

Detecting Decision Ambiguity from Facial Images

Pavel Jahoda², Antonin Vobecky¹, Jan Cech¹, and Jiri Matas¹

¹ Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague

² Faculty of Information Technology, Czech Technical University in Prague

Abstract—In situations when potentially costly decisions are being made, faces of people tend to reflect a level of certainty about the appropriateness of the chosen decision. This fact is known from the psychological literature. In the paper, we propose a method that uses facial images for automatic detection of the decision ambiguity state of a subject. To train and test the method, we collected a large-scale dataset from “Who Wants to Be a Millionaire?” – a popular TV game show. The videos provide examples of various mental states of contestants, including uncertainty, doubts and hesitation. The annotation of the videos is done automatically from on-screen graphics. The problem of detecting decision ambiguity is formulated as binary classification. Video-clips where a contestant asks for help (audience, friend, 50:50) are considered as positive samples; if he (she) replies directly as negative ones. We propose a baseline method combining a deep convolutional neural network with an SVM. The method has an error rate of 24%. The error of human volunteers on the same dataset is 45%, close to chance.

I. INTRODUCTION

Machines understanding whether a person is uncertain when important decisions are being made might have applications spanning educational environment, marketing, security, etc. This paper tests a conjecture that such state can be estimated by observing face images.

A human conversation is accompanied by several non-verbal signals, especially by facial expressions. Facial expressions facilitate transmitting affective or mental states of interlocutors. Communication of uncertainty/doubts is important to ensure mutual understanding throughout a dialogue, to detect and resolve possible misunderstandings [1].

In this paper, we train a classifier that takes a short video sequence of a face and provides automatic recognition of two states: (1) decision ambiguity, (2) decision unambiguous. For the first state, the uncertainty/doubts or a hesitation are present, while missing in the latter. We trained on a dataset constructed from popular TV quiz “Who Wants to be a Millionaire?” [2]. The dataset contains hundreds of videos with many identities that capture emotional and mental states of contestants. The decision ambiguity is made explicit when a contestant asks for help. See examples of highly scoring faces of both classes in Fig. 1.

Classifying six basic facial expressions that are connected with emotions (anger, disgust, fear, joy, sadness and surprise) [3] is a standard problem in computer vision with

The research was supported by the Czech Science Foundation project GACR P103/12/G084, by the Technology Agency of the Czech Republic TE01020415, and by the CTU student grant SGS17/185/OHK3/3T/13.



Fig. 1. Images correctly classified, with high confidence, into the positive (top) and negative (bottom) classes, respectively. The positive class covers decision ambiguity – the contestants asked for help. For the negative class, the help was available, but not used.

many solutions: using hand-crafted features [4] and more recently employing deep learning [5], [6] to name a few.

On the other hand, attempts to automatically recognize other cognitive or mental states from facial images are rather rare. For instance, there is work on detecting fatigue, e.g. [7], a depression [8], recognizing acceptance or rejection in negotiations [9], dialog engagement [10], agreement/disagreement and cluelessness [11], or puzzlement [12]. To the best of our knowledge, we are not aware of any work on automatic detection of uncertainty from facial images or sequences.

The reason might be the absence of a suitable dataset. There are standard datasets of the six basic expressions, e.g. [13], [14], usually acquired in a controlled laboratory setup and the emotions are posed by non-professional volunteers. The emotions are often exaggerated and thus not very realistic. Few exceptions exist: expression examples collected from movies [15] or social media [16]. We propose to use the TV quiz videos to detect the decision ambiguity.

Facial expressions are performed by contracting a subset of facial muscles. The Facial Action Coding System (FACS), a taxonomy of the muscle activations, was developed by Ekman and Friesen [17], [18] inspired by Hjorstjö [19]. Any facial expression is described by certain facial action units.

Signaling uncertainty/doubts is shown to be connected with typical facial expressions in [20]. The study concludes through a role playing procedure and manual labeling that a subset of action units are consistently triggered. In this paper, we examine these findings on our dataset.

To summarize, our contributions are: (1) We collected a dataset of about 1,600 videos of facial expressions from TV quiz “Who Wants to Be a Millionaire”. The annotation of the uncertainty is done automatically by localizing events when



Fig. 2. Frames from the original footage of the “Who Wants to Be a Millionaire” game show. Relevant events of the game were found automatically, analyzing on-screen graphics. Left: the contestant had asked for help (phone call). Right: the “answer locked” event.

a contestant asked for help (audience, friend, 50:50). (2) We created a *new baseline method* based on deep learning to recognize uncertainty. To assess the accuracy of the baseline, we compared it with human estimates. The proposed method outperformed the average human annotator significantly.

