Learning Grimaces by Watching TV

Sam Albanie
http://www.robots.ox.ac.uk/~albanie
Andrea Vedaldi
http://www.robots.ox.ac.uk/~vedaldi

Engineering Science Department
Univeristy of Oxford
Oxford, UK

Abstract

Differently from computer vision systems which require explicit supervision, humans can learn facial expressions by observing people in their environment. In this paper, we look at how similar capabilities could be developed in machine vision. As a starting point, we consider the problem of relating facial expressions to objectively-measurable events occurring in videos. In particular, we consider a gameshow in which contestants play to win significant sums of money. We extract events affecting the game and corresponding facial expressions objectively and automatically from the videos, obtaining large quantities of labelled data for our study. We also develop, using benchmarks such as FER and SFEW 2.0, state-of-the-art deep neural networks for facial expression recognition, showing that pre-training on face verification data can be highly beneficial for this task. Then, we extend these models to use facial expressions to predict events in videos and learn nameable expressions from them. The dataset and emotion recognition models are available at http://www.robots.ox.ac.uk/~vgg/data/facevalue.

1 Introduction

Humans make extensive use of facial expressions in order to communicate. Facial expressions are complementary to other channels such as speech and gestures, and often convey information that cannot be recovered from the other two alone. Thus, understanding facial expressions is often necessary to properly understand images and videos of people.

The general approach to facial expression recognition is to label a dataset of faces with either *nameable expressions* (*e.g.* happiness, sadness, disgust, anger, etc.) or *facial action units* (movements of facial muscles such as tightening the lips or raising an upper eyelid) and then learn a corresponding classifier, for example by using a deep neural network. In contrast, humans need not to be *explicitly told* what facial expressions means, but can learn that by associating facial expressions to how people react to particular events or situations.[1]

In order to investigate whether algorithms can also learn facial expressions by establishing similar associations, in this paper we look at the problem of *relating facial expressions to objectively-quantifiable contextual events in videos*. The main difficulty of this task is that there is only a weak correlation between an event occurring in a video and a person showing a particular facial expression. However, learning facial expressions in this manner has three important benefits. The first one is that it grounds the problem on objectively-measurable

[1] Generating certain facial expressions is an innate ability; however, recognizing facial expression is a learned skill.

Figure 1: ***FaceValue*** dataset. We study facial expressions from objectively-measurable events occurring in the "Deal or No Deal" gameshow. *Top:* detection of an event at round $t = 6$ in the game. Left: a box is opened, revealing to the contestant that her prize is *not* the one of value $x_t = £5$. Since this is a low amount, well below the expected value of the prize of $E_5 = £17,331$, this is a "good" event for the contestant. Right: the contestant's face, intuitively expressing happiness, is detected. Note also the overlay for $x_t = £5$ disappearing from a frame to the next; our system can automatically read such cues to track the state of the game. *Bottom:* four example tracks, the top two for "good" events and the bottom two for "bad" events, as defined in the text.

quantities, whereas labelling emotions or even facial action units is often ambiguous. The second benefit is that contextual information can often be labelled in videos fully or partially automatically, obviating the cost of collecting large quantities of human-annotated data for data-hungry machine learning algorithms. Finally, the third advantage is that the ultimate goal of face recognition in applications is not so much to describe a face, but to infer from it information about a situation or event, which is tackled directly by our study.

Concretely, our first contribution (Sect. 2; Fig. 1) is to develop a novel dataset, *FaceValue*, of faces extracted from videos together with objectively-measurable contextual events. The dataset is based on the "Deal or No Deal" TV program, a popular game where contestants can win or lose significant sums of money. Using a semi-automatic procedure, we extract significant events in the game along with the player (and public) reaction. We use this data to predict from facial expressions whether events are "good" or "bad" for the contestant. To the best of our knowledge, this is the first example of leveraging gameshows in facial expression understanding and the first study aiming to relate facial expressions to people's activities.

