Automatic face analysis tools for interactive digital games

Anonymised for blind review

.

Anonymous
Anonymous
Anonymous

ABSTRACT

Individuals with Autism Spectrum Condition (ASC) face a number of challenges during social interaction due to their limited ability to decode non-verbal cues. Interactive digital games can help as tools to teach such people how to express and recognise non-verbal cues and interpret their emotional meanings. In this paper, we present a set of interactive interfaces that can assist in teaching people with social communication difficulties about facial expressions and face touches. We describe the underlying technologies of three systems: Facial Affect Mapping Engine (FAME), age and gender estimation module and hand-over-face gesture interactive mimicking avatar. The three systems are fully automatic and run in real-time. We show their potential use as integral components in serious games that focus on social inclusion.

INTRODUCTION

Accurately reading non-verbal signals is essential in understanding, analysing and predicting human behaviour. People who have difficulty reading non-verbal cues, as in the case of people with autism spectrum conditions, face a number of challenges dealing with and integrating in the society [6]. Individuals on the autism spectrum find it difficult to recognise emotional cues, body language, jokes, subtle hints and other non-verbal communication signals. Interactive digital games can help as tools to teach such people how to express and recognise non-verbal cues and interpret their emotional meanings [11].

The face area is one of the main channels for displaying non-verbal signals [7]. In this paper, we present a set of interactive interfaces that were built on computer vision and machine learning technologies that analyse facial expressions and hand-over-face occlusions. We present prototypes of the following three interactive systems:

1. Facial Affect Mapping Engine (FAME), which is a framework for mapping and manipulating facial expressions in real-time across images and video streams.

2. Age and gender estimation module, which estimates the gender and age of the user. This module is integrated into FAME as it uses similar facial expression tracking techniques.

3. Face-touches mimicking avatar, which is an interactive interface that displays an avatar mimicking hand-over-face touches displayed by the user. The interface uses state-of-the-art hand-over-face occlusion detection techniques and works in real-time.

Our aim is to provide tools that can help individuals with limited interaction capabilities to recognise and express social signals such as facial expressions and hand-over-face gestures. Moreover, the underlying methodologies presented in this paper can be used as integral components in more complex serious games that focus on social interaction.

FACIAL AFFECT MAPPING ENGINE

First, we present the Facial Affect Mapping Engine (FAME). FAME is a framework for mapping and manipulating facial expressions across images and video streams. It allows for an interaction between an animator and an avatar/puppet. The animator and avatar can take form as both a video stream or an image.

Methodology

The Facial Affect Mapping Engine uses techniques from both facial puppetry and face swapping to transfer facial expressions through video streams. The main system architecture consists of three stages: face tracking, expression manipulation and video re-synthesis. Face tracking is done using Constrained Local Neural Field (CLNF) model [2] to detect the face and track facial landmarks. We extract and track facial expressions from both the puppet's and animator's video streams.

For expression manipulation, a few steps are employed. We first triangulate the face images with the points tracked, then apply a piecewise affine-warp to normalise the faces to a common reference frame. At this point, the expression in the video source has been separated into a texture (image) and shape (triangular mesh). Facial expressions and head pose produce changes in the face texture, due to changes in illumination, wrinkles and so on. The dynamic features (illumination, wrinkles) from the user needs to be combined with the static features (identity) from the puppet texture. To achieve

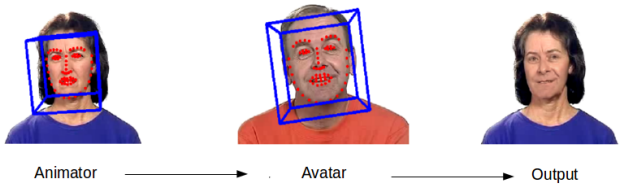

Figure 1. The role of animator and avatar in FAME.

this, we use the approach outlined by Liu *et al.* [12] to obtain an Expression Ratio Image (ERI). In order to counter the effects of mis-alignments in face tracking, we limit the domain of ERI to only pixels which get darker, such as shadows, wrinkles and frown-lines. Almost no visual expression information is lost by doing so.

To resynthesise the face, the expression avatar texture is warped to the modified shape of the animator. Hardware-assisted graphics is used to blend the new face and the background image. Then a real-time blending technique is used to preserve local lightening conditions and map the animated facial expressions. Full details on the methodology is described in Anonymous et al.[1].

Results

The system provides three modes of operation:

(a) A user live-animates a face from a source image. The reference face (avatar) is then blended onto the user's video-stream. The user's expression and head pose are transferred to the character in the source image (Figure 1).