The rest of the paper is structured as follows. The dataset is presented in Sec. II, the proposed method is described in Sec. III, the experiments are reported in Sec. IV. Sec. V concludes the paper.

II. THE MILLIONAIRE DATASET

We collected a dataset composed of hundreds of episodes of the popular television game show “Who Wants to Be a Millionaire.” The game show of British origin exists since 1998 and its franchise is still broadcast in many countries worldwide [2].

The idea of the game is summarized below. Each game has one contestant who is asked general knowledge questions increasing in both complexity and the monetary reward. Four possible answers are given (A,B,C,D) and the contestant chooses one of them. The game ends if the answer is wrong (a contestant loses the payout), or the contestant gives up and takes the payout. If a contestant is *uncertain* he/she has three options for help (from a friend on the phone, from audience voting, or 50:50 which means that two wrong answers disappear). Each of the help options can be used only once.

We downloaded 1,637 videos of total length 41,261 minutes that we found on YouTube. The videos are at least 20 minutes long and at least of 480p image resolution. Most of the videos we used are in English (from UK, US, Nigeria), namely 439 videos lasting 11,353 minutes together. The remaining videos, from Czech Republic, Vietnam, Turkey, India, Afghanistan and Romania, were not used in our study.

Although we use only the video, the dataset content is in fact multi-modal: the video and the audio streams, automatically generated subtitles from speech to text (from YouTube). Moreover, for every frame of the video, we detected faces by commercial multi-view detector [21], and detected and recognized scene texts by method [22].

Since the format of the game is almost identical in all its variants, and since the game states are apparent in on screen graphics, the annotation is done almost fully automatically with a minimum manual effort. The question with all possible answers is always shown as well as it is also shown graphically when the help option is taken, see Fig. 2. Hence, to

detect temporal events of game state changes, we proceed as follows. First the on-screen graphics masks are set up. Then it is scanned through all frames for events of: ‘game started’, ‘question posed’, ‘help taken’, ‘answer locked’, ‘correct answer revealed’, and ‘game ended’. It was done either by simply detecting the presence of the specific mask in a frame or by color matching (e.g. orange color for locked answer, green color for the correct answer). Then the sequence starting on the ‘question posed’ event ending on the ‘answer locked’ event lasting 10 seconds maximum was clipped out. The label was positive (ambiguous decision/uncertainty) if the help was asked, and negative (decision clear) if the help was not used while it was available.

To distinguish the contestant’s face from faces of other people (e.g. a host or a random audience member), we used the VGG Face descriptor. The descriptor is the response of the penultimate layer of the VGG Face network [23]. The host is the same for one or many series, so several examples of him were manually found. Then the host is represented by a component-wise median descriptor over the examples. Then within a clip, VGG descriptors of all faces were clustered into two clusters representing the contestant and the host plus possible outliers. The contestant is represented by the median of the cluster that is more distinct to the host’s descriptor. Contestant’s faces are found by thresholding L_2 distance between every detected face and the median descriptor. This simple heuristic turned out to produce no mismatches in the annotation.

For the uncertainty dataset, we erase all frames which capture a different face than of the contestant. As a result, the sequences are of various lengths mainly due to shots to the host. All sequences are 50–150 frames long (25 FPS). The data were randomly split into disjoint training, validation and test subsets, namely 1,628 training sequences, 90 validation and 90 test sequences. The training set contains 502 positive and 1,126 negative sequences. The validation and test sequences were balanced (same number of positive and negative sequences) and we made sure that no individual appears simultaneously in the training, validation or test sets.

We believe the samples collected this way represent challenging but realistic examples of the uncertainty/ambiguity in the decision making. In the remaining text we will show that a simple method that was trained on the dataset outperforms the average human estimate in the classification accuracy by a significant margin.

III. METHOD DESCRIPTION

We propose two baseline approaches for the uncertainty classification problem. We assume we are given a short video sequences of similar but not necessarily the same length to be discriminated into two classes: the positive class (uncertainty), and the negative class (background). The sequence-level labels are given, while the frame labels are not available.

