Multi-modal Gesture Recognition Challenge 2013: Dataset and Results

Sergio Escalera, UB & CVC & ChaLearn
Jordi Gonzàlez, UAB & CVC
Xavier Baró, UOC & CVC
Miguel Reyes, UB & CVC
Oscar Lopés, CVC
Isabelle Guyon, ChaLearn
Vassilis Athitsos, Texas Univ. & ChaLearn
Hugo J. Escalante, INAOE & ChaLearn

Challenge organization

Multi-modal ChaLearn Gesture Recognition Challenge and Workshop

http://gesture.chalearn.org/
sunai.uoc.edu/chalearn

Web of the competition
Data

The challenge features a **quantitative evaluation** of automatic gesture recognition from a multi-modal dataset recorded with **Kinect** (providing RGB images of face and body, depth images of face and body, skeleton information, joint orientation and audio sources), **including 13,858 Italian gestures from near 30 users.**

The emphasis of this edition of the competition will be on multi-modal automatic learning of a **vocabulary of 20 types of Italian anthropological/cultural gestures performed by different users**, with the aim of **performing user independent continuous gesture recognition combined with audio information**.

Gesture categories (1/2)

(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*

Gesture categories (2/2)

(11) *Ok*

(12) *Cosa ti farei*

(13) *Basta*

(14) *Le vuoi prendere*

(15) *Non ce ne piu*

(16) *Ho fame*

(17) *Tanto tempo fa*

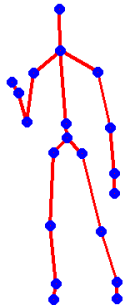
(18) *Buonissimo*

(19) *Si sono messi d'accordo*

(20) *Sono stufo*

Data and modalities



- Framerate 20FPS
- RGB: 640x480
- Depth: 640x480
- Audio: Kinect 20 michropone array
- Users: 27
- Italians: 81%

- Total number of sequences: 956 ϵ [1,2] min.
- Total number of gestures: 13,858
- Total number of frames: 1.720.800
- Noisy gestures

Data structure information: *S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopes, I. Guyon, V. Athistos, H.J. Escalante, "Multi-modal Gesture Recognition Challenge 2013: Dataset and Results", ICMI 2013.*

MMGR Challenge
Categories and data modalities

MMGR Challenge

Easy and challenging aspects of the data.

Easy
Fixed camera
Near frontal view acquisition
Within a sequence the same user
Gestures performed mostly by arms and hands
Camera framing upper body
Several available modalities: audio, skeletal model, user mask, depth, and RGB
Several instances of each gesture for training
Single person present in the visual field

Challenging
Within each sequence:
Continuous gestures without a resting pose
Many gesture instances are present
Distracter gestures out of the vocabulary may be present in terms of both gesture and audio
Between sequences:
High inter and intra-class variabilities of gestures in terms of both gesture and audio
Variations in background, clothing, skin color, lighting, temperature, resolution
Some parts of the body may be occluded
Different Italian dialects

Schedule

- **April 30th, 2013**: Beginning of the challenge competition, release of first data examples.
- **May 20th, 2013**: Full release of training and validation data. Training data with ground truth labels.
- **August 1st, 2013**: Encrypted Final evaluation data and ground truth labels for the validation data are made available.
- **August 15th, 2013**: End of the challenge competition. Deadline for code submission. The organizers start the code verification by running it on the final evaluation data and obtaining the team scores.
- **August 25th, 2013**: Deadline for fact sheets.
- **September 1st, 2013**: Release of the verification results to the participants for review.

