KOÇ UNIVERSITY

Use of Affect Based Interaction Classification for Continuous Emotion Tracking

Hossein Khaki and Engin Erzin
Multimedia, Vision and Graphics Lab (MVGL)
Department of Electrical and Electronics Engineering

42st IEEE International Conference on Acoustics, Speech and Signal Processing
ICASSP 2017
5-9 March 2017 – NEW ORLEANS, USA

Koç University
Istanbul, Turkey

Outline

- Related Studies and Motivation
- JESTKOD Database
- GIT-CER system
- Experimental Evaluations
- Conclusion and Future work

Related Studies and Motivation

Observations

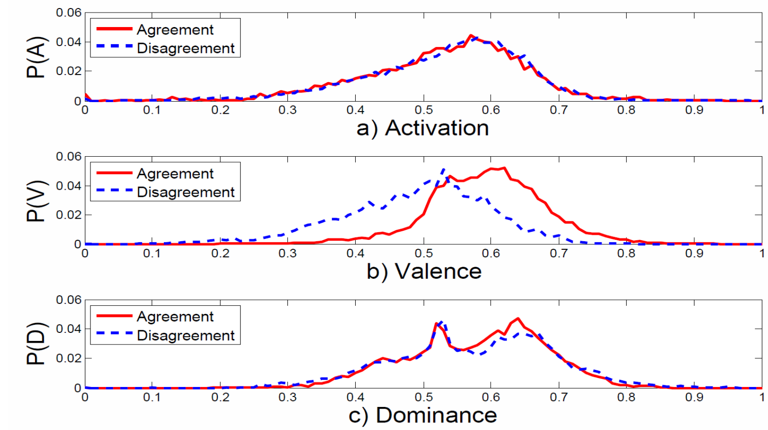
Dyadic interaction type helps Valence estimation [1]

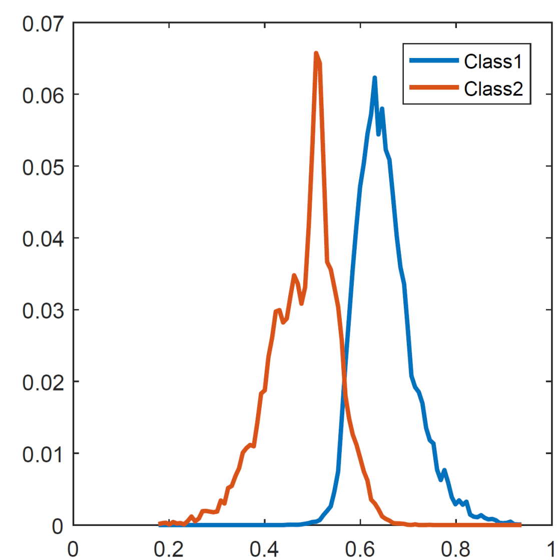
+

Locating temporal segmentation to cluster AVD in two general high/Low classes

Problem

Does general high/Low classes help to estimate AVD better?





[1]- H. Khaki and E. Erzin, “Use of agreement/disagreement classification in dyadic interactions for continuous emotion recognition,” in INTERSPEECH, 2016.

JESTKOD database

- A natural and affective dyadic interactions
- Equipment:
  - A high-definition video recorder
  - Full body motion capture system with 120 fps
  - Individual audio recorders
- 5 sessions, totally 56 agreement and 42 disagreement clips
- In each clips: 2 participants, around 2~4 minutes
- Totally 10 participants
  - 4 female/6 male, ages: 20 - 25
- Language: Turkish
- Annotation
  - Activation
  - Valence
  - Dominance