Our second contribution is to carefully assess the difficulty of this problem by establishing a human baseline and by extending the latter to existing expression recognition datasets for comparison (Sect. 3). We also develop a number of *state-of-the-art expression recognition models* (Sect. 4) and show that excellent performance can be obtained by transferring deep neural networks from face verification to expression recognition. Our final contribution is to extend such systems to the problem of recognising *FaceValue* events from facial expressions (Sect. 5). We develop simple but effective pooling strategies to handle face tracks, integrating them in deep neural network architectures. With these, we show that it is not only possible to predict events from facial expressions, but also to learn nameable expressions by looking at people spontaneously reacting to events in TV programs.

Dataset	Size	Labelling Technique	Expressions	Labels
FER	35,887 Faces	Internet search	Mixed	6+1 emotions
AFEW 5.0	1,426 Clips	Subtitles	Acted	6+1 emotions
SFEW 2.0	1,635 Faces	Subtitles	Acted	6+1 emotions
AM-FED	168,359 Faces	Human experts	Spontaneous	FACS
FaceValue (ours)	192,030 Faces	Metadata extraction	Spontaneous	Event Outcome

Table 1: **Comparison of emotion-based datasets of faces in challenging conditions.**

1.1 Related work

Facial expressions are a non-verbal mode of communication complementary to speech and gestures [1, 11]. They can be produced unintentionally [10], revealing hidden states of the actor in pain or deception detection [2]. Facial expressions are commercially valuable, attracting increasing investment from advertising agencies that seek to understand and manipulate the consumer response to a product [12] and corresponding regulatory attention [31].

Face-related tasks such as face detection, verification and recognition have long been researched in computer vision with the creation of several labelled datasets: FDDB [18], AFW [39] and AFLW [21] for face detection; and LFW [16] and VGG-Face [28] for face recognition and verification. Face detectors and identity recognizers can now rival the performance of humans [33]. Facial expression recognition has also received significant attention in computer vision, but it presents a number of additional subtleties and difficulties which are not found in face detection or recognition. The main challenge is the consistent labelling of facial expressions which is difficult due to the subjective nature of the task. A number of coding systems have been developed in an attempt to label facial expressions objectively, usually at the level of atomic facial movements, but even human experts are not infallible in generating such annotations. Furthermore, getting these experts to annotate a dataset is expensive and difficult to scale [27]. Another issue is the "authenticity" of facial expressions, arising from the fact that several datasets are acted [34], either specifically for data collection [25] [24] [14] or indirectly as data is extracted from movies [8]. Our *FaceValue* dataset sidesteps these problems by recording spontaneous reactions to objectively-occurring events in videos.

Examples of datasets which contain challenging variations in pose, lighting conditions and subjects are given in Table 1. Of these, two in particular have received significant research interest as popular benchmarks for facial expression recognition. The *Static Facial Expression in the Wild 2.0* (SFEW-2.0) data [7] (used in the *EmotiW* challenges [8]) consists of images from movies which collectively contain 1,635 faces labelled with seven emotions (this dataset was constructed by selectively extracting individual frames from AFEW-5.0 [9]). The *Facial Expression Recognition 2013* (FER-2013) dataset [13], which formed the basis of a large Kaggle competition, contains 35k images labelled with the same seven emotions. These datasets were used to develop several state-of-the-art emotion recognition systems. Among the top-performing ones, the authors of [37] and [19] propose ensembles of deep network trained on the FER and SFEW-2.0 data. There are also several commercial implementations of expression recognition, such as CMU's IntraFace [5] and the Affectiva face software.

2 *FaceValue*: expressions in context

In this section we describe the *FaceValue* dataset (Fig. 1) and how it was collected.