	# Sequences	# Gesture samples
Development	393	3362
Validation	287	7754
Test	276	2742

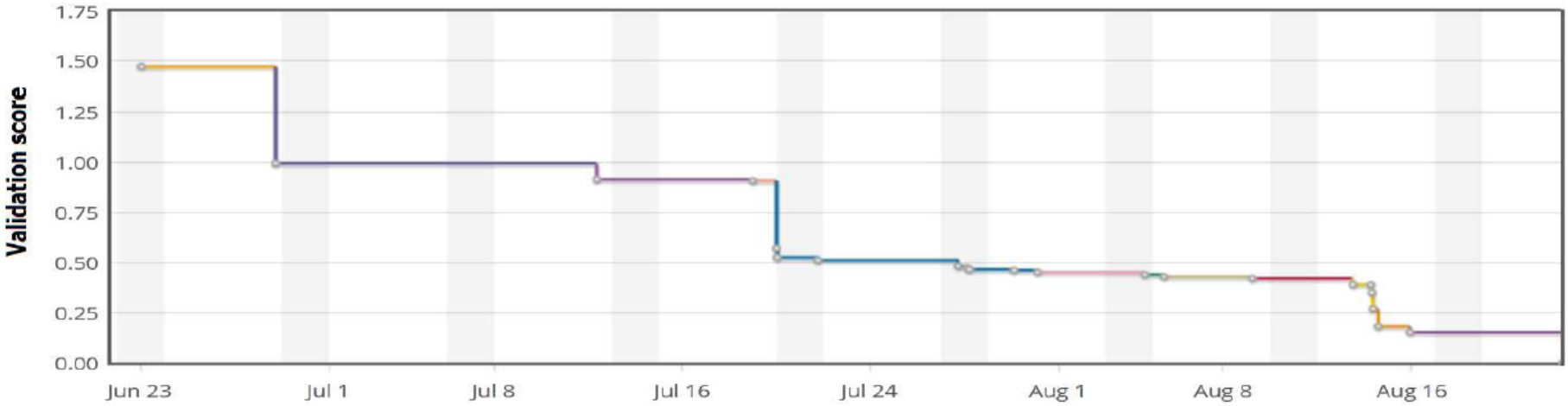
Evaluation metric and participant entries

- For each unlabeled video, the participants were instructed to provide an ordered list of labels R corresponding to the recognized gestures.
- We compared this list with the truth labels *T* i.e. the prescribed list of gestures that the user had to play during data collection.
- We computed the Levenshtein distance *L(R,T)*, that is the minimum number of edit operations (substitution, insertion, or deletion) that one has to perform to go from R to *T* (or vice versa).
- The overall score is the sum of the Levenshtein distances for all the lines of the result file compared to the corresponding lines in the truth value file, divided by the total number of gestures in the truth value file.

kaggle

$$L([124],[32]) = 2,$$
$$L([1],[2]) = 1,$$
$$L([222],[2]) = 2.$$

Evaluation metric and participant entries



Best public score obtained in the validation set during the Challenge.

kaggle

$$L([124],[32]) = 2,$$

$$L([1],[2]) = 1,$$

$$L([222],[2]) = 2.$$

Results

- Participation

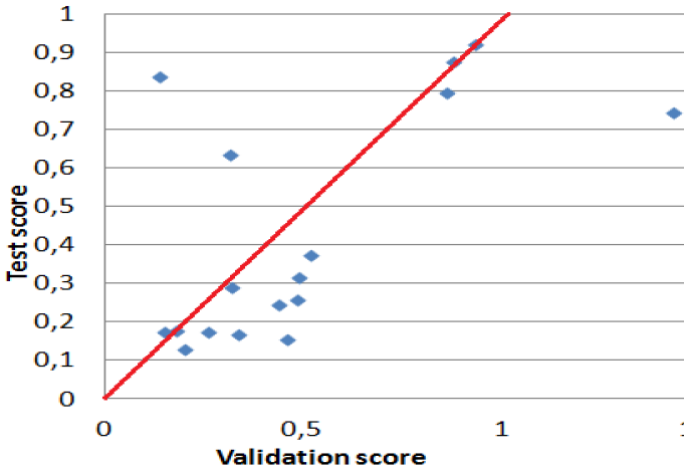
• The challenge attracted high level of participation, with a total of **54 teams and near 300 total number of entries**.
• Finally, **17 teams successfully submitted their prediction in final test set, while providing also their code for verification and summarizing their method by means of a fact sheet questionnaire**.
• After verifying the codes and results of the participants, the final scores of the top rank participants on both validation and test sets were made public.
• In the end, **the final error rate on the test data set was around 12%.**

