A Psychologically Inspired Approach Match-Score Fusion Model for Facial Expression Recognition and Analysis

Albert Cruz, Bir Bhanu, Songfan Yang

{acruz,bhanu,syang}@ee.ucr.edu

Center for Research in Intelligent Systems, University of California, Riverside, CA 92521

Summary

Affective Computing—the emergent field where computers interpret human emotion and respond in kind with emotions to facilitate non-verbal communication—has reached a bottleneck where systems cannot correctly interpret human facial expressions. Challenge workshops, such as the Audio/Visual Emotion Challenge 2011 (AVEC), are created for researchers to tackle this problem.

We propose a method that emulates the Human Visual System (HVS) for processing facial expressions that ranked 2nd video-analysis in the AVEC workshop.

1. Sensitive Artificial Learner (SAL)

- Affective Computing impacts intelligent tutoring systems, treatment of Aspergers, and counselor programs like SAL, in AVEC.
- A user (1) talks to SAL (2).
- The program may be belligerent—as pictured, or melancholy, causing emotional reactions.
- Given user video (1), the program must detect emotional state.

2. Technical Challenges

- Loading all video frames would require 65GB of memory for per-frame LPQ features. Need a method for intelligent frame selection
- State-of-the-art must be given a frontal face. However, face may be obscured, or out of frame, e.g. (3) and (4). A robust alignment algo. is required.

3. Proposed Method

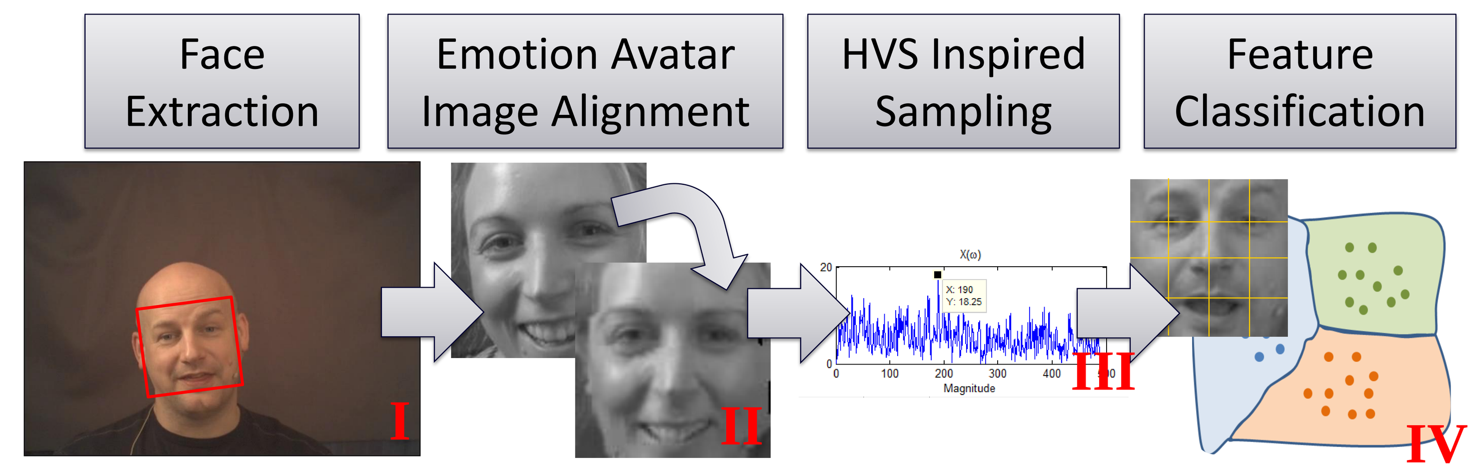

Figure 2: The proposed method.

I. Face ROI extracted with haar-like features.
II. Emotion Avatar Image hallucinates frontal face.
III. Vision and Attention Theory samples video dynamically.
IV. Frames classified using Local Phase Quantization and Linear Support Vector machines.

