29th EUSIPCO
European Signal Processing Conference
DUBLIN // IRELAND
23–27 AUGUST 2021

An Audiovisual Child Emotion Recognition System for Child-Robot Interaction Applications

Panagiotis P. Filntisis[1], Niki Efthymiou[1], Gerasimos Potamianos[2], and Petros Maragos[1]

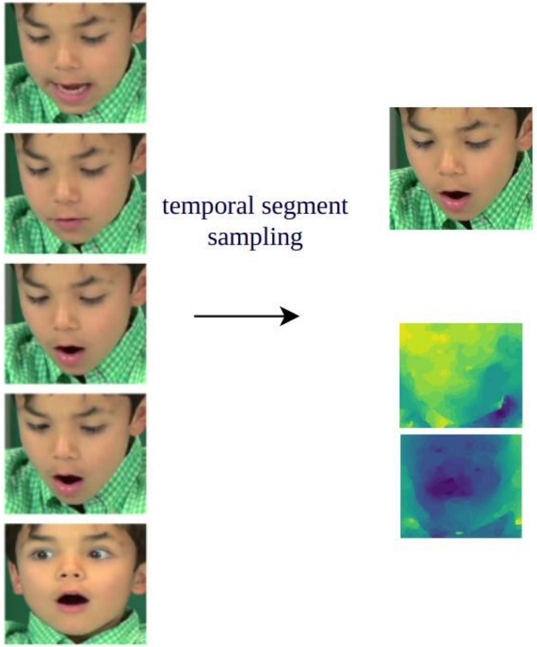
[1]School of ECE, National Technical University of Athens, 15773 Athens, Greece
[2]Department of ECE, University of Thessaly, 38221 Volos, Greece

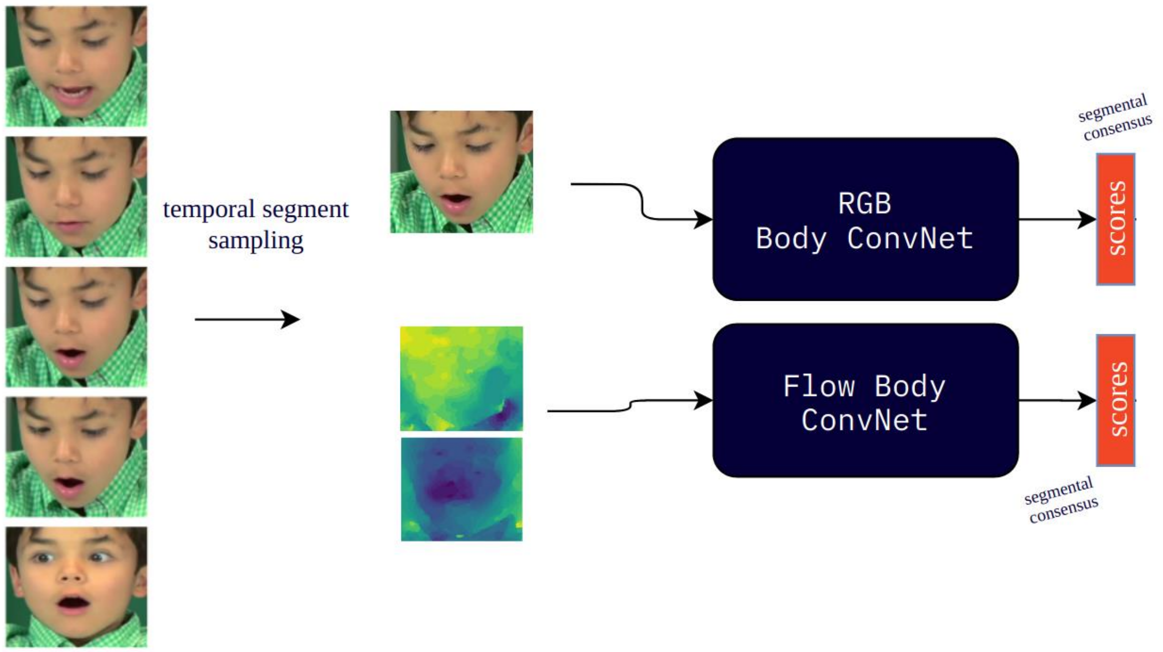
Motivation & Challenges

- Emotion is a fundamental aspect of human communication
- Empathic Robots create stronger bonds with children
- Children present different behavioral patterns compared to adults
  - Use of multiple modalities can help
- Human-robot communication requires real time performance