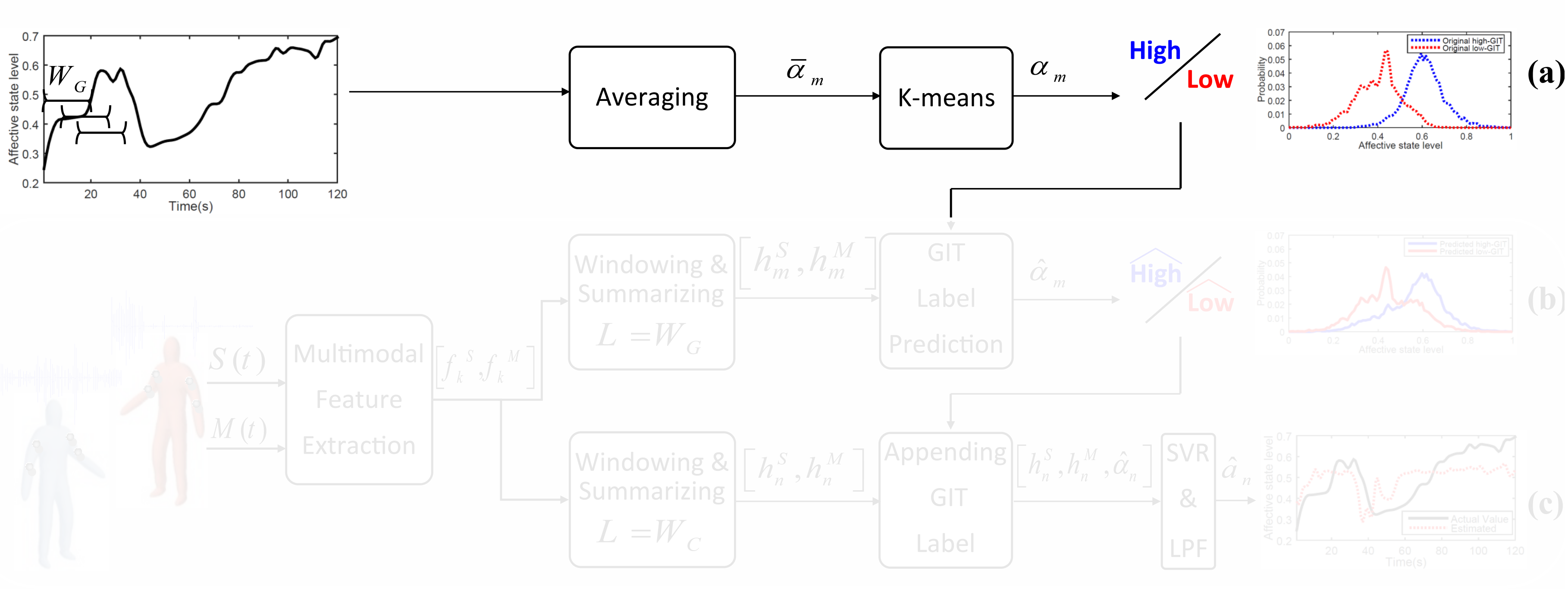
Mean Pearson's correlation between the consensus rating and individual annotations		
Activation	Valence	Dominance
0.5568	0.5638	0.7369

Topics in the JESTKOD database

Pair #	Agreement scenario	Num. clips	Disagreement scenario	Num. clips
1	Cinema, World cuisine, Holiday resorts, TV series	13	Football, Maths, Game consoles, PC Games	13
2	Football, World cuisine, Music, Cinema, Literature	13	Geography, Holiday resorts, PC Games, Theatre, Dance	16
3	Cinema, Sports, PC Games, Music, World cuisine	11	Cinema History, TV series, Animals, Education	17
4	World cuisine, Holiday resorts, Science-fiction, History, Theatre, Cities	16	Football, Cinema PC Games, TV series, Literature, Physics	17
5	Cinema, Languages, PC Games, Cities, Game consoles	13	Cinema, Sports, Holiday resorts, Nutrition, Musicals	16
Total		66		79

The JESTKOD database is available upon request for academic purposes.
http://mvgl.ku.edu.tr/databases

