Overview of the Multimedia Information Processing for Personality & Social Networks Analysis Contest

Gabriela Ramírez[1], Esaú Villatoro[1], Bogdan Ionescu[2], Hugo Jair Escalante[3,4], Sergio Escalera[4,7], Martha Larson[6], Henning Müller[7], and Isabelle Guyon[4,8]

[1] Universidad Autónoma Metropolitana Unidad Cuajimalpa (UAM-C), Mexico
[2] Multimedia Lab, University Politehnica of Bucharest, Romania
[3] Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico
[4] ChaLearn, Berkeley, California,USA
[5] MIR Lab, Delft University of Technology, Netherlands
[6] University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
[7] Computer Vision Center (UAB) & University of Barcelona,
[8] Université Paris-Saclay, France

Abstract. Progress in the autonomous analysis of human behavior from multimodal information has lead to very effective methods able to deal with problems like action/gesture/activity recognition, pose estimation, opinion mining, user tailored retrieval, etc. However, it is only recently that the community has been starting to look into related problems associated with more complex behavior, including personality analysis, deception detection, among others. We organized an academic contest co-located with ICPR2018 running two tasks in this direction. On the one hand, we organized an information fusion task in the context of multimodal image retrieval in social media. On the other hand, we ran another task in which we aim to infer personality traits from written essays, including textual and handwritten information. This paper describes both tasks, detailing for each of them the associated problem, data sets, evaluation metrics and protocol, as well as an analysis of the performance of simple baselines.

Keywords: Information fusion · personality analysis · social networks · handwritten recognition · multimedia information processing.

1 Introduction

Computer Vision and Multimedia information processing are fruitful research fields that have focused on several tasks, among them the analysis of human behavior. Although great advances have been obtained in the so-called Looking at People field (see e.g., [2,8]), it has only been recently that attention from this area is targeting problems that have to do with more complex and subconscious behavior. For instance, personality and social behavior are only starting to be

explored from computer vision and multimedia information processing perspectives [1]. This is the case due to a lack of data and benchmarks to evaluate such tasks.

Nevertheless, the availability of massive amounts of multimodal information together with the dominance of social networks as a fundamental channel where users interact have attracted the interest of the community in this direction of research. It should be noted that tools for effectively analyzing this behavior can have a major impact into everyone's life, with applications in health (e.g., support for mental disorders), security (e.g., forensics, preventive applications), human computer/machine/robot interaction (e.g., affective/interactive interfaces) and even entertainment (e.g., user-tailored systems).

We organize an academic contest collocated with the 2018 ICPR International Conference on Pattern Recognition[9], comprising two tracks on the analysis of non-obvious human behavior from multimodal and social media data. On one hand we focus on information fusion for social image retrieval and diversification (DivFusion Task) using multimodal information obtained from social networks. This is a follow up of past challenges on diversification organized as part of the MediaEval Benchmarking Initiative for Multimedia Evaluation[10]. On the other hand, we organize the first competition on recognizing personality from digitized written documents (HWxPI Task). A new data set is released comprising in addition to images the transcripts of documents. With this track we aim to set the basis for research on inferring personality from user handwriting (including taking into account errors and the type of writing). Both tracks are at the frontier of research on Looking At People and multimedia information processing.

In this paper we describe the proposed tasks, the associated data sets and evaluation protocol in detail. Results obtained by baseline methods are reported and future work directions motivated by the contest are discussed.