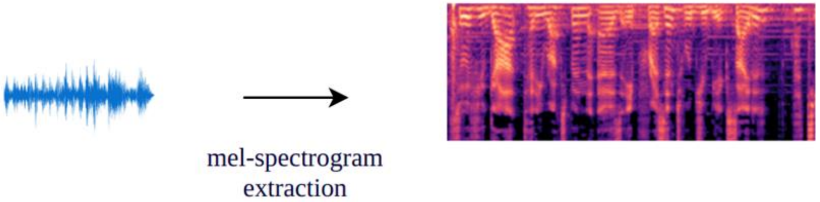

Method

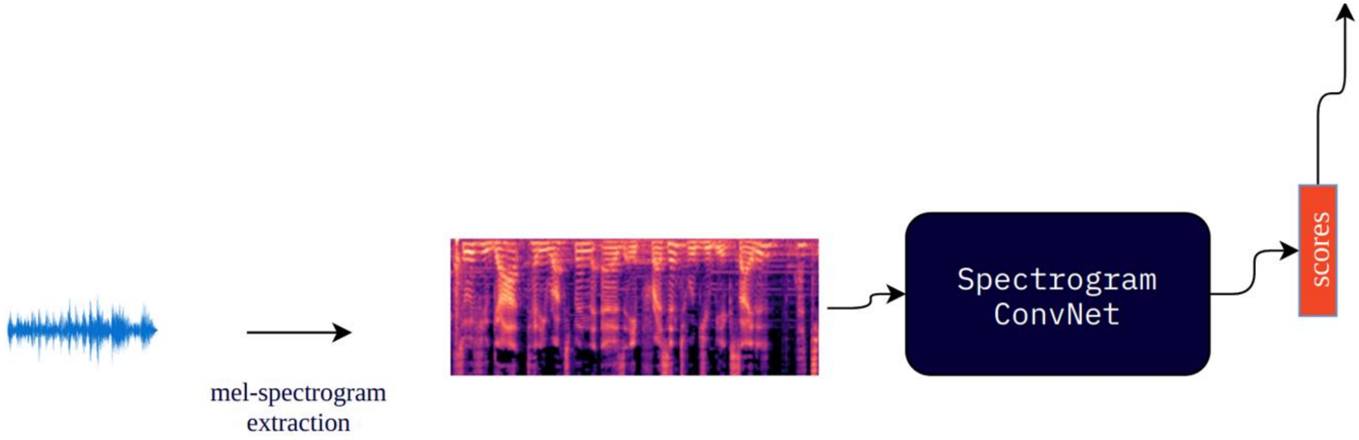
Visual Branch



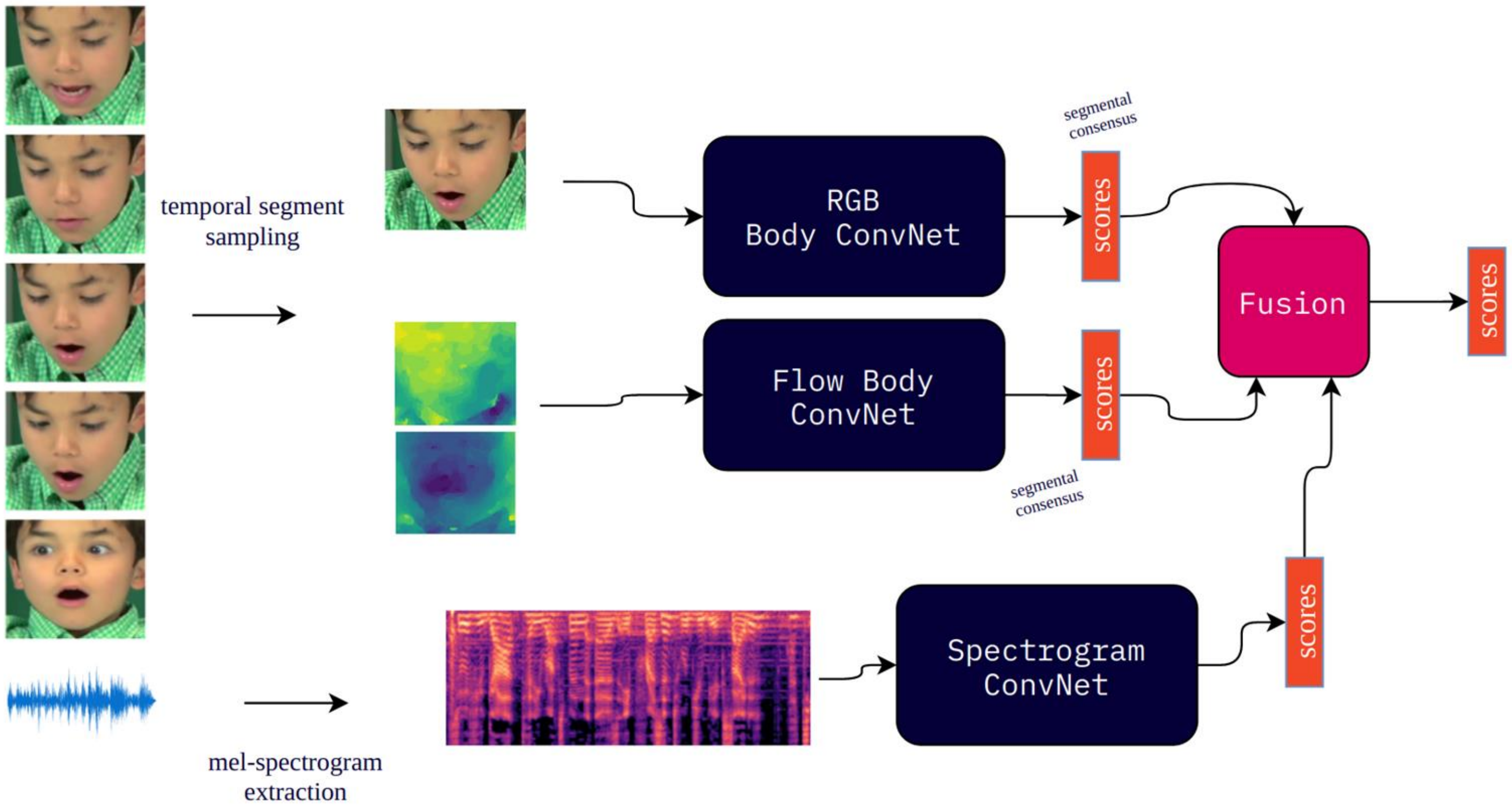

temporal segment sampling
RGB Body ConvNet
Flow Body ConvNet
segmental consensus
scores
scores
segmental consensus

Audio Branch



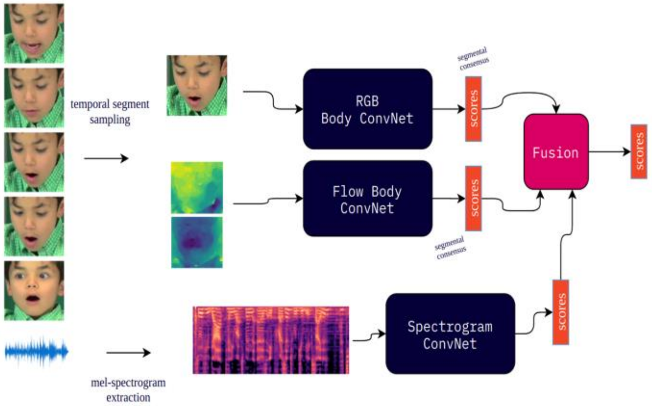


29th EUSIPCO
European Signal Processing Conference
DUBLIN // IRELAND
23–27 AUGUST 2021


temporal segment sampling
segmental consensus
RGB Body ConvNet
scores
Fusion
scores
Flow Body ConvNet
scores
segmental consensus
scores
Spectrogram ConvNet
mel-spectrogram extraction

