Automatic Long-Term Deception Detection in Group Interaction Videos

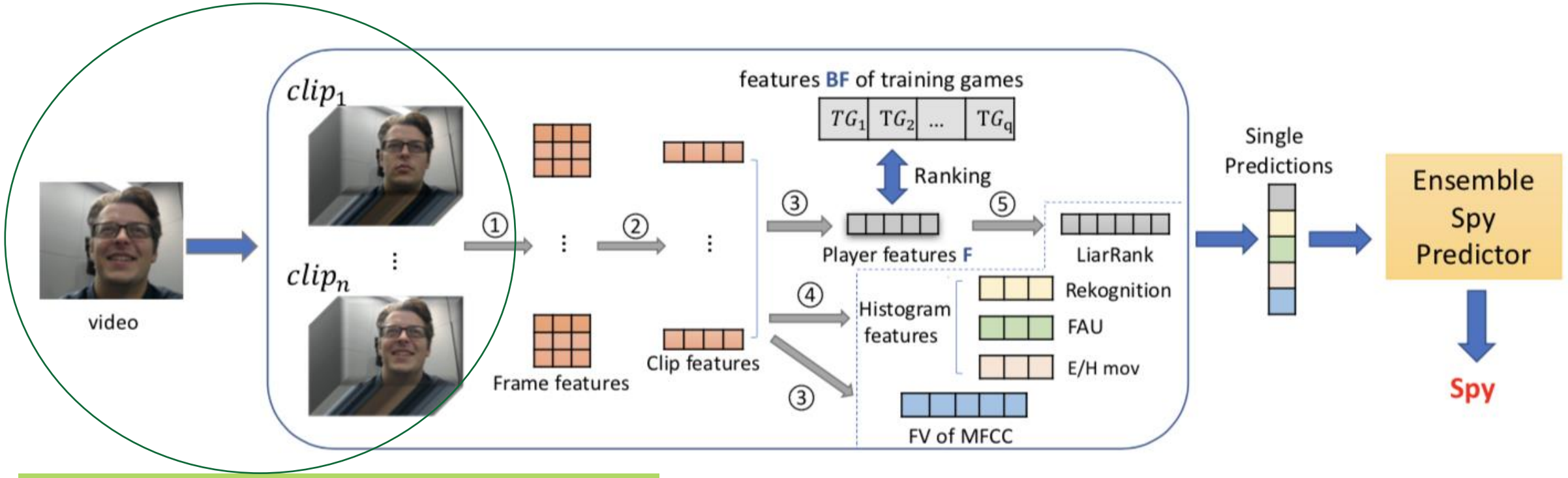
Chongyang Bai, Dartmouth College
Maksim Bolonkin, Dartmouth College
Judee Burgon, University of Arizona
Chao Chen, Dartmouth College
Norah Dunbar, UC Santa Babara
Bharat Singh, University of Maryland
Zhe Wu, University of Maryland
V.S. Subrahmanian, Dartmouth College

IEEE International Conference on Multimedia and Expo (ICME) 2019

Introduction

- A fully automated system (LiarOrNot) for predicting long-term deception in videos
- A class of histogram-based features
- A novel “meta-feature” called LiarRank that builds on the basic features
- An ensemble based prediction model
- Achieves an **AUC of 0.705** in predicting the role of a player in the game
- AUC for human prediction is 0.583