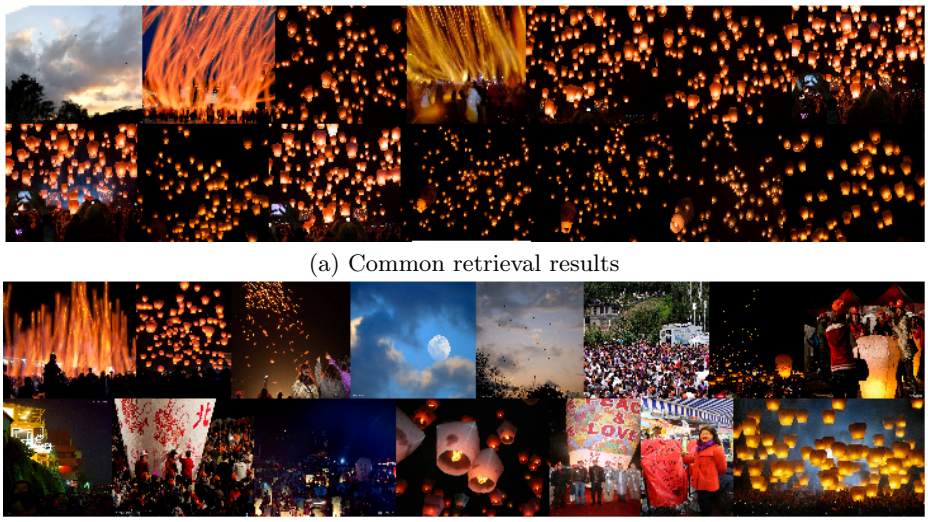
The remainder of the paper is organized as follows. Section 2 describes the DivFusion task. Section 3 describes the HWxPI task. Section 4 outlines preliminary conclusions and future work directions.

2 DivFusion task

2.1 Overview

Diversification of image search results is now a hot research problem in multimedia. Search engines such as Google Image Search are fostering techniques that allow for providing the user with a diverse representation of search results, rather than providing redundant information, e.g., the same perspective of a monument or location. The DivFusion task builds on the MediaEval Retrieving Diverse Social Images Tasks that were addressing specifically the diversification of image search results in the context of social media. Figure 1 illustrates a diversification example from MediaEval 2015.

[9] `http://www.icpr2018.org/`.
[10] `http://www.multimediaeval.org/`.



(a) Common retrieval results

(b) Results after diversification

Fig. 1. Example of retrieval and diversification results for query "Pingxi Sky Lantern Festival" (to 14 results): (a) Flickr initial retrieval results; (b) diversification achieved with the approach from TUW [11] (best approach from MediaEval 2015).

Participants receive a list of image search queries with up to 300 photos retrieved from Flickr and ranked with Flickr's default "relevance" algorithm[11]. The data are accompanied by various metadata and content descriptors. Each query comes also with a variety of diversification system outputs (participant runs from previous years), ranging from clustering techniques, greedy approaches, re-ranking, optimization and multi-system methods to human-machine-based or human-in-the-loop (i.e., hybrid) approaches. They are to employ fusion strategies to refine the retrieval results to improve the diversification performance of the existing systems even more.

The challenge reuses the publicly available data sets issued from the 2013-2016 MediaEval Retrieving Diverse Social Images tasks [7,6,5,4], together with the participant runs. The data consist of hundreds of Flickr image query results (>600 queries, both single- and multi- topic) and include: images (up to 300 results images per query), social metadata (e.g., description, number of comments, tags, etc.), descriptors for visual, text, social information (e.g., user tagging credibility), as well as deep learning features, expert annotations for image relevance and diversification (i.e., clustering of images according to the similarity of their content). An example is presented in Figure 1. The data are accompanied by 180 participant runs that correspond to the output of various image search retrieval

[11] `https://www.flickr.com/services/api/`.

diversification techniques (each run contains the diversification of each query from a data set). These allow to experiment with different fusion strategies.

2.2 Data set

The data consist of a set of 672 queries and 240 diversification system outputs and is structured as following, according to the development — validation — testing procedure:

- *devset* (development data): contains two data sets, i.e., devset1 (with 346 queries and 39 system outputs) and devset2 (with 60 queries and 56 system outputs);
- *validset* (validation data): contains 139 queries with 60 system outputs;
- *testset* (testing data): contains two data sets, i.e., seenIR data (with 63 queries and 56 system outputs, it contains the same diversification system outputs as in the devset2 data) and unseenIR data (with 64 queries and 29 system outputs, it contains unseen, novel diversification system outputs).

Overall the provided data consist of the following (which varies slightly depending on the data set, as explained below): *images Flickr* — the images retrieved from Flickr for each query; *images Wikipedia* — representative images from Wikipedia for each query; *metadata* — various metadata for each image; *content descriptors* — various types of content descriptors (text-visual-social) computed on the data; *ground truth* — relevance and diversity annotations for the images; *diversification system outputs* — outputs of various diversifications systems. For more information, see the challenge web page[12].