Data source. The "Deal or No Deal" TV game show[2] was selected as the basis for our data for a number of reasons. First, it contains a very significant amount of data. The show has been running nearly daily in the UK for the past eleven years, totalling 2,929 episodes. Each episode focuses on a different player and lasts for about forty minutes. Furthermore, the same or very similar shows are or were aired in dozens of other countries. Second, the game is based on simple rules and a sequence of discrete events that are in most cases easily identifiable as positive or negative for the player, and hence can be expected to induce a corresponding emotion and facial expression. Furthermore, these events are easily detectable by parsing textual overlays in the show or other simple patterns. Thirdly, since there is a single player, it is easy to identify the person that is directly affected by the events in the video and the camera tends to focus on his/her face.

An example of the in-game footage and data extraction pipeline is shown in Fig. 1. The rules of the game are easily explained. There are $n = 22$ possible cash prizes $\mathcal{X}_0 = \{p_1, p_2, \ldots, p_n\}$ where prizes $p_1 < p_2 < \cdots < p_n$ range from 1p up to £250,000. Initially the player is assigned a prize $x_0 \in \mathcal{X}_0$ but does not know its value. Then, at each round of the game the player can randomly extract (realised as opening a box, see Fig. 1 top-left) one of the prizes $x_t \neq x_0$ from $\mathcal{X}_t$ and reveal it, resulting in a smaller set $\mathcal{X}_t = \mathcal{X}_{t-1} - \{x_t\}$ of possible prizes. Through this process of elimination the player obtains information about his/her prize x_0. Occasionally the player is offered the opportunity to leave the game with a prize p_d ("deal") determined by the game's host or to continue playing ("no deal") and eventually leave with x_0.

The expected value E_t of the win x_0 at time t is $E_t = \operatorname{mean} \mathcal{X}_t$. When a prize x_t is removed from $\mathcal{X}_{t-1}$, the player perceives this as a "good" event if $E_t > E_{t-1}$, which requires $x_t < E_{t-1}$, and a "bad" event otherwise. In practice we conservatively require $E_t > E_{t-1} + \Delta$ for a good event, where $\Delta = £750$. Interestingly, the game is continued even after the player has taken a "deal"; in this case the roles of "good" and "bad" events are reversed as the player hopes that the accepted deal p_d is higher than the prize x_0 he/she gave up.

Dataset content. The data in *FaceValue* is defined as follows. Faces are detected right after a new prize x_t is revealed for about seven seconds. These faces are collected in a "face track" $\mathbf{f}_t$. Furthermore, the face track is assigned the binary label:

$$y_t = d_t \times \begin{cases} +1, & x_t + \Delta < E_{t-1}, \\ -1, & x_t + \Delta \geq E_{t-1}, \end{cases}$$

where d_t is $+1$ if the deal was *not* taken so far, and -1 otherwise. Note that there are several levels of indirection between y_t and a particular expression being shown in $\mathbf{f}_t$. For example, a player may not perceive a good or bad event according to this simple model, or could be responding to a stroke of bad luck with an ironic smile. The labels y_t themselves, however, are completely objective.

Data is extracted from 102 episodes of the show, resulting in 192,030 frames distributed over 2,118 labelled face tracks. Shows are divided into training, validation and test sets, which also means that mostly different identities are contained in the different subsets.

[2] Outside of computer vision, the interesting decision making dynamics of contestants in a high-stakes environment during the "Deal or No Deal" game show have attracted research by economists [30].

Data extraction. One advantage of studying facial expressions from contextual events is that these are often easy to detect automatically. In our case, we take advantage of two facts. First, when a prize is removed from the set $\mathcal{X}_t$, this is shown in the game as a box being opened (Fig. 1 top-left). This scene, which occurs systematically, is easy to detect and is used to mark the start of an event. Next, the camera moves onto the contestant (Fig. 1 top-middle) to capture his/her reaction. Faces are extracted from the seven seconds that immediately follow the event using the face detector of [20] and are stored as part of the face track $\mathbf{f} = (f_1, f_2, \ldots, f_T)$. Occasionally the camera may capture the reaction of a member of the public; while it would be easy to distinguish different identities (e.g. by using the VGG-Faces model of Sect. 4), we prefer not to as the public is sympathetic with the contestant and tends to react in a similar manner, improving the diversity of the collected data. Finally, the value of the prize x_t being removed can be extracted either from the opened box using a text spotting system or, more easily, by looking at which overlay is removed (Fig. 1 top-right). After automatic extraction, the data was fully checked manually for errors to ensure its quality.

