ChaLearn Joint Contest on Multimedia Challenges Beyond Visual Analysis: An overview

Hugo Jair Escalante[1,2], Víctor Ponce-López[3,4,5], Jun Wan[6], Michael A. Riegler[7], Baiyu Chen[1]1, Albert Clapés[4]
Sergio Escalera[3,4], Isabelle Guyon[1,8], Xavier Baró[3,5], Pål Halvorsen[7], Henning Müller[9] and Martha Larson[10]
[1] *ChaLearn*, California, USA, [2] *INAOE*, Puebla, Mexico, [3] *Computer Vision Center, UAB*, Barcelona, Spain,
[4] *Dept. Mathematics and Computer Science, UB*, Spain, [5] *EIMT/IN3, Open University of Catalonia*, Barcelona, Spain,
[6] *NLPR, Institute of Automation, Chinese Academy of Sciences*, China, [7] *Simula Research Laboratory*, Norway,
[8] *Université Paris-Saclay*, Paris, France, [9] *University of Applied Sciences Western Switzerland*, Sierre, Switzerland,
[10] *Multimedia Information Retrieval Lab Delft University of Technology*, Delft, Netherlands,
[10] *University of California Berkeley*, Berkeley, USA

***Abstract*—This paper provides an overview of the Joint Contest on Multimedia Challenges Beyond Visual Analysis. We organized an academic competition that focused on four problems that require effective processing of multimodal information in order to be solved. Two tracks were devoted to gesture spotting and recognition from RGB-D video, two fundamental problems for human computer interaction. Another track was devoted to a second round of the first impressions challenge of which the goal was to develop methods to recognize personality traits from short video clips. For this second round we adopted a novel collaborative-competitive (i.e., *coopetition*) setting. The fourth track was dedicated to the problem of video recommendation for improving user experience. The challenge was open for about 45 days, and received outstanding participation: almost 200 participants registered to the contest, and 20 teams sent predictions in the final stage. The main goals of the challenge were fulfilled: the state of the art was advanced considerably in the four tracks, with novel solutions to the proposed problems (mostly relying on deep learning). However, further research is still required. The data of the four tracks will be available to allow researchers to keep making progress in the four tracks.**

I. Introduction

Research advances in computer vision and pattern recognition have resulted in tremendous progress on several problems and applications. Because of this, several problems on visual analysis can be considered as more or less solved (e.g., face recognition), at least in certain scenarios and under specific conditins. Despite these important advances, there are still many open problems that are receiving much attention from the community because of their potential applications. We organized a contest around four human-centered multimedia analysis problems, which in order to be solved require of the effective processing of multimodal information (e.g., audio, RGB-D video, etc.). This paper provides an overview of the task and the outcomes of the evaluation. It also analyzes the results to identify future challenges.

A. Human-centered Multimedia Analysis

There are two types of human-centered multimedia analysis: One, often referred to as *looking at people*) [**?**], seeks to understand people (what they are doing, their underlying characteristics) through analysis of video or image content that depicts people ("People in the Content"). The other type analyzes not video content that depicts people, but videos or images that people watch ("Content used by People"). The multimedia and computer vision communities have in recent years realized the importance of these types of multimedia analysis, as evidenced by [**?**]. Both types of analyses present specific challenges, that make it necessary to go "Beyond visual analysis", as we do here for recognizing personality traits, gestures and video recommendations.

It is important to recognize that whenever humans are directly involved, subjectivity and diversity take central stage. These factors are important for simple, common-sense reasons: people do not interpret and image or video in the same way, or perform actions in exactly the same way. Cases involving human action or interpretation stand in direct contrast to the challenges that are conventionally addressed by multimedia analysis and computer vision. Consider the case of fish classification in the first chapter of Duda, Hart and Stork's classic *Pattern Classification* book [**?**]. Here, the problem is use an image of a fish taken at a conveyor belot in a fish-packing plant to classify a fish as either a sea bass or a salmon. This case is clearly not a human-centered problem: the type of fish can be objectively determined, and there is no room for diversity of example.

