High-level Features for Multimodal Deception Detection in Videos

Rodrigo Rill-García
Hugo Jair Escalante, Luis Villaseñor-Pineda, Verónica Reyes-Meza

General Objective

To develop a multimodal information fusion method, inspired by classifier ensemble techniques, for deception detection in videos using high-level features

Extracted Features

Modality	Visual (per frame)	Accoustic (per frame)	Textual (per video)
Views	AU Int	Voice	Char 1-grams
	AU Pres	Glottal Flow	Char 2-grams
	Eye LM	MCEP	Char 3-grams
	Facial LM	HMPDM	Char 4-grams
	Gaze	HMPDD	POS 1-grams
	Head		POS 2-grams
			POS 3-grams
			POS 1-grams
			BoW
			LIWC
			Syntax Info

Figure 1. The different views extracted for each of the 3 proposed modalities. For attributes extracted per frame, fixed size vectors were created using functional statistics.

Proposed Methods

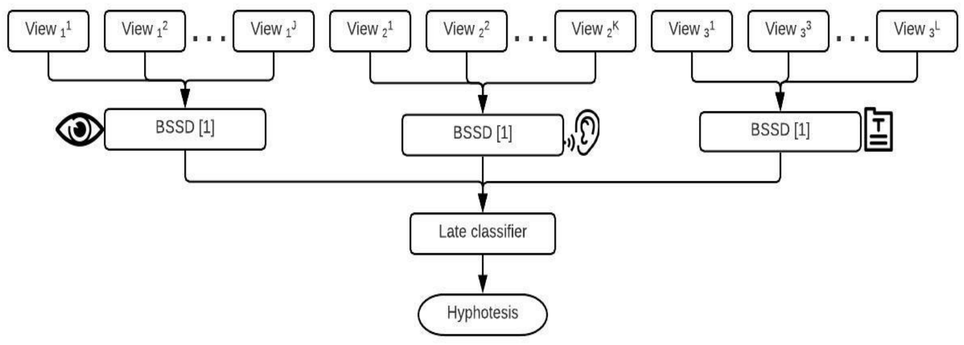

Figure 4. Block diagram of Hierarchical Boosting with Shared Sampling Distribution.

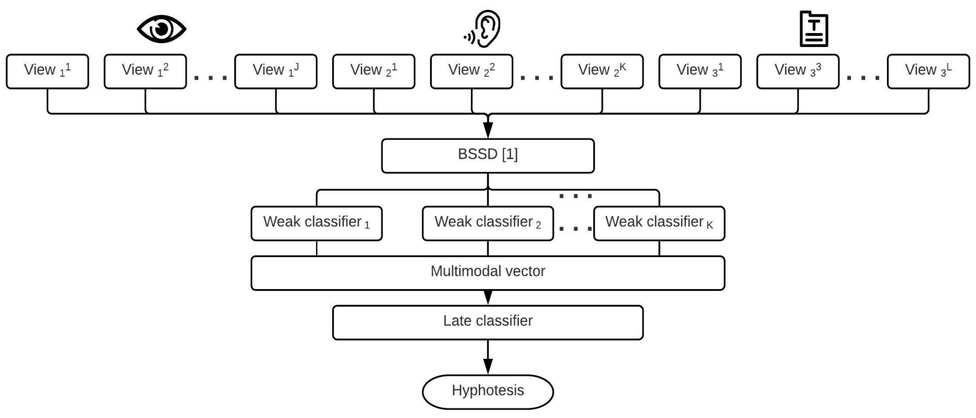

Figure 5. Block diagram of Stacked Boosting with Shared Sampling Distribution.

Databases

Database	Court Trial	Abortion/Friend Spanish
Deceptive/Truthful	61/60	22/21
Subjects	60	12

Table 1. Summary of the databases used.

Figure 2. Examples of Spanish videos.