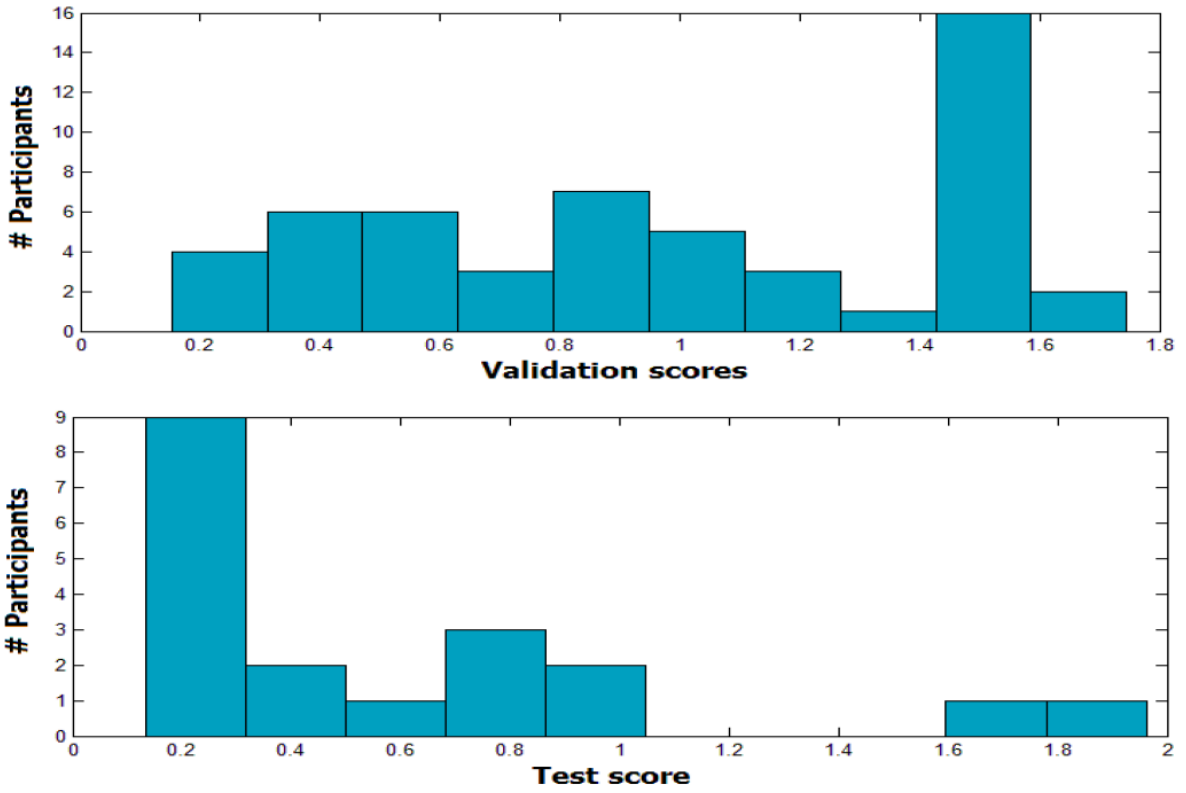
Top rank results on validation and test sets.

TEAM	Validation score	Test score
IVA MM	0.20137	**0.12756**
WWEIGHT	0.46163	**0.15387**
ET	0.33611	**0.16813**
MmM	0.25996	0.17215
PPTK	0.15199	0.17325
LRS	0.18114	0.17727
MMDL	0.43992	0.24452
TELEPOINTS	0.48543	0.25841
CSI MM	0.32124	0.28911
SUMO	0.49137	0.31652
GURU	0.51844	0.37281
AURINKO	0.31529	0.63304
STEVENWUDI	1.43427	0.74415
JACKSPARROW	0.86050	0.79313
JOEWAN	0.13653	0.83772
MILAN KOVAC	0.87835	0.87463
IAMKHADER	0.93397	0.92069

MMGR Challenge

Results





Validation and test scores histograms.