Architecture Benefits

1. models long-range temporal structure
2. temporal sampling
   a. avoids overfitting
   b. acts as a type of data augmentation
3. small computational cost compared to other architectures



The EmoReact Database

- 63 children, aged 4-14
- 1102 youtube videos (432 train, 303 val, 367 test), audio and visual expressions
- 8 emotions (curiosity, uncertainty, excitement, happiness, surprise, disgust, fear, frustration)

29th EUSIPCO
European Signal Processing Conference
DUBLIN // IRELAND
23–27 AUGUST 2021

Experimental Details

- Resnet50 CNN Backbone
- Visual backbone pretrained on the AffectNet Dataset
- 60 epochs
- 0.01 learning rate SGD (reduced at 20 and 40 epochs)
- evaluation with ROC-AUC

3 ablation studies

- computational burden and performance against segments
- feature fusion vs score fusion
- per emotion/modality performance

ROC AUC AND AVERAGE TIME ELAPSED PER EPOCH WITH VARYING NUMBER OF SAMPLED SNIPPETS.

Segments	ROC AUC		sec/train epoch	sec/val epoch
	Balanced	Unbalanced		
RGB				
1	0.685	0.773	11	7
3	0.713	0.786	27	20
5	0.709	0.787	40	26
10	0.715	0.788	73	51
Flow				
1	0.585	0.741	37	23
3	0.596	0.744	101	70
5	0.623	0.757	166	115
10	0.627	0.759	294	210

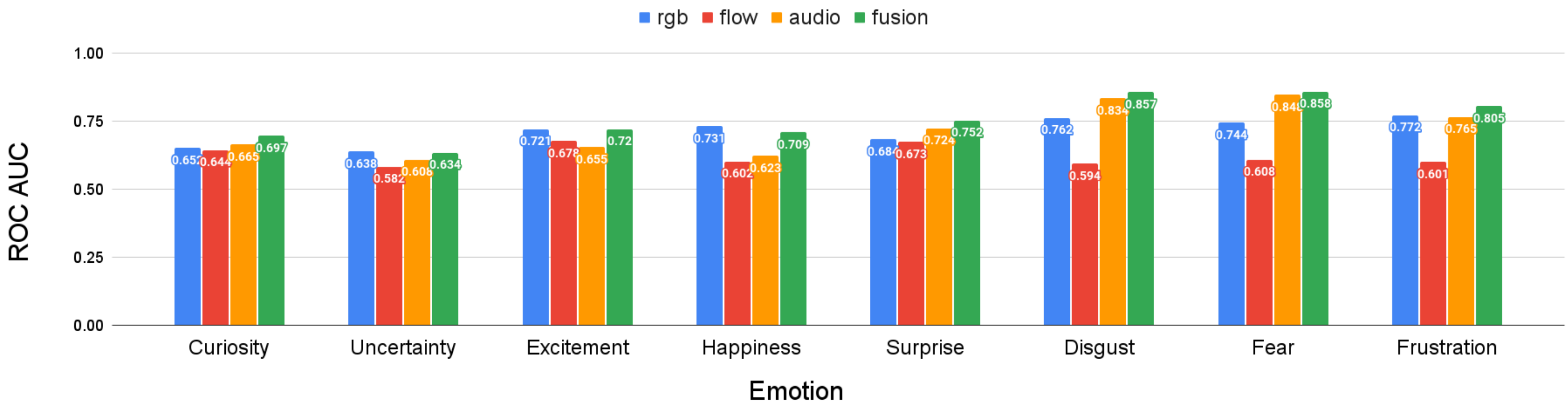
- increasing # segments above a threshold → small performance increase, large computation
    - choosing a trade-off at 3 RGB segments and 5 Flow segments

RESULTS ON THE EMOREACT DATASET FOR DIFFERENT FUSION AND TRAINING SCHEMES BETWEEN THE RGB-AUDIO AND FLOW-AUDIO MODALITIES.

