Intel Collaborative Research Institute

Computational Intelligence

Implicit Affective Video Tagging from Facial Expressions

Daniel Hadar[1,2] **Talia Tron**[1,3] **Daphna Weinshall**[1]

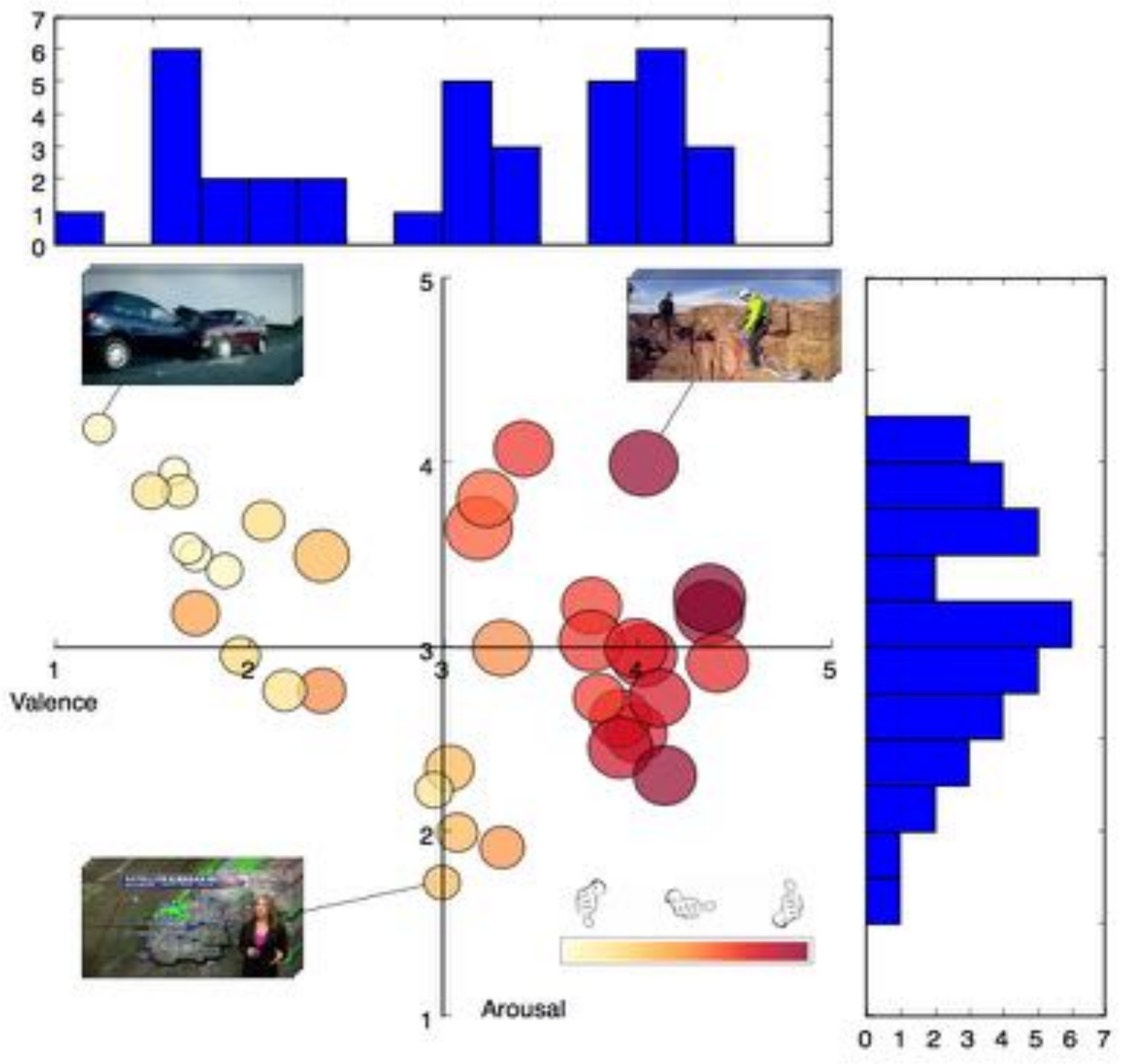
1- School of Computer Science and Engineering, The Hebrew University of Jerusalem, 2- Department of Cognitive Science
The Hebrew University of Jerusalem, 3- Center for Neural Computation, The Hebrew University of Jerusalem

Abstract

Affective Video Tagging describes the annotation of video clips, designed to characterize them in universal terms and to enable the consolidating of one's profile based on the tagging (*e.g.* for recommendation systems). As opposed to *explicit* affective video tagging (where one knowingly assigns tags to clips), *Implicit* tagging refers to utilizing one's non-verbal behavior (*e.g.* facial expressions) to derive tags. In this ongoing study we present an automatic model that derives affective video tagging from subjects facial expressions, using 3D video photography techniques, signal processing and machine learning tools.