MMGR Challenge

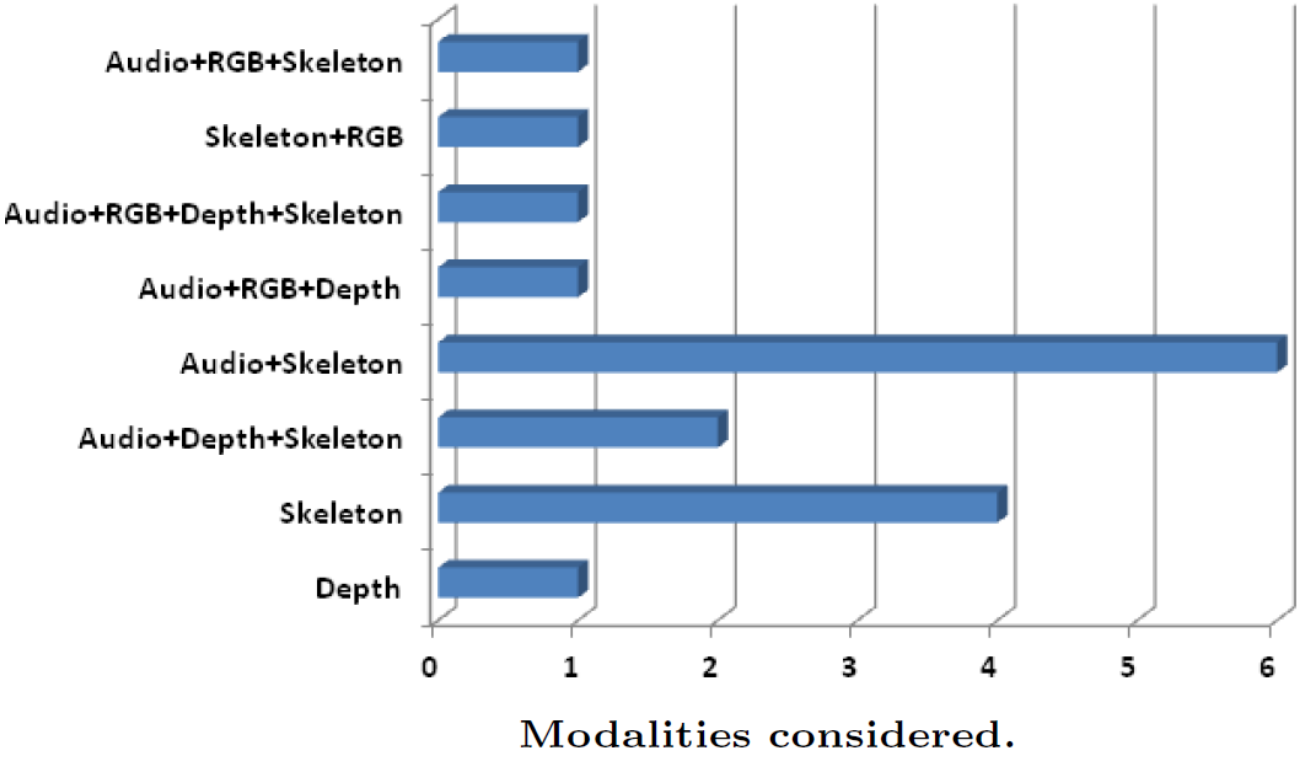
Results

Team methods and results. Early and late refer to early and late fusion of features/classifier outputs. HMM: Hidden Markov Models. KNN: Nearest Neighbor. RF: Random Forest. Tree: Decision Trees. ADA: Adaboost variants. SVM: Support Vector Machines. Fisher: Fisher Linear Discriminant Analysis. GMM: Gaussian Mixture Models. NN: Neural Networks. DGM: Deep Boltzmann Machines. LR: Logistic Regression. DP: Dynamic Programming. ELM: Extreme Learning Machines.

TEAM	Test score	Rank position	Modalities	Segmentation	Fusion	Classifier
IVA MM	0.12756	1	Audio,Skeleton	Audio	None	HMM,DP,KNN
WWEIGHT	0.15387	2	Audio,Skeleton	Audio	Late	RF,KNN
ET	0.16813	3	Audio,Skeleton	Audio	Late	Tree,RF,ADA
MmM	0.17215	4	Audio,RGB+Depth	Audio	Late	SVM,Fisher,GMM,KNN
PPTK	0.17325	5	Skeleton,RGB,Depth	Sliding windows	Late	GMM,HMM
LRS	0.17727	6	Audio,Skeleton,Depth	Sliding windows	Early	NN
MMDL	0.24452	7	Audio,Skeleton	Sliding windows	Late	DGM+LR
TELEPOINTS	0.25841	8	Audio,Skeleton,RGB	Audio,Skeleton	Late	HMM,SVM
CSI MM	0.28911	9	Audio,Skeleton	Audio	Early	HMM
SUMO	0.31652	10	Skeleton	Sliding windows	None	RF
GURU	0.37281	11	Audio,Skeleton,Depth	DP	Late	DP,RF,HMM
AURINKO	0.63304	12	Skeleton,RGB	Skeleton	Late	ELM
STEVENWUDI	0.74415	13	Audio,Skeleton	Sliding windows	Early	DNN,HMM
JACKSPARROW	0.79313	14	Skeleton	Sliding windows	None	NN
JOEWAN	0.83772	15	Skeleton	Sliding windows	None	KNN
MILAN KOVAC	0.87463	16	Skeleton	Sliding windows	None	NN
IAMKHADER	0.92069	17	Depth	Sliding windows	None	RF