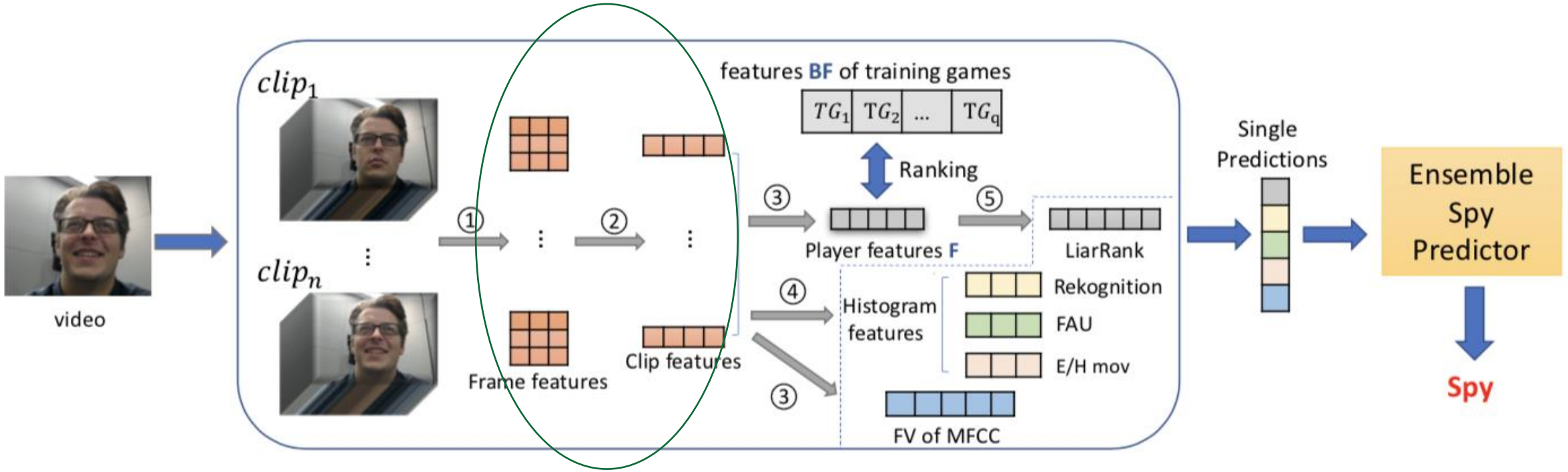
Architecture



Step 1: uniformly sample 10-second clips in every 30 seconds.

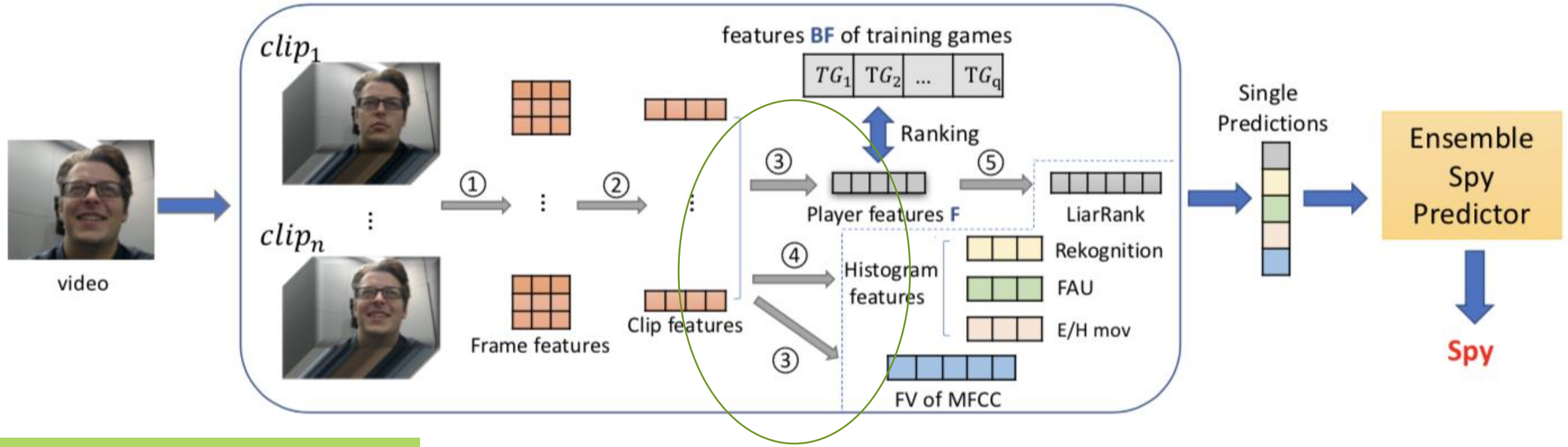
To resolve the challenge of long videos

Architecture



Step 2: Extract visual and audio features for frame and clips
1. VGG Face
2. Facial Action Units
3. Emotions (from Amazon Rekognition)
4. Eye/Head Movements
5. Mel-Frequency Cepstral Coefficients (MFCC)

