The Tenth International Conference on Intelligent Systems and Applications, INTELLI 2021, July 18 to July 22, 2021, Nice, France, https://www.iaria.org/conferences2021/INTELLI21.html

A Hybrid Model to Improve Occluded Facial Expressions Prediction in the Wild during Conversational Head Movements

Authors

Arvind Bansal

Arvind Bansal, Kent State University, Kent, Ohio, USA

Mehdi Ghayoumi, eCornell, Cornell University, Ithaca, New York, USA

Emails: akbansal@kent.edu and mg948@cornell.edu

Mehdi Ghayoumi

Presented by Arvind Bansal

KENT STATE UNIVERSITY

Short Biography of Arvind Bansal (Presenter)

- Full professor of Computer Science and Director of Masters in Artificial Intelligence program at Department of Computer Science at Kent State University, Kent, Ohio, USA
- PhD in 1988 from Case Western Reserve University, Cleveland, OH, USA
- Research contributions in
  - parallel logic programming; massive parallel knowledge bases; self healing and fault tolerant multi-agent systems; social robotics (facial expression recognition, gesture recognition and generation); ECG analysis; genomic and proteomics; multimedia languages and synchronization; high performance programming languages
- Authored two textbooks
  - Introduction to Programming Languages (2013), undergraduate, published by CRC press
  - Introduction to Computational Health Informatics (2020), graduate, published by CRC press

Research Interest of the Group

- Social Robotics
  - Facial expression analysis
  - Gesture analysis
  - Gesture generation in humanoids
  - Multimodal integration of human emotions
- Intelligent analysis of Biosignals
  - ECG analysis
  - Cardiac echogram analysis
- Intelligent analysis of micro-RNA targets to understand human disease

KENT STATE UNIVERSITY

Motivation

- Social robotics has an important role for elderly and health care due to negative population growth in developed countries
- Understanding emotions/pain is essential for empathy and care in social robots
- Facial expressions are a major aspect of involuntary expression of emotions
- Ekman's model - six basic facial expressions: anger, disgust, fear, happiness, sadness, and surprise derived using discriminatory facial expression points
- Facial discriminatory feature-points are occluded by external objects; hand gestures; head rotations during conversational gestures; multi-party interaction
- Current schemes to handle occlusion of facial expressions are limited to small obstructions on frontal head positions
- Popular CNN based techniques degrade 30% to 50% beyond partial occlusion

*The image is taken from Wikimedia commons and is in public domain

KENT STATE UNIVERSITY

Contribution

- A hybrid model integrating CNN + symmetry based geometric modeling
  - symmetry is used to reconstruct discriminatory feature-points
  - improvement is beyond partial occlusion: 8%(sadness) upto 21% (anger)
  - symmetry-based geometric modeling is rotation invariants and corresponds to FACS
- Symmetry-based geometric modeling provides temporal context
  - maps continuous motion accurately to the corresponding aligned CNN-based model
- Improves prediction for
  - facial expression in conversational gestures involving continuous extreme head rotations such as denial, argumentation, multi-party interactions
  - predicting during stochastic occlusion caused by bad lighting and shadows

Related Work

- Techniques for reducing partial occlusion by external objects / hand gestures
  - multiple fixed posed alignments (Seshadri et al. 2016)
  - building textures of occluded patches using nonoccluded space (Zhang et el. 2018)
  - sparse matrix representation and maximum likelihood estimation (Liu et al. 2014)
  - combination of Gabor filter and cooccurrence matrix (Li et al. 2015)
  - LSTM autoencoders (Zhao et al. 2018)
  - Bayesian networks (Miyakoshi and Kato, 2011)
- CNN based approaches
  - Gabor filter and dimension reduction + CNN (proposed by us in 2016)
  - CNN + LSTM + transfer learning for mapping to fixed alignment (T-H. S. Li et al. 2019)
  - CNN + local and global texture + attention (Y. Li et el. 2019)