The area of Human-centered Multimedia Analysis addresses issues that are characterized by the fact that the way in which people behave or in which they judge content depends on factors learned over a lifetime. In order to develop solutions to these questions, computer vision and multimedia analysis need to be able to adapt to a large variety of people in both scenarios. A clear example of this situation is the first impressions challenge [**?**], where participants devised systems that can learn to predict the apparent personality traits of people in very short videos.

B. Joint challenge organization

The contest we organized has been supported by three organizations with vast experience and prestige in the organi-

zation of academic contests, namely: Chalearn[1], MediaEval[2] and ImageCLEF [**?**]. The contest was also supported by the IAPR TC-12[3] on visual and multimedia information systems.

The involved organizations offer to the research community an opportunity to test their multimedia analysis technology on a standard formulation of a problem, using standard definitions and evaluation protocols. Such a set up is necessary in order to have a fair and accurate measure of the relative performance of algorithms. By measuring performance with benchmarks we bring the research field forward as a whole, since direct performance comparison allows us to know exactly when a new algorithm has succeeded in surpassing the state of the art (and should be pursued further) and when an established algorithm does not achieve the state of the art (and should be modified, or possibly abandoned).

Joint-challenges are an important aspect of sharing results and techniques. Benchmarks usually require a high degree of topical and technical focus from their participants. For this reason, there is a danger that benchmarking communities turn inward, leading to limited innovation and small modifications of existing techniques. Such an introversion can lead to missed opportunities to learn from each other. In particular, we are interested in overlaps between the approaches that are developed for human-centered multimedia analysis tasks and to sharing tools and code. We are also interested in exchanging experiences and best practices in designing and carrying out benchmarks, also for approaches such as EaaS (Evaluation as a Service [**?**]). Specific decisions about how the task is formulated and offered to participants can have a significant impact on benchmark success. It is also important for groups that develop and offer challenges to the research community be able to learn from each other.

Examples of cross pollination in the organized contest include the joint formulation of the tracks, the common usage of the CodaLab platform for the four tracks, and the common evaluation protocol (with particularities for each specific track, e.g., in terms of metrics). On the basis of these considerations, the success of this challenge can be greatly attributed to the joint organization. For this reason, we foresee such collaboration as fruitful to adopt (and adapt) for future challenges.

The rest of this paper is organized as follows. The next section provides a general overview of the contest. Then, the results of the four tracks are discussed in Sections **??**- **??**. Finally, the main findings and outcomes of the challenge are discussed in Section **??**.

II. Contest overview

This section provides a general overview of the contest, details on each track are discussed in the following sections.

A. Contest tracks

The contest comprised the following four tracks:

- **First impressions round 2 (Track 1).** To recognize apparent personality traits from 15-second videos, where big-five categories were considered. This is a follow up of the First impressions challenge at ECCV 2016 [**?**].
- **Isolated and continuous gesture recognition (Tracks 2 and 3)** To recognize gestures in either segmented or continuous video, starting from RGB-D data and considering a large number of categories and domains.
- **Context of experience (Track 4)** To determine whether videos are suitable to be shown in a certain context.

Figure **??** shows samples taken from the data sets used in the four tracks of the contest.

Fig. 1. Samples of data from the different tracks. From top to bottom: first impressions track, isolated and continuous gesture recognition, and context of experience tracks.

B. Protocol and contest duration

A common generic protocol was adopted for the four tracks on the contest (slight changes were adopted for the first track, see Section **??**). The four tracks used the CodaLab open source platform of Microsoft[4]. Participants had to submit prediction results during the challenge (see below). Track winners had to publicly release their source code and submit a fact sheet summarizing their methodologies. Please note that specific evaluation metrics were adopted for each track. The competition lasted 45 days. Two stages of the contest can be distinguished for the four tracks:

- **Development phase:** participants had access to labeled development (training) data for developing their systems; they also had access to unlabeled validation data. During this phase, participants could receive immediate feedback on their performance in validation data through the leader board in CodaLab.
- **Final phase:** participants were provided with unlabeled final (test) data, for which they had to send predictions.

