Football Action Classification using Face Emotion Features

Josep Reverter Sancho
Supervisors: Coloma Ballester,
Alejandro Cartas, Gloria Haro

upf. **Universitat Pompeu Fabra** *Barcelona*   Escola d'Enginyeria

Introduction and Motivation

- Football: more than 20 million players (in more than 120 countries) and more than 4 billion fans.

- Statistics, Artificial Intelligence and Computer Vision for tactical and communication purposes.

- Live written annotations, match highlights, automatic summaries...

- ✗ Field information │ ✓ Emotions of players and fans.
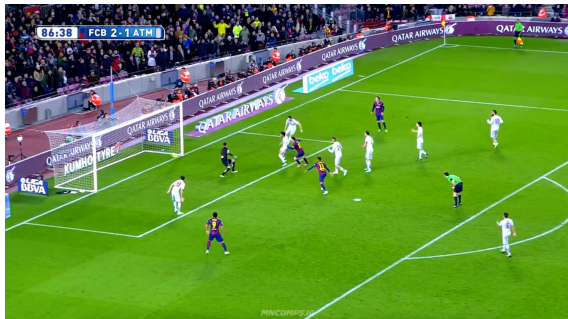
Objectives

1

Determine and implement an adequate face detector for the correct and accurate detection of faces.

2

Determine and implement an emotion feature extractor from faces.

3

Develop an action classification model that takes as input the extracted emotion features
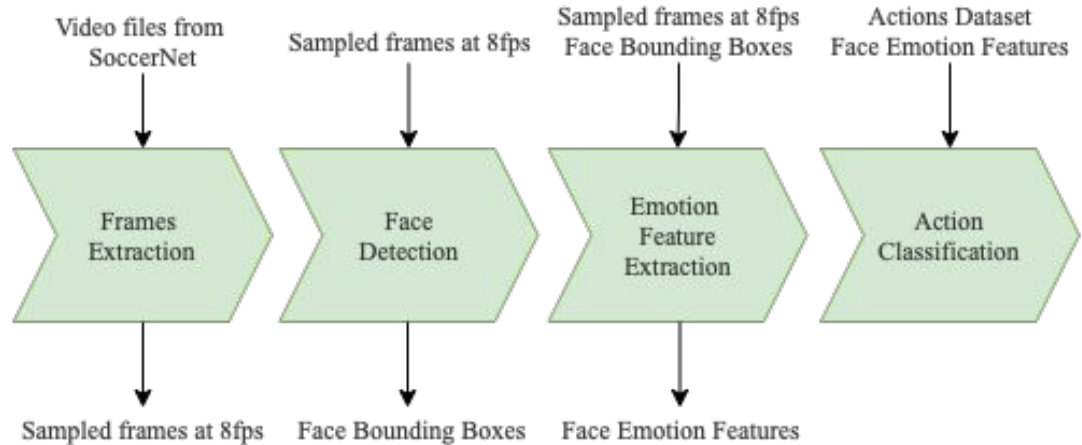
Proposed Methodology

Feature Extraction:

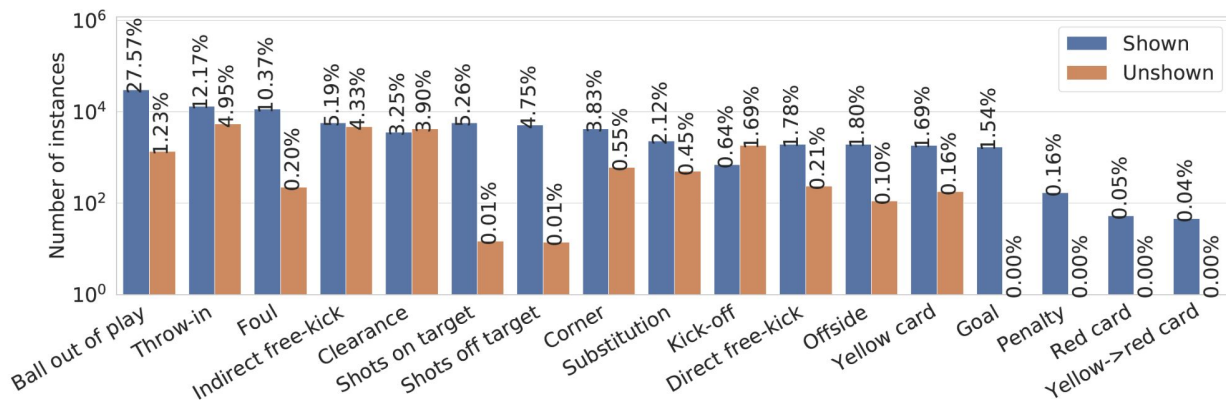- Frames Extraction
- Face Detection
- Emotion Feature Extraction

