# Results and Analysis of ChaLearn LAP Multi-modal Isolated and Continuous Gesture Recognition, and Real versus Fake Expressed Emotions Challenges

Jun Wan
NLPR, CASIA, China
jun.wan@ia.ac.cn

Sergio Escalera
CVC, UB, Spain
sergio@maia.ub.es

Gholamreza Anbarjafari
iCV, UT, Estonia
shb@icv.tuit.ut.ee

Hugo Jair Escalante,
INAOE, Puebla, Mexico
hugojair@inaoep.mx

Xavier Baró
UOC, CVC, Spain
xbaro@uoc.edu

Isabelle Guyon
U. Paris-Saclay, France & ChaLearn, USA
guyon@chalearn.org

Meysam Madadi
CVC, UAB, Spain
mmadadi@cvc.uab.es

Juri Allik
Institute of Psychology, UT, Estonia
juri.allik@ut.ee

Jelena Gorbova
iCV, UT, Estonia
lena@icv.tuit.ut.ee

Chi Lin, Yiliang Xie
MUST, Macau, China
{linantares,microos316}@gmail.com

## Abstract

*We analyze the results of the 2017 ChaLearn Looking at People Challenge at ICCV. The challenge comprised three tracks: (1) large-scale isolated (2) continuous gesture recognition, and (3) real versus fake expressed emotions tracks. It is the second round for both gesture recognition challenges, which were held first in the context of the ICPR 2016 workshop on "multimedia challenges beyond visual analysis". In this second round, more participants joined the competitions, and the performances considerably improved compared to the first round. Particularly, the best recognition accuracy of isolated gesture recognition has improved from 56.90% to 67.71% in the IsoGD test set, and Mean Jaccard Index (MJI) of continuous gesture recognition has improved from 0.2869 to 0.6103 in the ConGD test set. The third track is the first challenge on real versus fake expressed emotion classification, including six emotion categories, for which a novel database was introduced. The first place was shared between two teams who achieved 67.70% averaged recognition rate on the test set. The data of the three tracks, the participants' code and method descriptions are publicly available to allow researchers to keep making progress in the field.*

## 1. Introduction

The goal of the, so called, looking at people (LAP) computer vision subfield is to develop automated tools for the visual analysis of human behavior in all of its forms. There are many tasks that can be framed within LAP, most notably, human action recognition, pose estimation and face analysis. Methods for LAP are used in a number of applications, including, human computer interaction, security, health care and rehabilitation, entertainment, among many others. Therefore, research on this topic has impact in several domains and scenarios.

We organized a challenge around two landmark LAP problems: *gesture* and *emotion* recognition. Although both tasks have been studied extensively in the past (see, e.g. [1, 2, 3, 4, 5]), we consider two settings of practical importance that have not been studied deeply. On the one hand, we organize a challenge on large scale multimodal gesture recognition. Contrary to previous challenges on gesture recognition (see [1]), this competition aims to develop methods that can recognize hundreds of categories coming from quite diverse domains. Two tracks are considered on this task: gesture recognition (from segmented video) and spotting (from continuous video). This is a second round of challenges for both tasks. In a first round, impressive progress was obtained [6], this challenge further pushes the state of the art in this pretty much relevant topic.

On the other hand, we also approach a novel problem

within facial emotion recognition: the problem of determining whether a given emotion is fake or not. In contrast with previous work on emotion recognition targeting *apparent* emotions, we aim at recognizing whenever an emotion is genuine. Although this is a daunting task, results obtained by participants were promising, exceeding our initial expectations. To the best of our knowledge this is the first challenge of its kind.

This paper provides an overview of the challenge, including a detailed description of the approached tasks, data, evaluation protocol, summary of results and the main findings derived from the challenge. The challenge attracted many participants (132 for the three tasks). Impressive results were obtained for the gesture recognition tracks and promising results were achieved in the emotion recognition problem. The data sets used for evaluation have been published and will remain publicly available so they can become widely used benchmarks to push the state of the art in LAP.

The rest of this paper is organized as follows. The next section provides an overview of the different tracks of the challenge. Next, Section 3 presents the gesture recognition tracks of the challenge. Section 4 describes the emotion recognition challenge. Finally, Section 5 outlines conclusions derived from this work.