[1]http://chalearn.org
[2]http://www.multimediaeval.org/
[3]http://iapr-tc12.info/
[4]https://competitions.codalab.org/

The winners of the contest were determined by evaluating performance in this data set. Participants also had to send code and fact sheets describing their methods. Code of participants was verified and the winners were announced.

C. Participation

Table **??** shows a summary of the participation in the four tracks. The number of registered participants is close to 200, whereas a lower number of participants sent predictions for the final phase, as expected in academic challenges.

Track	Registered	Test pred.	Code	Fact sheet
First impressions	51	6	4	4
Isolated GR	51	8	7	7
Continuous GR	48	3	3	3
Context of Exp.	16	3	2	2
Total	166	20	16	16

TABLE I
Summary of participation for the four tracks of the contest.

III. First impressions: Collaborative competition

This section summarizes the results of the first impressions track. This is a second round of the challenge that implements a collaborative competition or "coopetition". The goal, as in the previous first round [**?**], has been to automatically recognize five "apparent" personality traits (the so-called "Big Five") from videos of subjects speaking in front of a camera, by using human judgment. A data set consisting of 10,000 shorts clips from YouTube videos was made available. The ground truth for personality traits was obtained from workers of Amazon Mechanical Turk (AMT). To alleviate calibration problems between workers, we used pairwise comparisons between videos, and variable levels were reconstructed by fitting a Bradley-Terry-Luce model with maximum likelihood [**?**]. The competition attracted 51 participants who are grouped in several teams. Four teams entered the final phase.

A. Coopetition setting

This part of the contest adopted a coopetition scheme (collaborative competition) to quantitatively evaluate the recognition of the apparent Big Five personality traits on multimodal audio+RGB data from YouTube videos. The winners had to publicly release their source code. The coopetition feature allowed participants to download other participant codes, and rank the quality of the downloaded code by using "like"/"unlike" buttons, and count the total number of downloads for a each public submission.

The coopetition had two phases:

- As in the first round of the challenge, a development phase during which the participants had access to 6,000 manually labeled continuous video sequences of 15 seconds each. Thus, 60% of the videos used for training are randomly grouped into 75 training batches. They could get immediate feedback on their prediction performance by submitting results on an unlabeled validation set of 2,000 videos. These 2,000 videos used for validation represent 20% over the total set of videos and are also randomly grouped into 25 validation batches.
- A final phase during which the competitors could submit their predictions on 2,000 new test videos (the remainder 20% over the total set of videos, also grouped into 25 test batches). The prediction scores on test data were not revealed until the end of the challenge.

B. Data

The data set consists of 10,000 clips extracted from more than 3,000 different YouTube high-definition (HD) videos of people facing and speaking in English to a camera. The people appearing are of different gender, age, nationality, and ethnicity, which makes the task of inferring apparent personality traits more challenging [**?**].

C. Metrics and evaluation

The participants of the different teams trained their models to imitate human judgments consisting in continuous target values in the range [0, 1] for each trait. Thus, their goal was to produce for each video in the validation or test set, 5 continuous prediction values in the range [0, 1], one for each trait.

For this task (similar in spirit to a regression task) the evaluation consisted in computing the **mean accuracy** over all traits and videos. Accuracy for each trait is defined as:

$$A = 1 - \frac{1}{N_t}\sum_{i=1}^{Nt}|t_i - p_i| / \sum_{i=1}^{Nt}|t_i - \bar{t}| \qquad (1)$$

where p_i are the predicted scores, t_i are the ground truth scores, with the sum running over the N_t test videos, and $\bar{t}$ is the average ground truth score over all videos[5]. Additionally, we also computed (but did not use to rank the participants) the coefficient of determination:

$$R^2 = 1 - \sum_{i=1}^{Nt}(t_i - p_i)^2 / \sum_{i=1}^{Nt}(t_i - \bar{t})^2 \ . \qquad (2)$$

We also turned the problems into classification problems by thresholding the target values at 0.5. In this way we obtained 5 binary classification problems (one for each trait).

