UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

softech-ict
Centro Interdipartimentale di Ricerca
Softech: ICT per le Imprese

AImageLab

EURECOM
Sophia Antipolis

Modeling Multimodal Cues in a Deep Learning-Based Framework for Emotion Recognition in the Wild

Stefano Pini

S. Pini[1], O. Ben Ahmed[2], M. Cornia[1], L. Baraldi[1], R. Cucchiara[1], B. Huet[2]

[1] name.surname@unimore.it, *University of Modena and Reggio Emilia, Italy*

[2] name.surname@eurecom.fr, *EURECOM, France*

Outline

- Our Goal
- Data Preprocessing
- Proposed Model
- Training Procedure
- Experimental Evaluation

Our Goal

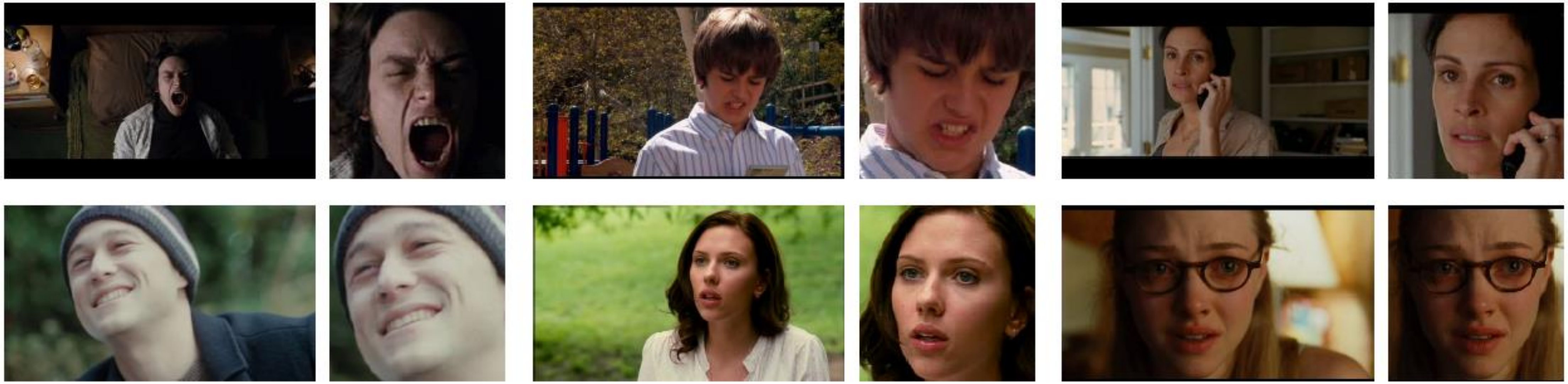
Recognize expressed emotions in unconstrained videos using:

- Multiple Modalities
  - Audio
  - Video
- Multiple Timescales
  - Single frame
  - Multiple frames
  - Video aggregation
- An End-to-End Deep Neural Network Architecture

Data Preprocessing

Data Preprocessing – Face Extraction

Face crops provided by the challenge contain many errors and they are not accurate, therefore we extract new face crops using the deep architecture called *Multi-Task Cascaded Convolutional Networks* proposed by Zhang *et al*[1].



Some examples of the extracted face bounding boxes.