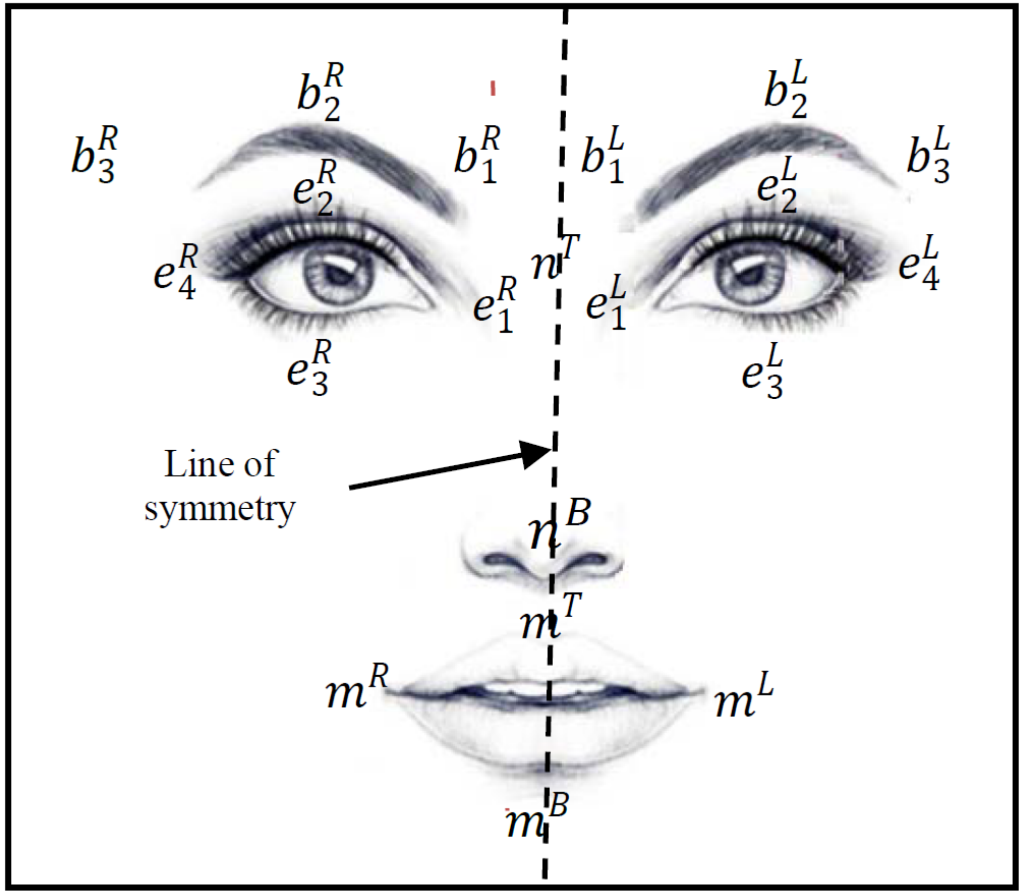
Current Limitations and This Approach

- Current limitations
  - facial symmetry is not fully exploited during extreme head-rotations
  - current schemes (including CNN approaches) are good for partial occlusion
  - beyond partial occlusion nonsymmetric approaches degrade 30% to 50% due to loss of discriminatory feature-points
- Approach in this research
  - combine rotation invariant geometric modeling corresponding to FACS with CNN
  - continuous line-segment change also provides temporal context
  - good for extreme head-rotations during conversational head-gestures and oblique line-of-view