Classification:

- Action Classification

Video files from SoccerNet

Sampled frames at 8fps

Sampled frames at 8fps
Face Bounding Boxes

Actions Dataset
Face Emotion Features

Frames Extraction

Face Detection

Emotion Feature Extraction

Action Classification
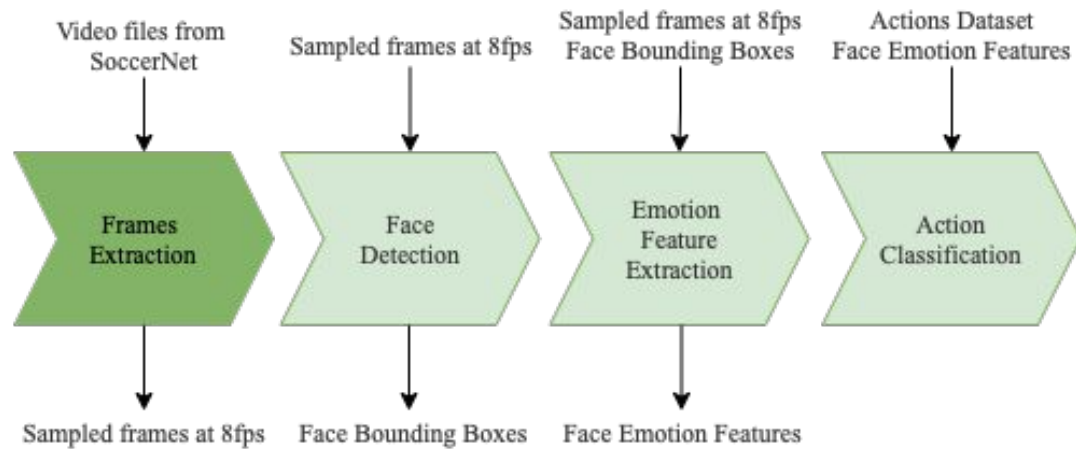
Sampled frames at 8fps

Face Bounding Boxes

Face Emotion Features

SoccerNet-v2

- 500 football matches from the 5 major European leagues + UEFA Champions League.

- 300,000 annotations.

- 17 classes.

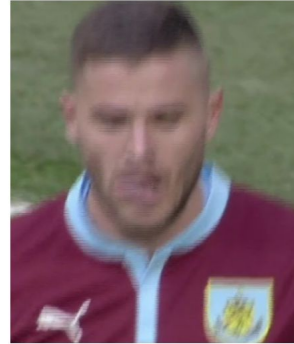- Detailed list of annotations per match.
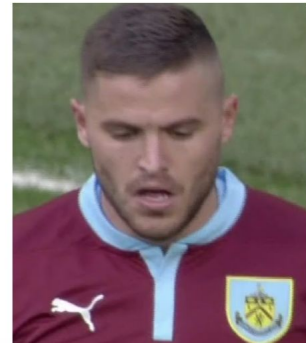
Frames Extraction

- ↑ fs │ ↑ quality │ ↑ time
- ↓ fs │ ↓ quality │ ↓ time

8fps: more than 21,600 frames per half

2 FPS   8 FPS

Video files from SoccerNet → **Frames Extraction** → Sampled frames at 8fps

Sampled frames at 8fps → **Face Detection** → Face Bounding Boxes

Sampled frames at 8fps
Face Bounding Boxes → **Emotion Feature Extraction** → Face Emotion Features