GIT-CER system

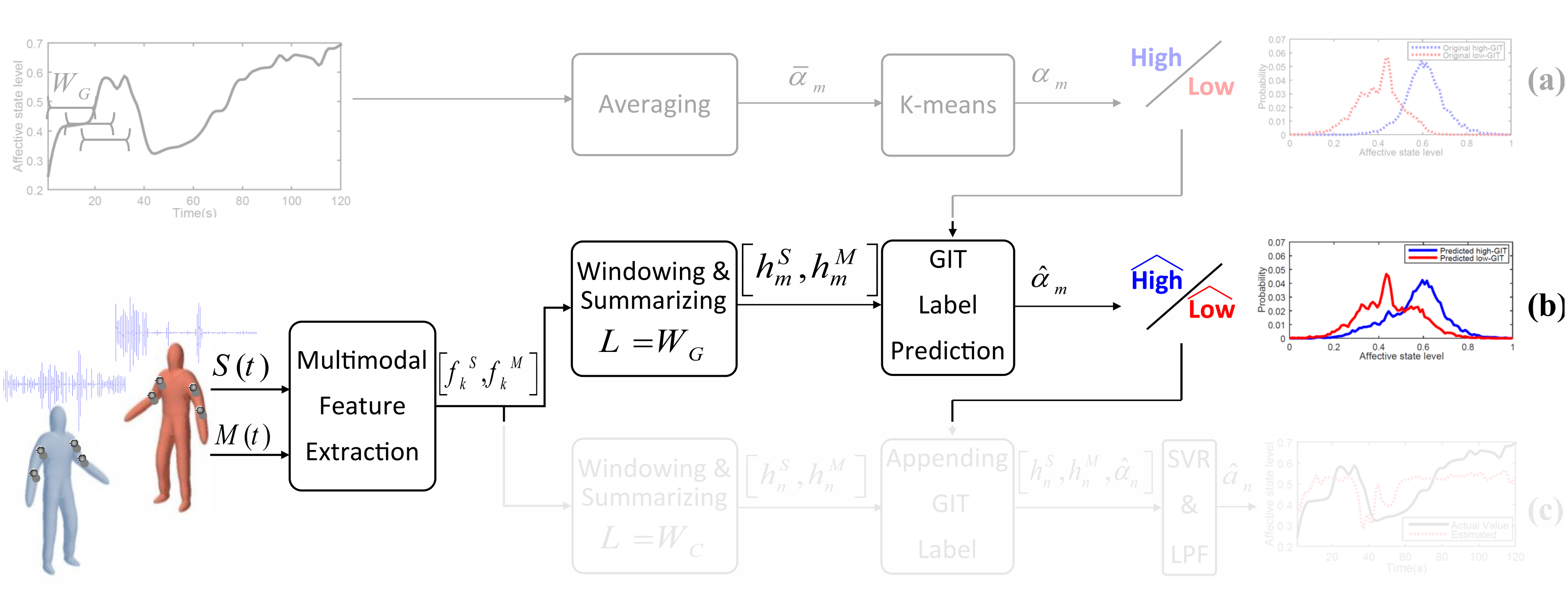


- a. Clustering:
  - Temporal segmentation
  - Splitting AVD into two general classes
  - Similar to DIT, Defining General DIT (GIT)