[1] *K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks", IEEE SPL, 2016.*

Data Preprocessing – Audio Extraction

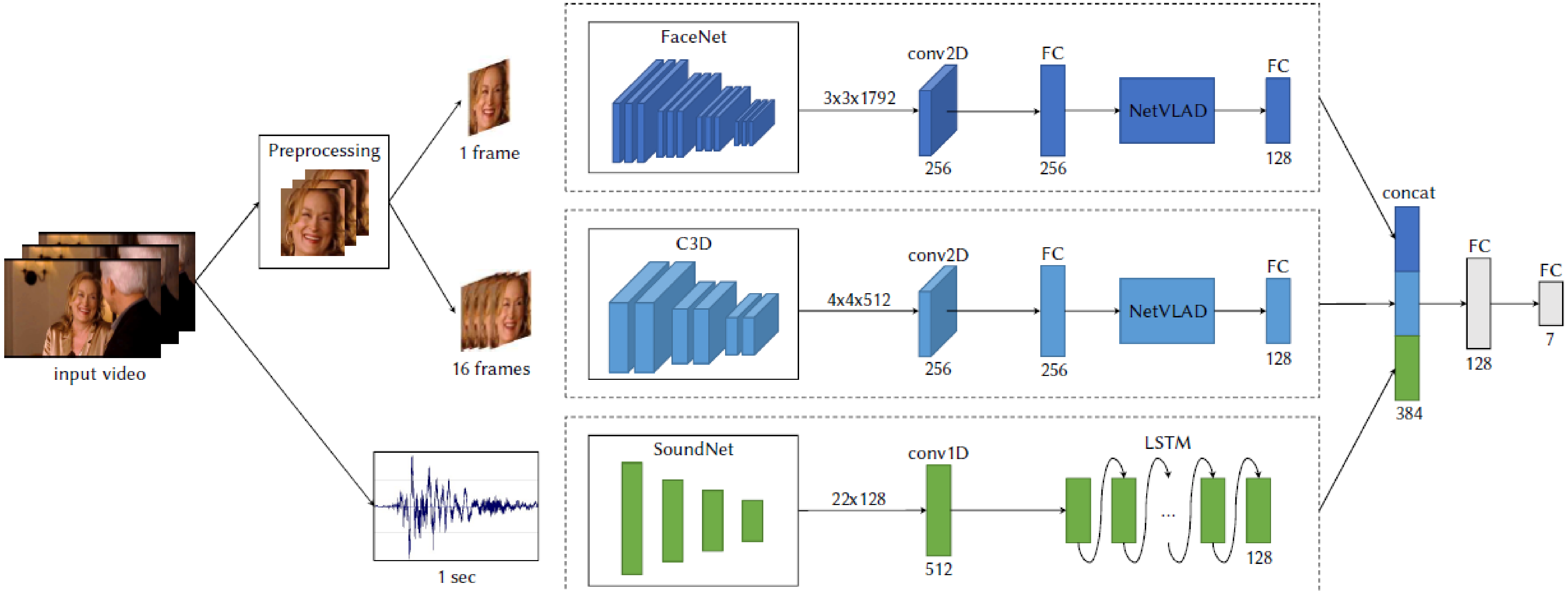
Regarding the audio data, we follow the process described in the work of *Aytar et al.*[1]:

- We extract the audio from each video with a sampling rate of 22,050 Hz
- We save it as mono-channel mp3 file
- We rescale it to be in the range [-256, 256]

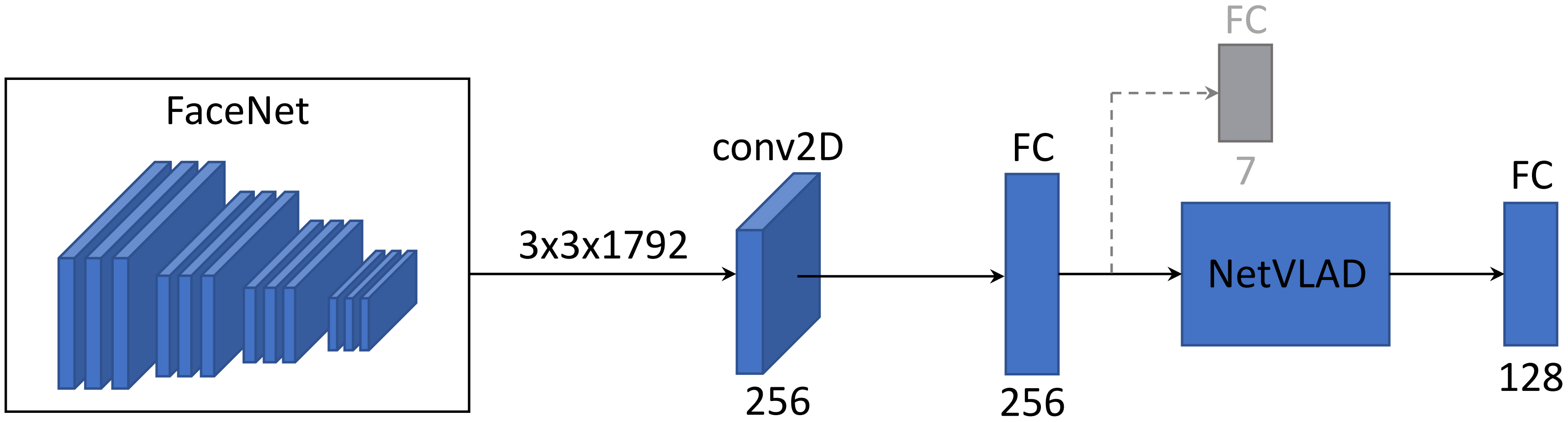
[1] *Y. Aytar, C. Vondrick, A. Torralba, "SoundNet: Learning Sound Representations from Unlabeled Video", NIPS, 2016.*