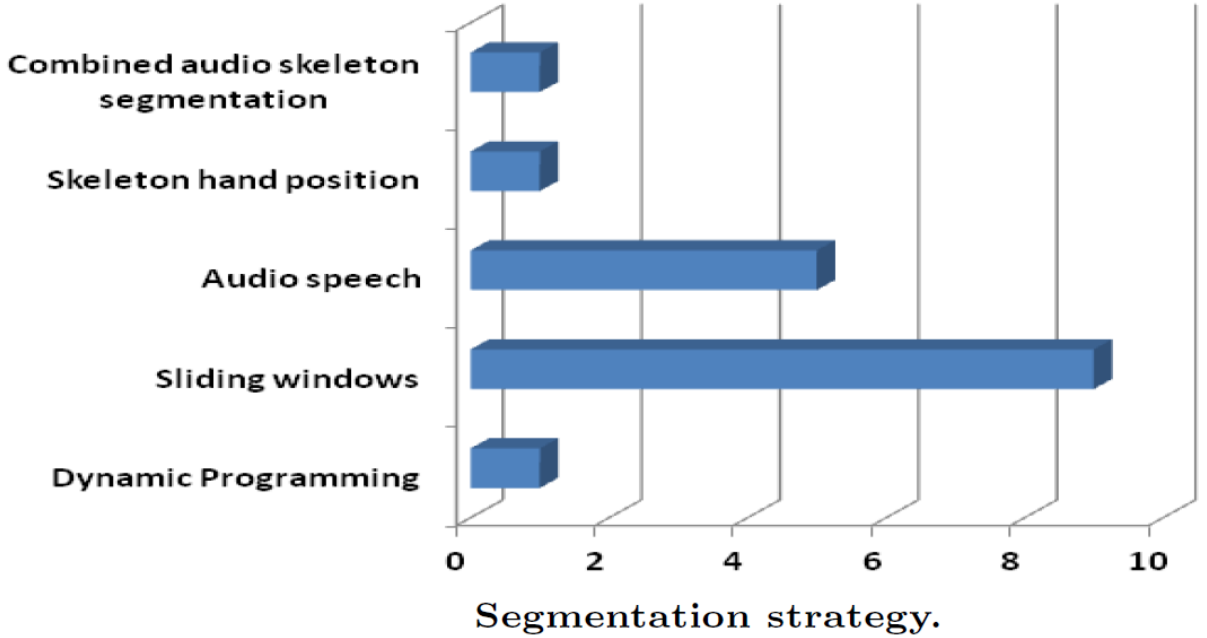
Results

- Fact sheets statistics

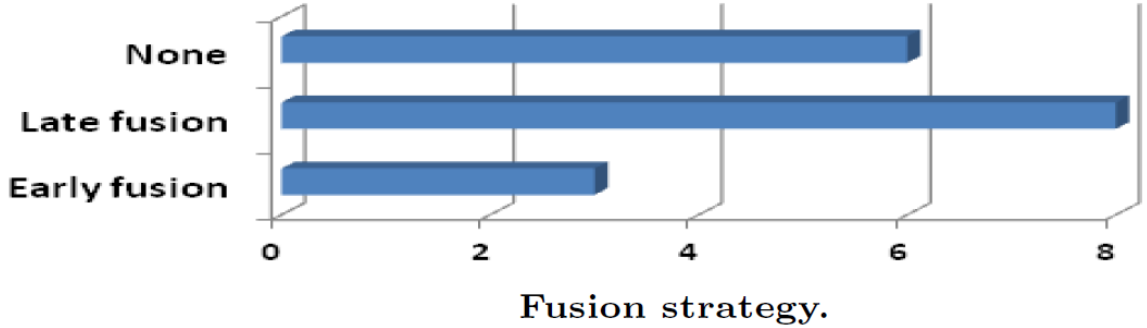