The problem is similar in spirit to Multiple Instance Learning (MIL) problem [24], which is a variant of semi-supervised learning where instances belong to bags and only

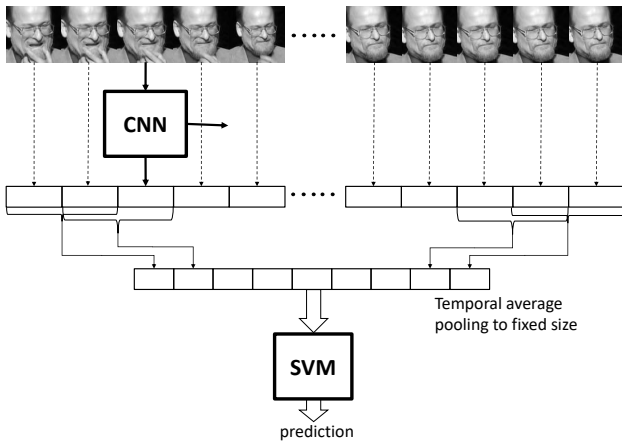


Fig. 3. The proposed architecture for sequence classification.

bag labels are given while instance labels are not accessible. The original MIL problem strictly requires that negative bags does not contain any positive instances and a positive bag contains at least a single positive instance.

In theory, we could assume that the positive sequence contains at least a single frame with a face expression of uncertainty. However, we could hardly assume there are absolutely no uncertainty frames in a negative sequence. In reality, traces of uncertainty expressions may appear in the negative bags especially in high stake situations, but the distributions of facial expressions over the sequence differ between the classes. Moreover, unlike MIL bags, our sequences have a structure. The expression evolves over the time after a question is known to a contestant.

First, we trained the naive classifier that assigns identical sequence labels to all their frames, see Subsec. III-A. This is clearly suboptimal, since certain facial expressions are common for both classes, e.g. a neutral expression. Nevertheless, we then use the output of the frame classifier to classify sequences, see Subsec. III-B.

A. Frame Classifier

We adopted a standard convolutional network with architecture similar to the VGG net [23]. The network takes 100×100 pixel gray scale images as an input. It has only seven convolutional and two fully connected layers. The simpler architecture allows real-time execution on a laptop CPU, which also implies faster training on GPU than the original VGG net. The architecture was borrowed from [25]. The dropout was used for fully connected layers and the batch normalization was inserted before every ReLU.

To initialize the network, we first trained it in a multi-task fully supervised setup to predict age, gender and localize facial landmark points. The training was done using about 80k images from several datasets (Morph [26], LFW [27], 300W [28]), and achieved a competitive prediction accuracy.

Then the uncertainty expression frame classifier was “fine-tuned” starting from the multi-task network with the output layer changed to a single scalar with a logistic loss on top. The pre-training on the large dataset is known to facilitate

training of the novel problem [29].

The frame classifier was trained from 76k labeled samples with equal distribution of positive and negative frames. A standard SGD optimization was used and the training converged after 5 epochs. We selected the parameters of the network with the lowest validation error.

B. Sequence Classifier

The trained network described in the previous section is executed on every frame of training sequences. To classify the sequences we exploit the distribution of the network score over the time. Then a fixed-length representation for sequences of various length is implemented by the temporal pooling. To produce k -dimensional vectorial output, k aggregation windows are evenly deployed over the signal of the network score of the sequence frames. Each of the vector component is calculated by averaging over the window. The windows are of the same length, with a small overlap. Note, the temporal pooling is analogy to the spatial pooling [30] used in object detection.

After that, each sequence is represented by k -dimensional vector and a standard kernel SVM was trained. Fig. 3 shows a diagram of our architecture.

The kernel type, and all the meta-parameters were selected based on the minimum error on the validation set. The lowest validation error occurred for $k = 13$ with a polynomial kernel SVM.

IV. EXPERIMENTS

To demonstrate the validity of the proposed approach, we first present a study with human estimates in Subsec. IV-A before showing results of the proposed classifier in Subsec. IV-B. Moreover, we evaluate accuracy of a classifier trained from automatically detected facial action units as another baseline in Subsec. IV-C.

All evaluations were done on test dataset of 90 sequences (45 positive, 45 negative examples) randomly selected from the dataset with no overlap with the training or validation sets. The classifiers were fit on the training set, all meta-parameters including the trained network epoch were selected based on the minimum error on the validation set. The test set was completely unseen during the training. Subjects present in the test sets do not appear in the training or validation sets.