Architecture

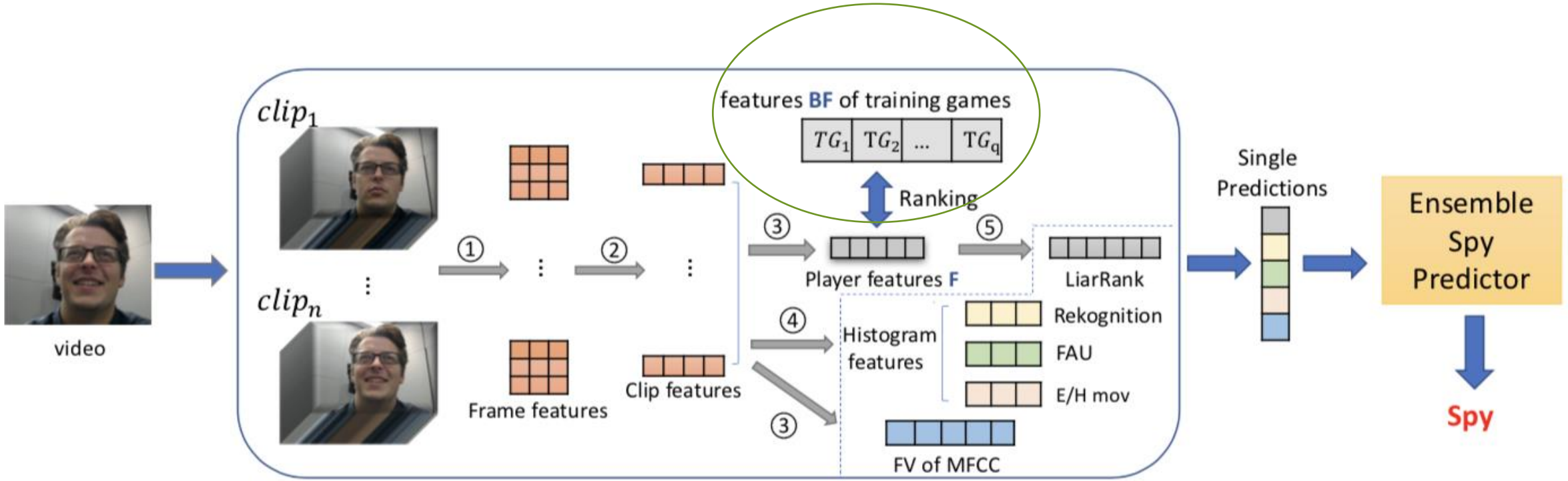


Step 3: Feature aggregation
1. Fisher Vector
2. Histogram

Different games have different number of clips and frames, so their feature vectors may be of different lengths.

We use these 2 aggregation methods to normalize these to a single length feature for each player.