Actions Dataset
Face Emotion Features → **Action Classification**
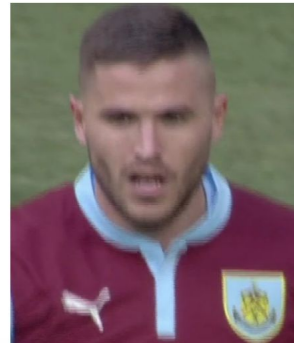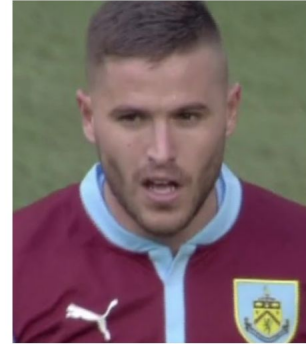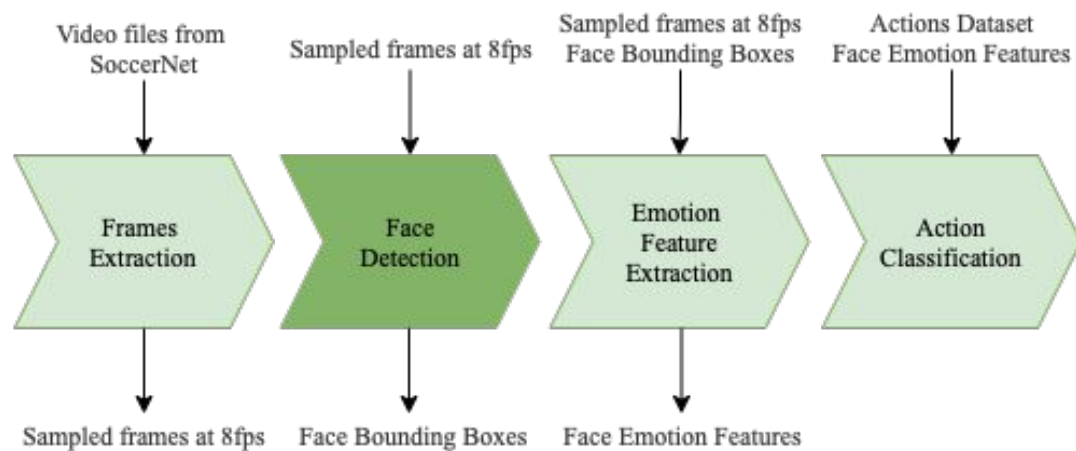
Face Detection

RetinaFace[1]:

- Learning based on 4 parts: face classification → face box regression → facial landmarks regression → and dense face regression

- Multi-task loss (for each part).

- Combination of a Feature Pyramid Network (FPN) and a Deformable Convolutional Network (DCN).
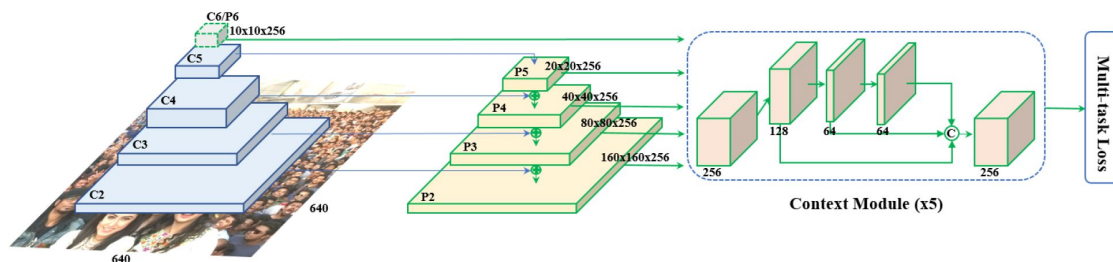
Implementation and Results

- Detect useful faces around actions.

- Time Window: 20s (smaller time windows can be chosen when we train the model).

- 5 classes with much more faces, and classes with a big faces per action ratio.

- Average distribution never exceeds ±60 frames.

