High-level Features for Multimodal Deception Detection in Videos

Rodrigo Rill-García*¹
Hugo Jair Escalante¹,², Luis Villaseñor-Pineda¹
Verónica Reyes-Meza³

*rodrigo.rill@{inaoep.mx, hotmail.com}
1. INAOE, Mexico
2. CINVESTAV, Mexico
3. UAT, Mexico
Motivation

• An “optimal” decision can be harmful if it is based on inaccurate (or wrong) data

• Purposely spreading inaccurate/wrong information is a way to mislead people
  – Doing so for personal gain is the definition of deceiving
Problem Description

• Deception detection is a **hard task** for humans
  - Untrained people have an average accuracy ~54% [1]

• Research supports that there is a **difference** in the way **liars** communicate in contrast with **truth tellers**
  - Furthermore, such difference **can be pointed out using Machine Learning**
Problem Description (2)

- There are many available sources of **cues of deception interpretable** by humans
  - Eye movements
  - Facial expressions
  - Voice
  - Speech
  - Etc.

- Recent research suggests **multimodal analysis** can **improve the performance** of analyzing different modalities separately
Objective

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

• “Detecting deceptive behavior via integration of discriminative features from multiple modalities” [2]
  – Physiological features, thermal videos and transcriptions
  – Early fusion
  – Fused non-invasive features surpassed physiological ones

• “Deception detection using real-life trial data” [3]
  – Videos (image) and transcriptions
  – Early fusion
  – Best performance with fused features

• “Deception detection in videos” [4]
  – Videos (image and audio) and transcriptions
  – Late fusion
  – Best performance with fused features

• “Toward End-to-End Deception Detection in Videos” [5]
  – Videos (image and audio)
  – Early fusion
  – Best performance with fused features

*No focus on multimodal fusion strategies*
Datasets

<table>
<thead>
<tr>
<th>Database</th>
<th>Court Trial</th>
<th>Abortion/Friend Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deceptive/Truthful</td>
<td>61/60</td>
<td>22/21</td>
</tr>
<tr>
<td>Subjects</td>
<td>60</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 1. Summary of the databases used.

Figure 1. Examples of Spanish videos.

Figure 2. Examples of court videos [3].

Rodrigo Rill-García, INAOE, Mexico
Feature Extraction

**Figure 3.** The different views extracted for each of the 3 proposed modalities.

<table>
<thead>
<tr>
<th>Modality</th>
<th>Visual</th>
<th>Acoustic</th>
<th>Textual</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU Int</td>
<td>Voice</td>
<td>Char 1-grams</td>
<td></td>
</tr>
<tr>
<td>AU Pres</td>
<td>Glottal Flow</td>
<td>Char 2-grams</td>
<td></td>
</tr>
<tr>
<td>Eye LM</td>
<td>MCEP</td>
<td>Char 3-grams</td>
<td></td>
</tr>
<tr>
<td>Facial LM</td>
<td>HMPPM</td>
<td>Char 4-grams</td>
<td></td>
</tr>
<tr>
<td>Gaze</td>
<td>HMPDD</td>
<td>POS 1-grams</td>
<td></td>
</tr>
<tr>
<td>Head</td>
<td>POS 2-grams</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>POS 3-grams</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>POS 1-grams</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BoW</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LIWC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Syntax Info</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4.** Creation of a fixed size vector from a number-of-frames-dependent matrix.

- Maximum
- Minimum
- Median
- Mean
- Standard deviation
- Variance
- Kurtosis
- Skewness
- 25th percentile
- 50th percentile
- 75th percentile

Fixed Length: $11^*N$

* OpenFace
** COVAREP
*** IBM Watson ASR, Google SyntaxNet, Python NLTK

Rodrigo Rill-García, INAOE, Mexico
Experimental settings

- **N feature sets** (views) are extracted per video
  - Textual modality is not extracted for Spanish
    - Lack of a Mexican Spanish ASR system

- Metric: *AUC ROC* of the *Deceptive* class
  - 10-folds cross-validation
    - *No subject seen in training* is contained *in the validation* set
Single views

- Court (Sklearn, LinearSVC)

**Figure 5.** Results for single views/modalities in the court database.

Rodrigo Rill-García, INAOE, Mexico
Single Views (2)

- Spanish (Sklearn, SVC: kernel=poly, C=0.01)

**Figure 6.** Results for single views/modalities in the Spanish database.
Complementarity

**Figure 7.** Complementarity measures for the court database.

There is diversity in the errors committed by each view.

The correct predictions from different views predict the whole datasets.

**Figure 8.** Complementarity measures for the Spanish database.

Rodrigo Rill-García, INAOE, Mexico
Proposed Methods (2)

**Algorithm 1: Boosting With Shared Sampling Distribution (BSSD)** [5]

1. Input: $z_0^j = \{x_i^j, y_i\}_{i=1}^n, j = 1, \cdots, M$.
2. Initialization: $W_1 = \{w_1(i) = \frac{1}{n}\}_{i=1}^n$.
3. For $k = 1$ to $k_{max}$
   - (a) Sample $z_k^j$ from $z_0^j$ using the distribution $W_k$.
   - (b) Compute hypothesis $h_k^j$ from $z_k^j$ for each view $j$.
   - (c) Calculate error $\epsilon_k^j$: $\epsilon_k^j = P_{i \sim W_k}[h_k^j(x_i^j) \neq y_i]$.
   - (d) If for each view: $\{\epsilon_k^j\}_{j=1}^M \leq 0.5$, select $h_k^*$ corresponding to $\epsilon_k^* = \min_j \{\epsilon_k^j\}$.
   - (e) Calculate $\alpha_k^* = \frac{1}{2} \ln \left( \frac{1 - \epsilon_k^*}{\epsilon_k^*} \right)$.
   - (f) Update $w_{k+1}(i) = \frac{w_k(i)}{Z_k} \times e^{-h_k^*(x_i^*)y_i\alpha_k^*}$, where $Z_k^*$ is a normalizing factor.
4. Output: $F(x) = \sum_{k=1}^{k_{max}} \alpha_k^* h_k^*(x^*)$.
5. Final hypothesis: $H(x) = \text{sign}(F(x))$.

---

**Figure 9.** Block diagram of Hierarchical Boosting with Shared Sampling Distribution.

**Figure 10.** Block diagram of Stacked Boosting with Shared Sampling Distribution.
Fusion Results Court

Figure 11. Results of fusion methods using all the views (left) and the best two views per modality (right) from the court database.

Best view: Gaze direction (0.683)
Early fusion: 0.623
Figure 12. Results of fusion methods using all the views (left) and the best two views per modality (right) from the Spanish database.

**Best view:** MCPE (0.856)

**Early fusion:** 0.700
Conclusions

- Despite language, context and topic differences, there are **views useful for deception detection** in both datasets
  - *Action units, eye landmarks, gaze direction* (visual)
  - *MCEP, glottal flow* (acoustical)

- **Fundamental frequency and voiced/unvoiced intervals** seem useful to **detect deception on uninterrupted speech**

- Complementarity analysis suggest it is *useful to fuse* features to improve performance
  - Fusion is *not trivial*
  - *Alternatives to concatenating* the multimodal features can improve the performance of a simple early fusion
Future work

- To explore **LSTM** networks for temporal analysis of features

- To use *boosting methods with tuned hyperparameters* per view

- To study pure **NN** approaches preserving high-level features

- To expand the **Spanish dataset**
References