D. Results and summary of participants methods

Table **??** summarizes the various approaches of the teams who participated in the final phase, uploaded their models, and provided a survey about methods.

All of the approaches use both audio and visual cues. *BU-NKU* did not use audio information in the first round, but after the coopetition they implemented the approach of *pandora* and obtained the best overall result on this second round. For the visual cues, the dominant approach is to learn the representations by means of Convolutional Neural

[5]This definition is slightly different from what we used on the leaderboard. The leaderboard accuracy is not normalized $A = 1 - \frac{1}{N_t}\sum_{i=1}^{Nt}|t_i - p_i|$. This change does not affect the ranking.

Networks [**?**]. Most teams also made semantic assumptions about the data by separately processing face and background. Another common approach is to use pre-trained deep models and fine-tune on the dataset provided by the challenge. In order to combine the different modalities, the teams used an early fusion scheme before being fed to different regression methods. Fully-connected Neural Networks or Support Vector Regressors were used for this purpose. A notable exception is the method proposed by team *evolgen*, which includes the temporal structure by partitioning the video sequences in segments and sequentially feeding the learned audio-video representation to a recurrent Long Short Term Memory architecture [**?**]. Readers are referred to Table **??** for a synthesis of the main characteristics of the methods that have been submitted to this challenge. Next, we summarize the three winning methods.

First place: the *BU-NKU* team uses both visual (face and scene) and audio modalities. They first estimate facial landmarks in order to perform face alignment. Then, they obtain both image-level deep features and Local Gabor Binary Patterns from Three Orthogonal Planes (LGBPTOP) from these face crops over video frames. The per-video facial representation consists hence of a set of functionals computed over the per-frame face features (e.g., mean, std, offset, etc). The deep architecture used for computing face features is VGG-Face [**?**] pre-trained on FER-2013 dataset. Similarly, scene features are extracted from the initial frame by using VGG-VD-19 [**?**] trained on ILSVRC 2012. Regarding the audio modality, a large pool of low level descriptors (LLD) are generated using the openSMILE toolbox [**?**]. Finally, Kernel Extreme Learning Machines (KELM) are used for each cue and individual scores are fused for final trait recognition.

Second place: the *evolgen* team proposed a multimodal LSTM architecture. The input video sequences are split into 6 non-overlapping partitions. From each partition, they extract the audio representation using classical spectral features and statistical measurements, forming a 68-dimensional feature vector. For the video representation, the authors propose selecting a frame from the partition, locating the face, and centering it through face alignment. The preprocessed data is passed to a Recurrent CNN, trained end-to-end, which uses a separate pipeline for audio and video. Each partition frame is processed with convolutional layers, afterwards applying a linear transform to reduce the dimensionality. The audio features of a given partition go through a linear transform and are concatenated with the frame features. The Recurrent layer is sequentially fed with the features extracted from each partition.

Third place: the *pandora* team uses Deep Convolutional Networks to focus on leveraging visual information from faces and supplementary information from background, whereas an ensemble of Decision Tree Regressors performs prediction on the acoustic features. The authors separately model grayscale and colored faces, which seem to laverage complementary information to one another. Apparent personality traits are predicted at frame-level. Then, the frame-wise predictions over 15 frames are concatenated to obtain a fixed-length representation per video, which gives a final descriptor of size $15 \times$ #traits $\times$ #models. The final prediction is done using a regressor over the former representation.