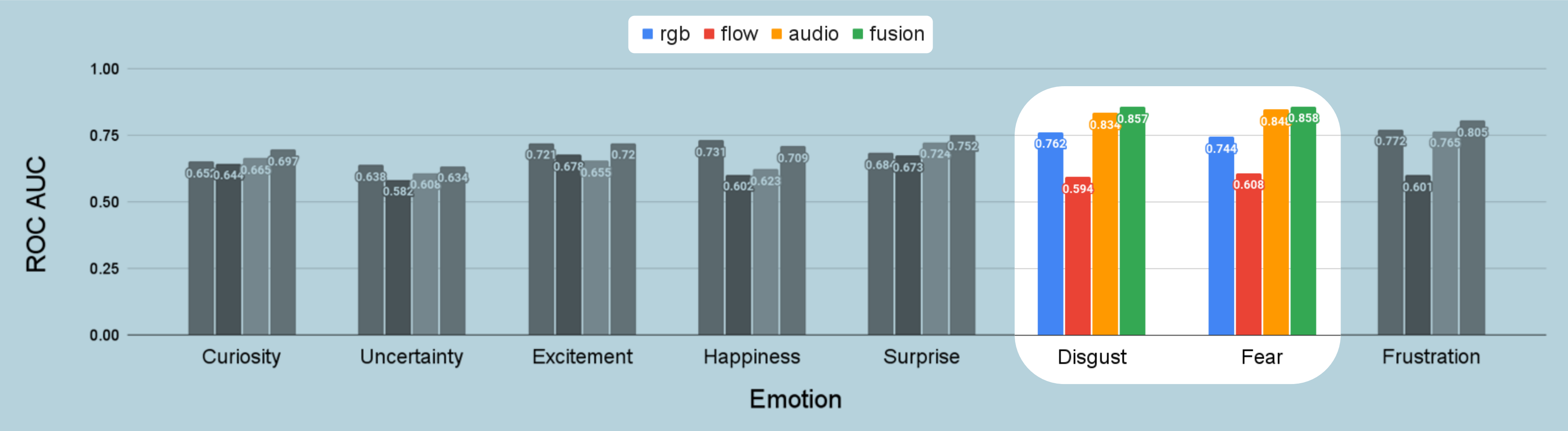
Fusion	**Training**	**ROC AUC**	
		Balanced	Unbalanced
Single Modality	Audio	0.715	0.750
	Visual (RGB)	0.713	0.786
	Visual (Flow)	0.623	0.757
Score Fusion RGB-audio	Joint Training	0.720	0.756
	Independent Training	**0.747**	**0.799**
Score Fusion Flow-audio	Joint Training	0.719	0.746
	Independent Training	0.725	0.787
Feature Fusion RGB-audio	Joint Training	0.719	0.769
Feature Fusion Flow-audio	Joint Training	0.707	0.744