*The image is taken from Wikimedia commons and is in public domain

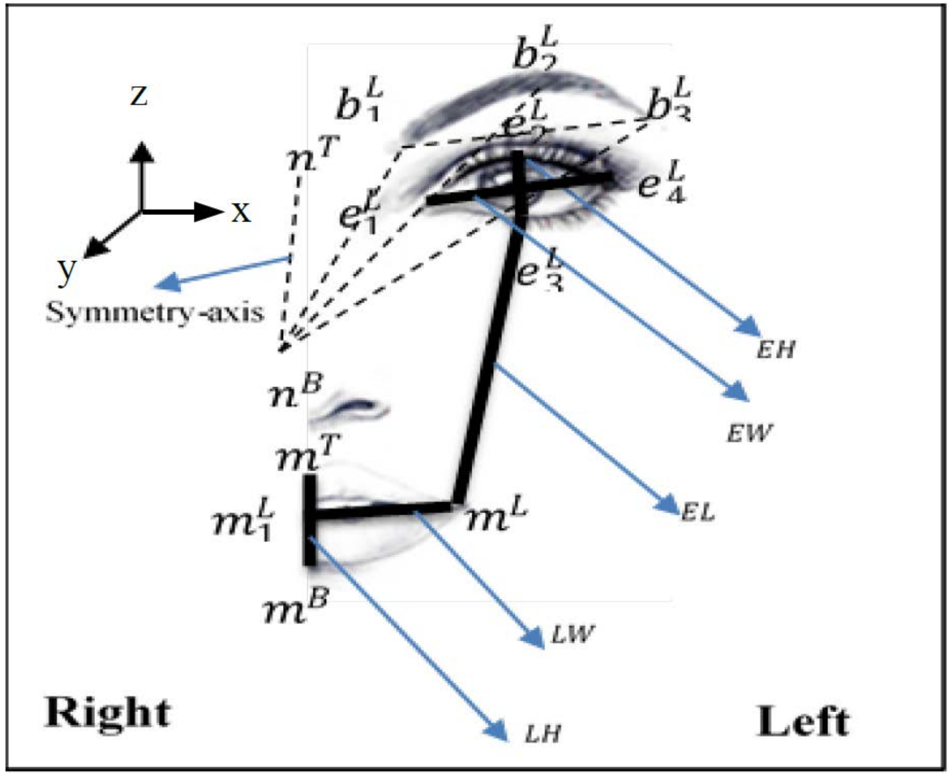
Facial Symmetry in Feature-points

- Two types
  - fixed – do not move with facial expressions
  - active – move with facial expressions
- Fixed points act as reference
  - two ends of the left and right eyes
  - bottom of a nose n^B
  - middle point between eye-brows n^T
- Active points predict facial expressions
  - three points on each brow for brow movement
  - two middle points of lips m^T and m^B
  - two endpoints of the mouth m^R and m^L
  - two middle points in each eye for eye-lid movement