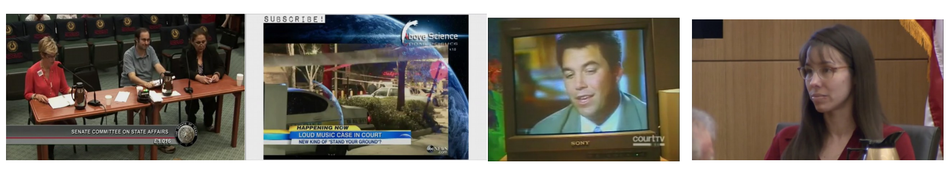

Figure 3. Examples of court videos [2].

Single Views

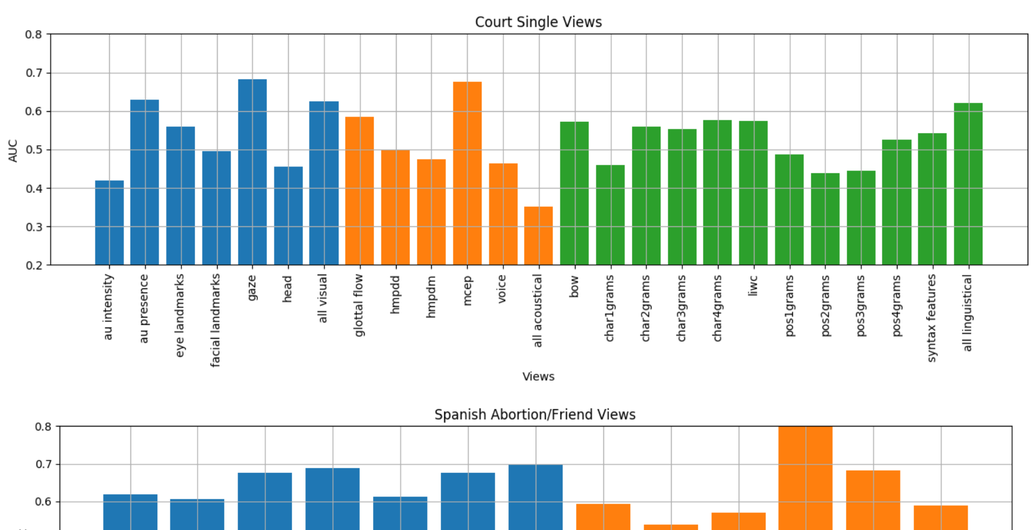

Figure 6. Comparison of single-view performances in the court (top) and the Spanish (bottom) databases. The best single views are gaze direction (0.683) for court and MCEP (0.856 -out of upper limit-) for Spanish.