E. Track conclusions

All three winning methods applied neural networks on visual cues. Moreover, all of them also used some kind of data pre-processing, such as face detection and alignment. Background information, when used was fed into separate network streams from the face stream, as it was the case of first and third place participants. The second method used end-to-end training, fusing the audio and video streams with fully-connected layers. The coopetition feature of this second round, although not applied to weight final prize ranking score, was used in order to allow participants to download the code between different teams and rank them per usability. *BU-NKU* clearly benefited from this fact by incorporating the audio features from the *pandora* team.

IV. Isolated and continuous gesture recognition

Tracks 2 and 3 of the contest were associated with the ChaLearn Looking at people (LAP) 2016 Large-scale Isolated and Continuous Gesture Recognition Challenges, respectively. The approached problems were recognizing gestures from either segmented or continuous RGB-D videos, respectively. The focus of both challenges was on "large-scale" learning and "user independent" gesture recognition,.

A. Data

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

- **Chalearn LAP RGB-D Isolated Gesture Dataset (IsoGD).** 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 2: 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).** 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 3, 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, whereby the subsets are mutually exclusive. For more information about these two data sets, please refer to [**?**].

B. Metrics and evaluation

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

$$r = \frac{1}{n}\sum_{i=1}^{n}\delta(p_l(i), t_l(i)) \tag{3}$$

	Pretraining	Preprocessing	Modality				Fusion
			Audio		Video		
			R[1]	L[2]	R[1]	L[2]	
BU-NKU	VGG-face (FER2013), VGG-VD-19 (ILSVRC12)	face alignment	LLD[8]	-	CNN(face/scene), LGBPTOP (face)	KELM[6]	early
evolgen	-	face alignment	spectral	RCNN[10]	RCNN[10]	RCNN[10]	early
pandora	VGG-Net	face alignment	LLD[8]	Bagged Regressor	CNN(face/scene)	CNN	early
Pilab	-	-	spectral	RF regressor	-	-	-

[1] R = Representation [2] L = Learning Strategy [3] logfbank = Logarithm Filterbank Energies [4] PSLR = Partial Least Square Regressor [5] SVR = Support Vector Regression [6] KELM = Kernel Extreme Learning Machine [7] FER = Facial Expression Recognition Dataset [8] LLD = Low Level Descriptor [9] MFCC = Mel Frequency Cepstral Coefficient [10] RCNN = Recurrent CNN.

TABLE II
Overview of the team methods comparing pre-training (topology and data), preprocessing if performed, representation, learning strategy per modality, and fusion.

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 [**?**], [**?**]. The Jaccard index measures the average relative overlap between true and predicted sequences of frames for a given gesture. Metric description details for this track can be found in [**?**],

C. Results and methods

1) Isolated gesture recognition challenge: In the final testing phase, eight teams submitted predictions. The summary of the method features is shown in Table **??**. Six teams were able to outperform the baseline method [**?**]. Next, we summarize the methods of the top 3 ranked participants.

Rank	Team	recognition rate r	Method
1	FLiXT	56.90%	C3D + RGB-D
2	AMRL	55.57%	CNN + depth
3	XDETVP-TRIMPS	50.93%	Pyramidal C3D + RGB-D
4	ICT_NHCI	46.80%	appearance model +RNN+RGB-D
5	XJTUfx	43.92%	CNN+MHI+depth
6	TARDIS	40.15%	dense trajectory+ fish vector encoding + SVM
-	baseline [**?**]	24.19%	-
7	NTUST	20.33%	-
8	Bczhangchen	0.45%	-

TABLE III
Summary of the results in the isolated gesture challenge.

First place: the *FLiXT* team recognizes gestures by employing both RGB and depth videos and learning with a 3D CNN model. Authors preprocessed the inputs and convert them into 32-frame videos. Since variations in background, clothing, skin color and other external factors may disturb the recognition, they employed saliency video to concentrate the gestures. The features of the videos were learned by the C3D model [**?**] in order to learn spatiotemporal features. This is also combined with RGB, depth and saliency features to boost final performance.