Development data set *Devset1* contains single-topic, location related queries. For each location, the following information is provided: location name — is the name of the location and represents its unique identifier; location name query id — each location name has a unique query id code to be used for preparing the official runs; GPS coordinates — latitude and longitude in degrees; link to the Wikipedia web page of the location; a representative photo retrieved from Wikipedia in jpeg format; a set of photos retrieved from Flickr in jpeg format (up to 150 photos per location); an xml file containing metadata from Flickr for all the retrieved photos; visual and textual descriptors; ground truth for both relevance and diversity. *Devset2* contains single-topic, location related queries. For each location, the following information is provided: location name — is the name of the location and represents its unique identifier; location name query id — each location name has a unique query id code to be used for preparing the official; GPS coordinates — latitude and longitude in degrees; link to the Wikipedia web page of the location; up to 5 representative photos retrieved from Wikipedia in jpeg format; a set of photos retrieved from Flickr in jpeg format (up to 300 photos per location); an xml file containing metadata from Flickr for all the retrieved photos; visual, text and credibility descriptors; ground truth for both relevance and diversity.

[12] `http://chalearnlap.cvc.uab.es/challenge/27/track/28/description/`.

Validation data set Contains both single- and multi-topic queries related to locations and events. For each query, the following information is provided: query text formulation — is the actual query formulation used on Flickr to retrieve all the data; query title — is the unique query text identifier, this is basically the query text formulation from which spaces and special characters were removed; query id — each location name has a unique query id code to be used for preparing the official runs; GPS coordinates — latitude and longitude in degrees (only for one-topic location queries); link to the Wikipedia web page of the query (when available); up to 5 representative photos retrieved from Wikipedia in jpeg format (only for one-topic location queries); a set of photos retrieved from Flickr in jpeg format (up to 300 photos per query); an xml file containing metadata from Flickr for all the retrieved photos; visual, text and credibility descriptors; ground truth for both relevance and diversity.

Test data set *SeenIR* data contains single-topic, location related queries. It contains the same diversification system outputs as in the *devset2* data. For each location, the provided information is the same as for *devset2*. No ground truth is provided for this data. *UnseenIR* data contains multi-topic event related and general purpose queries. It contains unseen, novel diversification system outputs. For each query, the following information is provided: query text formulation — is the actual query formulation used on Flickr to retrieve all the data; query title — is the unique query text identifier, this is basically the query text formulation from which spaces and special characters were removed; query id — each query has a unique query id code to be used for preparing the official runs; a set of photos retrieved from Flickr in jpeg format (up to 300 photos per query); an xml file containing metadata from Flickr for all the retrieved photos; visual, text and credibility descriptors. No ground truth is provided for this data.

Ground truth The ground truth data consists of relevance ground truth and diversity ground truth. Ground truth was generated by a small group of expert annotators with advanced knowledge of the query characteristics. For more information on ground truth statistics, see the recommended bibliography on the source data sets [7,6,5,4].

Relevance ground truth was annotated using a dedicated tool that provided the annotators with one photo at a time. A reference photo of the query could be also displayed during the process. Annotators were asked to classify the photos as being relevant (score 1), non-relevant (score 0) or with a "do not know" answer (score -1). A definition of relevance was available to the annotators in the interface during the entire process. The annotation process was not time restricted. Annotators were recommended to consult any additional information about the characteristics of the location (e.g., from the Internet) in case they were unsure about the annotation. Ground truth was collected from several annotators and final ground truth was determined after a majority voting scheme.

Diversity ground truth was also annotated with a dedicated tool. The diversity is annotated only for the photos that were judged as relevant in the previous

step. For each query, annotators were provided with a thumbnail list of all the relevant photos. The first step required annotators to get familiar with the photos by analysing them for about 5 minutes. Next, annotators were required to re-group the photos into similar visual appearance clusters. Full size versions of the photos were available by clicking on the photos. A definition of diversity was available to the annotators in the interface during the entire process. For each of the clusters, annotators provided some keyword tags reflecting their judgments in choosing these particular clusters. Similar to the relevance annotation, the diversity annotation process was not time restricted. In this particular case, ground truth was collected from several annotators that annotated distinct parts of the data set.