Video Database

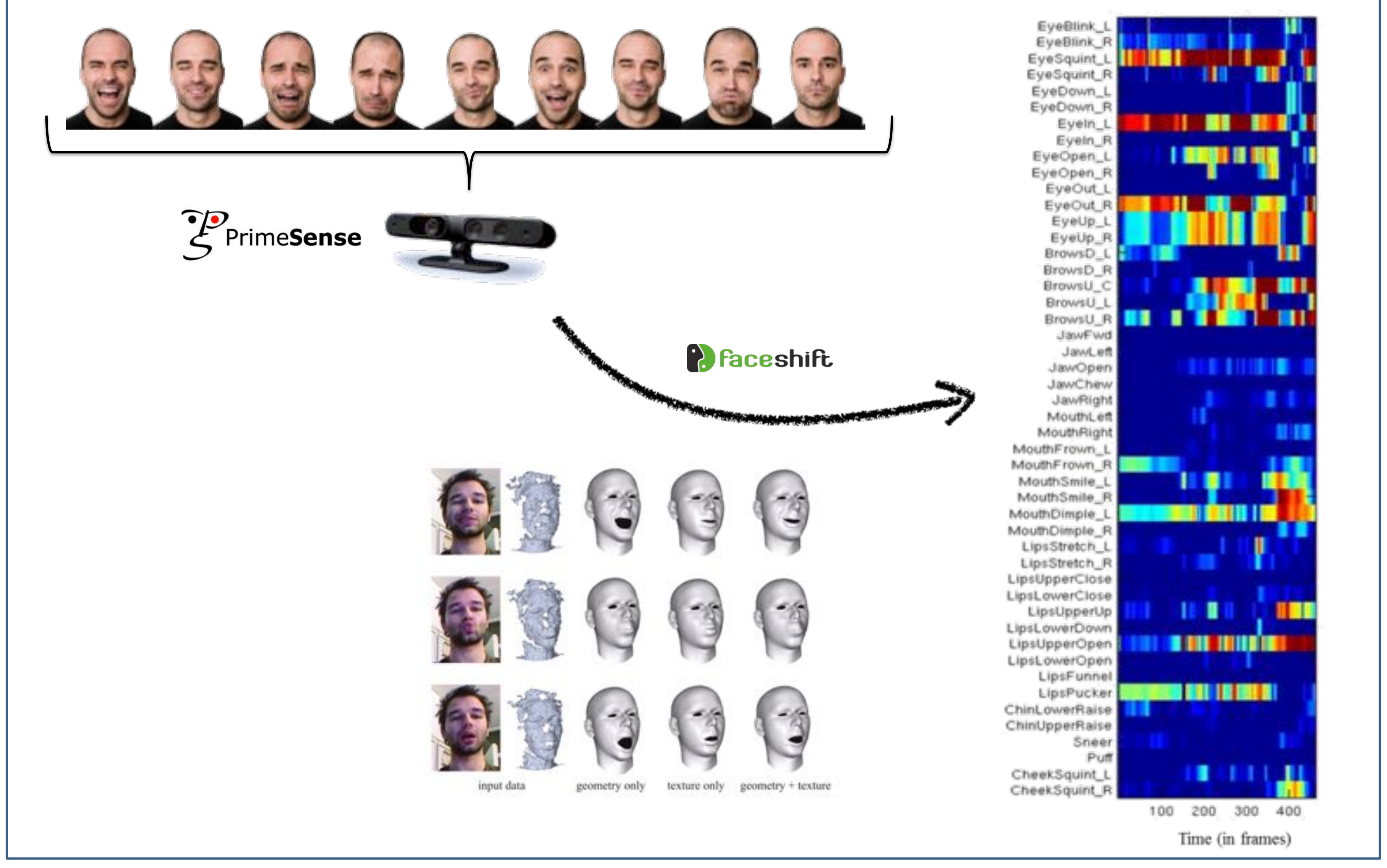
To measure affective response, we had built an Emotion Eliciting Video Clips Database (EEVDB), including 36 short publicly available video clips (6-30 seconds, μ = 20sec.). Each clip was rated on 4 qualities (**Valence**, **Arousal**, **Likeability** and **Desire to Watch Again**) in 1-5 scale, and then described using free text in Hebrew, by 26 independent raters.