## 2. Contest Overview

This section provides generic information about the three tracks which belong to the series of Chalearn LAP events [1]. Common to the three tracks is the evaluation protocol. For each track, training, validation and training data sets were provided. Training data were released labeled, validation data were used to provide feedback to participants in a leaderboard and test data were used to determine the winners. Note that each track had its own evaluation metrics. The three tracks were run in the CodaLab platform[2]. Top three ranked participants for each track will be eligible for prizes. The baseline methods and scores for all the tracks are also provided.

The challenge comprised two stages: development and final phases.

- **Development Phase:** Participants had access to labeled development (training) and validation data, with ground-truth labels in track 1 and 2 (gesture recognition challenges, round 2), while emotion challenges provided training data and unlabeled validation data. During this phase, participants could receive immediate feedback on their performance on validation data through the leaderboard in CodaLab.
- **Final Phase:** The unlabeled final (test) data were provided for all 3 tracks. The winners of the contest were determined by evaluating performances on these 3 datasets. The participants also had to send code and fact sheets describing their methods to challenge organizers. All the code of participants was verified and replicated prior to announcing the winners.

To be eligible for prizes, the winners had to publicly release their code and fact sheet.

## 3. Large-scale Isolated and Continuous Gesture Recognition Challenges

Tracks 1 and 2 focused on the problems of isolated and continuous gesture recognition, respectively (round 2), where the focus was on recognizing gestures from either segmented or continuous RGB-D videos. The first round of both challenges was previously held in conjunction with the ICPR 2016 contest program (see [6] for results and findings). It attracted 12 and 5 participating teams on the learning and final evaluation stages for track 1 and 2, respectively. And there are 8 teams' performances are better than our baseline method or the best performance of the first round (5 teams for track 1 and 3 teams for track 2). In total 79 participants were registered for both challenge tracks.

### 3.1. Data

Associated with these tracks we recently released two large-scale gesture recognition data sets [7]:

- **Chalearn LAP RGB-D Isolated Gesture Dataset (IsoGD)[3].** It includes 47933 RGB-D gesture videos. Each RGB-D video represents one gesture only, and there are 249 gesture labels performed by 21 different individuals. This data set was used for track 1: isolated gesture recognition, and the goal was to recognize the categories of gestures in pre-segmented RGB-D videos.
- **Chalearn LAP RGB-D Continuous Gesture Dataset (ConGD)[4].** It comprises 47933 RGB-D gestures in 22535 RGB-D gesture videos. Each RGB-D video may represent one or more gestures, and there are 249 gesture labels performed by 21 individuals. This data set was used for track 2, and the focus was on segmenting and recognizing gestures from continuous video (gesture spotting).

Both the IsoGD and ConGD databases were divided into three sub-data sets for evaluation (recorded by Microsoft Kinect 1, 320×240, 10fps), whereby the subsets are mutually exclusive. For more information about these two data

[1] http://chalearnlap.cvc.uab.es/
[2] https://competitions.codalab.org/
[3] http://www.cbsr.ia.ac.cn/users/jwan/database/isogd.html
[4] http://www.cbsr.ia.ac.cn/users/jwan/database/congd.html

Figure 1. Examples of gestures from the IsoGD and ConGD.

sets, please refer to [7]. Some examples are presented in Figure 1.

### 3.2. Metrics and Evaluation

For the isolated gesture recognition challenge, we used the recognition rate $r$ as the evaluation criteria:

$$r = \frac{1}{n}\sum_{i=1}^{n}\delta(p_l(i), t_l(i)) \quad (1)$$

where $n$ is the number of samples; $p_l$ is the predicted label; $t_l$ is the ground truth; $\delta(j_1, j_2) = 1$, if $j_1 = j_2$, otherwise $\delta(j_1, j_2) = 0$.

For continuous gesture recognition, we used the Jaccard Index (the higher the better), similarly to previous ChaLearn Looking at People challenges [8, 9]. The Mean Jaccard Index (MJI) measures the average relative overlap between true and predicted sequences of frames for a given gesture. Metric description details for both tracks can be found in [7].

### 3.3. Results and Methods

In the following, we first report the details of isolated and continuous gesture challenges respectively, and then give a brief conclusion for each track.