$$\bar{\alpha}_m = \frac{1}{W_G} \sum_{k=1+(m-1)R_G}^{W_G+(m-1)R_G} a_k,$$

$$\alpha_m = \mathcal{Q}(\bar{\alpha}_m)$$

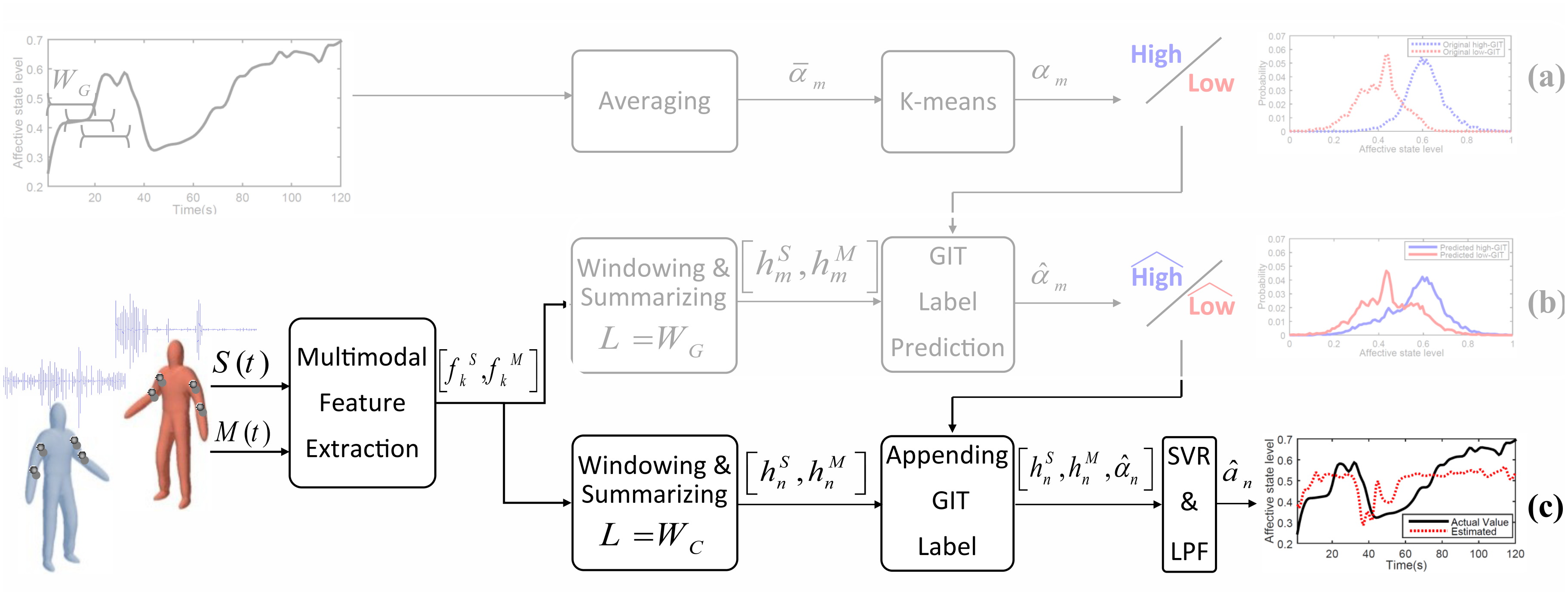
GIT-CER system



- b. GIT Label Prediction:
  - Linear SVM unimodal and multimodal system.

$$\hat{\alpha}_m = \Phi(h_m^S, h_m^M) \qquad 8 \leq W_G \leq 30\ sec$$

GIT-CER system



- c. Appending GIT label and Continuous Emotion Recognition

$$\hat{a}_n = \Psi(h_n^S, h_n^M, \hat{\alpha}_n) \qquad W_c = 1.5\ sec$$

Experimental Evaluations (parameters)

- **Feature extraction:**
  - **Speech**: 16.66 ms win with 8.33 ms frame shifts $\Rightarrow$ 39D $=$ (E $+$ 12MFCCs) $+ \Delta + \Delta\Delta$
  - **Motion**: 24D $= (\varphi, \theta, \psi)$ of the arm & forearm joints with their derivatives
- **Training and testing strategy:**

  Leave-one-session-out => Speaker independent

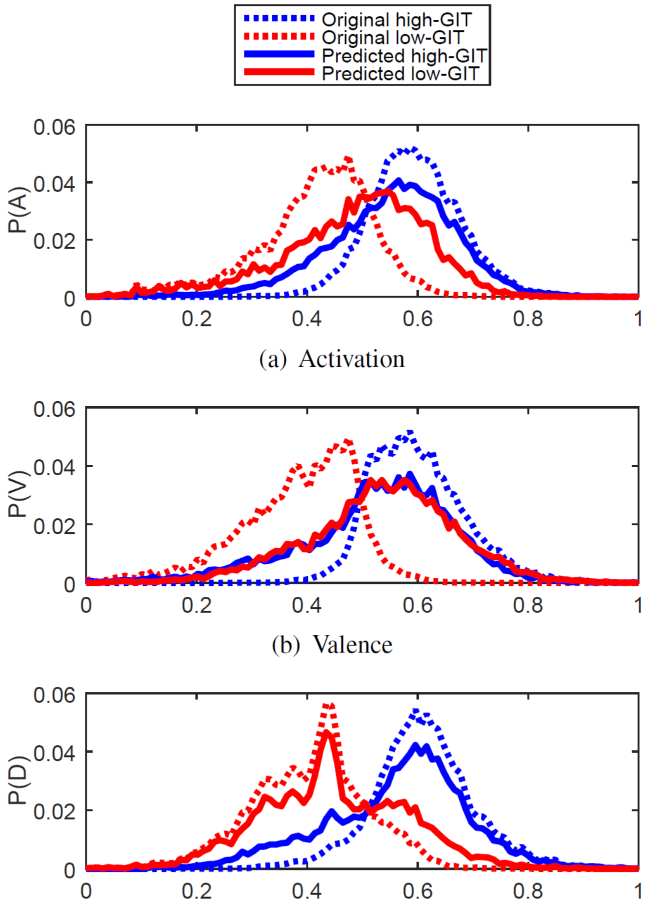
- **Feature Summarization:**
  - **Statistical functions:** Adjust the PCA output dimension to preserve 90% of the total variance
- **Prediction and regression:**
  - **GIT prediction:** Linear kernel SVM
  - **CER:** RBF kernel SVR
- **Performance metric:** The average Pearson correlation between consensus ratings and their estimation