3 Benchmark data and human baselines

As *FaceValue* defines a new task in facial expression interpretation, in this section we establish a human baseline as a point of comparison with computer vision algorithm performance. In order to compare *FaceValue* to existing facial expression recognition problems we establish similar baselines for two standard expression recognition datasets, FER and SFEW 2.0, introduced below.

Benchmark datasets: FER and SFEW 2.0. The FER-2013 data [13] contains 48×48 pixel images obtained by querying Google image search for 184 emotion-related keywords. The dataset contains 35,887 images divided into 4,953 "anger", 547 "disgust", 5,121 "fear", 8,989 "happiness", 6,077 "sadness", 4,002 "surprise" and 6,198 "neutral" further split into training (28,709), public test (3,589) and private test (3,589) sets. Goodfellow *et al*. [13] note that this data is likely to contain label errors. However, their own human study obtained an average prediction accuracy of $65 \pm 5\%$, which is comparable to the $68 \pm 5\%$ performance obtained by expert annotators on a smaller but manually-curated subset of 1,500 acted images.

The SFEW-2.0 data [7] contains selected frames from different videos of the *Acted Facial Expressions in the Wild* (AFEW) dataset [6] assigned to either: 225 "angry", 75 "disgust", 124 "fear", 256 "happy", 228 "neutral", 234 "sad" and 150 "surprise". The training, validation and test splits are provided as part of the EmotiW challenge [8] and are adopted here. The AFEW data was collected by searching movie close captions for emotion-related keywords and then manually curating the results, generating a smaller number of labelled instances than FER.

Human baselines. For each dataset we consider a pool of annotators, most of which are not computer vision experts, and ask them to predict the label associated with each face. In order to motivate annotators to be as accurate as possible, we pose the annotation process as a challenge. The goal is to guess the ground-truth label of an image and a score displaying the annotators' prediction accuracy is constantly updated. Ultimately, annotators performances are entered in a leaderboard. We found that this simple idea significantly improved the annotators' performance.

The dataset instances selected for the annotation tasks were constructed as follows. From FER, a random sample of 500 faces was extracted from the Public Test set. From SFEW 2.0, the full Validation set (383 samples) was used (faces were extracted from each image as described in section 4). From *FaceValue*, a random sample of 250 face tracks was extracted from the validation set, each of which was transformed into an animated GIF to allow annotators to see the face motion. Performance on each dataset was evaluated by partitioning into five folds, each of which was annotated by a separate pool. Every face instance across the three datasets received at least four annotations.

On FER, our annotators achieved lower performance than results previously reported in [13] (58.2% overall accuracy vs 65%). However, we also noted a significant variance between annotators ($\pm 8.0\%$), which means that at least some of them were able to match or exceed the 65% mark. The unevenness of the annotators shows how difficult or ambiguous this task can be even for motivated humans. The annotators found SFEW-2.0 a more challenging task, obtaining an average accuracy of $53.0 \pm 9.4\%$ overall. One possible reason for this difference is the manner in which the datasets were constructed. FER faces were retrieved using Internet search queries which likely returned fairly representative examples of each expression; in contrast SFEW images were extracted from movies. On *FaceValue*, the average annotator accuracy was $62.0 \pm 8.1\%$. Since the classification task was binary, to facilitate a comparison with algorithmic approaches, the ROC-AUC was also computed for each annotator, resulting in an annotator average of $71.0 \pm 5\%$. The relatively low scores of humans on each dataset illustrate the particularly challenging nature of the task. This difficulty is underlined by the low levels of inter-annotator agreement (measured using *Fleiss' kappa*) on the three datasets of 0.574, 0.424 and 0.491 respectively.