score average fusion with independent of modalities training gives best result

Per Emotion Performance

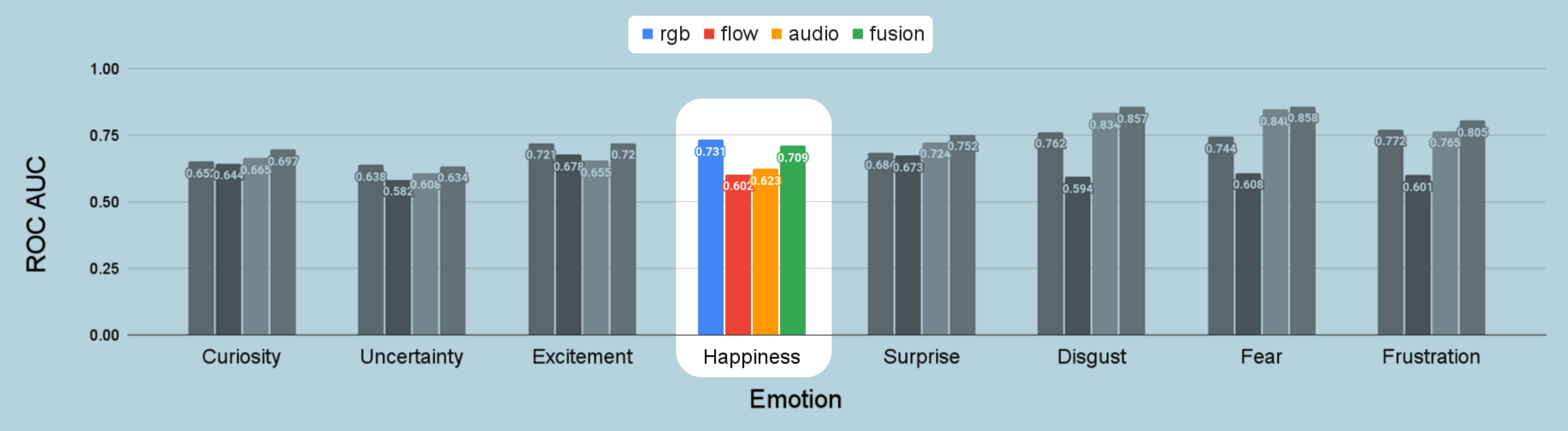


Per Emotion Performance



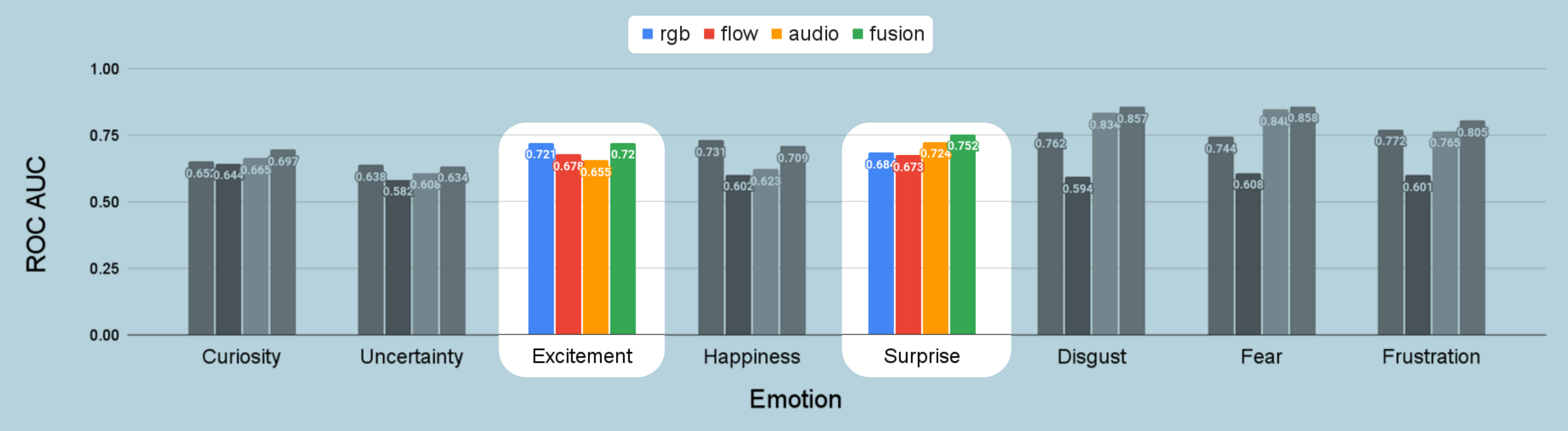
fear & disgust best identified through speech