(b) The source video animates the user's face. For non-extreme head movements, that animated face is again overlaid on the user's video-stream, so that the source video's expression and head pose are transferred to the user.

(c) The user's facial expression and head pose is exaggerated, attenuated or altered dynamically, then transferred back to the user's video stream. Only one video-stream is used; the identity and expression are both from the same person.

The system produces real-time near-photorealistic results in a wide range of cases, especially if the face is not occluded and with reasonable lightening variations. Figure 2 shows an example of the GUI of FAME in action with a live user as a puppeteer and an unlabelled image of a celebrity as a puppet.

AGE AND GENDER ESTIMATION INTERFACE

The second system we describe in this paper is a computer vision system capable of identifying the gender and age of a person. The system works in real-time on either a video feed or on a simple image and is capable of accurately detecting the age and the gender of a person in the video or an image. We present results of the system when integrated with FAME framework.

Methodology

Databases

For training we used three datasets listed in this section.

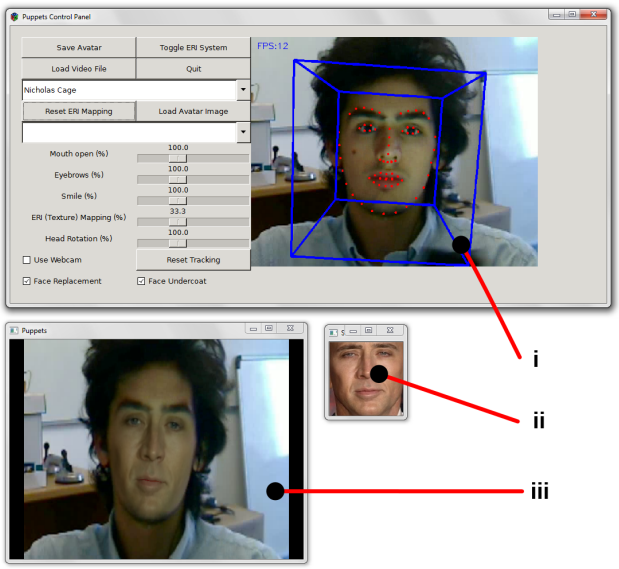

Figure 2. Facial Action Mapping Engine interface showing an example of mapping a user's facial expressions into an avatar of a celebrity. (i) is the input video stream and expression control, (ii) is the puppet avatar, and (iii) is the expression-mapped output.

The first dataset used was the *Images of Groups Dataset* [8]. It contains 28,231 faces labeled with age (categories: 0-2, 3-7, 8-12, 13-19, 20-36, 37-65, and 66+) and gender. Many images of faces are of very low resolution, people often wear sun-glasses, face occlusions, or unusual facial expressions. It is a difficult and *in-the-wild* dataset.

The second dataset used was Multi-PIE [9], a database of more than 750,000 images of 337 people recorded in up to four sessions over the span of five months. We used only close to frontal images from this dataset. It provides labels of subject gender and age.

The final dataset used was MORPH [16] that contains 55,000 images of more than 13,000 individuals. Ages range from 16 to 77 with a median age of 33.

Approach

The system works by first detecting (and tracking in case of video) facial landmarks. This is performed using the Constrained Local Neural Field [2] face tracker. The face is then aligned to a common reference frame using piece-wise-affine warping (similar to that of FAME).

We approached the task of age and gender estimation using Support Vector Machines. First a Principal Component Analysis was performed on the aligned texture images of the face (similar to appearance model in Active Appearance Models [4]).

After the dimensionality reduction a Suport Vector Machine was trained for gender and a Support Vector Regressor for age estimation. For this we used the LibSVM toolkit [3]. In both cases the Radial Basis Function kernel was used.

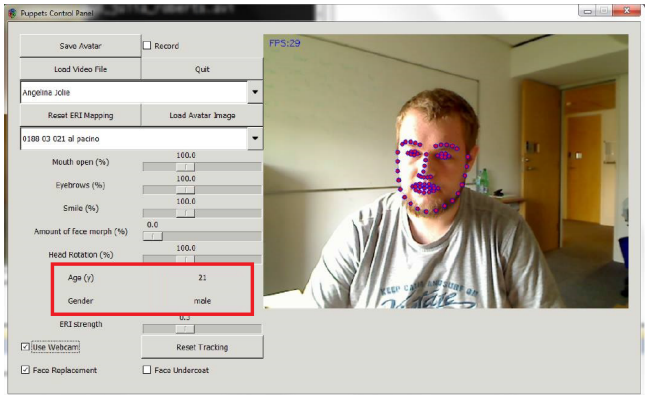

Figure 3. An example of age and gender estimation detection integrated into FAME interface. Preliminary results shows that the system work in real time and is generalisable to new faces.