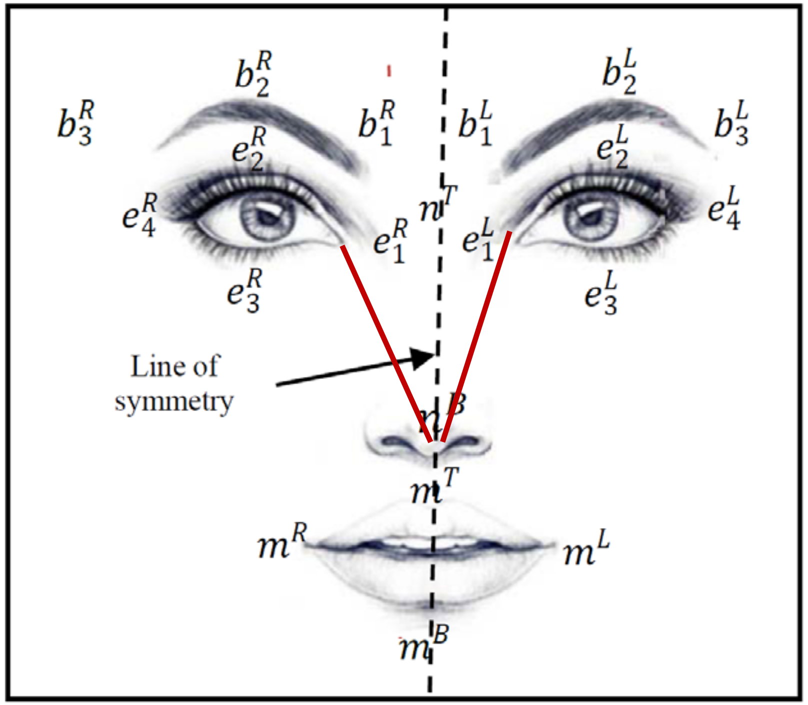
Symmetry-based Geometric Model

- Line-segments for both left and right
  - six nose bottom n^B to three eyebrow points
  - two mouth middle to mouth end
  - one joining two middle points of lips
  - two mouth end to eye middle
- Normalized head-rotation invariant ratios
  - division by $n^B n^T$ for vertical ratio
  - division by eye-width for horizontal ratio
  - lip height, lip width, eye-to-lip; brow-width; inner-brow-height; outer-brow-height; middle-brow-height; eye-height



KENT STATE UNIVERSITY

Handling Occlusion



- Multiple alignments every 15 degree
- Ratios $n^Be_1^L / n^Be_1^R$ and $n^Te_1^L / n^Bn^T$ are used to measure angle of rotation
- Angles maps to nearest alignment
- Frontal pose: $n^Be_1^L / n^Be_1^R = 1 \pm \varepsilon$
- Ratio deviates with head-rotations
- Beyond $\pm$ 45° symmetry based geometric model is used

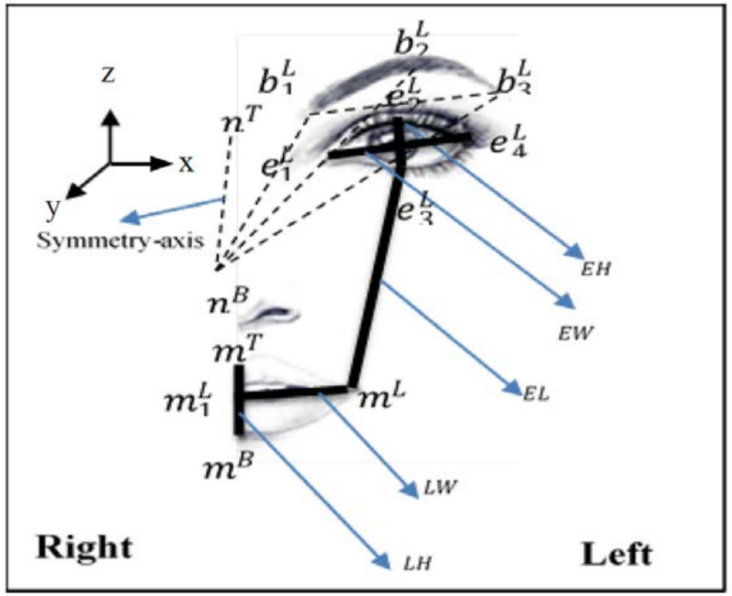
KENT STATE UNIVERSITY

FAU Correspondence

- 17 major FAUs responsible for basic six facial expressions
  - #1, #2, #4, #5, #6, #7, #8, #10, #12, #15, #16, #17, #20, #23, #26, #27, #41
- Line –segments are normalized using $n^B n^T$ *(vertical) and EW (horizontal)*
- Increase and decrease in the normalized ratios corresponds to FAU movements
- Combinations of increase/decrease of normalized ratios correspond to FAUs

