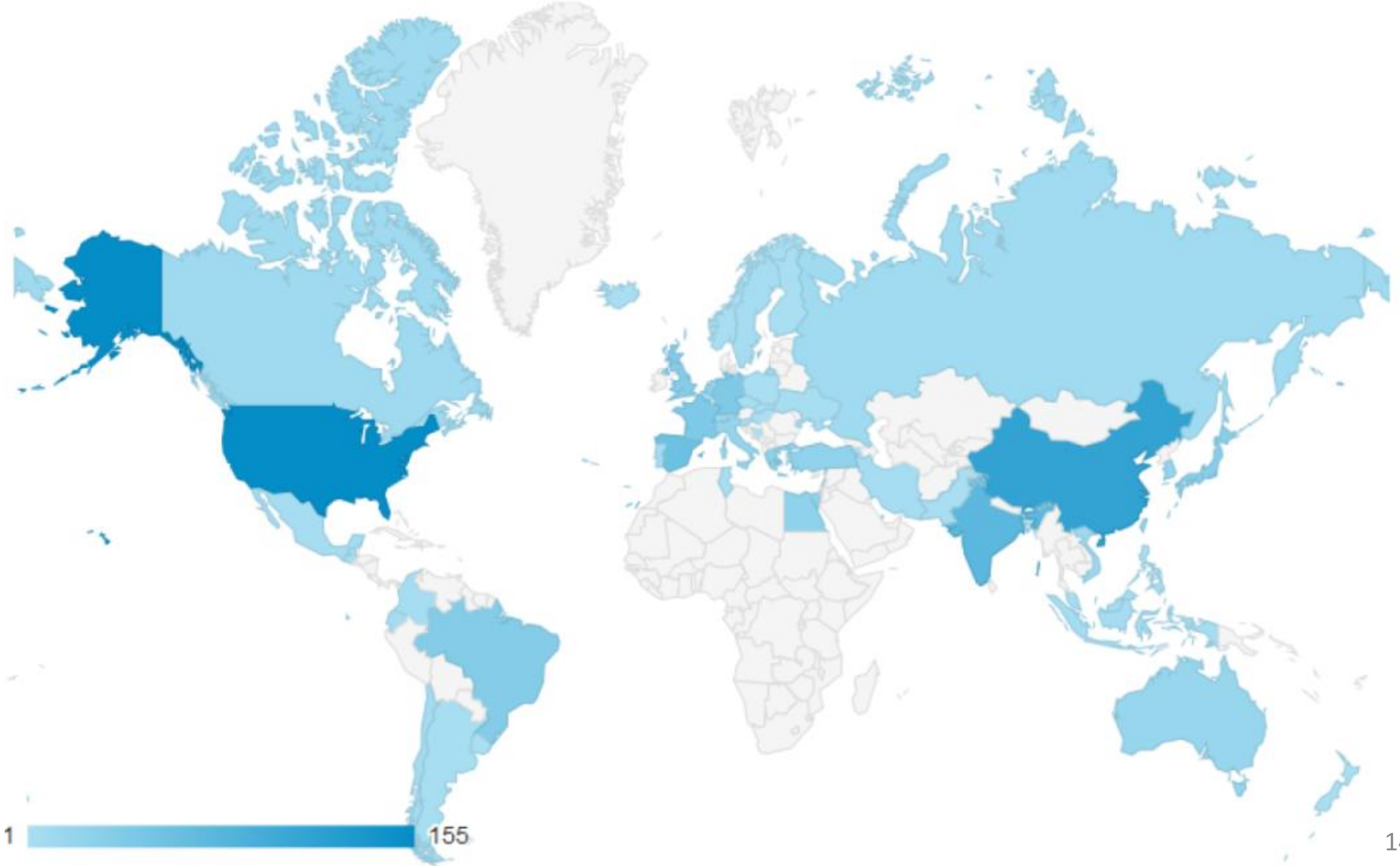
Participation

- 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

http://gesture.chalearn.org/

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



Participation

http://gesture.chalearn.org/

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

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%)

Results

- **Track1 results**

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

Track 1 Pose Recovery results.

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.

[*] Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE TPAMI (2013)

Results

- **Track2 results**

Team name	Accuracy	Rank	Features
CUHK-SWJTU	**0.507173**	1	Improved trajectories [*]
ADSC	**0.501164**	2	Improved trajectories [*]
SBUVIS	**0.441405**	3	Improved trajectories [*]
DonkeyBurger	0.342192	4	MHI, STIP
UC-T2	0.121565	5	Improved trajectories [*]
MindLAB	0.008383	6	MBF

Team name	Dimension reduction	Clustering	Classifier	Temporal coherence	Gesture representation
CUHK-SWJTU	PCA	-	SVM	Sliding windows	Fisher Vector
ADSC	-	-	SVM	Sliding windows	-
SBUVIS	-	-	SVM	Sliding windows	-
DonkeyBurger	-	Kmeans	Sparse code	Sliding windows	-
UC-T2	PCA	-	Kmeans	Sliding windows	Fisher Vector
MindLAB	-	Kmeans	RF	Sliding windows	BoW