Each circle is a video clip on V-A plain; color and size correspond to likeability level and the desire to watch again

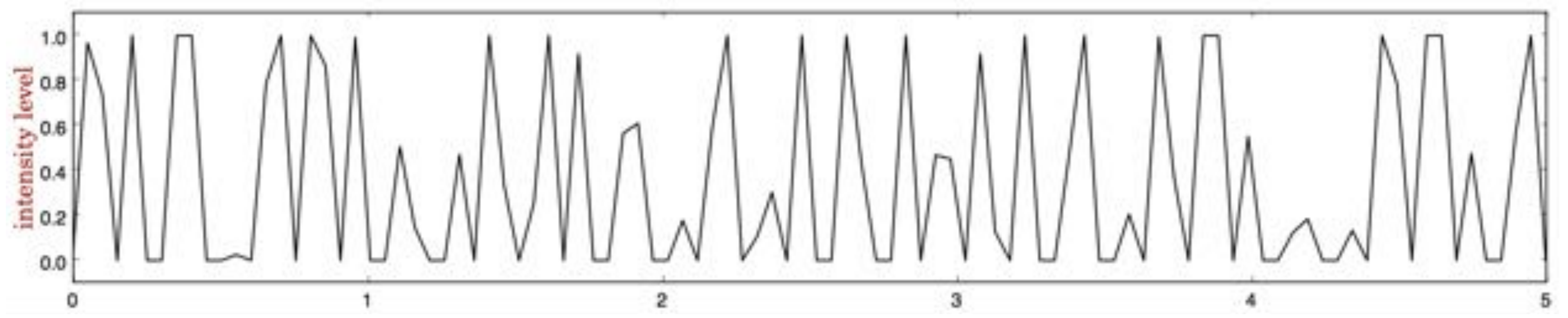
Method and Technology

While watching the video clips, facial activity was recorded (n=26 participants) using a Structured light 3D camera (Carmine 1.09):



Data Analysis

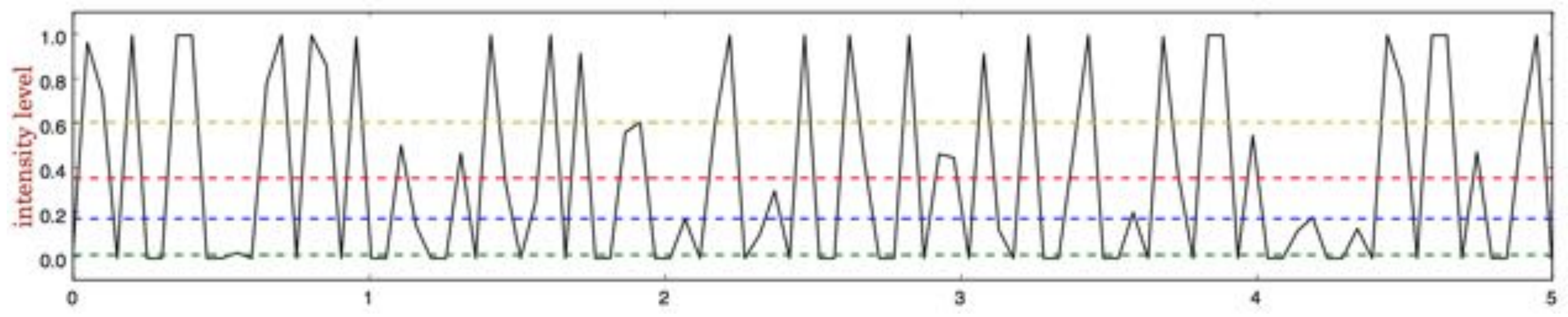
Every facial muscle (*FM*) for each subject and each clip provides a signal of intensity level across time:



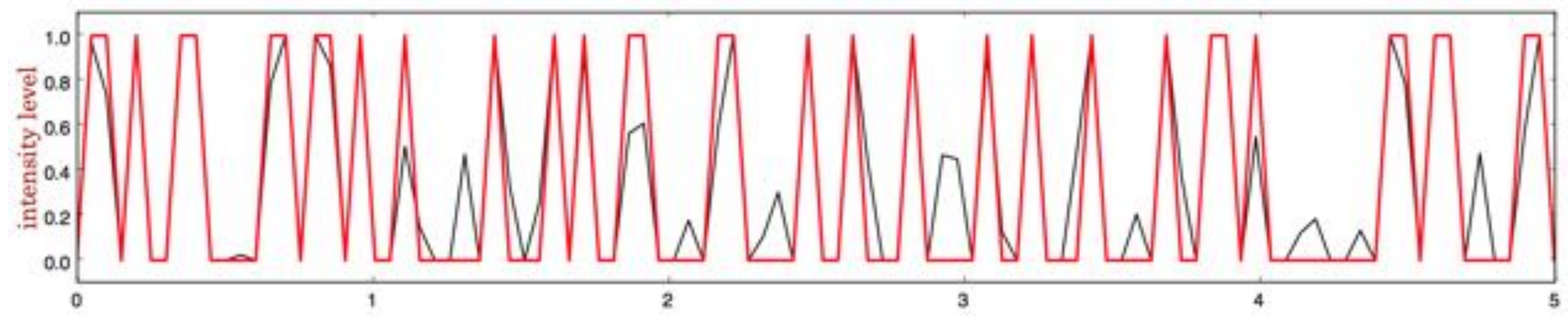
Several features were computed for each subject (for each clip):

Moments

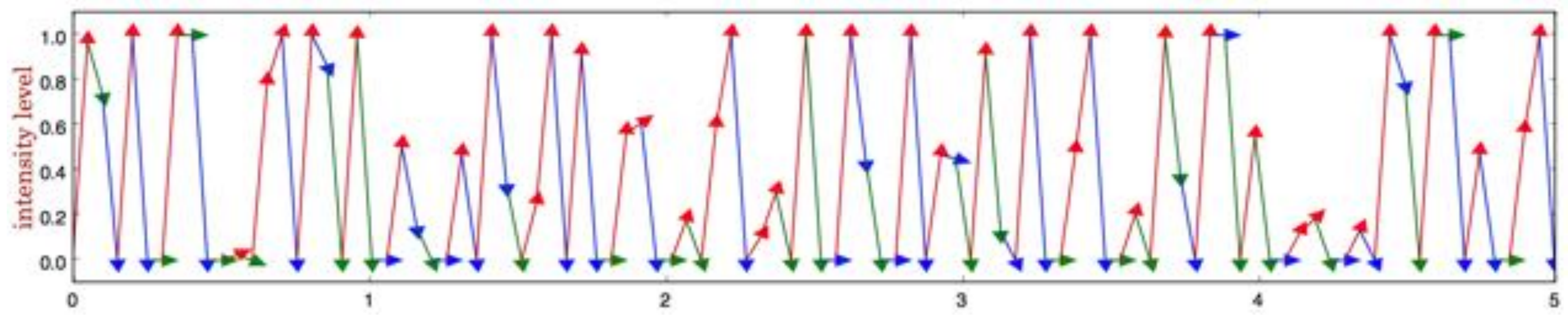
(mean, variance, skewness & kurtosis)



Quantized



Dynamic Features



Model

A model was learned for each subject:

$$f_{subj} : FEATURES \rightarrow (V, A, L, D)$$

where: $FEATURES \equiv \{MOMENT_i\}_{i=1}^{4\times FM} \cup \{QUANTIZED_i\}_{i=1}^{3\times FM} \cup \{DYNAMIC_i\}_{i=1}^{3\times FM}$

A total of ~460 descriptors for each subject.

Results

	Pearson's r	p-value	Learner	Features Used
Valence	.51	*<.0001*	SVR (linear kernel)	PCA(6) over each feature set
Arousal	.57		Ridge	PCA(3) over each feature set
Likeability	.53		SVR (linear kernel)	PCA(6) over each feature set
Rewatch	.56		SVR (linear kernel)	PCA(7) over each feature set

Moment Features were taken over the entire signal, while Quantized and Dynamic Features were taken over **highlight times** – segments with maximum facial activity.

This is an ongoing research thus the results are intermediate.