Line Segment and FAU Correspondence

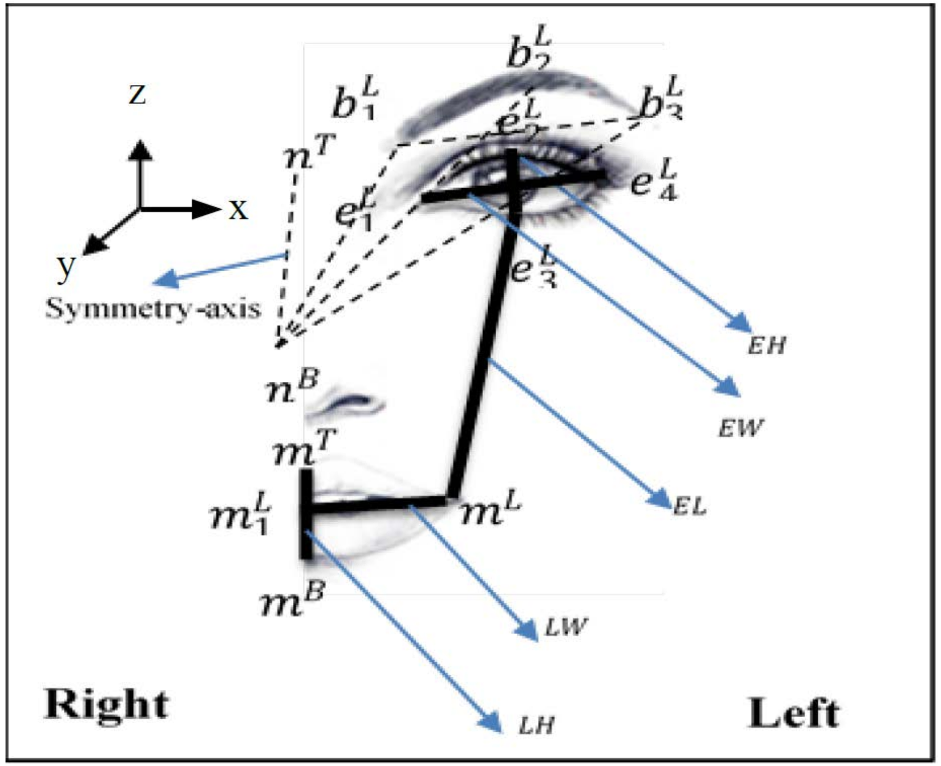
Line seg.	FAU Subset	Basic Emotions
LH	8, 10, 16, 17, 23, 26, 27	anger, disgust, fear, sad, surprise
LW	6, 12, 15, 16, 20, 23	happiness and sadness
EL	6, 15	disgust, fear, happiness, sadness
EH	5, 7	anger
$\lvert b_1^L\ b_3^L \rvert_x$	4	anger, disgust, fear, sadness
$\lvert n^B\ b_1^L \rvert_z$	1, 4, 9	anger, disgust, fear, sadness, surprise
$\lvert n^B\ b_2^L \rvert_z$	4, 5	fear and surprise
$\lvert n^B b_3^L \rvert_z$	2	fear
$n^B n^T$	vertical reference	used for vertical normalizations
EW	horizontal reference	invariant with head-rotation

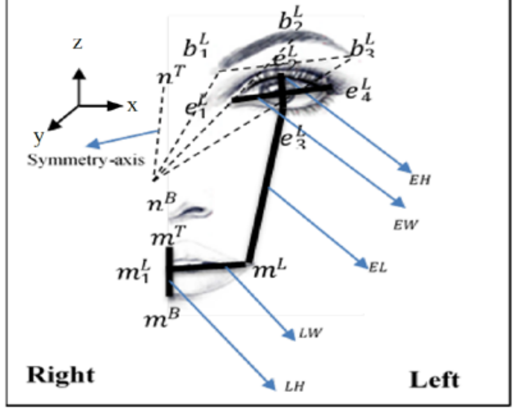


KENT STATE UNIVERSITY

Line Ratios Corresponding to FAUs

Line Ratio	Normalized Ratios	Description
R^{LH}	$\lvert LH \rvert / \lvert n^B n^T \rvert$	lip height ratio
R^{LW}	$\lvert LW \rvert_X / \lvert EW \rvert$	lip-width ratio
R^{EL}	$\lvert EL \rvert_Z / \lvert n^B n^T \rvert$	eye-to-lip ratio
R^{BW}	$\lvert b_1^L b_3^L \rvert_X / EW$	brow-width ratio
R^{IBH}	$\lvert n^B b_1^L \rvert_Z / \lvert n^B n^T \rvert$	inner brow-height ratio
R^{MBH}	$\lvert n^B b_2^L \rvert_Z / \lvert n^B n^T \rvert$	mid-brow height ratio
R^{OBH}	$\lvert n^B b_3^L \rvert_Z / \lvert n^B n^T \rvert$	outer-brow height ratio
R^{EH}	$\lvert EH \rvert / \lvert n^B n^T \rvert$	eye-height ratio