#### 3.3.1 Isolated Gesture Recognition Challenge

Table 1 shows the final ranking of the isolated gesture recognition challenge, where results of five teams/participants and a new baseline [10] have been reported. For completeness, we report in that table the performances obtained in rounds 1 & 2. Compared with the performances of the first round, the best recognition rate $r$ obtained in round 2 improved considerably (from 56.90% to 67.71% on the test set). We notice that the new baseline [10] also achieved the second best performance. This baseline uses multiple modalities (RGB, depth, optical flow and saliency streams) and a spatio-temporal network architecture, with a consensus-voting strategy (see [10] for details).

| Rank by test set | Team | $r$ (valid set) | $r$ (test set) |
|---|---|---|---|
| ROUND 2 | | | |
| 1 | ASU | 64.40% | 67.71% |
| 2 | SYSU_ISEE | 59.70% | 67.02% |
| 3 | Lostoy | 62.02% | 65.97% |
| 4 | AMRL | 60.81% | 65.59% |
| 5 | XDETVP | 58.00% | 60.47% |
| - | baseline [10] | 49.17% | 67.26% |
| ROUND 1 | | | |
| 1 | FLiXT [11] | 49.2% | 56.90% |
| 2 | AMRL [12] | 39.23% | 55.57% |
| 3 | XDETVP-TRIMPS [13] | 45.02% | 50.93% |
| - | baseline [7] | 18.65% | 24.19% |

Table 1. Summary of the results in the isolated gesture recognition challenge (Rounds 1 & 2).

Table 2 shows a brief summary of each participants/teams' methodology. It can be seen that most participants used C3D [14] and/or LSTM neural networks using as input modalities RGB-D, flow and/or skeleton. In the remainder of this section we summarize the methods of the top ranking participants.

**First place (ASU):** This method includes four parts. First, a data enhancement strategy based on RGB and depth data is used, which are retinex for unifying the illumination of RGB video and median filter for eliminating noise in depth videos. Additionally, optical flow information is generated as another modality of data, which capture the gesture motions. Then, two different sampling strategies are adopted. One is uniform sampling and the other is sectional weighted sampling. After that, the C3D model [14] and Temporal Segment Network [15] (TSN) are used for feature extraction. Later, features extracted from the same modality are fused in terms of canonical correlation analysis and features from different modalities are fused by stacking. To train and test the models, it took us about 19.4 hours (using a graphic card with 10G memory) for C3D and 14.2 hours for TSN (using a graphic card with 4-6G memory). Clas-

| Team | Pre-trained | Pre-process | Modality | Data | Fusion or Classify |
|---|---|---|---|---|---|
| ASU | C3D[1] (Sports-1M) | data enhancement | C3D, TSN[2] | RGB-D, flow | SVM |
| SYSU_ISEE | VGG16 (UCF-101) | Rank Pooling[3], RMPE[4] | LSTM, VGG16 | RGB-D, flow, skeleton | Score fusion |
| Lostoy | C3D (Sports-1M) | openpose for hand cropping | C3D, ResNet-18 | RGB-D | Score fusion |
| AMRL | ResNet-50 (ImageNet, SKIG) | – | ConvLSTM, Resnet-50, C3D | RGB-D | Score Fusion |
| XDETVP | – | – | LSTM, C3D | RGB-D, flow | SVM |

1. C3D [14]: 3d convolutional networks; 2. TSN [15]: Temporal segment networks; 3. Rank Pooling [16]; 4 RMPE [17]: Regional Multi-person Pose Estimation;

Table 2. Overview of the team methods in the isolated gesture recognition challenge (Round 2).

sification is performed by a linear-SVM classifier to limit the complexity of the final stage. The experiments are processed on a PC with Intel Core i7-6700 CPU @ 3.40GHz, 16 GB RAM and Nvidia TITAN X GPU.

**Second place (SYSU_ISEE):** The SYSU_ISEE team considered modeling both dynamic and static action cues for gesture recognition. For the dynamic cues, the method learned discriminative motion features from RGB-D videos, optical flow sequences, and skeletons. The skeleton information was estimated via the Regional Multi-person Pose Estimation [17] (RMPE) algorithm. For the static action cues, it employed the rank pooling method [16] to represent all the optical flow frames and depth frames. All of them (except skeletons) were entered into the VGG-16 network separately to fuse information. The skeletons were processed separately by deep LSTM network to learn the temporal dependencies. Robust recognition results were attained by a late fusion of the VGG-16 and LSTM network prediction scores. The basic model used in this method is VGG-16 and the count of parameter is about 135 millions.

**Third place (Lostoy):** Participants argued that CNN based models can easily overfit to background, clothing etc. for gesture recognition (like the IsoGD dataset). Thus, this team proposed a masked C3D method for gesture recognition, which is simple to implement and yet provide useful guidance for CNN. It applied the pose estimation method to detect the hand locations and regions outside of hand bounding boxes are set to 0. Then, the masked RGB-D images are used to learn C3D model [14] for classification. The whole system was implemented with Pytorch. The training stage was carried out on a 4 x Titan X(Maxwell) GPUs with 6GB GPU memory footage for each GPU. Each training stage cost 6 hours. The testing time was about 1-2 minutes.