4. Emotion Avatar Image

Figure 3: (7) Result of Avatar Image Registration when warping (6) onto (5). Images warped spatially with an objective function similar to optical flow:

$$E(w) = \sum_p \min(\|s_A(p) - s_i(p + w)\|_1) + \sum_p u^2(p) + v^2(p)/\sigma^2 + \sum_{(p,q)\in N_q} (\min(\alpha|u(p) - u(q)|) + \min(\alpha|v(p) - v(q)|))$$

where p is a pixel in the image, $w(p)$ the motion vector between query and target where $w(p)=(u(p),v(p))$, S_A and S_i are SIFT features of target and the query respectively and is the N_q 4-member neighborhood about p.

5. Vision and Attention Theory

- The HVS processes natural scenes with a 50ms-1s latency that increases as the scene is unchanged (Buswell 1935, Viviani 1990).
- Rate of processing visual information increases proportionally with the rate of change of visual information.

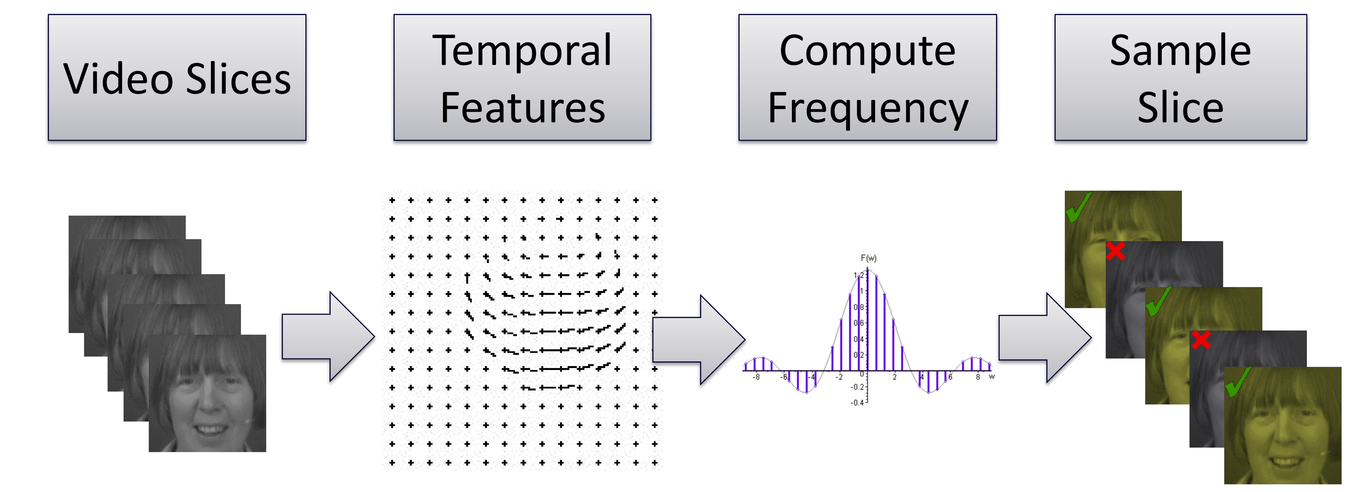

Figure 4: Overview of sampling approach.

- Videos reduced to slices.
- For each slice, visual Information quantified using Optical Flow via:

$$v(t) = \Sigma_p \|g(f_t, f_{t-1})\|$$

where $\|g(.,.)\|$ is the optical flow between two frames.
- The slice is sampled at the dominant frequency of the Discrete-Time-Fourier-Transform of $v(t)$.
- Decimates frames needed a factor of 20.

6. Competition Results

	(%)	Arousal	Expectancy	Power	Valence	Average
Development	Proposed	69.3	65.6	59.9	68.8	65.9
	Schuller *et al.* 2011	60.2	58.3	56	63.3	59.7
Testing	Proposed	56.5	59.7	48.5	59.2	55.9
	Schuller *et al.* 2011	42.2	53.6	36.4	52.5	46.2

Figure 5: 10% improvement over state-of-the-art Schuller *et. al 2011* for the difficult testing section.

7. Conclusion

- Used aspects of Vision and Attention Theory to make problem feasible in terms of computation size.
- Avatar Image Registration robust enough for SAL.
- Proposed approach ranked 2nd for AVEC video challenge.
- Support for work was provided by NSF IGERT: Video Bioinformatics Grant DGE 0903667.