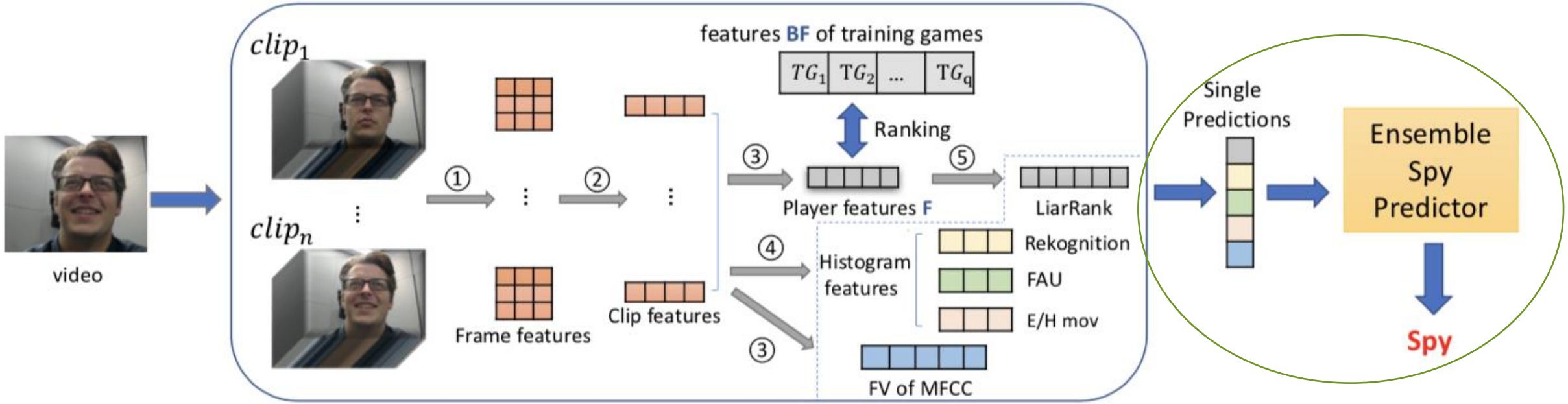
Architecture



Step 4: LiarRank meta features

Capture the game-level information for a player comparing to all games in the training set

Architecture

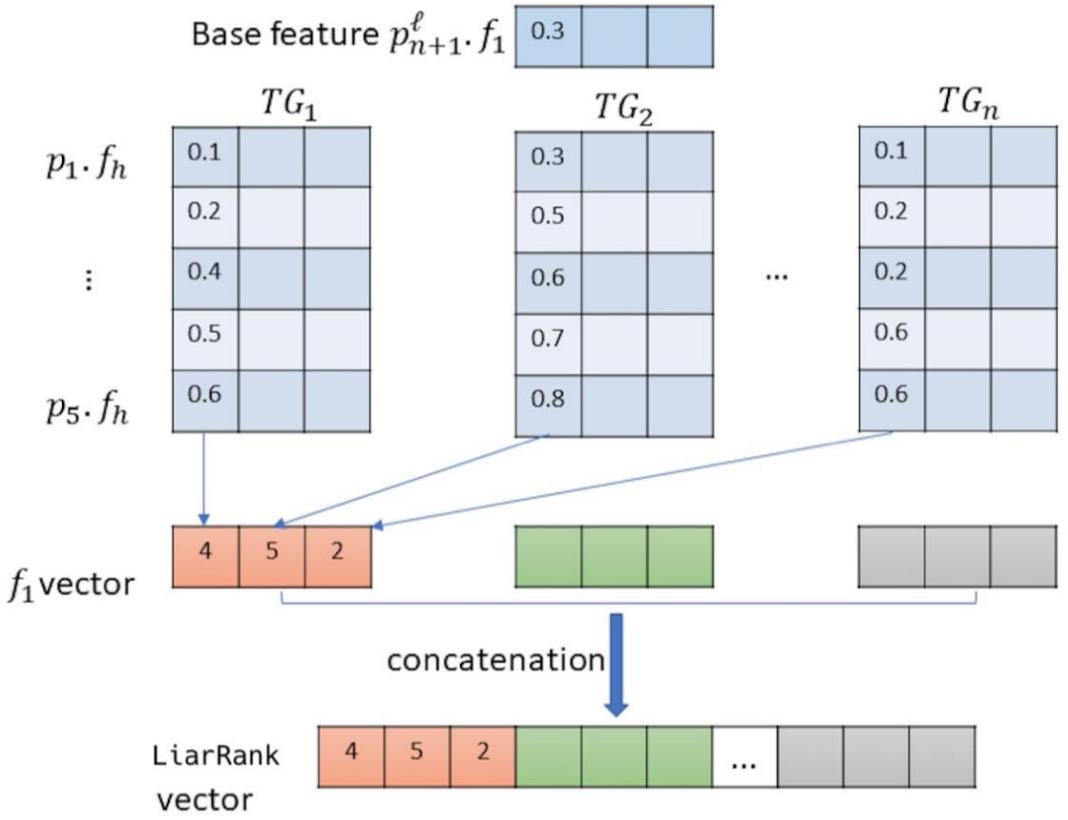


Step 5: Ensemble prediction

Optimize weights of 5 predictors (each from a kind of features)
Final prediction is the weighted sum of the 5 predictors.