Results

- Fact sheets statistics



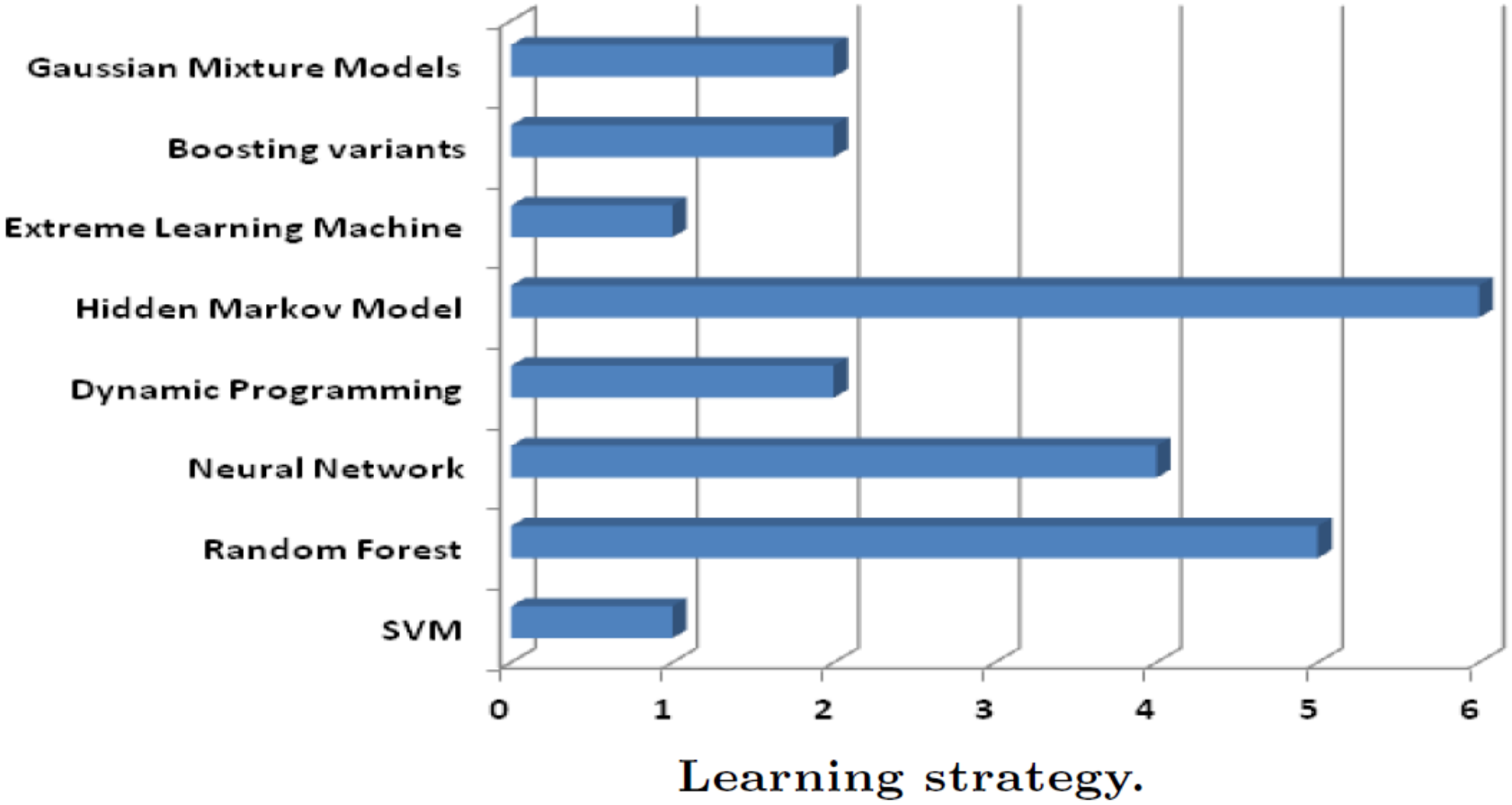
Segmentation strategy.



Fusion strategy.

