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

Differently from computer vision systems which require explicit supervision, humans can learn facial expressions by simply observing other humans in their environment. In this paper, we consider the problem of developing similar capabilities in machine vision. As a starting point, we look at the problem of relating facial expressions to objectively measurable events occurring in videos and make four contributions towards this goal. Firstly we construct and make available *FaceValue*, a dataset of facial expressions labelled with events to facilitate the study of this problem. Second, we evaluate existing emotion recognition CNN architectures on standard benchmarks and demonstrate the value of pre-training on face related tasks to compensate for a scarcity of labelled training data for emotion recognition. Third, we provide human baselines for the difficulty of emotion recognition in general, and specifically the difficulty of predicting events from expressions on our new dataset. Finally, we extend the standard emotion recognition architectures to predict events in videos and learn nameable expressions from them.

FaceValue Dataset

The *FaceValue* dataset comprises 192,030 faces collected from 102 episodes of the TV gameshow "Deal or No Deal" which provides a diverse source of facial expressions and game events. We associate event labels with face tracks, where an event label consists of a sum of money that has been removed from the contestant's potential prizes (see Figure 1 for examples), together with its position in the sequence of events that have taken place in the game.

Emotion Recognition Models and Human Baselines

We train and evaluate a number of CNN architectures on the FER and SFEW 2.0 emotion recognition benchmarks and show that pre-training for the task of face verification produces a substantial jump in performance (+6% average relative improvement on FER vs Imagenet pre-training), and single model test accuracies of 72.89% (FER) and 59.41% (SFEW 2.0).

Figure 1: ***FaceValue*** dataset. *Top row:* detection of an event in the game, and the corresponding reaction of the contestant's face. *Bottom:* four example tracks, the top two for "good" events and the bottom two for "bad" events (see paper for details)

Learning Expressions

We adapt the CNN architectures for emotion recognition to our primary task of learning expressions from events in the *FaceValue* dataset. Despite its challenging nature, we show that CNNs can perform well at the task of predicting event labels directly from expressions. Similarly to the FER and SFEW 2.0 benchmarks, the best model marginally outperforms the accuracy of a committee of human annotators.

Conclusions

Experimental results show that learning facial expressions from contextual events rather than directly labelled data is challenging, but feasible. The dataset and emotion recognition models are available at `http://www.robots.ox.ac.uk/~vgg/data/facevalue`.