Proposed Model

Proposed Model



Proposed Model – CNN Branch



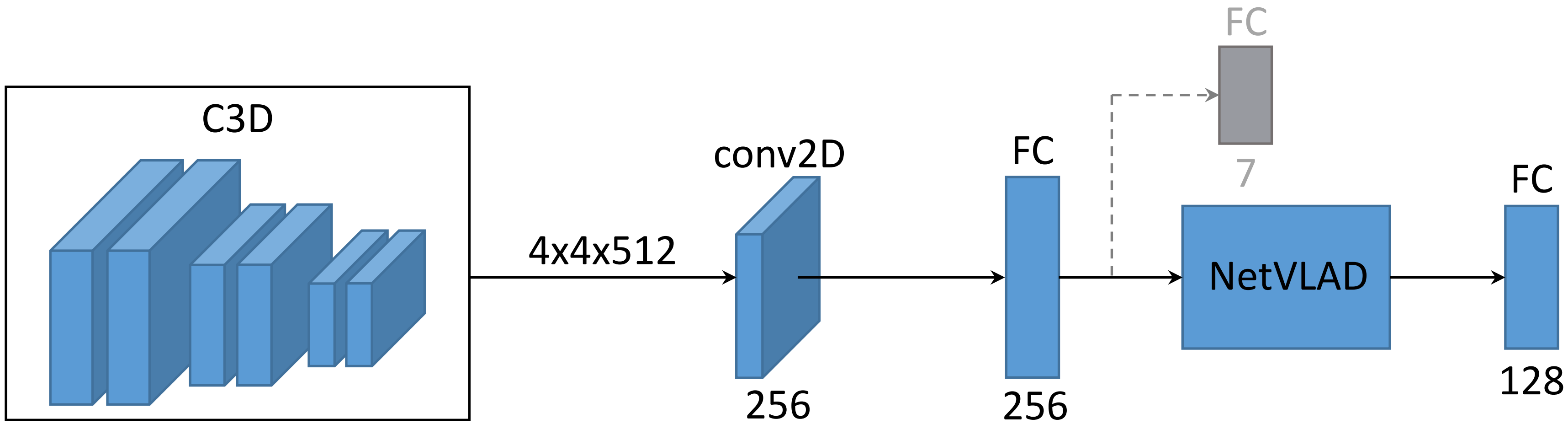
- Input: single video frames
- Feature extractor: FaceNet[1]-like architecture, pre-trained with MS-Celeb-1M[2]
- Network composed by: a 2D convolutional layer and a fully connected layer
- Training through: an additional 7-unit fully-connected layer
- Temporal aggregation: NetVLAD[3] layer

[1] *F. Schroff, D. Kalenichenko, J. Philbin, "FaceNet: A unified embedding for face recognition and clustering", CVPR, 2015.*

[2] *Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, "MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition", ECCV, 2016.*