Boxplot grouped by label

relative_frame

Video files from SoccerNet → **Frames Extraction** → Sampled frames at 8fps

Sampled frames at 8fps → **Face Detection** → Face Bounding Boxes

Sampled frames at 8fps / Face Bounding Boxes → **Emotion Feature Extraction** → Face Emotion Features

Actions Dataset / Face Emotion Features → **Action Classification**
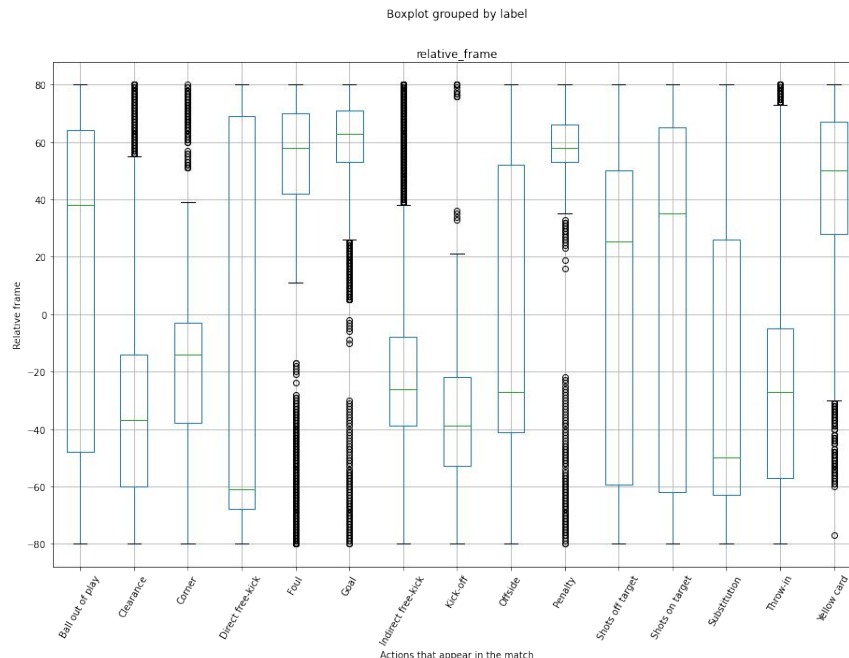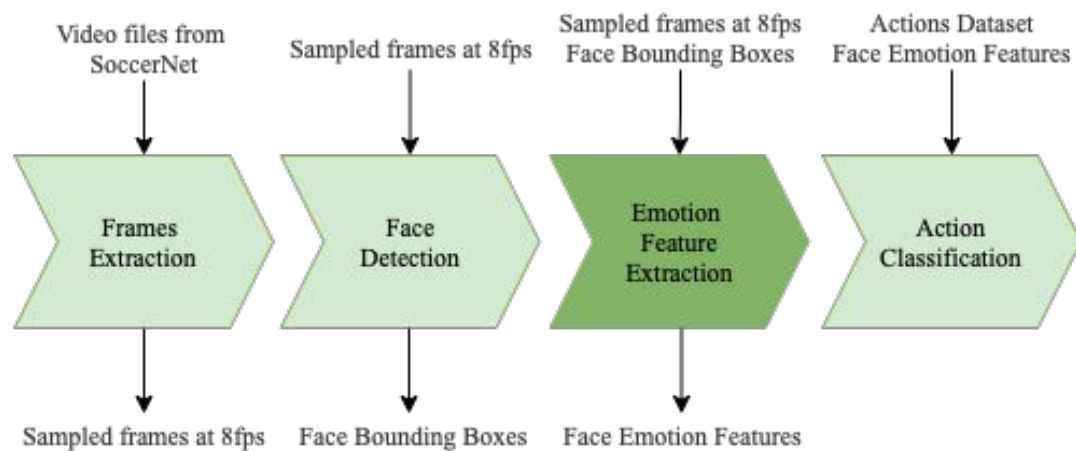
Emotion Feature Extraction

DeepFace[2]

- Convolutional Neural Network.

- 5 convolutional layers → flatten → 3 fully-connected layers.

- 7 emotion classification.

- Output of the flatten layer as the Face Emotion Feature.