2.3 Evaluation

Performance is assessed for both diversity and relevance. We compute Cluster Recall at X (CR@X) — a measure that assesses how many different clusters from the ground truth are represented among the top X results (only relevant images are considered), Precision at X (P@X) — measures the number of relevant photos among the top X results and F1-measure at X (F1@X) — the harmonic mean of the previous two. Various cut off points are to be considered, e.g., X=5, 10, 20, 30, 40, 50. Official ranking metrics is the CR@20. This metric simulates the content of a single page of a typical web image search engine and reflects user behavior, i.e., inspecting the first page of results in priority. Metrics are to be computed individually on each test data set, i.e., seenIR data and unseenIR data. Final ranking is based on overall mean values for CR@20, followed by P@20 and then F1@20.

2.4 Baseline

To serve as reference, each of the provided data sets is accompanied by a baseline system consisting of the Flickr initial retrieval results obtained with text queries. These are obtained with the Flickr's default relevance retrieval system. The testset baseline is also provided. It achieves the following performance: average metrics — CR=0.3514, P=0.6801, and F1=0.4410; metrics on SeenIR data — CR=0.3419, P=0.8071, F1=0.4699; metrics on UnseenIR data — CR=0.3609, P=0.5531, F1=0.4122.

2.5 Discussion

15 teams registered to the task but none of them managed to finish the competition by the deadline. One reason for this is the large and very rich data. Therefore, it is difficult to maneuver in the time alloted by the competition. Even though no system results were analysed, the provided evaluation data and tasks remain open and anyone interested in benchmarking fusion techniques can take advantage of this framework.

3 HWxPI task

3.1 Overview

According to Pennebaker, language is a good indicator of our personality, since through language one can express its way of thinking and feeling [9]. The personality can be determined by stable patterns of behavior surfaced in any particular situation. In other words, the personality is defined by the characteristics that do not change, and that are independent of the situation in which a person is [12]. Consequently, an automatic method can be use for extracting these patters in any production of a subject. In this fashion, using more views of the same subject can lead to identify complementary aspects of her or his personality.

In this task we aim to provide a standardized multimodal corpus for the personality identification problem. Particularly, the HWxPI task consists of estimating the personality traits of users from their handwritten texts and the corresponding transcripts. The traits correspond to the big five personality model used in psychology: extroversion, agreeableness, conscientiousness, emotional stability, and openness to experience.

The challenge comprises two phases: development and evaluation. For the first phase, the participants were encouraged to develop systems using a set of development pairs of handwritten essays (including image and text) from 418 subjects. Each subject has an associated class 1 and 0, corresponding to the presence of a strong or a weak presence of a specific personality trait. Thus, participants had to develop five binary classifiers to predict the pole of each trait.

For the final evaluation phase, an independent set of 293 unlabeled samples were provided to the participants. The provided predictions used the models trained on the development data.

The complete schedule of our challenge was managed through Codalab[13]. Particularly, the first phase started February 20th and during the following three and a half month participants could follow their performance on the validation set. A total of 9 participants submitted predictions for the validation set. The final phase started on June 13th and same as in the previous phase, participant were able to see the results of their approach, this time, on the test set. For this phase, only two participants submitted predictions.

3.2 Data set

The corpus used in this task consists of handwritten Spanish essays from Mexican undergraduate students. A subset of this corpus and the gathering methodology is described in [10]. The textual modality is a transcription of the essays marked with seven tags of some handwritten phenomena: <D:description> (drawing), <IN> (insertion of a letter into a word), <MD> (modification of a word, that is a correction of a word), <DL> (elimination of a word), <NS> (when two words

[13] https://competitions.codalab.org/competitions/18362

Table 1. Participants in the corpus divided by gender and average age per partition (train-validation-test).

Partition	Female	Male	Total
Train	209	209	418
Validation	125	293	293
Test	59	66	125

Table 2. Number of subjects by class per trait.