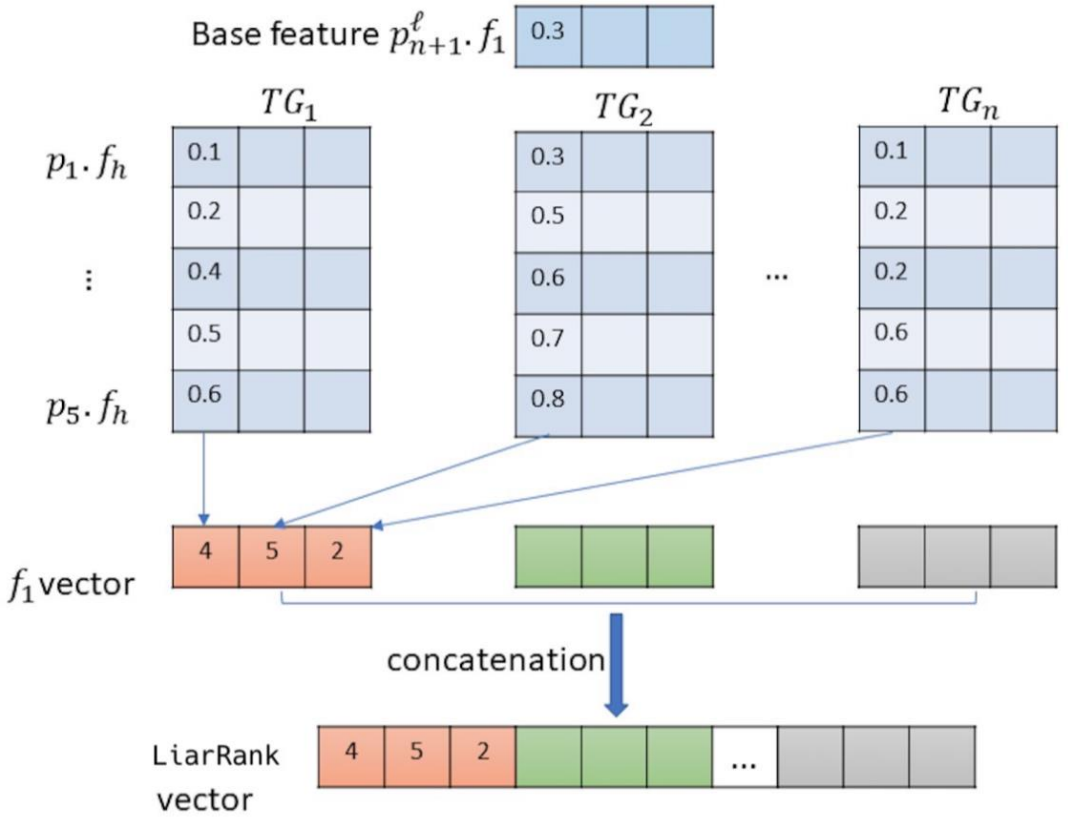
LiarRank meta-feature

- In the training data, we know who are spies
- n games in training set denoted by TG
- Given a player p_{n+1} in a clip, pretend he was in all training games and compare each of his features with those of the villagers and spies.
- LiarRank of p_{n+1} is the rank of a base feature f's value in a game
- Resulting to |F| * |G| features for each player
  - |F| is the dimension of basic features
  - |G| is the number of games in training set



LiarRank meta-feature

- Build upon any base feature
- Example on the right:
  - 3 games TG_1, TG_2, TG_3 in training set
  - 3 base features f_1, f_2, f_3
  - For f_1 of a given player, its rank is 4, 5, 2 in the three games in training set.
  - Also generate ranks for f_2 (green) and f_3 (gray)



Result: Single feature models (Fisher vectors)

Features	RF	L-SVM	NB	LR	KNN
Average VGG Face (baseline)	0.516	0.533	0.549	0.546	0.50
VGG Face clip-level voting	0.503	0.520	0.550	0.527	0.479
FV of VGG Face	0.468	0.573	0.502	0.584	0.502
FV of VGG Face + FS	0.506	0.470	0.491	0.467	0.522
LiarRank of FV of VGG Face + FS	0.639	0.647	**0.663**	0.652	0.603
FV of MFCC frame-level	0.606	0.395	0.56	0.608	0.579
FV of MFCC clip-level	0.586	0.441	0.533	0.579	0.595

LiarRank meta feature boosts the performance of VGG Face + Fisher Vector
LiarRank is robust across all classifiers

Result: Single feature models (Histogram vectors)