Convolutional + ReLU
Max Pooling 2D
Average Pooling 2D
Fully Connected + ReLU
Fully Connected + SoftMax

Face Emotion Vector

44 x 44 x 64
20 x 20 x 64
18 x 18 x 64
16 x 16 x 64
7 x 7 x 64
5 x 5 x 128
3 x 3 x 128
1 x 1 x 128
1 x 1 x 1024
1 x 1 x 1024
1 x 1 x 7

Video files from
SoccerNet

Sampled frames at 8fps

Sampled frames at 8fps
Face Bounding Boxes

Actions Dataset
Face Emotion Features

Frames
Extraction

Face
Detection

Emotion
Feature
Extraction

Action
Classification

Sampled frames at 8fps

Face Bounding Boxes

Face Emotion Features

Action Classification

Model Scheme:

- 2 average pooling layers.
- 2 fully-connected layers.

Dataset Preprocessing

- 135 complete matches:

  - Frames that have no faces: 60% (949,104/1,562,545)
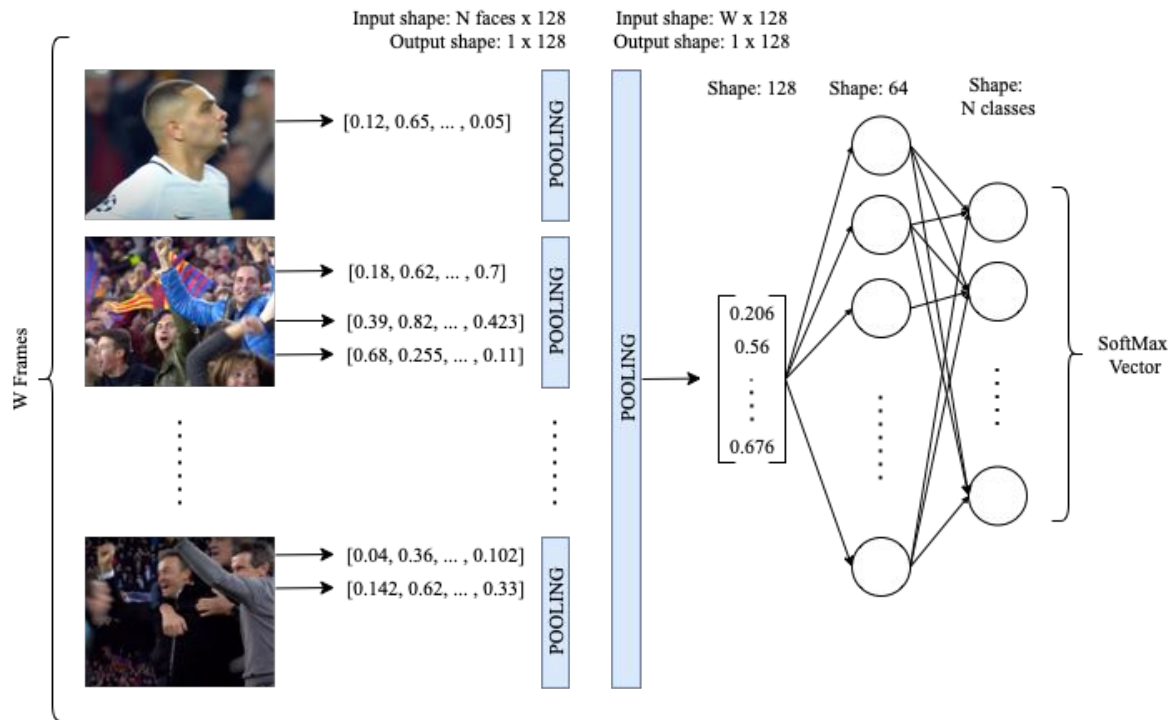 - Actions that have no faces: 5.7% (1,290/22,327)

Model Training

- SGD optimizer (momentum of 0.9).
- Cross-Entropy Loss.
- Early Stop of 15 iterations.