**Fourth place (AMRL):** The AMRL team proposed a multimodal gesture recognition method based on heterogeneous networks. Convolutional neural networks (CNNs) and convolutional LSTM networks [18] (ConvLSTM) are used to construct a heterogeneous network that combines the representation capability of ConvLSTM and CNNs in the temporal and spatial domain. Firstly, the proposed method represents the RGB and depth image sequence into body dynamic image and hand dynamic image as the inputs of CNNs respectively through bidirectional rank pooling. Then it learns short-term spatiotemporal features of gestures through 3D convolutional neural network, and learns long-term spatiotemporal features based on the extracted short-term spatiotemporal features. To learn fine-grained levels spatiotemporal features, the Faster R-CNN [19] is used to detect the hand part. This proposed method based on heterogeneous network can learn different levels of complementary spatiotemporal features.

**Fifth place (XDETVP):** The XDETVP team presented a multimodal gesture recognition method based on 3-D convolutional neural networks and convolutional Long-Short-Term-Memory (LSTM) networks. First, it learns the short-term and long-term spatiotemporal features with 3DCNN and convLSTM networks [20]. Then, the CNN networks are applied to recognize gestures based on learned 2D spatio-temporal feature maps. The features of the three modalities (RGB, Depth, Flow) obtained by the temporal pooling layer are combined to construct feature vectors to train and test SVM classifiers. For training the networks, it costs about three days on TITAN X (GPU) for a single modality.

### 3.3.2 Continuous Gesture Recognition Challenge

The final ranking of three teams/participants that entered the final phase for the continuous gesture recognition challenge is reported in the Table 3. As before, we report results for rounds 1 and 2. The table shows that the best Mean Jaccard Index (MJI) has improved considerably (from 0.2869 to 0.6103 on the test set) in the second round, compared with the performances of the first round. Additionally, Table 4 shows a brief summary of each participant/team. In the remainder of this section we summarize their methodologies.

**First place (ICT_NHCI):** First, the RGB and depth image frames are calibrated and hand regions are detected via

| Team | Pre-trained | Pre-process | Modality | Data | Fusion or Classify |
|---|---|---|---|---|---|
| ICT_NHCI | C3D (Sports-1M), VGG (ImageNet) | face and hand detection | Faster-RCNN, C3D | RGB-D | SVM |
| AMRL | ResNet-50 (ImageNet, SKIG) | – | Conv. LSTM, C3D, Resnet-50 | RGB-D | Score Fusion |
| PaFiFA | – | – | 3D CNN [22] | RGB-D | Score Fusion |
| Deepgesture | – | – | bidirectional LSTM, CNN | RGB | Softmax |

Table 4. Overview of the team methods in the continuous gesture recognition challenge (Round 2).

| Rank by test set | Team | MJI (valid set) | MJI (test set) |
|---|---|---|---|
| ROUND 2 | | | |
| 1 | ICT_NHCI | 0.5163 | 0.6103 |
| 2 | AMRL | 0.5957 | 0.5950 |
| 3 | PaFiFA | 0.3646 | 0.3744 |
| 4 | Deepgesture | 0.3190 | 0.3164 |
| ROUND 1 | | | |
| 1 | ICT_NHCI [21] | 0.2655 | 0.2869 |
| 2 | TARDIS [22] | 0.2809 | 0.2692 |
| 3 | AMRL [23] | 0.2403 | 0.2655 |
| - | baseline [7] | 0.0918 | 0.1464 |

Table 3. Final ranking in the ConGD dataset (Rounds 1 & 2).

a two-streams Faster R-CNN method. Thus, the continuous gesture sequence can be segmented into several isolated gestures via the temporal segmentation. In order to represent each gesture by the hand posture and location information, the face region is located and the relative hand locations are encoded into the 3D convolution features. The face region only is considered in the RGB image while in the depth channel, the face region is not added because of the coarse calibration. Then the hand spatiotemporal features were extracted by the C3D model [14]. Lastly, RGB and depth features are fused and provided to a SVM classifier to recognize gestures. It took about 5 hours to perform temporal segmentation using MATLAB, 80 hours to train the RGB and hand detection models, 60 hours to detect hands (in one TITAN X GPU), 4 hours to detect faces, 50 hours to fine-tune the C3D model, 1.5 hours for extracting the last layer features, and 20 minutes to train the SVM classifier. In the testing stage for the whole test set, it took about 15 hours for hand detection (one TITAN X GPU), 0.5 hours for face detection, 0.5 hour for temporal segmentation, 0.5 hours for feature extraction, and 5 minutes to get the recognition results.