Second place: the *AMRL* team proposes three simple, compact yet effective representations from depth sequences for gesture recognition in the context of CNNs. The three representations are called Dynamic Depth Image (DDI), Dynamic Depth Normal Image (DDNI) and Dynamic Depth Motion Normal Image (DDMNI). They are all based on bidirectional rank pooling, converting the depth sequences into images. Such representations enable the use of existing CNN models directly on video data applying fine-tuning without introducing many parameters to learn. The 3 representations model the posture and motion cues in different levels of abstraction, complementing each other in order to improve final gesture recognition performance.

Third place: the *XDETVP-TRIMPS* team proposes a pyramidal 3D CNN. First, each video is segmented into three parts which may overlap in some degree according to the frame count of the video file. Then, sixteen frames are sampled from each part and the whole video file via uniform sampling with temporal jitter. Finally, four sixteen-frame batches are used to train the C3D model [**?**] on the RGB and depth modalities. Gesture prediction is obtained by fusing the outputs of both modalities.

2) Continuous gesture recognition challenge: Three teams submitted predictions in the final stage of the challenge. The performance of all the methods improved the provided baseline. Results and methods are shown in Table **??**.

Rank	Team	Mean Jaccard Index $\overline{J_S}$	Method
1	ICT_NHCI	0.2869	appearance model +RNN+RGB-D
2	TARDIS	0.2692	C3D + sliding windows + RGB-D
3	AMRL	0.2655	QOM+CNN+depth
-	baseline [**?**]	0.1464	-

TABLE IV
Summary of the results in the continous gesture challenge.

First place: the *ICT_NHCI* team transforms the continuous gesture recognition problem into the isolated recognition problem with an accurate gesture segmentation. For segmentation, it is considered that the subject puts the hands down after performing each gesture. Therefore, they used a face detector [**?**] and a hand detector [**?**] to estimate the distances

between each pair of three points (one face, two hands). For gesture recognition, the two streams Recurrent Neural Network (RNN) method [**?**] is applied. It first extracts HOG and Skeleton features from RGB and depth videos. On each separated channel, the hand shape and position features are fused by concatenation. Then, features from different channels are fused by the RNN model.

Second place: the *TARDIS* team trained an end-to-end deep network for gesture recognition (jointly learning both the feature representation and the classifier). The network performs three-dimensional (i.e. space-time) convolutions to extract features related to both the appearance and motion. Space-time invariance is encoded via pooling layers. Before being adapted to the task of gesture recognition, the earlier stages of the network are partially initialized using C3D method [**?**]. In order to perform spotting, the deep-volume features are computed on a sliding spatio-temporal volume. The output predictions are then refined via two stages of majority voting filtering.

Third place: the *AMRL* authors approach the problem in two stages: segmentation and recognition. For segmentation, they obtain the begin and end frames of each gesture based on quantity of movement (QOM) and then propose compact representations for depth sequences, called Improved Depth Motion Maps (IDMM), which convert each depth sequence into an image in order to recognize the gestures using ConvNets. This method enables the use of existing CNN models directly on video data with ne-tuning, without introducing a large set of parameters to be learned.

D. Track conclusions

In agreement with the state of the art in computer vision, deep learning solutions (CNNs, C3D and RNN) dominated both gesture recognition challenge tracks. Interestingly, there was only one team that approached the spotting problem directly, as opposed to the other teams that segmented first and then recognized. Although participants did a great progress in both tasks, achieving almost 60% of recognition performance when hundreds of categories are considered in the isolated track, and getting close to 30% of overlap in he continuous case, results still suggest that there is much room for improvement in both challenges.