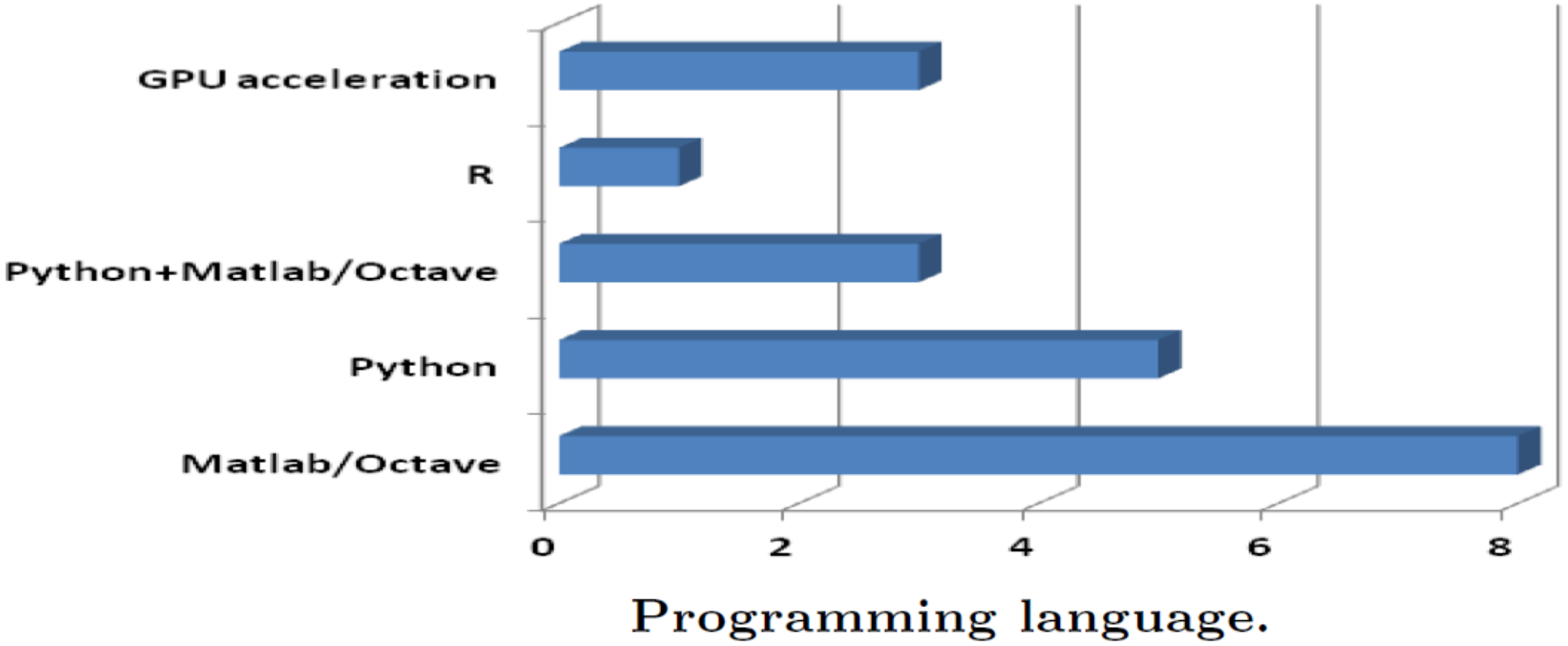
Results

- Fact sheets statistics



Results

- Fact sheets statistics



Results

- Winner methods

TEAM	Validation score	Test score
IVA MM	0.20137	**0.12756**
WWEIGHT	0.46163	**0.15387**
ET	0.33611	**0.16813**

TEAM	Test score	Rank position	Modalities	Segmentation	Fusion	Classifier
IVA MM	0.12756	1	Audio,Skeleton	Audio	None	HMM,DP,KNN
WWEIGHT	0.15387	2	Audio,Skeleton	Audio	Late	RF,KNN
ET	0.16813	3	Audio,Skeleton	Audio	Late	Tree,RF,ADA

The first ranked team IVAMM

•Feature vector based **on audio and skeletal** information and applied **late fusion** to obtain final recognition results.

•A simple **time-domain end-point detection algorithm based on joint coordinates is applied to segment continuous data sequences into candidate** gesture intervals.

•A **Gaussian Hidden Markov Model is trained with 39-dimension MFCC** features and generates confidence scores for each gesture category.

•A **Dynamic Time Warping based skeletal feature classifier** is applied to provide complementary information.

•The confidence scores generated by the two classifiers are firstly normalized and then combined to produce a **weighted sum**.

•A single **threshold approach** is employed to classify meaningful gesture intervals from meaningless intervals caused by false detection of speech intervals.

Fusing Multi-modal Features for Gesture Recognition, Jiaxiang Wu, Jian Cheng, Chaoyang Zhao and Hanqing Lu, ChaLean MMGR, ICMI, 2013.

Results

- Winner methods

TEAM	Validation score	Test score
IVA MM	0.20137	**0.12756**
WWEIGHT	0.46163	**0.15387**
ET	0.33611	**0.16813**

TEAM	Test score	Rank position	Modalities	Segmentation	Fusion	Classifier
IVA MM	0.12756	1	Audio,Skeleton	Audio	None	HMM,DP,KNN
WWEIGHT	0.15387	2	Audio,Skeleton	Audio	Late	RF,KNN
ET	0.16813	3	Audio,Skeleton	Audio	Late	Tree,RF,ADA

The second ranked team WWEIGHT

•**Audio and skeletal information**, using both joint spatial distribution and joint orientation.

•The method first searches for regions of time with **high audio-energy to define 1.8-second-long windows of time that potentially contained a gesture.**

•Feature vectors are then defined using a **log-spaced audio spectrogram and the joint positions and orientations above the hips.**

•There were **1593 features total** (9 time samples × 177 features per time sample). Since some of the detected windows can contain distracter gestures, an extra 21st label is introduced, defining the 'not in the dictionary' gesture category.

•An **ensemble of randomized decision trees** (ExtraTreesClassifier, 100 trees, 40% of features) **and a K-Nearest Neighbor model (7 neighbors, L1 distance). The posteriors from these models are averaged with equal weight.**

ExtraTreesClassifier Feature Importance.