	Train		Validation		Test	
Trait	high	low	high	low	high	low
Openness	239	179	71	54	192	101
Consciousness	171	247	61	64	137	156
Extroversion	212	206	168	125	68	57
Agreeableness	177	241	61	64	142	151
Emot. Stability	186	232	73	52	148	145

were written together; e.g., "Iam" instead of "I am") and, SB (syllabification). The image modality was captured by scanner without edition of any type to the jpg file.

Each pair of text plus image is labeled with the personality of its authors. Accordingly, the big five model of personality was used with traits: extroversion, agreeableness, conscientiousness, emotional stability, and openness to experience. Using the TIPI questionnaire [3] with answers of each subject, we could determine the class for each trait. The instrument provided a series of norms to decide the direction of each trait among four classes: high, medium high, medium low and low. For the HWxPI task, we binarize the classes to 1 for subjects with high and medium high and 0 for low and medium low traits.

The total number of instances in the corpus is 836, divided into three subsets: training, validation and test. The participants were given the training subset with the corresponding solution, the validation set was also available for tuning if necessary, while the test partition was used for the evaluation. The distribution of users (also referred to as subjects) per partition is shown in Table 1. And the information about the users per trait is shown in Table 2.

Finally, an example of a pair of image and text is given in Table 3. The image we provide has the complete letter sheet with considerable blank spaces.

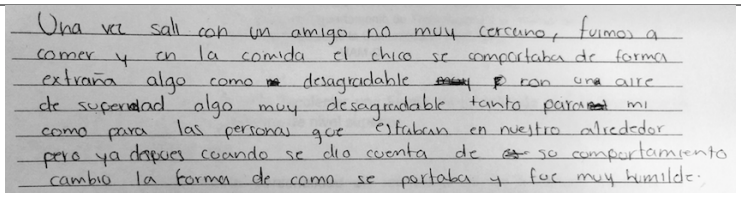
3.3 Evaluation

We used Area under the ROC curve (AUC metric) to measure the performance of classification methods.

3.4 Baseline and preliminaries results

Since we have two modalities we use two baselines: one for text only and one for images only. For the first baseline (B_text) we preprocessed the transcriptions

Table 3. Example of handwritten essay in Spanish, its manual transcription with added tags and its corresponding English translation

	
Manual transcription	Una vez sali <FO:salí> con un amigo no muy cercano, fuimos a comer y en la comida el chico se comportaba de forma extraña algo como <DL> desagradable <DL> <DL> con un <MD> aire de superioridad <MD> algo muy desagradable tanto para <DL> mi <FO:mí> como para las personas que estaban en nuestro alrededor pero ya despues <FO:después> cuando se dio cuenta de <DL> su comportamiento cambio <FO:cambió> la forma de como <FO:cómo> se portaba y fue muy humilde.
English translation	Once I went out with a friend not so close to me, we went to eat and while eating the guy was acting a little weird kind of rude as he was superior to me, it was rude for me as for the people around us but after he realized his behavior he changed the way he was acting and he was humble.

removing symbols and numbers. We keep the tags without the correction (for FO tag) without the description of the drawing (for the D tag). Then, we represented the text using character tri-grams with the TF weighting and an SVM classification.

For the second baseline (B_image) we relied on visual information only. We extracted histograms of oriented gradient (HoG) from the handwritten images and used them as input to the classification model. Nine bins and a cell size of 32×32 pixels was considered. SVM was used for classification.

Participant teams. During the development phase of our challenge we evaluated the methods of two participants over the validation set. The results per trait are shown in Figure 2. We can see that while the overall performance of the text baseline (B_text) is better than the participants, for some traits the participants performed better, particularly for agreeableness (*agr* in the figure) and openness (*ope* in the figure).

The first team called P_JR used the scanned image (visual information only) of the hand-written essay divided into patches and a convolutional neural network (CNN) as the classifier. From the 418 images in color in the train set, they obtained 216 patches of each scan in gray scale and then binarized them resulting in approximately 90,000 images. The CNN consists of five convolutional layers to extract features of the patches and 3 fully connected layers to perform the classification.