**Second place (AMRL):** The AMRL team first segmented isolated gestures from the depth sequence based on quantity of movement (QOM) [12], then used the heterogeneous networks to recognize gestures, which were introduced in Sec. 3.3.1 for the fourth place of isolated gesture recognition challenge.


Figure 2. In the IsoGD, some gesture classes are easy to fused. (a) Gesture label (static): 11; (2) Gesture label (dynamic): 26.

**Third place (PaFiFA):** An end-to-end deep neural network was proposed based on raw RGB video pixels with temporal convolutions and bidirectional LSTM networks [24]. The model used 20 non-linearity layers with 824,233 parameters and was trained without depth images nor external data. In the preprocessing stage, RGB was converted to gray-scale and the preceding frame was subtracted. The depth images were not used. The model uses residual connections [25], ELU non-linearities [26], temporal convolutions and recurrence (LSTM) [24], batch normalization [27] and data augmentation. For evaluation, a sliding window of 32 frames was used with a stride of 16 for each 32 input frames the middle 16 predictions are used. Finally, a post-processing technique was used to smooth out predictions over the frames. The statistical mode over a window of 39 frames was selected for each frame.

### 3.4. Conclusions: tracks 1 and 2

In agreement with the state of the art in computer vision, deep learning solutions (CNNs, C3D and LSTM) dominated both gesture recognition challenge tracks. Interestingly, in the second round, the performance of both challenge tracks improved significantly, and the estimated skeleton information has improved to be effective for gesture recognition (i.e. SYSU_ISEE, Lostoy). Participants did a great progress in both tasks, achieving 67% of recognition performance when hundreds of categories are considered in the isolated track, and getting 61% of overlap in the continuous case.

Besides, we also analysis the confusion matrix of the participants. There are some gesture classes easy to confused

for all teams, such as the label 11 (Gesture: Mudra2/Anjali. Description: Joint both hands-static gesture) and label 26 (Gesture: ItalianGestures/Madonna. Description: Join both hands together, fingers touching, hands pointing away from you.) in the IsoGD.

## 4. Real Versus Fake Expressed Emotions Challenge

In the third challenge track participants focused on the recognition of fakeness and trueness for 6 basic emotions. Within Real Versus Fake Expressed Emotions Challenge, a novel RGB video data-set for the task was released. This track attracted 9 participating teams on the learning stage and 5 teams for final evaluation stage. In total 52 participants were registered for this challenge track.

### 4.1. Data

For training, validation and test sets 480, 60 and 60 RGB videos were provided respectively. The whole dataset contains videos of 50 different subjects. For each subject, there are 12 videos about 3-5 seconds long representing 6 basic emotions (Anger, Happiness, Sadness, Disgust, Contempt, Surprise) for real and fake expressions. Some dataset examples are presented in Figure 3.

During the recording subjects were asked to watch a video, which should provoke a certain emotion. For the real emotion set subjects were supposed to express the same emotion which was provoked by the shown video. In the second case the expressed emotion and stimulated emotion were contrasted (e.g to record a faked surprise we've shown a calling disgust video and asked to act surprise) [28, 4]. For each video in all of training, validation and test sets were previously announced which of the 6 emotions is displayed, so that participants only had to predict whether the specific emotion is faked or real.

### 4.2. Metrics and Evaluation

To evaluate the performance the percentage of correctly classified videos (real or fake) was calculated for each emotion class and the average of calculated percentages $r$ was taken as final performance rate:

$$r = \frac{1}{6}\sum_{i=1}^{6} 10 \times \left( \sum_{j=1}^{10} \delta(p_l(j), t_l(j)) \right) \qquad (2)$$

where $p_l$ and $t_l$ are predicted labels and ground truth respectively, if $p_l(j) = t_l(j)$ then $\delta = 1$, otherwise $\delta = 0$.