[3] *R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition", CVPR, 2016.*

Proposed Model – C3D Branch



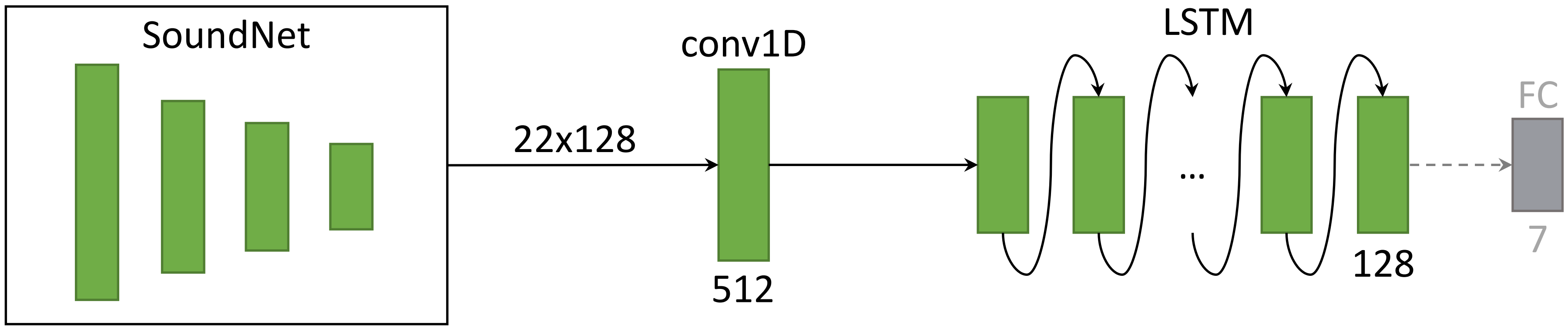
- Input: video slices of 16 frames
- Feature extractor: C3D network architecture[1], pre-trained with YouTube Sports-1M[2]
- Network composed by: a 2D convolutional layer and a fully connected layer
- Training through: an additional 7-unit fully-connected layer
- Temporal aggregation: NetVLAD[3] layer

[1] *D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks", ICCV, 2015.*

[2] *A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, F. Li, "Large-Scale Video Classification with Convolutional Neural Networks", CVPR, 2014.*

[3] *R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition", CVPR, 2016.*

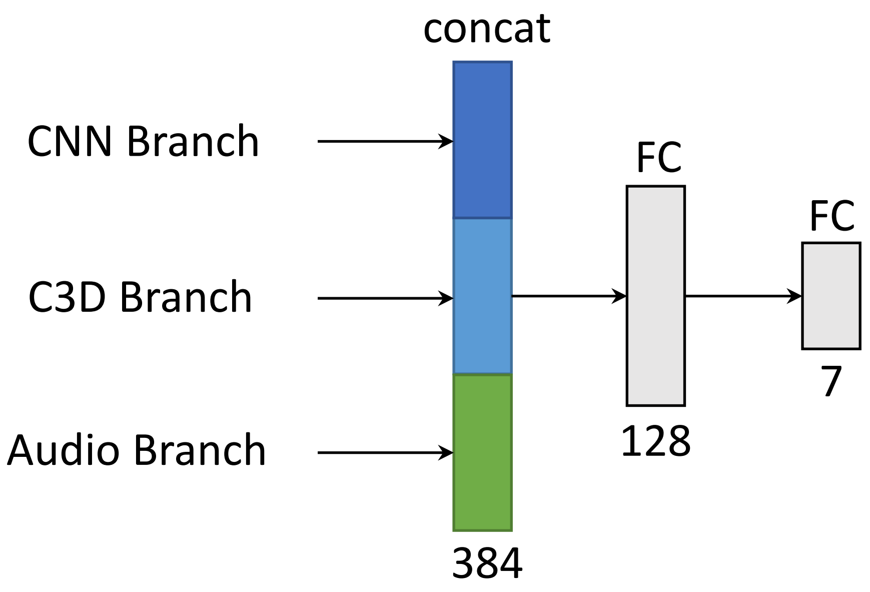
Proposed Model – Audio Branch



- Input: 1-second length audio waveforms
- Feature extractor: SoundNet[1] network, pre-trained with 2M unlabeled videos
- Network composed by: a 1D convolutional layer and a dual-level LSTM layer
- Training through: an additional 7-unit fully-connected layer

[1] *Y. Aytar, C. Vondrick, A. Torralba, "SoundNet: Learning Sound Representations from Unlabeled Video", NIPS, 2016.*

Proposed Model – Multimodal Fusion



- Input: Concatenation of the outputs of the three branches
- Network composed by: two fully connected layers
- Training includes: the NetVLAD and the following fully connected layers
- Output: final classification of the input video