FAUs and Normalized Ratio Conditions

FAU	Condition (n = m + k and k > 0)
1	$R_n^{IBR} > R_m^{IBR}$
2	$R_n^{OBR} > R_m^{OBR}$
4	$R_n^{IBR} < R_m^{IBR} \wedge R_n^{MBR} < R_m^{MBR} \wedge R_n^{OBR} < R_m^{OBR}$
5, 27	$R_n^{EH} > R_m^{EH}$
6, 12	$R_n^{LH} < R_m^{LH} \wedge R_n^{EL} < R_m^{EL}$
7, 41	$R_n^{EH} < R_m^{EH}$
8	$R_n^{LH} < R_m^{LH}$

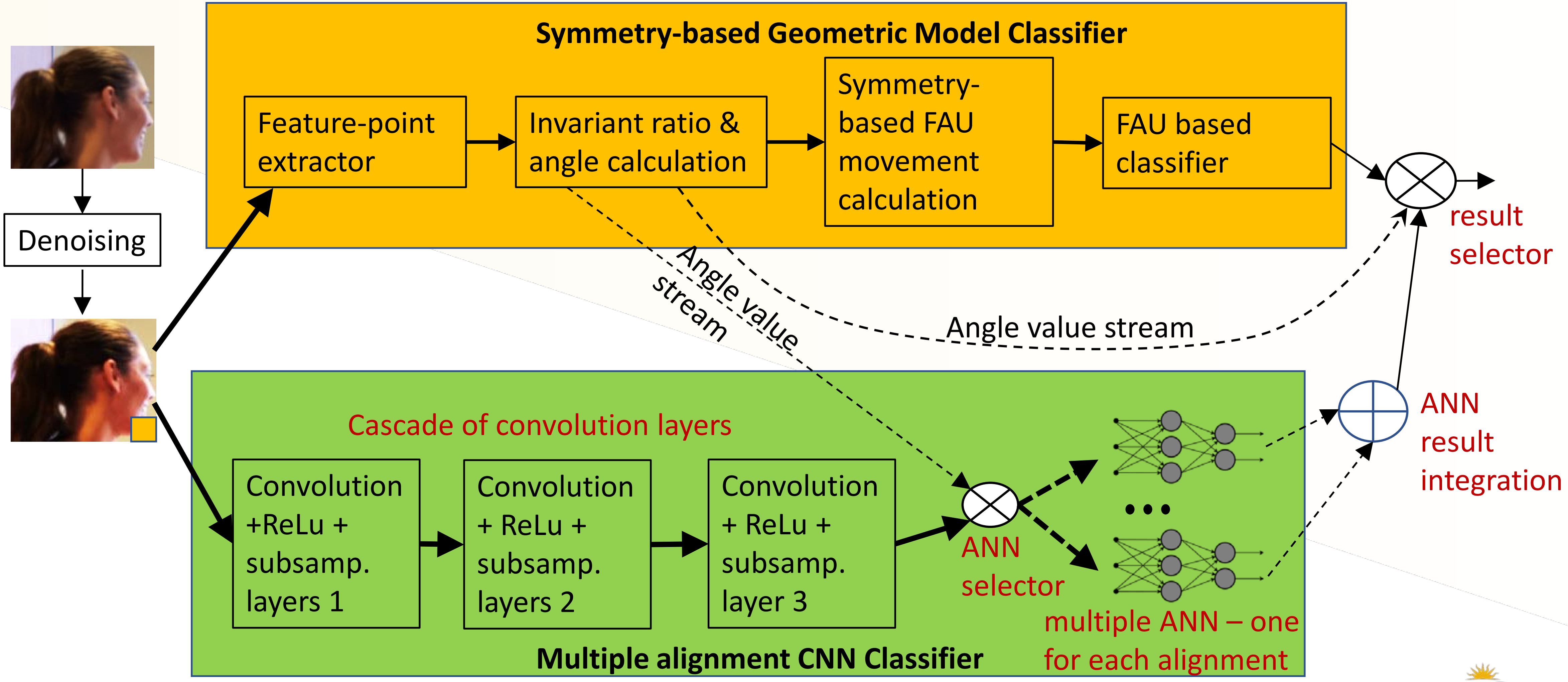
FAU	Condition (n = m + k and k > 0)
10	$R_n^{LH} > R_m^{LH}$
15	$R_n^{EL} > R_m^{EL} \wedge R_n^{EW} > R_m^{EW}$
16	$R_n^{LH} < R_m^{LH} \wedge R_n^{EL} > R_m^{EL}$
17	$R_n^{EL} < R_m^{EL}$
20	$R_n^{LW} < R_m^{LW}$
23	$R_n^{LW} > R_m^{LW}$
26	$R_n^{EL} > R_m^{EL}$