### 4.3. Results and Methods

The recognition rates for validation and test sets calculated by equation (2) are presented in the Table 5 and as shown here the NIT-OVGU and HCILab teams obtained the highest performance rate on final evaluation stage. In Table 6 are presented percentages of correctly classified patterns for each emotion class based on final evaluation predictions. The standard deviation for HCILab and NIT-OVGU teams are 18.8 and 24.8 respectively. Hence, the predictions submitted by HCILab team are more consistent accross emotion classes.

| rank by test set | Team name | rate (validation set) | rate (test set) |
|---|---|---|---|
| 1 | NIT-OVGU | 76.7 | 66.7 |
| 1 | HCILab | 71.7 | 66.7 |
| 3 | TUBITAK UZAY-METU | 61.7 | 65.0 |
| 4 | BNU CIST | 53.3 | 61.7 |
| 5 | faceall Xlabs | 58.3 | 51.7 |

Table 5. Final ranking in the emotion track.

**First place (NIT-OVGU team):** The method proposed by the NIT-OVGU team consists of three steps. Firstly the authors estimate the intensity of facial action units (AU) as it described in [31]. For each video frame the method applies face detection, facial landmark localization, face registration, Local Binary pattern (LBP) feature extraction, and finally predicts AU intensities with Support Vector Regression (SVR) ensembles. Next they condense the obtained time series to descriptors as it is proposed in [33]. The time series are smoothed with first order Butterworth filter. After that the second derivative is calculated and from repeatedly smoothed time series 17 statistics are extracted. In total a 440-dimensional feature space are obtained on this stage. Finally authors classify the videos with Rank-SVM [34]. For a pair of videos the Rank-SVM decides which of the videos shows a more real emotion than the other one.
The obvious advantage of the proposed method is that the number of model parameters to optimize during training is very low in compared to e.g standard deep learning methods. The time needed for all stages including face detection, features extraction, training and predicting labels for test set is around 3.5 hours and it's requires about 800 MiB of CPU RAM and 3400 MiB GPU RAM.

**First place (HCILAB team):** the method proposed by HCILAB team modifies the model described in [35], which is based on the properties of mirror neurons. Firstly facial landmarks from each frame were extracted using the DLIB library. Next the authors trained a LSTM-PB network for each emotion class. The LSTM-PB network is a modification of network described in [35], where the Recurrent Neural Network (RNN) is replaced with Long short-term memory (LSTM). For learning a two-stage training procedure was used: finding the optimal weights of LSTM-PB network by a back-propagation algorithm, and learning of the optimal values of parametric bias by accumulating gradients


Figure 3. Examples of faked and real expression from third challenge track.

| | Happiness (%) | Sadness (%) | Disgust (%) | Contempt (%) | Surprise (%) | Anger (%) |
|---|---|---|---|---|---|---|
| NIT-OVGU | 40 | 100 | 100 | 60 | 60 | 40 |
| HCILab | 40 | 60 | 60 | 80 | 100 | 60 |
| TUBITAK UZAY-METU | 50 | 70 | 70 | 50 | 80 | 70 |
| BNU CIST | 70 | 70 | 70 | 40 | 70 | 50 |
| faceall Xlabs | 40 | 50 | 50 | 70 | 50 | 50 |

Table 6. Percentage of correctly classified videos in each emotion class (final evaluation stage)

of the previous stage. In proposed method gradient boosting is used to train a Real/Fake discrimination in parametric bias space. As in the method proposed by NIT-OVGU team the algorithm detects pair of videos with the same subject per each emotion class. The algorithm requires 32 Gb RAM and in total it takes about 3 days for training and about a hour for prediction on test-set running on 12 Gb GPU.

**Third place:** the algorithm is build on the assumption, that brief emotional changes in eyes and mount movements can be distinct indicators for real/fake emotions recognition. The proposed method contains two stages: features extraction and classification. On the first stage the robust micro-emotional visual descriptors for each emotion type is obtained. To compute descriptors from small temporal windows (i.e. 150 ms) of the videos, the authors used the robust video representation method [36] with the long short-term memory model. For emotion detection high-level convolutional features were used. To obtain one global representation for each video, the computed descriptors were pooled with Compact Bilinear Pooling (CBP) [37]. Finally a SVM classifier was applied to get final predictions.
One of the highest contributions of this method is the novel video representation method, which can boost visual pooling by partially retaining sequential information in the representation. In this method face detection and emotion feature extraction steps consume most of the time. Other steps such as feature learning and classifier training have relatively lower complexity and can be done in a few minutes.