Training Procedure

Training Steps

Training is executed in two sequential steps:

1. Training of the Single Branches
   - The three branches are independently trained with the categorical cross-entropy loss function
   - Various datasets are used to increase the training data:
     - CNN branch (single frames) > AFEW[1] and FER[2] dataset
     - C3D branch (16-frame slices) > AFEW[1] dataset
     - Audio branch (1-second length audio) > AFEW[1] and eNTERFACE[3] dataset

2. Multimodal Fusion Network Training
   - The NetVLAD layers, the following fully connected layers, and the multimodal network are trained
   - Two different loss functions were tested
   - Only the AFEW[1] dataset is used during training

[1] *A. Dhall, R. Goecke, S. Lucey, T. Gedeon, "Collecting Large, Richly Annotated Facial-Expression Databases from Movies", IEEE MM, 2012.*

[2] *I. J. Goodfellow et al., "Challenges in Representation Learning: A Report on Three Machine Learning Contests", ICONIP, 2013.*

[3] *O. Martin, I. Kotsia, B. M. Macq, I. Pitas, "The eNTERFACE'05 Audio-Visual Emotion Database", ICDE Workshops, 2006.*

Multimodal Fusion Network Loss Functions

1. Standard Categorical Cross-Entropy (CCE) Loss Function

   Common categorical cross-entropy loss function:

$$L = -\sum_{i=1}^{N} \mathbf{t}_i^T \log(\mathbf{p}_i)$$

2. Weighted Categorical Cross-Entropy (CCE) Loss Function

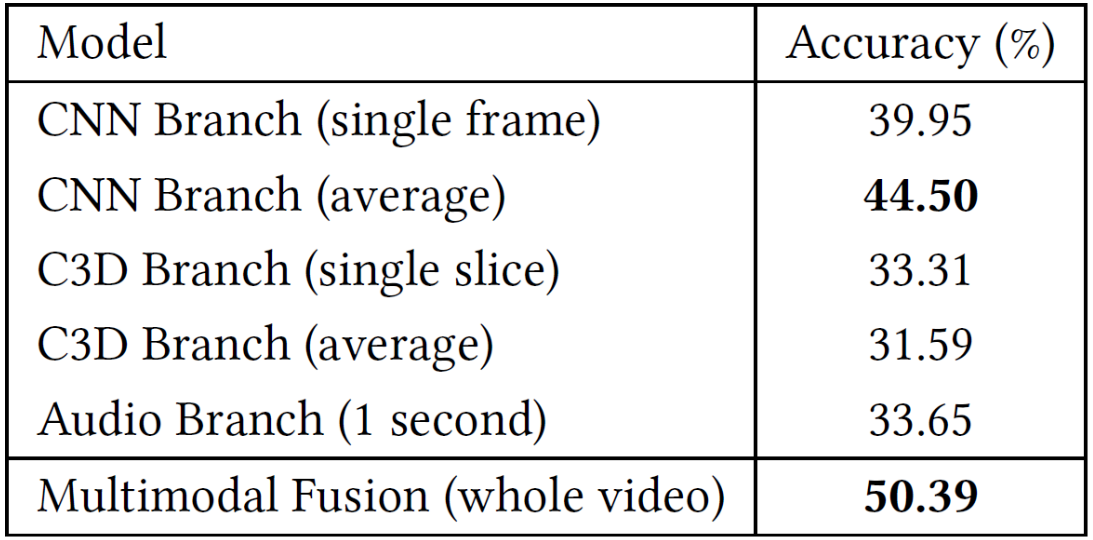
   The standard CCE loss value is multiplied by a regularizing parameter based on the distribution of the classes in the training set.

   The goal is to increase the importance of the most frequent classes and to reduce the importance of the less frequent ones.