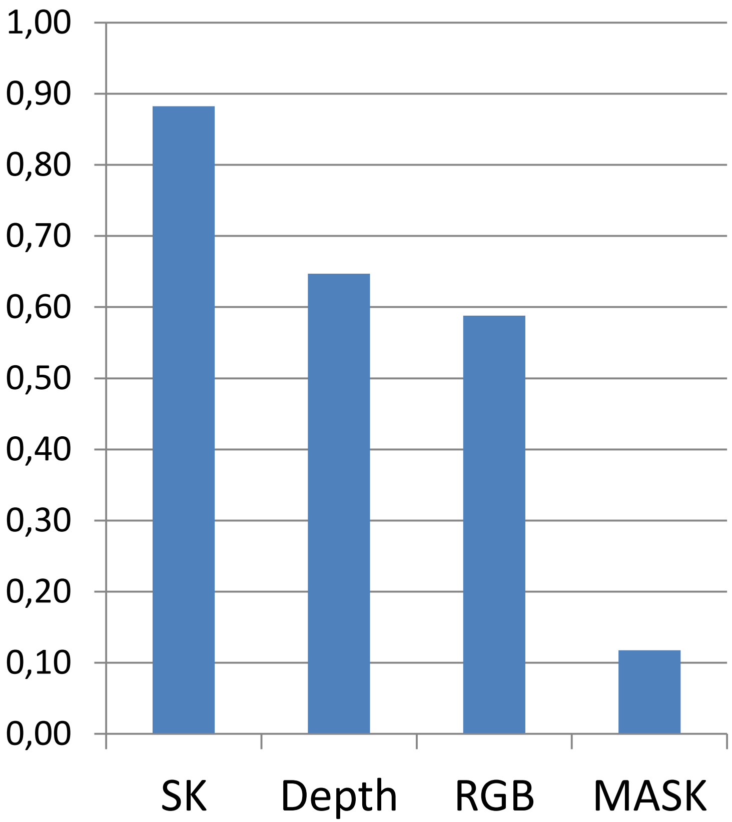
* Wang, H., Schmid, C.: Action recognition with improved trajectories. ICCV (2013)

Results

- **Track3 results**

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

Percentage of methods using each independent modality

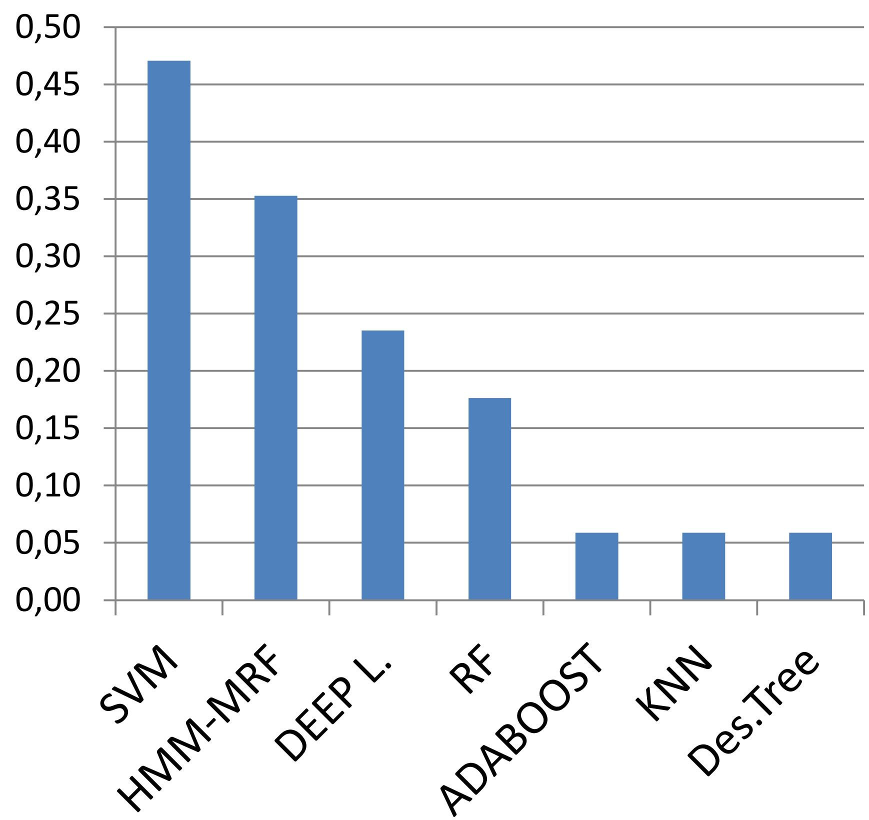


Results

- **Track3 results**

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

Percentage of methods using each gesture classification strategy



Results

- **Track3 results**

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

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.

Program

http://gesture.chalearn.org/

Organizers

HuPBA

Universitat de Barcelona

Microsoft Research

CVC Computer Vision Center

UAB Universitat Autònoma de Barcelona

ChaLearn

UOC Universitat Oberta de Catalunya www.uoc.edu

INAOE

Sponsors

Microsoft Research

Universitat de Barcelona

ChaLearn

Organizers


Sergio Escalera


Jordi Gonzàlez


Xavier Báró


Miguel Reyes


Víctor Ponce


Meysam Madadi


M. Ángel Bautista


Isabelle Guyon


Hugo J. Escalante


Jamie Shotton

http://gesture.chalearn.org/

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

We wait for your contributions!

Guest Editors

Dr. Sergio Escalera, University of Barcelona & Computer Vision Center
Dr. Jordi Gonzàlez, 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

Thank you and hope to see you in our next event!