V. Context of experience

The Context of Experience track has as goal to predict the multimedia content that users find most fitting to watch in specific viewing situations (contexts). We focus on the case of viewers watching movies on an airplane. Here, viewers can be considered largely to have the same aim (i.e., viewing intent). They want to pass the time, and keep themselves occupied in the small space of an airplane cabin, and minimize the impact of the limitations of the technology (e.g., screen size), and the environment (e.g., background noise, interruptions, presence of strangers). This common aim leads us to assume that people will want to watch will depend on the context in which they are experiencing the multimedia content, and not exclusively their personal preferences. The objective of the task is to predict which videos allow viewers to achieve this goal, given the context, which includes the limitations of the technology (e.g., screen size), and the environment (e.g., background noise, interruptions, presence of strangers). Airplanes provide the basis for a later study of other stressful contexts include hospital waiting rooms, and dentists offices, where videos are shown during treatment.

A. Data

The challenge provided participants with a list of movies (including links to descriptions and video trailers), and requires them to classify each movie into +goodonairplane/-goodonairplane classes. The dataset includes movies, metadata, extracted audio and visual features and links to movie trailers, and is described in detail in [**?**]. For this challange, the development set contains 146 movies and has further been split into a training set with 96 movies and a validation set with 50 movies. This has been done to make it possible for the participants to test their approaches before the final challenge data is made available (receiving immediate feedback in the CodaLab leader board). The test set for the final evaluation consists of 175 movies. The ground truth of the task is derived from two sources. First, actual movie lists used by a major airline, and second user judgments on movies that are collected via a crowdsourcing tool.

B. Metrics and evaluation

For the evaluation we use the standard metrics Precision, Recall and F1 score. Negative and positive classes in both data sets are balanced as good as possible. Participants are asked to submit a predicted class for each movie in the test data set. The metrics are then calculated and provided to the participants. For a transparent and fair procedure, the labels used for the evaluation will be released together with the results. We also provide a random baseline in the leaderboard for the challenge phase. The random baseline is the average of ten random classification predictions. The values for the random baseline are F1 score of 0.594, precision of 0.618 and recall of 0.572.

C. Results and methods

Three participants submitted to the challenge. An overview can be found in table **??**. Two submitted the fact sheet and their code. An overview about the used features of the participants can be found in table **??**. Only one participant used a multimodal approach to tackle the task. The other two relied on metadata. The last participant did not provide fact sheet or code and just stated that they used a regression for the classification on all the provided metadata. For this reason, this team *asm* was not included in the final ranking. In the following, we briefly describe the methods of the top ranked participants.

First place: the *itec-aau* team performed a simple metadata approach where they used some of the provided metadata and created a new self-created feature based on the metadata that they call hotness. Hotness is higher the closer the movie

Team	Audio	Visual	Textual	Metadata
itec-aau	no	no	no	yes
tud-mmc	yes	yes	yes	yes
asm	no	no	no	yes

TABLE V
Summary of the features used by the participants.

release year is to the actual date. For the 90s, 80s and 70s (and older) one hotness score is used. To classify the data they used the Weka Library and the LMT classier with minor adjustments.

Second place: the *tud-mmc* team proposed a meta-learning approach that can be divided into three stages: classifier selection, feature selection and classifier stacking. Classifier selection is used to filter the classifier on different models based on their performance (only consider classifier has a better performance than random guess). Feature selection is used to select features for varies classifiers that is able to achieve the best performance on F1 score. Based on the predictions of selected classifiers and selected feature subspace, they trained a second level classifier to predict the final label.

Rank	Team	F1	Precision	Recall
1	itec-aau	0.676	0.623	0.739
2	tud-mmc	0.641	0.569	0.733
3	baseline	0.594	0.618	0.572
4	asm	0.697	0.547	0.958

TABLE VI
Summary of the results context of experience challenge.

D. Track conclusions

We had 16 teams that were interested in the task. For a task with a rather unconventional idea this is a good start. Nevertheless, only 3 participants submitted in the final challenge phase. After having a closer look into why this was the case we found two main problems. First, it seemed that some participants did not have enough time to solve the task until the required deadline. Second, for some of the participants the task was too complex and they could not manage or where not interested to process other data beside of the image data which is important for this type of task.