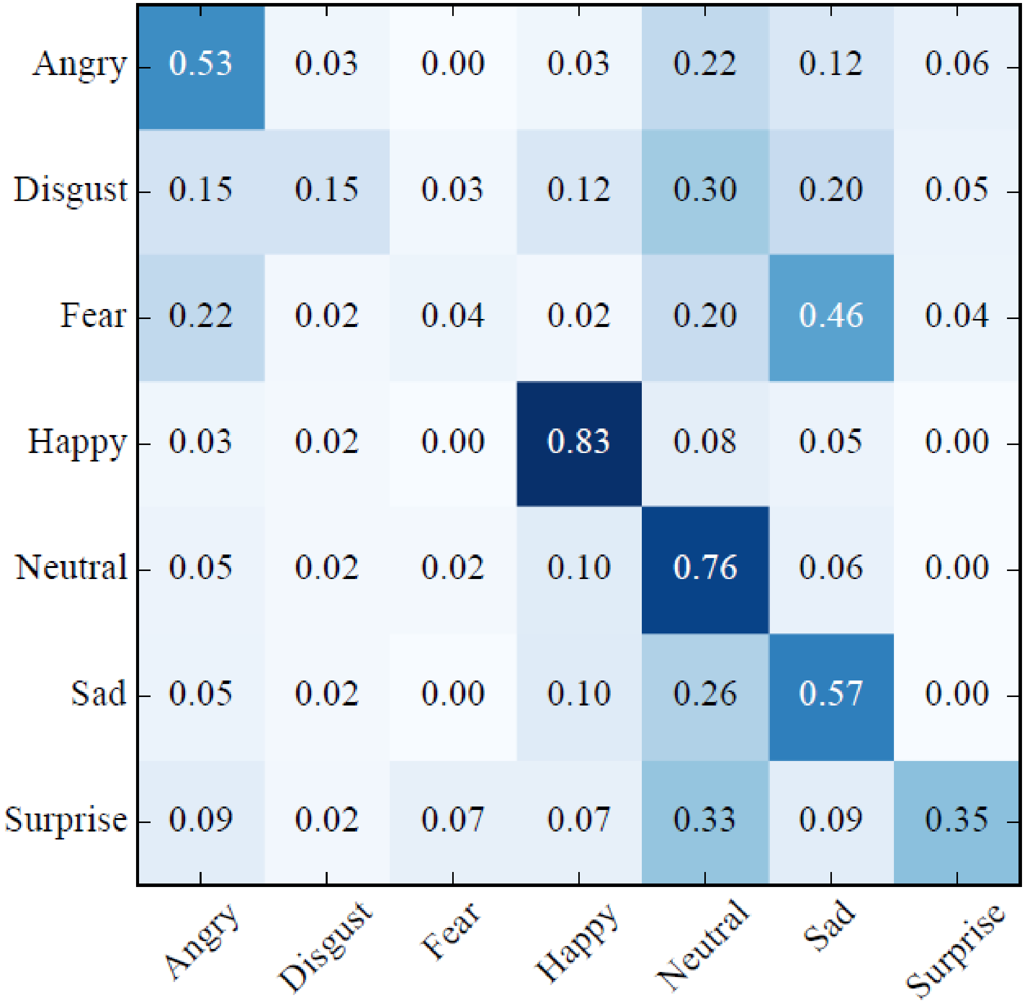
$$L' = -\sum_{i=1}^{N} \lambda_{c_i} \mathbf{t}_i^T \log(\mathbf{p}_i)$$

Experimental Evaluation

Results on Validation Set



Model	Accuracy (%)
CNN Branch (single frame)	39.95
CNN Branch (average)	**44.50**
C3D Branch (single slice)	33.31
C3D Branch (average)	31.59
Audio Branch (1 second)	33.65
Multimodal Fusion (whole video)	**50.39**