4 Expression recognition networks

In this section we develop state-of-the-art models for facial expression recognition in the two popular emotion recognition benchmarks of Sect. 3, namely FER and SFEW 2.0. Deep networks are currently the state-of-the-art models for emotion recognition, topping two of the last three editions of the *Emotion recognition in the Wild* (EmotiW) contest [23]. While the standard approach is to learn large ensembles of deep networks [19, 37], here we show that a single network can in fact be competitive or better than such ensembles if trained effectively. In order to do so we expand the available training data by pre-training models on other face recognition tasks, and in particular face identity verification, using the recent VGG-Faces dataset [29].

Architectures and training. We base our models on four standard CNN architectures: AlexNet [22], VGG-M [3], VGG-VD-16 [35] and ResNet-50 [15]. AlexNet is used as a reference baseline and is pre-trained on the ImageNet ILSVRC data [32]. VGG-VD-16 is pre-trained on a recent dataset for face verification called VGG-Faces [29]. This model achieves near state-of-the-art verification performance on the LFW [16] benchmark; however, it is also extremely expensive. Thus, we train also a smaller network, based on the VGG-M configuration. All models are trained with batch normalization [17] and are implemented in the MatConvNet framework [36].

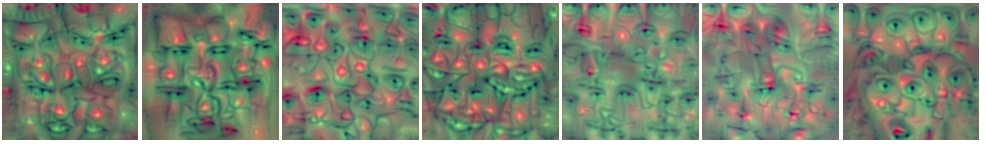
Statistics such as image resolution and the usage of colour in the target datasets, and FER in particular, differ substantially from LFW and VGG-Faces. Nevertheless, we found that simply rescaling the smaller FER images to the higher VGG-Faces resolution together with duplicating the grayscale intensities for the three colour channels produced excellent results.

Model	Pretraining	Test (Public)	Test (Private)
AlexNet	ImageNet	62.44%	63.28%
VGG-M	ImageNet	66.04%	67.57%
Resnet-50	ImageNet	67.79%	69.02%
VGG-VD-16	ImageNet	66.92%	70.38%
AlexNet	VGGFaces	70.47%	71.44%
VGG-M	VGGFaces	71.08%	72.08%
Resnet-50	VGGFaces	69.23%	70.33%
VGG-VD-16	VGGFaces	**72.05%**	**72.89%**
HDC⋆ [19]	-	-	70.58%
HDC †† [19]	-	-	72.72%

Table 2: Accuracy on FER-2013 of different CNN models and training strategies.

Model	Pretraining	Val	Test
AlexNet	VGGFaces	37.67%	-
VGG-M	VGGFaces	42.90%	-
Resnet-50	VGGFaces	47.48%	-
VGG-VD-16	VGGFaces	43.58%	-
AlexNet	VGGFaces+FER	38.07%	50.81%
VGG-M	VGGFaces+FER	47.02%	53.49%
Resnet-50	VGGFaces+FER	50.91%	45.97%
VGG-VD-16	VGGFaces+FER	**54.82%**	59.41%
CMU⋆ [37]	FER combined	52.29%	58.06%
HDC⋆ [19]	FER + TFD	52.50%	57.3%
CMU †† [37]	FER combined	55.96%	61.29%
HDC †† [19]	FER + TFD	52.80%	**61.6%**

Table 3: Accuracy on SFEW-2.0 of different CNN models and training strategies



Anger Disgust Fear Happiness Neutral Sadness Surprise

Figure 2: Visualizations of the FER emotions for the VGG-VD-16 model.