Results

The system was trained and validated on a combined dataset constructed from the *Images of Groups Dataset*, half of the subjects from Multi-PIE, and half of the subjects from MORPH. Our approach was tested on Multi-PIE and MORPH datasets in a person independent manner.

The system performed with 97% accuracy on the Multi-PIE and 95% on the MORPH dataset for the task of gender estimation. For age estimation the model achieved Mean Absolute Error of 6.5 years on Multi-PIE and of 8.3 on MORPH datasets.

To test if our system is generalizable, we integrated it into FAME framework, described in the previous section, to estimate the age and gender of users in real-time. Figure 3 shows the age and gender estimation boxes integrated into FAME interface. Preliminary testing showed that the system can generalise to new faces.

FACE-TOUCHES MIMICKING AVATAR

The third system we present is an interactive avatar that mimics hand-over-face touches of the user in real-time. Hand-over-face touches are a common and an important communication cue. Recent studies showed evidence that some hand-over-face touches can be considered as affective cues and can be mapped to cognitive mental sates [14]. Based on the hand-over-face occlusion detection methodology described by Mahmoud et al. [15], we build an interactive interface, where an avatar mimics hand-over-face touches displayed by the user in real-time.

Methodology

Database

For training, we used the same set of videos of natural expressions from Cam3D video corpus [13], which includes naturalistic hand-over-face occlusions. Cam3D has natural expressions and does not restrict the video collection to faces. We used 350 video segments (all 177 videos that include hand-over-face occlusion + 173 videos with no occlusion).

Approach

For feature extraction, we extracted - for simplicity - only spatial features: Histograms of Oriented Gradients (HOGs) [5] and likelihood values of facial landmarks detection. Facial landmarks detection likelihoods were part of the CLNF [2] model. Principal Component Analysis (PCA) was then used to reduce the dimensionality of HOG features.

Using the extracted features, we trained binary linear Support Vector Machine (SVM) classifiers for the following classification tasks: 1) Occlusion/ no-Occlusion, 2) Chin occlusion, 3) Lips occlusion, and 4) Middle face (nose/cheek) occlusion. Upper face occlusion was not included as there were not enough samples of this type of occlusion in the training dataset.

The classification is done at the frame level. However, since hand position is not supposed to change on a one-frame level, a majority vote technique was employed to aggregate the output of the classification in a sliding window of 10 frames. This helps in recovering any noise in the output sequence.

GUI

We built a simple avatar with a stylised face and a hand using Blender: an open source 3D animation suite [10]. With the designed avatar head, we created a set of images representing different hand-over-face positions to display them according to the output of the classifiers. The system outputs an appropriate avatar image corresponding to the classification results. The output is displayed with a delay of 5 frames to convey the impression of mimicry.

Results

Using a leave-one-user-out validation approach, we evaluated our system on Cam3D dataset. We managed to get classification results comparable to the results presented in [13] when spacial features (HOGs and likelihoods only) are used, with basic hand occlusion detection rate of 80%. This was expected as we did not use any temporal features to speed-up system performance. Since real-time performance is very important in an interactive interface, these results were satisfying.

The system manages to generalise well to new faces, with lighting and background conditions different from Cam3D. Figure 4 displays sample screen shots of the system working in real-time. The system outputs two windows: One window is displaying the captured webcam image of the user, with the head and facial landmarks highlighted, and the second window is displaying the output avatar image corresponding to the hand-over-face gesture detected.

CONCLUSION AND FUTURE WORK

In this paper, we presented the main building blocks of three interactive interfaces that analyse facial expressions and face touches in real-time. We presented prototypes and performance results of the three systems in action, showing their potential use as interactive tools to teach facial expression recognition and expression. Our future goal is to integrate