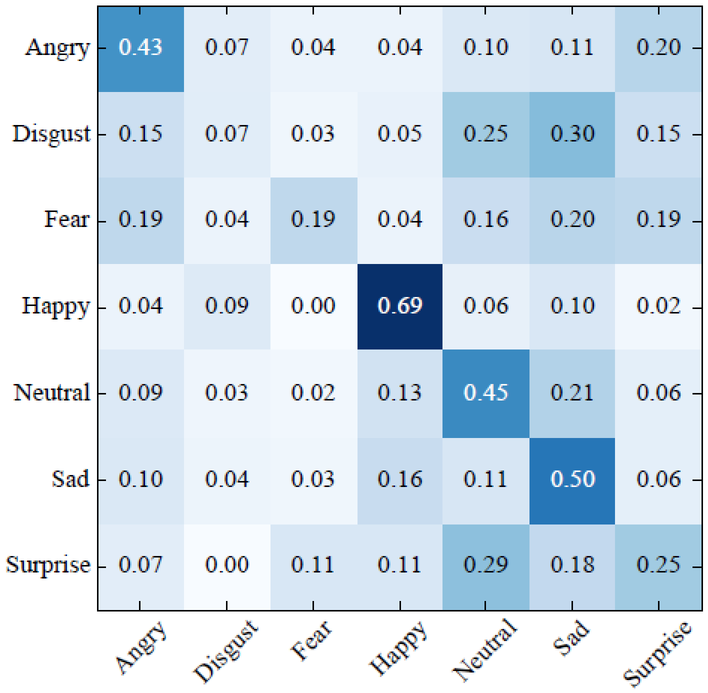
	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.53	0.03	0.00	0.03	0.22	0.12	0.06
Disgust	0.15	0.15	0.03	0.12	0.30	0.20	0.05
Fear	0.22	0.02	0.04	0.02	0.20	0.46	0.04
Happy	0.03	0.02	0.00	0.83	0.08	0.05	0.00
Neutral	0.05	0.02	0.02	0.10	0.76	0.06	0.00
Sad	0.05	0.02	0.00	0.10	0.26	0.57	0.00
Surprise	0.09	0.02	0.07	0.07	0.33	0.09	0.35

Best Challenge Submissions – Results on Test Set

Submission 3: Standard CCE
Submission 4: Weighted CCE

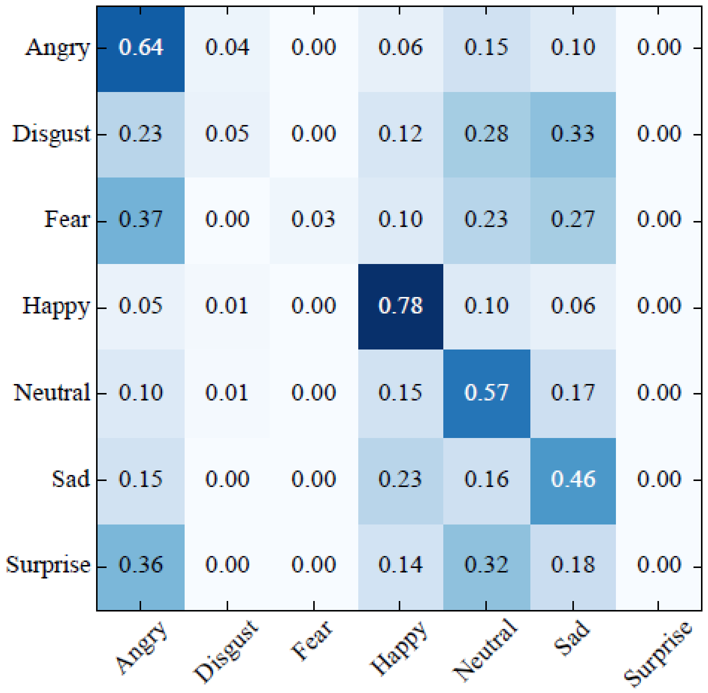
Submission	Cross Validation (%)	Test Set (%)
3	**53.49**	44.56
4	48.30	**49.92**

AFEW validation set is split in five subject-independent folds and the models are trained five times, following the k-fold cross validation technique.



	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.43	0.07	0.04	0.04	0.10	0.11	0.20
Disgust	0.15	0.07	0.03	0.05	0.25	0.30	0.15
Fear	0.19	0.04	0.19	0.04	0.16	0.20	0.19
Happy	0.04	0.09	0.00	0.69	0.06	0.10	0.02
Neutral	0.09	0.03	0.02	0.13	0.45	0.21	0.06
Sad	0.10	0.04	0.03	0.16	0.11	0.50	0.06
Surprise	0.07	0.00	0.11	0.11	0.29	0.18	0.25

Submission 3



	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.64	0.04	0.00	0.06	0.15	0.10	0.00
Disgust	0.23	0.05	0.00	0.12	0.28	0.33	0.00
Fear	0.37	0.00	0.03	0.10	0.23	0.27	0.00
Happy	0.05	0.01	0.00	0.78	0.10	0.06	0.00
Neutral	0.10	0.01	0.00	0.15	0.57	0.17	0.00
Sad	0.15	0.00	0.00	0.23	0.16	0.46	0.00
Surprise	0.36	0.00	0.00	0.14	0.32	0.18	0.00

Submission 4

Conclusion

We propose an End-to-End Deep Neural Network Architecture for emotion recognition in the wild.