We also experimented with the other approach of pretraining by reducing the resolution and removing colour information from VGG-Faces; while this resulted in very competitive and more efficient networks, the full resolution models were still a little more accurate and are used in the rest of the work.

After pre-training, each model is trained on the FER or SFEW 2.0 training set with a fine tuning ratio of 0.1. This is obtained by retaining all but the last layer, performing N-way classification, where N is the number of possible facial expression classes.

Results. Table 2 compares the different architecture and the state-of-the-art on FER. When reporting ensemble models, ⋆ denotes the best single CNN and †† denotes the ensemble. The best previous results on FER is 72.72% accuracy, obtained using the hierarchical committee of deep CNNs described in [19], combining more than 36 different models. By comparison, VGG-VD-16 pre-trained on VGG-Faces achieves a slightly superior performance at 72.89%. VGG-M achieves nearly the same performance (-0.8%) at a substantially reduced computational cost. We also note the importance of choosing a face-related pre-training set, as pre-training in ImageNet loses 3-4% of performance.

Table 3 reports the results on the SFEW-2.0 dataset instead. Since the dataset itself consists of labelled scene images, we use the faces extracted by the accurate face detection pipeline described in [37] which applies an ensemble of face detectors [4, 38, 39]. As SFEW is much smaller than FER, pre-training is in this case much more important. The best result achieved by any of the four models pre-trained with ImageNet only was 31.19%. Pre-training on VGG-Faces produced substantially better results (+10%), and pre-training on VGG- Faces and FER-Train produced better still (+18%). The best single model, VGG-VD-16, achieves better performance than existing single and ensemble networks (+2.5%) on the validation set, and better performance than all but the ensembles of [19, 37] on the test

Model	Pre-training	Method	Val.	Test
VGG-M	VGGFace+FER	voting	0.656	0.592
VGG-VD	VGGFace+FER	voting	0.653	0.618
VGG-M	VGGFace	pooling arch.	0.764	0.691
VGG-VD	VGGFace	pooling arch.	0.726	0.671
VGG-M	VGGFace+FER	pooling arch.	**0.794**	**0.722**
VGG-VD	VGGFace+FER	pooling arch.	0.741	0.675

Table 4: ROC-AUC on *FaceValue*

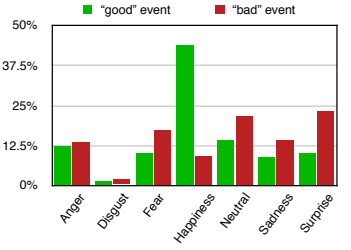


Figure 3: FER expressions from *FaceValue*.

set (-2%).

Visualizations. While CNNs perform well, it is often difficult to understand what they are learning given their black-box nature. Here we use the technique of [26] to visualize the the best FER/SFEW model. This technique seeks to find an image I which, under certain regularity assumptions, maximizes the CNN confidence $\Phi_c(I)$ that I represents emotion c. Results are reported in Fig 2 for the VGG-VD-16 model trained on the FER dataset. Notably, the reconstructed pictures are mosaics of parts representative of the corresponding emotions.

5 Relating facial expressions to events in videos

In this section we focus on the main question of the paper i.e. whether facial expressions can be used to extract information about events in videos.

Baselines: individual frame prediction and simple voting. As baseline, a state-of-the-art emotion recognition CNN Φ is applied to each frame in the face track. The T faces in a face track $\mathbf{f} = (f_1, \ldots, f_T)$ are individually classified by $\Phi(f_t)$ and results are pooled to predict whether the event is positive $y = +1$ or negative $y = -1$. Positive emotions (happiness) vote for the first case, negative emotions (sadness, fear, anger, disgust) for the second and neutral/surprise emotions are ignored. The label with the largest number of votes in the track wins.