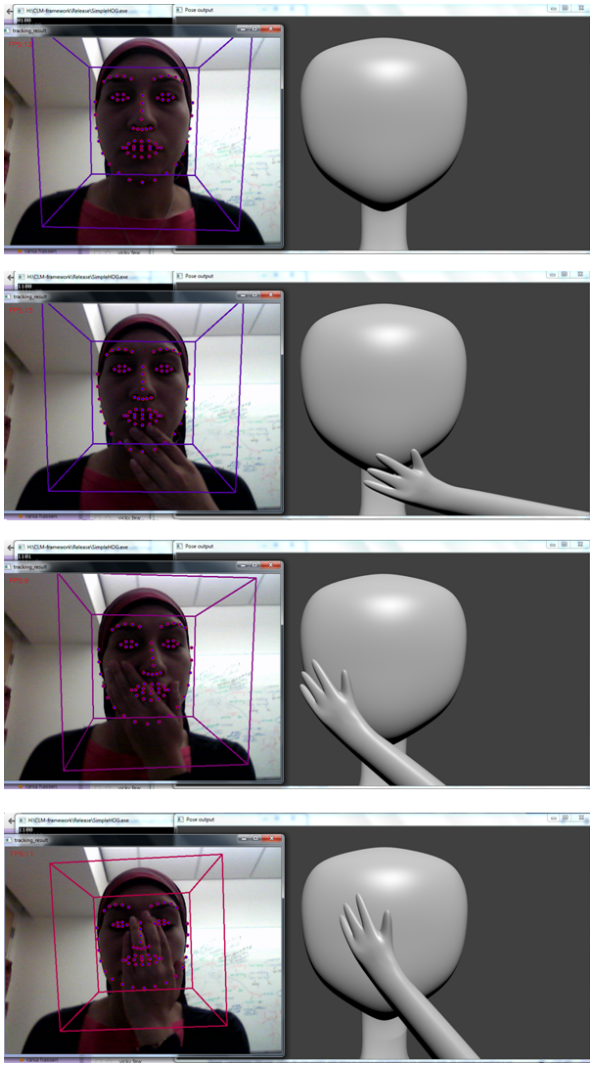

Figure 4. Examples of hand-over-face mimicking avatar in action. The avatar head - shown on the right window - mimics the face touches expressed by the user, whose webcam image is shown in the widow on the left. Examples show scenarios of : no occlusion, chin occlusion, cheek occlusion and middle face occlusion, respectively (from top to bottom).

our systems in more complex serious games that focus on social inclusion of individuals with limited social capabilities, such as the work by Schuller et al.[17].

REFERENCES

1. Anonymous, A. Anonymised for blind review. In *Anonymised for blind review.* (2014).
2. Baltrusaitis, T., Morency, L.-P., and Robinson, P. Constrained local neural fields for robust facial landmark detection in the wild. In *IEEE International Conference on Computer Vision Workshops* (2013).
3. Chang, C.-C., and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2* (2011), 27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
4. Cootes, T. F., Edwards, G. J., and Taylor, C. J. Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*, 6 (2001), 681–685.
5. Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, IEEE (2005), 886–893.
6. Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., and Iacoboni, M. Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nature neuroscience 9*, 1 (2006), 28–30.
7. Ekman, P., Friesen, W. V., and Ellsworth, P. *Emotion in the Human Face*, second ed. Cambridge University Press, 1982.
8. Gallagher, A., and Chen, T. Understanding images of groups of people. In *Proc. CVPR* (2009).
9. Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. Multi-pie. *IVC 28*, 5 (2010), 807 – 813.
10. Hess, R. *The essential Blender: guide to 3D creation with the open source suite Blender*. No Starch Press, 2007.
11. Lillard, A. Pretend play as twin earth: A social-cognitive analysis. *Developmental Review 21*, 4 (2001), 495–531.
12. Liu, Z., Shan, Y., and Zhang, Z. Expressive expression mapping with ratio images. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, ACM (2001), 271–276.
13. Mahmoud, M., Baltrušaitis, T., Robinson, P., and Riek, L. 3D corpus of spontaneous complex mental states. In *Affective computing and intelligent interaction*. Springer, 2011.
14. Mahmoud, M., and Robinson, P. Interpreting hand-over-face gestures. In *Affective Computing and Intelligent Interaction*. Springer, 2011.
15. Mahmoud, M. M., Baltrušaitis, T., and Robinson, P. Automatic detection of naturalistic hand-over-face gesture descriptors. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ACM (2014).
16. Ricanek, K., and Tesafaye, T. Morph: A longitudinal image database of normal adult age-progression. In *IEEE 7th International Conference on Automatic Face and Gesture Recognition* (2006).
17. Schuller, B., Marchi, E., Baron-Cohen, S., O'Reilly, H., Pigat, D., Robinson, P., and Daves, I. The state of play of asc-inclusion: an integrated internet-based environment for social inclusion of children with autism spectrum conditions. *arXiv preprint arXiv:1403.5912* (2014).