Experimental Evaluations – GIT Prediction

- GIT predication phase:
  - Search over different $8 \leq W_G \leq 30$ sec to maximize the statistical difference between predicted high and low GIT
  - Statistical difference measure:
    Kullback-Leibler divergence (KLD)
  - GIT prediction from speech and motion

$D_{KL}(P_H, P_L)$ / (W_G)		
Activation	Valence	Dominance
0.31/(13)	0.11/(29)	0.81/(13)

- Dominance: Well separated ☺
- Activation Medium separated! 😐
- Valence: Not separated ☹



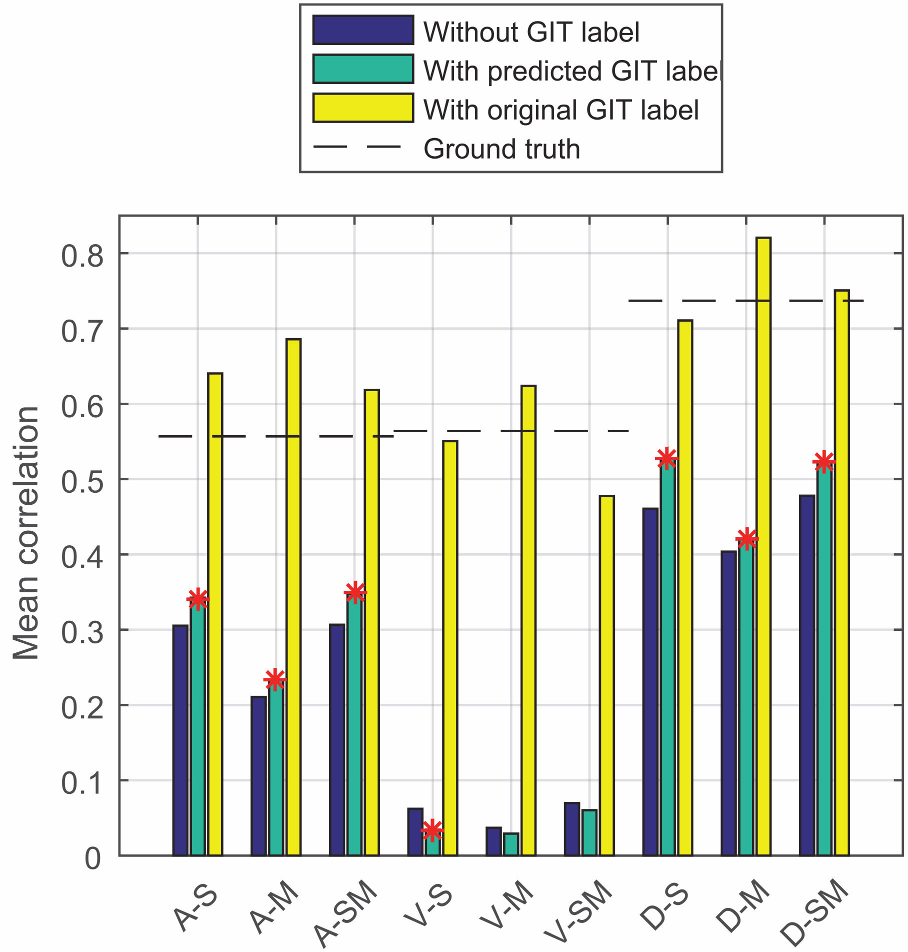
(a) Activation

(b) Valence

(c) Dominance

Experimental Evaluations – Emotion Recognition

- Continuous Emotion Recognition
  - From **S**peech, **M**otion and multimodal speech & motion (**SM**)
  - For **A**ctivation, **V**alence and **D**ominance
- Observations
  - Multimodal tests have almost always the highest correlation
  - Predicted GIT (green bars) improves dominance and activation
  - Valence regression is always poor (No facial expression data)
  - Yellow bars: Theoretical upper bounds



Star signs indicate the statistically significant ($p<0.05$) difference between CER without GIT labels and CER with predicted GIT labels

Conclusions and Future work

- Conclusions
  - Our hierarchical continuous emotion recognition system consist of:
    - Temporal clustering of AVD to form GIT label
    - Predict GIT Label with multimodal feature set
    - Append predicted GIT to multimodal feature for continuous emotion recognition
  - GIT labels provide useful discrimination for the activation and dominance attributes in the JESTKOD dataset
  - GIT labels introduce side information for CER problem

- Future work
  - Use of affect context, such as GIT, for continuous emotion recognition

Thanks.

!?QUESTIONS?!