- Weighted Loss:
  - Inverse Number of Samples
  - Opposite of the Class Ratio

$$W = \frac{\frac{1}{\text{Number of samples of each class}}}{\sum(\frac{1}{\text{Number of samples of each class}})} * \text{Number of classes}$$

$$W = 1 - \frac{\text{Number of samples of each class}}{\text{Total number of samples}}$$

Experimental Results

Weight Equation	$T = 10$ ($W = 81$)	$T = 16$ ($W = 129$)	$T = 20$ ($W = 161$)
Not using Weighted Loss	28.7%	28.6%	28.5%
INS	**25.2%**	**18.4%**	**20.9%**
OCR	28.7%	28.6%	28.5%

C	WE	TW	TA	Pen	Ko	Goal	Sub	Son	Sof	Cl	Foul	Dfk	Cor
Chosen	**INS**	**10**	**41%**	-	-	**0%**	**76%**	-	-	**21%**	**64%**	-	**2%**
Chosen	**INS**	**16**	**22%**	-	-	**74%**	**47%**	-	-	**21%**	**16%**	-	**14%**
Chosen	**OCR**	**10**	**50%**	-	-	**0**	**18%**	-	-	**63%**	**70%**	-	**2%**
Chosen	OCR	16	46%	-	-	0%	0%	-	-	48%	74%	-	0%
Em	**INS**	**10**	**18%**	**0%**	-	**0%**	**80%**	**56%**	**13%**	-	**0%**	**0%**	**16%**
Em	**INS**	**16**	**17%**	**0%**	-	**60%**	**60%**	**31%**	**12%**	-	**4%**	**17%**	**14%**
Em	OCR	10	26%	0%	-	0%	0%	0%	0%	-	100%	0%	0%
Em	OCR	16	36%	0%	-	0%	0%	0%	0%	-	100%	0%	0%
M.FpA	INS	10	59%	0%	0%	0%	18%	-	-	93%	-	-	-
M.FpA	INS	16	59%	0%	0%	0%	29%	-	-	89%	-	-	-
M.FpA	OCR	10	61%	0%	0%	0%	22%	-	-	95%	-	-	-
M.FpA	OCR	16	61%	0%	0%	0%	20%	-	-	94% -	-	-	-

Confusion Matrix of the *Chosen actions*
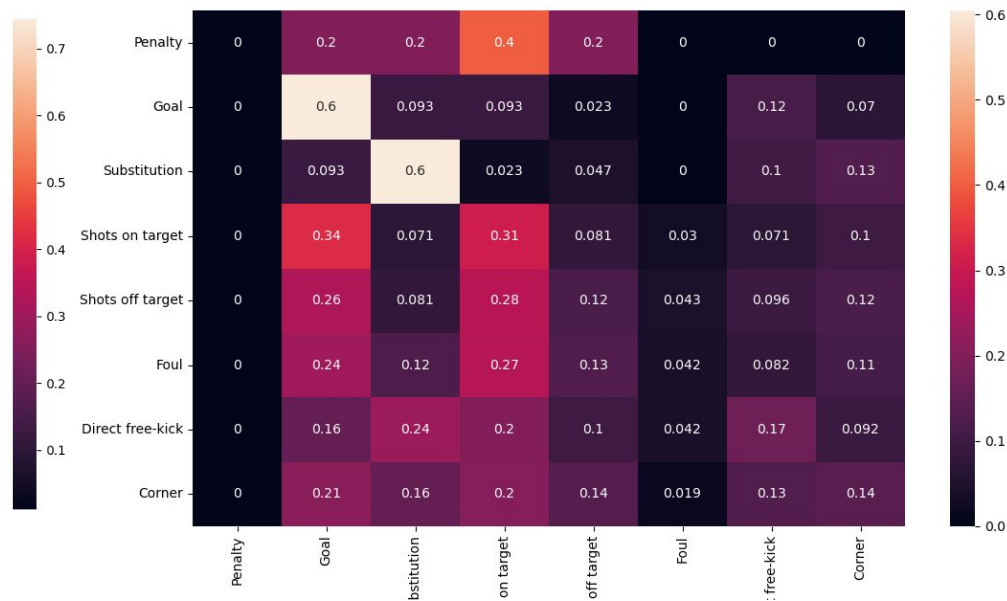
Confusion Matrix of the *Emotion actions*

Conclusions and Future Work

- INS and Time window of 16 seconds.

- Initial approach for understanding football games based on emotions, as we have completed the main objective of detecting football actions using face emotion features.

- Combine our model with a model that use audio emotion features.

- Implement cameras that focus on the public all the time.

- Considering fan's and player's team.

Thank you for your attention
Questions?