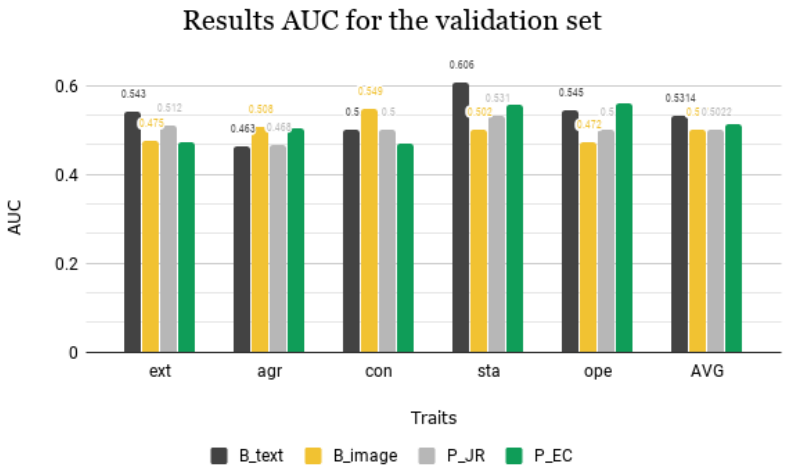


Fig. 2. AUC performance for baselines (B_text and B_image) and participants (P_JR and P_EC) in the development phase of the challenge.

The second team called P_EC used the transcriptions (text) by using a Bag of Words (BoW) and LIWC features for each subject representation. Then they concatenate these vectors with features extracted from the essay images (i.e., slant, size, space, etc.). The image features were extracted at character level, for the character segmentation task they trained a region proposal network over the EMNIST and Chars74k (handwritten) data sets. The features mentioned above were extracted using the character bounding box and the extreme points of the character contour.

For the final phase, results can be seen in Figure 3. Similar to the development, the average performance for the text baselines is slightly better than the results of the participants.

3.5 Discussion

Identifying a subject's personality given a small sample of handwritten text (transcriptions and/or images) is a very difficult task. We can see from the results of this challenge that relying only on the images can be useful to identify traits such as conscientiousness (*con*). Using only text from the transcriptions of each essay can help to identify emotional stability (*sta*) in both subsets as well.

As the results show, both participants have slightly lower performances than our baselines. However, the team called P_EC has a consistently better performance on openness. This team used a combination of both modalities (text and visual information). More experiments need to be done to determine the pertinence of use of both modalities for other traits.

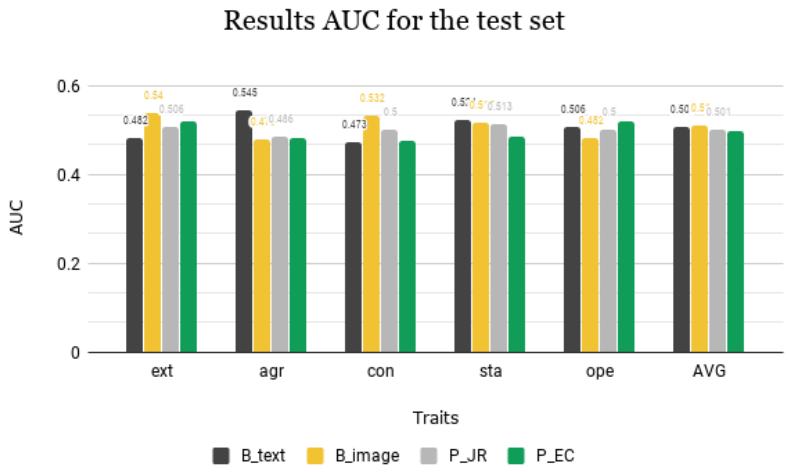


Fig. 3. AUC performance for baselines (B_text and B_image) and participants (P_JR and P_EC) in the final phase of the challenge.

4 Discussion

The two challenges describes in this paper make two large and interesting data sets available to the scientific community. Despite the limited participation the resources can clearly be useful in future research and with a longer time available to experiment this can lead to much more interesting results than in the short time available for the competition. Both resources will remain accessible for research and should help in the multimodal analysis of data about people, in this cases the need for diversity in the retrieval of images from social networks and on the detection of personality traits from handwritten texts.