Multimodal Fusions

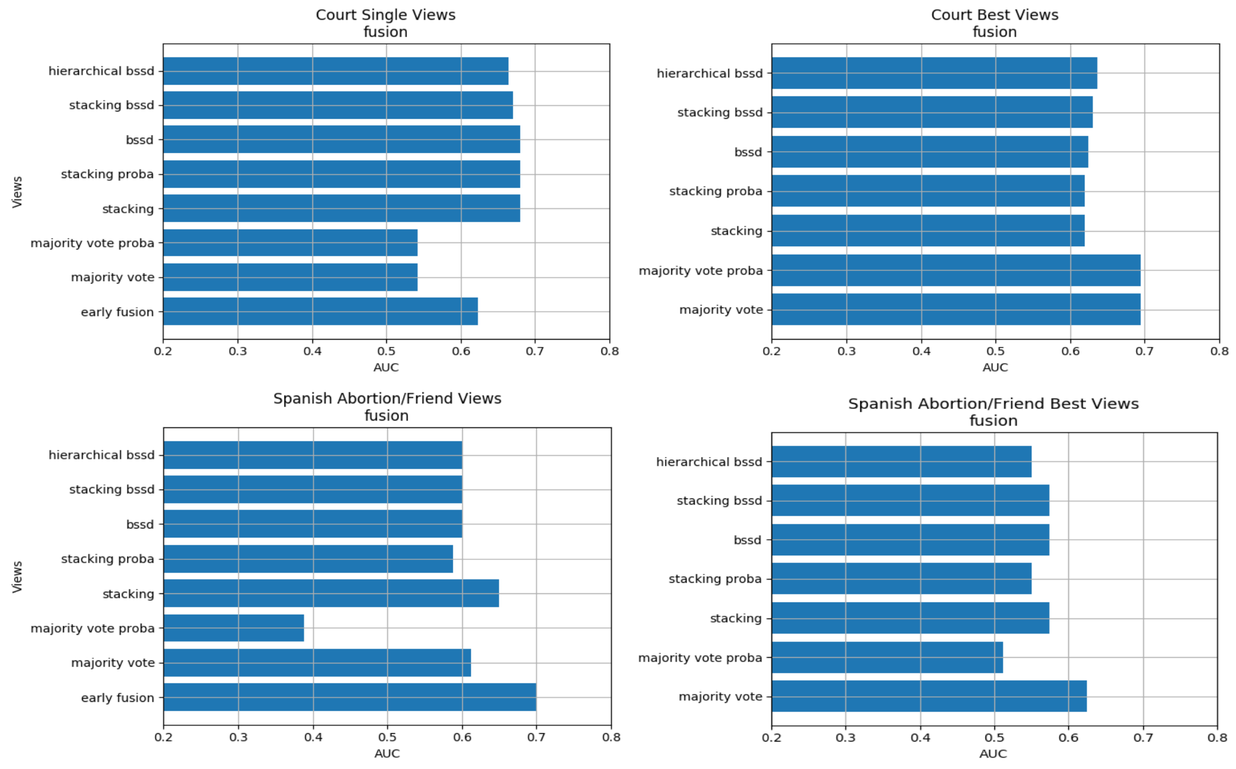

Figure 7. Comparison of fusion methods using all the available views (left) and the best scored ones (right); top graphs are from the court database, while the bottom ones are from the Spanish one.

Research Questions Answered So Far [3]

✓ ***Is it possible to automatically extract high-level features that are useful for automatic deception detection?***

Yes, features such as ***gaze direction, MCEP, eye landmarks and glottal flow*** were useful for deception detection in two different datasets

✓ ***How such features should be analyzed in order to deal with different length speeches?***

Functional statistics seem to be able to capture cues for deception on different length videos

✓ ***Is there complementarity between the features that can can be extracted within and across modalities?***

- ***CFD shows a complementarity*** between the predictions done by different features sets
- ***MPA suggests a possible improvement*** by fusing such predictions

» ***What is a proper fusion method to take advantage of the strengths of each feature set?***

- ***Methods based on multiples views*** seem to take advantage of the multimodal diversity
- ***Late fusion approaches*** tend to work better than early concatenation
- Proposed ***boosting methods*** are competitive with traditional ones with the advantage of ***automatic feature set selection and weighting*** for making predictions

Future Work

➔ To explore LSTM networks for temporal analysis of features

➔ To use boosting methods with tuned hyperparameters per view

➔ To study NN approaches preserving high-level features

➔ To expand the Spanish dataset

References

[1] Barbu, Costin, Jing Peng, and Guna Seetharaman (2010). "Boosting information fusion". In: 2010 13th International Conference on Information Fusion. IEEE, pp. 1–8.
[2] Pérez-Rosas, Verónica et al. (2015). "Deception detection using real-life trial data". In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, pp. 59–66.
[3] Rill-García, Rodrigo et al. (2019). "High-level features for multimodal deception detection in videos". In: Chalearn Looking at People series Face Spoofing Attack Workshop and Challenge at CVPR2019

Coordinación de Ciencias Computacionales
CTBC Centro Tlaxcala de Biología de la Conducta
Instituto Nacional de Astrofísica Óptica y Electrónica, Luis Enrique Erro # 1, Tonantzintla, Puebla, México
Universidad Autónoma de Tlaxcala, Av. Universidad #1, Tlaxcala, Tlaxcala, Mexico
CONACYT