A. Human estimates

We prepared on-line form where 20 videos were shown (10 positive, 10 negative examples) randomly taken from our test set (90 sequences). The sequences were without the sound track showing only the contestant’s face. Volunteers were asked to guess if a contestant in each video asked for help in the game or replied directly. The volunteers could play the videos as many times as they wanted before making their guess. We instructed the volunteers thoroughly about the problem, making a special emphasis on their understanding which situation of the game the sequences capture. We make no manipulation with the videos (besides cropping the face),

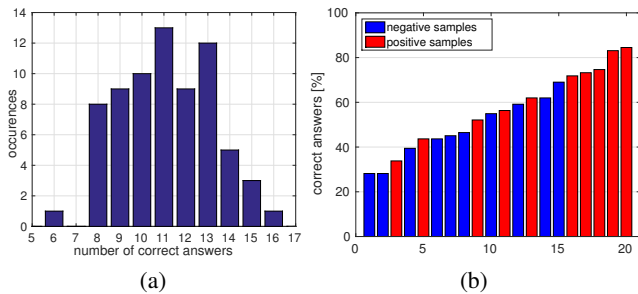


Fig. 4. Decision ambiguity estimation by humans on 20 sequences. The histogram of the number of correct estimates for 72 volunteers (a). The percentage of correct estimates per video-sequence (b).

they were played in the original frame rate. 72 volunteers participated in our study. Each volunteer filled the form only once. The volunteers were mostly students of computer science.

The results are shown in Fig. 4a as a histogram of number of correctly decided sequences. Average and median number of correct estimates was 11.1 and 11 respectively out of 20, which implies the expected error of a human estimate is 45%. The task showed to be challenging for human annotators. Note that certain videos were easier than others, which is shown in Fig. 4b as a percentage of correct estimates per video. The differences between the classes were in general subtle, while for certain videos more obvious.

B. Evaluation of the proposed baseline

The frame classifier based on the deep CNN described in Subsec. III-A achieved the test error 36.5%. This is rather surprising considering the complex mix of facial expressions a contestant makes during the sequence. Predicting the sequence labels by simply averaging the frame scores over the sequence achieved error rate 36.3% only.

The other sequence classifier, Subsec. III-B, which takes into account the temporal structure, further improved the test error to 24.4% on the entire set of 90 sequences and 25% on test subset for the human accuracy (5 wrong classifications out of 20 sequences). The result is significantly superior to human estimates. The reason indicates that the classifier trained from a large dataset is able to exploit subtle statistical differences between the classes and outperforms intuition of an average human.

For the sake of introspection, we calculated activation maps of the trained network by method [31]. The activation map shown in Fig. 5 was averaged over test frames. The locations that are activated correspond mostly to important facial features, especially to the mouth, eyes and eyebrows. Note that work [20] reports which facial action units are activated to express the uncertainty: AU15 (Lip Corner Depressor) + AU17 (Chin Raiser) are always activated, sometimes together with AU1 (Inner Brow Raiser) + AU 2 (Outer Brow Raiser). These conclusions are consistent with our findings.

C. Experiments with Facial Action Units

To further compare the proposed method with more traditional approach to facial expression analysis, we carried

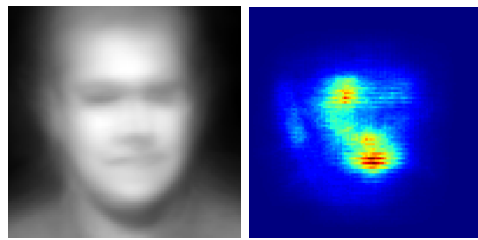


Fig. 5. Average image and the activation map of the trained CNN.

out an experiment with Facial Action Units (FAU) [18]. The FAUs were automatically detected by OpenFace [32], which extracts activation score of 17 FAUs from a face image.

We trained an SVM per frame independently, which results in 48.3% error on the test set. To classify the sequences, the score signals of FAU activations over all sequence frames were re-sampled by moving average into length 30, the value selected on the validation set. Again the SVM classifier taking 510 inputs (17×30) was trained and achieved the error rate 41.8% on the test set.

The result is not favorable for the FAU method. Assuming the FAUs were extracted accurately enough for our dataset, the results suggests that the deep approach provides a better representation, capturing phenomena like hands over the face or a head tilting as shown in Fig. 1.

V. CONCLUSION

In this paper, we introduced the problem of automatic recognition of decision ambiguity from image sequences. For evaluation and training, a dataset from the popular television game show was collected. We implemented a baseline classifier that combines deep net with SVM. The average human error was 45% and the classifier achieved an error rate 25% on the same test set of 20 sequences, and 24% on the test set of 90 sequences. The activation maps of the trained neural network can be interpreted as triggering the same Facial Action Units as reported by psychologists [20].

Several types of decision ambiguity/certainty exists, e.g. a contestant is certain about either the correct answer or the decision to ask for help. Our study concludes that it is indeed possible to predict with a fair accuracy whether a contestant will take a help, which is an expression of one kind of decision ambiguity. It would be interesting to explore the generalization of the method in a completely different setup.