- We employ deep features regarding:
  - Static visual appearance
  - Visual appearance evolution through time
  - Audio information

- We combine multiple modalities with a Multimodal fusion network.

- We outperform the challenge baseline on:
  - Validation set > 11.52% absolute gain
  - Test set > 9.45% absolute gain

UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

softech-ict
Centro Interdipartimentale di Ricerca
Softech: ICT per le Imprese

AImageLab

EURECOM
Sophia Antipolis

Thank you for your attention

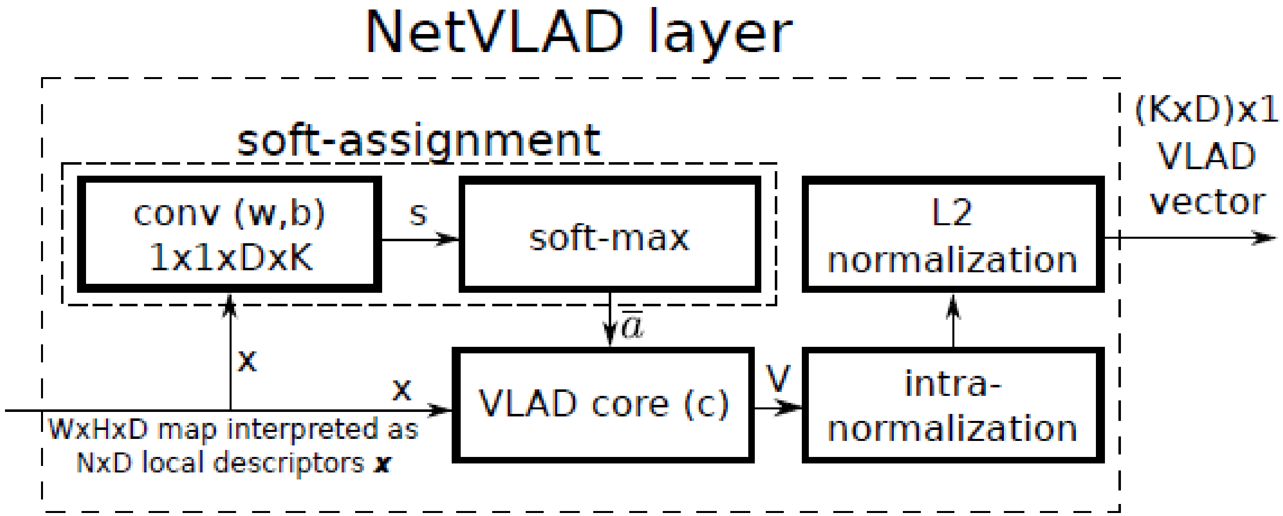
Any question?

imagelab.ing.unimore.it

www.eurecom.fr

Extra Slides

NetVLAD



$$V(j,k) = \sum_{i=1}^{N} a_k(\mathbf{x}_i)\,(x_i(j) - c_k(j))$$

$$\bar{a}_k(\mathbf{x}_i) = \frac{e^{\mathbf{w}_k^T \mathbf{x}_i + b_k}}{\sum_{k'} e^{\mathbf{w}_{k'}^T \mathbf{x}_i + b_{k'}}}$$

[1] *H. Jégou, M. Douze, C. Schmid, P. Pérez, "Aggregating local descriptors into a compact image representation", CVPR, 2010.*

[2] *R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition", CVPR, 2016.*

[3] *A. Miech, I. Laptev, J. Sivic, "Learnable pooling with Context Gating for video classification", CVPR Workshop, 2017.*