Amazon Rekognition					
Frame hist.		Clip hist.		Combined	
Disgusted, Surprised	0.630	Smile, Angry, Disgusted	0.634	Smile, Angry, Disgusted	**0.676**
Surprised	0.622	Smile , Angry	0.623	Smile, Disgusted	0.647
Calm	0.622	Smile, Disgusted, Calm	0.618	Angry	0.638
All features	0.557	All features	0.544	All features	0.563
Facial Action Units					
Frame hist.		Clip hist.		Combined	
AU07+AU10+AU12	**0.621**	AU06+AU14	0.609	AU07+AU09+AU10	**0.621**
AU12+AU23+AU25	0.614	AU07+AU09+AU10	0.606	AU07+AU10+AU23	0.617
AU09+AU10+AU12	0.612	AU07+AU14+AU45	0.603	AU12+AU25	0.611
All features	0.592	All features	0.577	All features	0.608
Eye/Head movement					
Frame hist.		Clip hist.		Combined	
3+8	0.632	1+6+8	**0.671**	1+3+4+5+6+8	0.643
3	0.624	1+6	0.642	1+3+5+8	0.627
3+7	0.615	1+3+6+8	0.636	1+3+5+6+8	0.625
All features	0.591	All features	0.560	All features	0.618

For expression features, the combination of Smile, Angry and Disgusted gave the highest AUC: 0.676

For Facial Action Units, the combination of AU07(Lid tightener), AU09(Nose wrinkler) and AU10(Upper lip raiser) gave the highest AUC: 0.621

For Eye/Head movements, the combination of horizontal eye movements, and x, z head movements gave the highest AUC: 0.671

Result: Ensemble model

- **Ensemble: 0.705 AUC**
- **Ablation test**

Removed feature	AUC
MFCC	0.703
E/H Movement	0.703
FAUs	0.702
Amazon Rek.	**0.688**
LiarRank	**0.688**

Emotion features and LiarRank are the most important features in this task

Demo

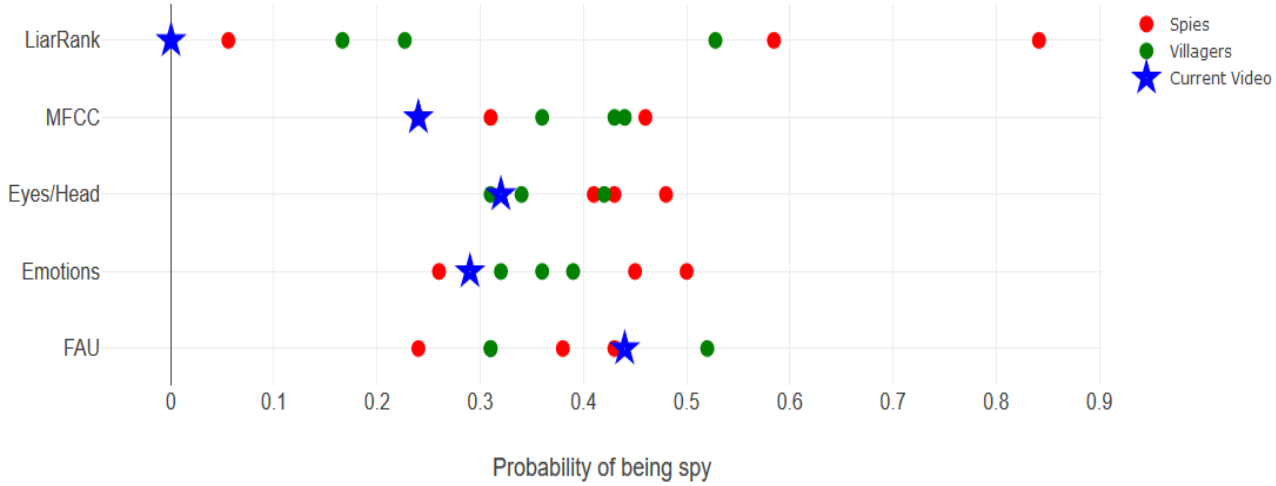
Demo available at
https://cs.dartmouth.edu/dsail/demos/liar-or-not

Demo

Human annotators answers:
Worker 1: SPY
Worker 2: SPY
Worker 3: SPY

LiarOrNot answer:
VILLAGER

Ground truth:
VILLAGER



Demo available at
https://cs.dartmouth.edu/dsail/demos/liar-or-not