KENT STATE UNIVERSITY

Implementation

- Multiple alignment CNN model with angle information upto partial occlusion
  - CNN classifier has convolution layer cascade
  - CNN classifier has multiple ANN layer – one for each angular alignment
  - Geometric model passes angle information to CNN classifier
- Symmetry-based Geometric model beyond partial occlusion
- CNN cascade has
  - Three CNN layers: conv-32 layer; conv-64 layer; conv-128 layer
  - Each CNN later has convolution filter + ReLu + pooling layer
  - followed by Softmax layer
- Uses RaFD database for training the CNN
- Epochs of 200 continuous facial images in wild for facial expression recognition

Implementation Architecture



Experimental Results

- Recall TP/(TP + FN) in trained database is much higher than in wild
  - even in frontal pose recall degrades by 6% – 22%
- In CNN model using RaFD controlled database, recall degrades by
  - 11%- 15% for partial occlusion (upto $\pm$ 45°)
  - 27%-35% for beyond partial occlusion (> $\pm$ 45°)
- Confusion matrix for CNN based model in wild shows
  - predicting negative emotions get mixed with higher error percentage: fear, sadness and anger
  - predicting neutral face gets mixed with anger, fear and happiness
- Hybrid model performs much better beyond partial occlusion in wild than CNN model even in controlled RAFD database
  - improvement is 8% (sadness) – 21% (anger) over CNN model
  - In the wild, degradation from the frontal pose to completely occluded state is 6% (anger) and 18% (sadness)

Discussion

- Reasons for deterioration of facial expression detection in wild
  - mixing of facial muscles and feature-points for *sadness*, *fear* and *anger*
  - variations in expressed intensity level of the intended facial expressions in real-time
  - continuous random head-motions during real-time causing noise
  - uneven ambient lighting conditions and shadows obscuring feature-points
  - video-frame may not correspond to the apex of facial-expression (Cruz et al., 2014)
- Reasons for mixing of negative facial expressions
  - mixing of facial muscles and feature-points for *sadness*, *fear* and *anger*
  - mixing of facial expressions in real-time
  - improper labeling during emotion transition
  - uncontrolled involuntary thoughts affecting involuntary facial expressions

Conclusion

- Conversational head-gestures / multi-party interactions cause extreme occlusions
- Current schemes are limited to patches of partial occlusion using many methods
- Popular CNN based model degrades significantly beyond partial occlusion
- Symmetry-based methods can reconstruct discriminative feature-points
- Hybrid model integrating CNN with rotation-invariant symmetry-based model improves recall in the wild beyond partial occlusion significantly
- Future work involves DBN to smooth spurious facial-expression predictions embedded flanked by the same facial expression.