Results

- Winner methods

TEAM	Validation score	Test score
IVA MM	0.20137	**0.12756**
WWEIGHT	0.46163	**0.15387**
ET	0.33611	**0.16813**

TEAM	Test score	Rank position	Modalities	Segmentation	Fusion	Classifier
IVA MM	0.12756	1	Audio,Skeleton	Audio	None	HMM,DP,KNN
WWEIGHT	0.15387	2	Audio,Skeleton	Audio	Late	RF,KNN
ET	0.16813	3	Audio,Skeleton	Audio	Late	Tree,RF,ADA

The third ranked team ET

•The features considered were **skeleton information and audio signal**.

•**Combined the output decisions** of two designed approaches.

•In the first approach, they look for **gesture intervals (unsupervised) using the audio files and extracts features from this intervals (MFCC)**. Using these features, authors train a **random forest and gradient boosting classifier**.

•The **second approach uses simple statistics** (median, var, min, max) on the first 40 frames for each gesture to build the training samples.

•The prediction phase uses a **sliding window**. The authors create a **weighted average of the output of these two models**.

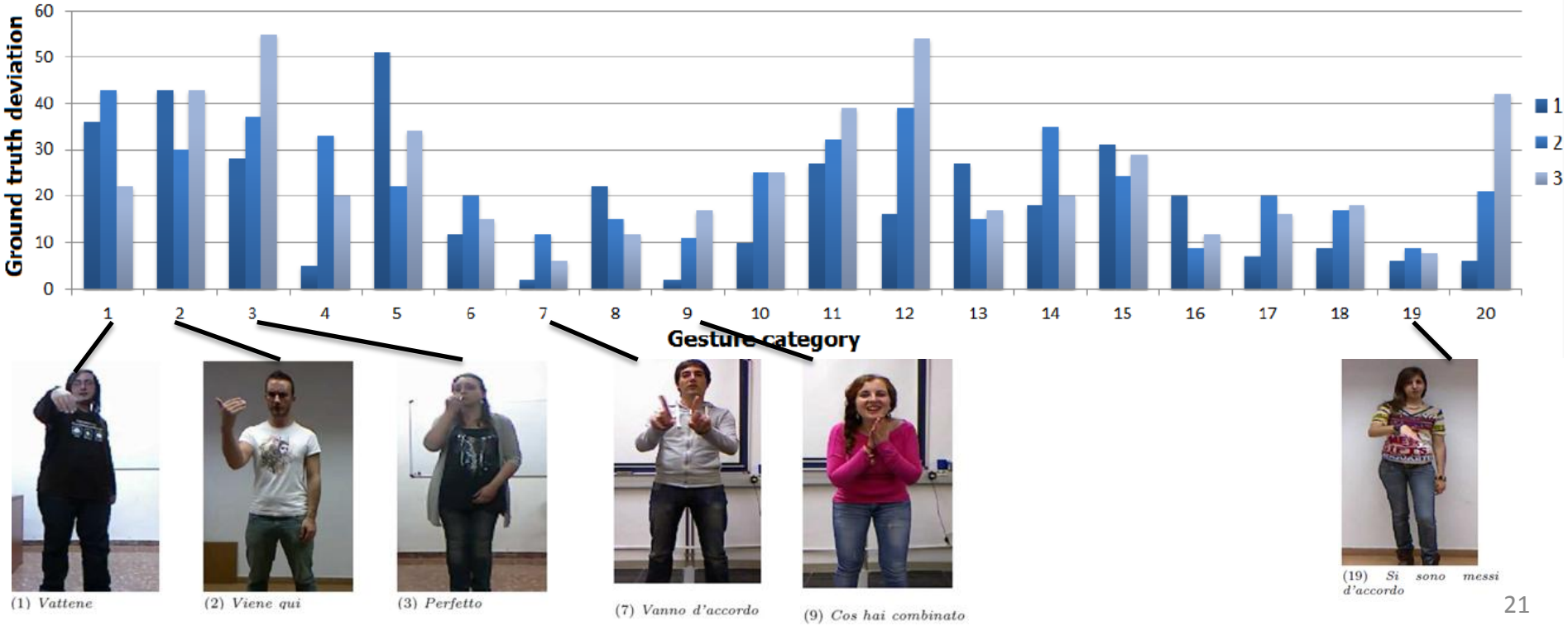
A Multi Modal Approach to Gesture Recognition from Audio and Video Data, Immanuel Bayer and Thierry Silbermann, ChaLearn MMGR, ICMI 2013.

Results

- Winner methods

TEAM	Validation score	Test score
IVA MM	0.20137	**0.12756**
WWEIGHT	0.46163	**0.15387**
ET	0.33611	**0.16813**

Correlation of the number of recognized gestures per category and GT gestures (averaged among all sequences)



Thank you

Organizers

Sponsors