5 Acknowledgements

Gabriela Ramirez and Esaú Villatoro would like to thank the UAM-C for the facilities provided in this project. Their research was partially supported by CONACYT-Mexico under project grant 258588 and under the Thematic Networks program (Language Technologies Thematic Network, project 281795). Bogdan Ionescu's work was supported by the Romanian Ministry of Innovation and Research, UEFISCDI, project SPIA-VA, agreement 2SOL/2017, grant PN-III-P2-2.1-SOL-2016-02-0002. We would like also to acknowledge the contribution of the task co-organizers: Andrei Jitaru and Liviu Daniel Stefan, University Politehnica of Bucharest, Romania. Hugo Jair Escalante was supported by INAOE. Sergio Escalera's work has been partially supported by the Spanish project TIN2016-74946-P (MINECO/FEDER, UE) and CERCA Programme / Generalitat de Catalunya.

References

1. Escalera, S., Baró, X., Guyon, I., Escalante, H.J.: Guest editorial: Apparent personality analysis. IEEE Transactions on Affective Computing **Forthcoming** (2018)
2. Escalera, S., Gonzàlez, J., Escalante, H.J., Baró, X., Guyon, I.: Looking at people special issue. International Journal of Computer Vision **126**(2-4), 141–143 (2018). https://doi.org/10.1007/s11263-017-1058-y, https://doi.org/10.1007/s11263-017-1058-y
3. Gosling, S.D., Rentfrow, P.J., Jr., W.B.S.: A very brief measure of the big-five personality domains. Journal of Research in Personality **37**(6), 504 – 528 (2003). https://doi.org/https://doi.org/10.1016/S0092-6566(03)00046-1, http://www.sciencedirect.com/science/article/pii/S0092656603000461
4. Ionescu, B., Gînscă, A.L., Zaharieva, M., Boteanu, B.A., Lupu, M., Müller, H.: Retrieving diverse social images at MediaEval 2016: Challenge, dataset and evaluation. In: MediaEval 2016 Workshop (2016)
5. Ionescu, B., Gînscă, A.L., Boteanu, B., Lupu, M., Popescu, A., Müller, H.: Div150multi: A social image retrieval result diversification dataset with multi-topic queries. In: International Conference on Multimedia Systems. pp. 46:1–46:6 (2016)
6. Ionescu, B., Popescu, A., Lupu, M., Gînscă, A.L., Boteanu, B., Müller, H.: Div150cred: A social image retrieval result diversification with user tagging credibility dataset. In: ACM Multimedia Systems Conference. pp. 207–212 (2015)
7. Ionescu, B., Radu, A.L., Menéndez, M., Müller, H., Popescu, A., Loni, B.: Div400: A social image retrieval result diversification dataset. In: ACM Multimedia Systems Conference. pp. 29–34 (2014)
8. Moeslund, T.B., Hilton, A., Krüger, V., Sigal, L. (eds.): Visual Analysis of Humans - Looking at People. Springer (2011). https://doi.org/10.1007/978-0-85729-997-0, https://doi.org/10.1007/978-0-85729-997-0
9. Pennebaker, J.W.: The Secret Life of Pronouns: What Our Words Say About Us. Bloomsbury Press, New York, 1st edn. (2011)
10. Ramírez-de-la-Rosa, G., Villatoro-Tello, E., Jiménez-Salazar, H.: TxPI-u: A resource for personality identification of undergraduates. Journal of Intelligent & Fuzzy Systems **34**(5), 2991–3001 (May 2018). https://doi.org/10.3233/JIFS-169484, https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs169484
11. Sabetghadam, S., Palotti, J., Rekabsaz, N., Lupu, M., Hanbury, A.: TUW @ MediaEval 2015 Retrieving diverse social images task. In: MediaEval 2015 Workshop (2015)
12. Vinciarelli, A., Mohammadi, G.: A survey of personality computing. IEEE Transaction on Affective Computing (2014)

View publication stats