**Fourth place:** The method based on the combination of the sequential texture and geometric features. On the preprocessing stage the OpenFace open-source was used to detect facial landmarks and HOG features. To aggregate HOG features of a face-image sequences authors use the temporal attention gated model from [38]. The selected model automatically learns the attention weights of each frame, and update the hidden states according to the attention gate. The auto-encoder LSTM was used to learn to encode the facial landmarks sequences into fixed length vector. The aggregated HOG and encoded landmark features are concatenated as final video representation. The whole algorithm takes about a hour running on GeForce GTX Titan GPU.

**Fifth place:** Authors use a pretrained CNN network VGG-16 on FER2013 dataset. Then, the VGG-16 is treated as a feature extractor and 4096 fc7 features are extracted from each video. Per each video 128 key-frames were selected to represent video on feature level. Before to train the LSTM network with obtained features, authors apply the Principal Component Analysis to reduce the features dimension to 1024. At the final stage 128-frame sequences representing each video are trained to LSTM network.

### 4.4. Conclusions: track 3

The final rank in Table 5 is based on averaged performance rate on final evaluation stage. Since the NIT-OVGU and HCILab teams had the equal performance rate 67.7 on final stage it was decided to split the first place between

| Team | Preprocessing | Pretrained | Features | Classification |
|---|---|---|---|---|
| NIT-OVGU | face detection | face detection model [29], Kazemi -Sullivan model [30], face recognition model dlib (VGG), AU Intensity Estimation model [31] | activity descriptors from 7 AUs intensity time series | Rank-SVM |
| HCILab | face detection | - | facial landmarks (dlib) | LSTM-PB |
| TUBITAK UZAY-METU | face detection | pretrained CNN emotion model [32] | high-level emotional features (conv5) | SVM |
| BNU CIST | OpenFace | - | HOG, facial landmarks | LSTM |
| faceall Xlabs | resizing (ratio 0.5) face detection | CNN network vgg16 | fc7 | LSTM |

Table 7. Overview of team methods in real vs faked emotions challenge.

these two teams. In order to keep the top-3 concept the TUBITAK UZAY-METU got the third place with recognition rate 65.0. Therefore there's no second place in faked vs true emotion challenge track.

## 5. Discussion

We organized three track contests on face and gesture recognition problems in order to solve: (1) a second round on large-scale RGB-D isolated and continuous gesture recognition challenge were launched; and (2) a real versus fake expressed emotions challenge was hold. Overall, the contest attracted many participants and has achieved good performances on three tracks. In general terms, the state of the art was advanced in related recognition problems (gesture recognition, and real vs fake expressed emotion recognition).

## Acknowledgment

This work was partially supported by the National Key Research and Development Plan (Grant No.2016YFC0801002), the Chinese National Natural Science Foundation Projects ♯61502491, ♯61473291, ♯61572501, ♯61572536, ♯61673052, Science and Technology Development Fund of Macau (No. 112/2014/A3), the Spanish projects TIN2016-74946-P and TIN2015-66951-C2-2-R (MINECO/FEDER, UE) and CERCA Programme / Generalitat de Catalunya, the Estonian Research Council grant ♯PUT638 and ♯IUT2-13. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. We are grateful for the support of Azure for Research for computation resources provided to the Codalab platform.

## References

[1] S. Escalera, I. Guyon, and V. Athitsos, *Gesture Recognition*. Springer, 2017. 1

[2] S. Escalera, X. Baró, H. J. Escalante, and I. Guyon, "Chalearn looking at people: A review of events and resources," in *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, pp. 1594–1601, 2017. 1

[3] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2009. 1

[4] I. Lüsi, J. C. J. Junior, J. Gorbova, X. Baró, S. Escalera, H. Demirel, J. Allik, C. Ozcinar, and G. Anbarjafari, "Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases," in *FG*, pp. 809–813, IEEE, 2017. 1, 6

[5] C. Loob, P. Rasti, I. Lüsi, J. C. J. Junior, X. Baró, S. Escalera, T. Sapinski, D. Kaminska, and G. Anbarjafari, "Dominant and complementary multi-emotional facial expression recognition using c-support vector classification," in *FG*, pp. 833–838, IEEE, 2017. 1

[6] H. J. Escalante, V. Ponce-López, J. Wan, M. A. Riegler, B. Chen, A. Clapés, S. Escalera, I. Guyon, X. Baró, P. Halvorsen, *et al.*, "Chalearn joint contest on multimedia challenges beyond visual analysis: An overview," in *ICPR*, pp. 67–73, 2016. 1, 2

[7] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li, "Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition," in *CVPR Workshops*, pp. 56–64, 2016. 2, 3, 5