There are several limitations of the current baseline. For instance, the faces are not precisely aligned and the training is not done in an end-to-end fashion. Handling these issues might further improve the accuracy.

The dataset will be published with the paper. The annotation includes all relevant events in the game show and bounding boxes of the faces of contestants. The results of automatic scene text recognition in all frames are included, despite not being used in the current study. We believe that the richness of emotions and the variety of cultures where the show was aired, together with the data multi-modality, may enable studies beyond the research reported in the paper.

REFERENCES

- [1] M. Stone and I. Oh, "Modeling facial expression of uncertainty in conversational animation," in *Modeling Communication with Robots and Virtual Humans*, 2008, INCS 4930.
- [2] https://en.wikipedia.org/wiki/Who_Wants_to_Be_a_Millionaire.
- [3] P. Ekman, *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life*. Henry Holt and Company, 2007.
- [4] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, 2009.
- [5] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proc. ICMI*, 2015.
- [6] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, "Dexpression: Deep convolutional neural network for expression recognition," 2015, arXiv:1509.05371.
- [7] A. Sahayadhas, K. Sundaraj, and M. Murugappan, "Detecting driver drowsiness based on sensors: A review," *MDPI open access: sensors*, no. 12, 2012.
- [8] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padiilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009.
- [9] S. Park, S. Scherer, J. Gratch, P. J. Carnevale, and L.-P. Morency, "I can already guess your answer: Predicting respondent reactions during dyadic negotiation," *IEEE Trans. on Affective Computing*, vol. 6, no. 2, 2015.
- [10] D. Klotz, J. Wienke, J. Peltason, B. Wrede, S. Wrede, V. Khalidov, and J.-M. Odobez, "Engagement-based multi-party dialog with a humanoid robot," in *SIGDIAL*, 2011.
- [11] D. W. Cunningham, M. Kleiner, C. Wallraven, and H. H. Bühlhoff, "The components of conversational facial expressions," in *Applied perception in graphics and visualization*, 2004.
- [12] J. Wang, X. Ma, J. Sun, Z. Zhao, and Y. Zhu, "Puzzlement detection from facial expression using active appearance models and support vector machines," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 7, no. 5, 2014.
- [13] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete expression dataset for action unit and emotion-specified expression," in *CVPR Workshop on Human Communicative Behavior Analysis*, 2010.
- [14] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *ICME*, 2005.
- [15] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Trans. on Multimedia*, vol. 19, no. 3, 2012.
- [16] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe, "The more the merrier: Analysing the affect of a group of people in images," in *International Conference and Workshops on Automatic Face and Gesture Recognition.*, 2015.
- [17] P. Ekman and W. Friesen, *Facial action coding system: Investigator's Guide*. Consulting Psychologists Press, 1978.
- [18] P. Ekman, W. Friesen, and J. Hager, "Facial action coding system (FACS): Manual and investigators guide," in *A Human Face*, 2002.
- [19] C.-H. Hjorstjö, *Man's face and mimic language*. Studentlitteratur, Lund, Sweden, 1970.
- [20] P. E. Ricci Bitti, L. Bonfiglioli, P. Melani, R. Caterina, and P. Garotti, "Expression and communication of doubt/uncertainty through facial expression," *Journal of Theories and Research in Education*, vol. 9, no. 1, 2014.
- [21] J. Sochman and J. Matas, "Waldboost – learning for time constrained sequential detection," in *Proc. CVPR*, 2005. Commercial implementation by Eyedea Recognition, Inc. www.eyedea.cz.
- [22] M. Busta, L. Neumann, and J. Matas, "Deep textspotter: An end-to-end trainable scene text localization and recognition framework," in *ICCV*, 2017.
- [23] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.
- [24] S. Andrews, I. Tsochantaris, and T. Hofmann, "Support vector machines for multiple-instance learning," in *NIPS*, 2002.
- [25] V. Franc and J. Cech, "Face attribute learning from weakly annotated examples," in *Automatic Face and Gesture Recognition Workshops, Biometrics in the Wild*, 2017.
- [26] K. J. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *IEEE 7th International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, April 2006, pp. 341–345.
- [27] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.
- [28] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing (IMAVIS)*, 2016, special Issue on Facial Landmark Localisation "In-The-Wild".
- [29] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *CVPR*, 2014.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. on PAMI*, vol. 37, no. 9, 2015.
- [31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Why did you say that? visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.
- [32] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision*, 2016.