Pooling architectures. There are two significant shortcomings in the baseline. First, it assumes a particular map between emotions in existing datasets and positive and negative events in *FaceValue*. Second, it integrates information across frames using an ad-hoc voting procedure which may be suboptimal. In order to address these shortcomings we learn on *FaceValue* a new model that explicitly pools information across frames in a track. A pre-trained network $\Phi = \Phi_1 \circ \Phi_2$ is split in two parts. Then, the first part is run independently on each frame, the results are pooled by either average or max pooling across time and the result is fed to Φ_2 for binary classification: $\Phi(\mathbf{f}) = \Phi_2 \circ \text{pool}(\Phi_1(f_1), \ldots, \Phi_1(f_T))$. The resulting architecture is fine-tuned on the *FaceValue* training set.

In practice, we found that the best results were obtained by using the emotion recognition networks such as VGG-VD-16 trained on the FER data (Sect. 4). All layers up to fc7, producing 4,096 dimensional feature vectors, are retained in Φ_1. The best pooling function was found to be averaging followed by L^1 normalization of the 4,096 dimensional features. The last layer Φ_8 is fully connected (in practice, this layer is a linear predictor). CNNs are trained using hinge loss, which generally performs better than softmax for binary classification.

Results. Table 4 reports the performance of different model variants on *FaceValue*. Similarly to Table 3, pre-training on VGG-Face+FER is preferable than pre-training on VGG-Face

Table 5: Comparison of human vs machine performance across benchmarks

Dataset	Metric	Human	Human Committee	Machine
FER (public test)	Accuracy	0.57	0.66	0.72
SFEW 2.0 (val)	Accuracy	0.53	0.63	0.56 [8]
FaceValue (val)	ROC-AUC	0.71	0.78	0.79

only. This is required for the voting classifier, but beneficial also when fine-tuning a pre-trained pooling architecture, which handily outperforms voting. VGG-M is in this case better than VGG-VD (+5.3%), probably due to the fact that VGG-VD is overfitted to the pre-training data. Finally, temporal average pooling is always better than max pooling.

Learning nameable facial expressions from events in videos. So far, we have shown that it is possible to predict events in videos by looking at facial expressions. Here we consider the other direction and ask whether nameable facial expressions can be learned by looking at people in TV programs reacting to events. To answer this question we applied the VGG-M pooling architecture to the FER images after pre-trained it on VGG-Faces (a verification task) and fine-tuning it on *FaceValue*. In this manner, this CNN is never trained with manually-labelled emotions. Fig. 3 shows the distribution of FER nameable expressions for faces associated to "good" and "bad" *FaceValue* events by this model. There is a marked difference in the resulting distributions, with a significant peak for *happiness* for predicted "good" events and *surprise* and negative emotions for "bad" ones. This suggests that it is indeed possible to learn nameable expressions from their weak association to events in video without explicit and dedicated supervision as commonly done.

Comparison with human baselines. Table 5 compares the performance of humans and of the best models on the three datasets FER, SFEW 2.0, and *FaceValue*. Remarkably, in all cases networks outperform individual humans by a substantial margin (*e.g.* +15% on FER and +8% on *FaceValue*). While this result is perhaps surprising, we believe the reason is that, in such ambiguous tasks, machines learn to respond as humans would on *average* whereas the performance of *individual* annotators, as reflected in Table 5, can be low due to poor inter-annotator agreement. To verify this hypothesis, we combined multiple human annotators in a committee and found that this gap either closes or disappears. In particular, on *FaceValue* the performance of the committee is just a hair's breadth lower than that of the machine (78% vs 79%).