[8] S. Escalera, X. Baró, J. Gonzalez, M. Á. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, J. Shotton, and I. Guyon, "Chalearn looking at people challenge 2014: Dataset and results.," in *ECCV Workshops*, pp. 459–473, 2014. 3

[9] X. Baró, J. Gonzalez, J. Fabian, M. A. Bautista, M. Oliu, H. Jair Escalante, I. Guyon, and S. Escalera, "Chalearn looking at people 2015 challenges: Action spotting and cultural event recognition," in *CVPR Workshops*, pp. 1–9, 2015. 3

[10] J. Duan, J. Wan, S. Zhou, X. Guo, and S. Li, "A unified framework for multi-modal isolated gesture recognition," in *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM),(Accept)*, 2017. 3

[11] Y. Li, Q. Miao, K. Tian, Y. Fan, X. Xu, R. Li, and J. Song, "Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model," in *ICPR*, pp. 25–30, 2016. 3

[12] P. Wang, W. Li, S. Liu, Z. Gao, C. Tang, and P. Ogunbona, "Large-scale isolated gesture recognition using convolutional neural networks," in *ICPR*, pp. 7–12, 2016. 3, 5

[13] G. Zhu, L. Zhang, L. Mei, J. Shao, J. Song, and P. Shen, "Large-scale isolated gesture recognition using pyramidal 3d convolutional networks," in *ICPR*, pp. 19–24, 2016. 3

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, pp. 4489–4497, 2015. 3, 4, 5

[15] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *ECCV*, pp. 20–36, 2016. 3, 4

[16] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *CVPR*, pp. 3034–3042, 2016. 4

[17] H. Fang, S. Xie, and C. Lu, "Rmpe: Regional multi-person pose estimation," *arXiv preprint arXiv:1612.00137*, 2016. 4

[18] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *NIPS*, pp. 802–810, 2015. 4

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, pp. 91–99, 2015. 4

[20] G. Zhu, L. Zhang, P. Shen, and J. Song, "Multimodal gesture recognition using 3d convolution and convolutional lstm," *IEEE Access*, 2017. 4

[21] X. Chai, Z. Liu, F. Yin, Z. Liu, and X. Chen, "Two streams recurrent neural networks for large-scale continuous gesture recognition," in *ICPR*, pp. 31–36, 2016. 5

[22] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Using convolutional 3d neural networks for user-independent continuous gesture recognition," in *ICPR*, pp. 49–54, 2016. 5

[23] P. Wang, W. Li, S. Liu, Y. Zhang, Z. Gao, and P. Ogunbona, "Large-scale continuous gesture recognition using convolutional neural networks," in *ICPR*, pp. 13–18, 2016. 5

[24] L. Pigou, A. van den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," *International Journal of Computer Vision*, pp. 1–10, 2015. 5

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016. 5

[26] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015. 5

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, pp. 448–456, 2015. 5

[28] I. Ofodile, K. Kulkarni, C. A. Corneanu, S. Escalera, X. Baro, S. Hyniewska, J. Allik, and G. Anbarjafari, "Automatic recognition of deceptive facial expressions of emotion," *arXiv preprint arXiv:1707.04061*, 2017. 6

[29] King, "Easily create high quality object detectors with deep learning," 2016. 8

[30] Kazemi and Sullivan, "One millisecond face alignment with an ensemble of regression trees," *IEEE CVPR*, 2014. 8

[31] A.-H. Werner, Saxen, "Handling data imbalance in automatic facial action intensity estimation," *BMVC*, 2015. 6, 8

[32] Ng, V. Nguyen, and Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," *ACM ICMI*, 2015. 8

[33] Werner, L.-E. Al-Hamadi, G. Walter, and Traue, "Automatic pain assessment with facial activity descriptors," *IEEE Transactions on Affective Computing*, p. 99, 2016. 6

[34] Joachims, "Optimizing search engines using clickthrough data," *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142, 2002. 6

[35] J. Tani, M. It, and Y. Sugita, "Self-organization of distributedly represented multiple behavior schemata in a mirror system: reviews of robot experiments using rnnpb," *Elsevier*, 2004. 6

[36] H. Z Xu, Yang, "A discriminative cnn video representation for event detection," *IEEE CVPR*, 2015. 7

[37] Gao, Z. Beijbom, and Darrell, "Compact bilinear pooling," *IEEE CVPR*, 2016. 7

[38] W. Pei, T. Baltrušaitis, D. M. Tax, and L.-P. Morency, "Temporal attention-gated model for robust sequence classification," *IEEE CVPR*, 2016. 7