Per Emotion Performance



happiness best identified through RGB

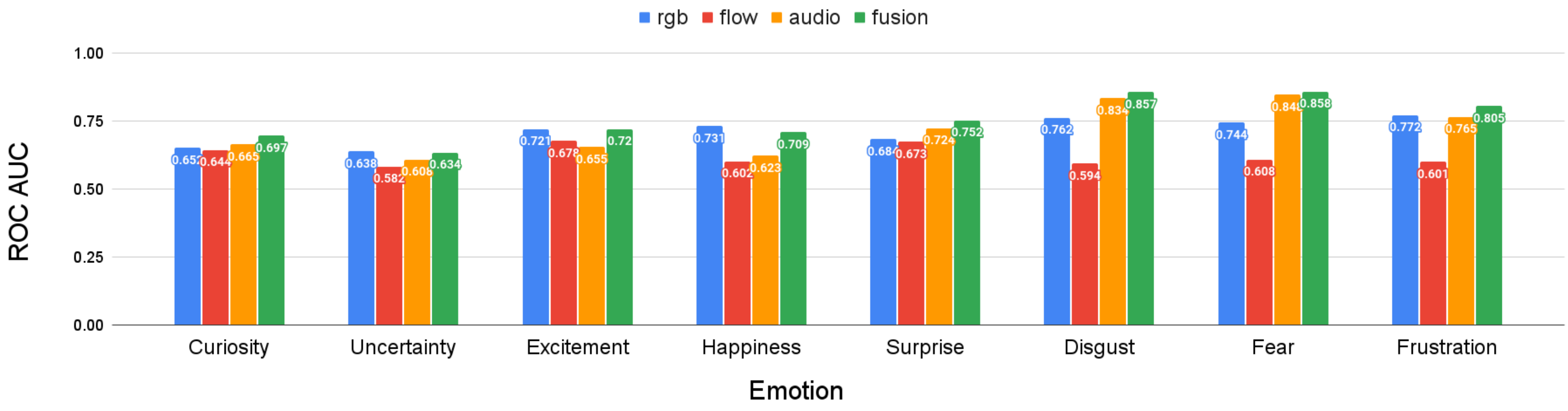
Per Emotion Performance



flow high performance in excitement and surprise

29th EUSIPCO
European Signal Processing Conference
DUBLIN // IRELAND
23–27 AUGUST 2021

Per Emotion Performance



fusion results in higher performance in most cases

FINAL ROC AUC RESULTS ON THE EMOREACT DATASET.

	ROC AUC	
	Balanced	Unbalanced
Audio		
audio features + SVM [1]	0.610	-
dnn ensemble + SVM [2]	0.718	-
Ours (End-to-End)	0.715	0.750
Visual		
openface + SVM [1]	0.620	-
Ours (Flow)	0.623	0.757
Ours (RGB)	0.713	0.786
AudioVisual		
[1]	0.640	-
Ours (RGB+Audio+Flow)	**0.754**	**0.809**

[1] B. Nojavanasghari, T. Baltrusaitis, C. E. Hughes, and L.-P. Morency, “EmoReact: a multimodal approach and dataset for recognizing emotional responses in children,” in Proc. ICMI, 2016
[2] B. Nagarajan and V. R.M. Oruganti, “Cross-domain transfer learning for complex emotion recognition,” in Proc. TENSYMP, 2019.

Conclusions

- We proposed a novel multimodal emotion recognition system for CRI
- Tackles CRI challenges: small datasets, fast inference, low-cost training
- Ablation studies on various aspects of the system
- Achieved state-of-the-art results on the EmoReact dataset

An Audiovisual Child Emotion Recognition System for Child-Robot Interaction Applications

Panagiotis P. Filntisis[1], Niki Efthymiou[1], Gerasimos Potamianos[2], and Petros Maragos[1]

[1]School of ECE, National Technical University of Athens, 15773 Athens, Greece
[2]Department of ECE, University of Thessaly, 38221 Volos, Greece

Email contact: filby@central.ntua.gr

ACKNOWLEDGMENTS

This research is carried out/funded in the context of the project "Intelligent Child-Robot Interaction System for designing and implementing edutainment scenarios with emphasis on visual information" (MIS 5049533) under the call for proposals "Researchers' support with an emphasis on young researchers- 2nd Cycle". The project is co-financed by Greece and the European Union (European Social Fund- ESF) by the Operational Programme Human Resources Development, Education and Lifelong Learning 2014-2020.

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

For more information:
http://cvsp.cs.ntua.gr/
https://robotics.ntua.gr/