6 Summary

In this paper we have investigated the problem of relating facial expressions with objectively-measurable events that affect humans in videos. We have shown that gameshows are a particularly useful data source for this type of analysis due to their simple structure, easily detectable events and emotional impact on the participants and have constructed a corresponding dataset *FaceValue*.

In order to analyze emotions in *FaceValue*, we have trained state-of-the-art neural networks for facial expression recognition in existing datasets showing that, if pre-trained on face verification, single models are competitive or better than the multi-network committees commonly used in the literature. Then, we have shown that such networks can successfully understand the relationship between certain events in TV programs and facial expressions

better than individual human annotators, and as well as a committee of several human annotators. We have also shown that networks trained to predict such events from facial expressions correlate very well to nameable expressions in standard datasets.

Acknowledgements

The authors gratefully acknowledge the support of the ESPRC EP/L015897/1 (AIMS CDT) and the ERC 677195-IDIU. We also wish to thank Zhiding Yu for kindly sharing his preprocessed SFEW dataset.

References

[1] Salvatore Attardo, Jodi Eisterhold, Jennifer Hay, and Isabella Poggi. Multimodal markers of irony and sarcasm. *Humor*, 16(2):243–260, 2003.

[2] Lana DS Besel and John C Yuille. Individual differences in empathy: The role of facial expression recognition. *Personality and Individual Differences*, 49(2):107–112, 2010.

[3] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. 2014.

[4] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *European Conference on Computer Vision*, pages 109–122. Springer, 2014.

[5] Fernando de la Torre, Wen-Sheng Chu, Xuehan Xiong, Francisco Vicente, Xiaoyu Ding, and Jeffrey Cohn. Intraface. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

[6] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Acted Facial Expressions in the Wild Database. Technical report, Australian National University, 2011.

[7] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Proc. ICCV Workshop*, 2011.

[8] Abhinav Dhall, O.V. Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proc. ACM Int. Conf. on Multimodal Interaction*, 2015.

[9] Abhinav Dhall et al. Collecting large, richly annotated facial-expression databases from movies. 2012.

[10] Paul Ekman and Wallace V Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106, 1969.

[11] Paul Ekman and Wallace V Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1):49–98, 1969.

[12] Rana El Kaliouby, Andrew Edwin Dreisch, Avril England, and Evan Kodra. Affect based concept testing, December 27 2012. US Patent App. 13/728,303.

[13] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59 – 63, 2015.

[14] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016.

[16] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[17] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, 2015.

[18] Vidit Jain and Erik G Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. *UMass Amherst Technical Report*, 2010.

[19] Bo-Kyeong Kim, Jihyeon Roh, Suh-Yeon Dong, and Soo-Young Lee. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, pages 1–17, 2016.

[20] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[21] Martin Koestinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[23] Gil Levi and Tal Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proc. ACM Int. Conf. on Multimodal InteractionP*, 2015.

[24] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[25] Michael Lyons, Shota Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205. IEEE, 1998.

[26] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. 2016.

[27] Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 881–888, 2013.

[28] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[29] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[30] Thierry Post, Martijn J Van den Assem, Guido Baltussen, and Richard H Thaler. Deal or no deal? decision making under risk in a large-payoff game show. *The American economic review*, 98(1):38–71, 2008.

[31] Michael E. Rep. Capuano and Walter B. Jr. Rep. Jones. We Are Watching You Act, H.R.1164, Introduced in US House of Representatives, 02/27/2015.

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2014.

[33] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[34] Nicu Sebe, Michael S Lew, Yafei Sun, Ira Cohen, Theo Gevers, and Thomas S Huang. Authentic facial expression analysis. *Image and Vision Computing*, 25(12):1856–1863, 2007.

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. 2015.

[36] Andrea Vedaldi and Karel Lenc. Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 689–692. ACM, 2015.

[37] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 435–442. ACM, 2015.

[38] Cha Zhang and Zhengyou Zhang. Improving multiview face detection with multi-task deep convolutional neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1036–1041. IEEE, 2014.

[39] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.