VI. Discussion and future work

We organized a four track contest on problems that require going beyond visual analysis in order to be solved: (1) a second round on the first impressions challenge was run in a coopetition scheme; (2-3) two challenges on large scale gesture spotting and recognition were launched; and (4) a novel competition on video recommendation. Overall, the contest attracted near 200 participants, 20 of which participated until the final stages. In general terms, we can say that the state of the art was advanced in four directions. Thus, we can conclude that the contest was a success. Much of this success is due to the joint organizational efforts by the involved organizations: ChaLearn, ImageCLEF and MediaEval, with support of the IAPR TC12.

Acknowledgments

We would like to thank the following people: Marc Oliu, Ciprian Corneanu. The following initiatives supported the challenge: IAPR TC12, MediaEval, ImageCLEF, ChaLearn, CASIA. We are also grateful with the sponsors of this challenge: Universitat de Barcelona, INAOE and ChaLearn. This work has been partially supported by the Spanish projects TIN2013-43478-P, TIN2015-66951-C2-2-R and the European Comission Horizon 2020 granted project SEE.4C under call H2020-ICT-2015 as well as EC FP7 CrowdRec 610594.

References

[1] S. Escalera, J. Gonzalez, X. Baro, H. J. Escalante, and I. Guyon, "Chalearn looking at people events," *IAPR NewsLetter*, vol. 37, no. 4, pp. 13–15, 2015.

[2] H. Gunes and H. Hung, "Emotional and social signals: A neglected frontier in multimedia computing?" *IEEE MultiMedia*, vol. 22, no. 2, pp. 76–85, 2015.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[4] V. Ponce-López, B. Chen, M. Oliu, C. Cornearu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn LAP 2016: First round challenge on first impressions," in *ECCV*, 2016.

[5] H. Müller, P. Clough, T. Deselaers, and B. Caputo, Eds., *ImageCLEF – Experimental Evaluation in Visual Information Retrieval*, ser. Springer International Series On Information Retrieval. Springer, 2010, vol. 32.

[6] A. Hanbury, H. Müller, K. Balog, T. Brodt, G. V. Cormack, I. Eggel, T. Gollub, F. Hopfgartner, J. Kalpathy-Cramer, N. Kando, A. Krithara, J. Lin, S. Mercer, and M. Potthast, "Evaluation–as–a–service: Overview and outlook," *ArXiv*, vol. 1512.07454, 2015.

[7] B. Chen, S. Escalera, I. Guyon, V. Ponce-López, N. Shah, and M. Oliu, "Overcoming calibration problems in pattern labeling with pairwise ratings: Application to personality trait," in *ECCV*, 2016.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, vol. 1, no. 3, 2015, p. 6.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[13] J. Wan, S. Z. Li, Y. Zhao, S. Zhou, I. Guyon, and S. Escalera, "Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition," in *CVPRW*, 2016.

[14] S. Escalera, X. Baro, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce, H. J. Escalante, J. Shotton, and I. Guyon, "Chalearn looking at people challenge 2014: Dataset and results." in *Proceedings of the ECCV14 ChaLearn Workshop on Looking at People*, 2014.

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

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[17] S. Wu, M. Kan, Z. He, S. Shan, and X. Chen, "Funnel-structured cascae for multi-view face detection with alignment awareness," in *Neurocomputing (Under review)*, 2016.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[19] K. Greff, R. Kumar Srivastava, J. Koutnk, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," in *arXiv.org*, 2015.

[20] M. Riegler, M. Larson, C. Spampinato, P. Halvorsen, M. Lux, J. Markussen, K. Pogorelov, C. Griwodz, and H. Stensland, "Right inflight?: A dataset for exploring the automatic prediction of movies suitable for a watching situation," in *Proc. of 7th International Conference on Multimedia Systems*. ACM, 2016, pp. 45:1–45:6.