High-level Features for Multimodal Deception Detection in Videos

Rodrigo Rill-García[1], Hugo Jair Escalante[1,2,3], Luis Villaseñor-Pineda[1], Verónica Reyes-Meza[4]
[1] Instituto Nacional de Astrofísica, Óptica y Electrónica, Tonanzintla, Puebla, Mexico
[2] CINVESTAV, Mexico, [3] ChaLearn, CA, USA
[4] Centro Tlaxcala de Biología de la Conducta, Universidad Autónoma de Tlaxcala, Mexico
rodrigo.rill@inaoep.mx

Abstract

Deception (the action of deliberately cause someone to believe something that is not true) can have many different repercussions in daily life. However, deception detection is an inherently complex task for humans. Due to this, not only there is uncertainty on which features should be used as cues for automatic deception detection, but labeled data is scarce. In this paper, we explore typical features that can be extracted from videos for affective computing and study their performance for deception detection in videos. Additionally, we perform a study of different multimodal fusion methods meant to improve the results obtained by using the different sets of extracted features separately, including a novel set of methods based on boosting. For this study, high-level features are extracted with open automatic tools for the visual, acoustical and textual modalities, respectively. Experiments are conducted using a real-life trial dataset for deception detection, as well as a novel Mexican deception detection dataset using Spanish as the spoken language.

1. Introduction

Decision making is a process that requires the analysis of available data. However, an "optimal" decision can be harmful if such data is inaccurate -not to say strictly wrong. Spreading inaccurate or wrong information purposely is a way to mislead people's decisions for our own convenience. According to the Oxford dictionary, that is the action of deceiving: "deliberately cause (someone) to believe something that is not true, especially for personal gain". Job interviews, court trials, police investigations... there are many cases where believing in someone who is actually lying can imply severe consequences.

Although deception detection is a hard task for ordinary people, previous research [7, 13, 15, 14] supports a well-known assumption that a difference exists in the way liars communicate in contrast with truth tellers. Particularly, evidence suggests that such difference can be pointed out using machine learning.

Furthermore, there are many available sources of cues for deception: eye movements, facial expressions, voice, speech, etc. Recent research [2, 1, 17] points out that multimodal analysis of videos is useful to achieve better results in the deception detection task, rather than using different modalities independently such as visual cues, thermal images, voice analysis or text analysis.

Inspired by such evidence, this work aims to explore high-level features, extracted from different modalities, that can be interpreted by humans while being useful for the automatic detection of deception in videos. Furthermore, a study is conducted on diverse methods inspired by classifier ensembles to fuse together such features (multimodal fusion), including a set of novel methods based on boosting to deal with data from different media (multimedia fusion). Two datasets are used for these experiments: a real-life trial dataset for deception detection as well as a novel Mexican deception detection dataset using Spanish as the spoken language.

Summarizing, this paper aims to: 1) present a study on high-level (interpretable by humans) feature sets that can be automatically extracted from videos for the deception detection task; 2) analyze the complementarity between such features to provide evidence on the benefits that could be obtained from fusing them; 3) present a study on first attempts to perform such fusion by using methods inspired in classifier ensembles; 4) perform a comparison for both single feature sets and fusions on two datasets with different language and contexts, including a novel Mexican database.

2. Related Work

Deception detection from videos is of particular interest as a non-invasive method, unlike traditional ones like the polygraph which are based on physiological data. Typical non-invasive sources of information include RGB videos, thermal videos, audio recordings and speech transcripts.

Abouelenien et al. [1] presented a database consisting of both physiological features (heart rate, blood volume pulse, respiration rate, skin conductance) and thermal videos, as well as transcriptions from videos. Thermal images are analyzed by face regions, while traditional linguistic features such as POS tags, unigrams, LIWC embeddings, etc. are extracted from transcriptions. They tested multiple modal combinations (early fusion approach) using decision trees, concluding that following a multimodal approach outperformed relying solely on single modalities. Furthermore, the fusion of features extracted from videos outperformed the results obtained by physiological features.

However, the above mentioned database was constructed by the cooperation of test subjects under controlled circumstances (e.g. participants may not really be motivated to lie). For studying deception in a more real context, Pérez-Rosas et al. [14] presented a novel dataset of real court trial videos. A multimodal approach is used again consisting on unigrams and bigrams from transcriptions and manual annotation of facial displays and hand gestures. Results obtained from individual features are compared to those from early fusions, reaching the highest score when using all the modalities together.

An automated deception detection system trained with this dataset is presented by Wu *et al.* [17]. The approach is multimodal again extracting a new modality, using features extracted from transcriptions, audio stream and video. Facial gestures (treated as micro-expressions) are used again, extracted by a trained classifier; additionally, video sequences are analyzed employing IDT (Improved Dense Trajectories). MFCC (Mel-frequency Cepstral Coefficients) are extracted and encoded from the audio modality. Finally, transcriptions are analyzed by using Glove (Global Vectors for Word Representation). As the aforementioned works, they performed experiments on single features as well as their combinations using a simple late fusion approach, reaching the best results when combining all the modalities.

Exploring an end-to-end framework, Karimi et al. [10] used a Deep Learning approach for automatic feature extraction from the video and audio channel, respectively. Unlike previous works, these are low level features; those are used again individually and early fused in the court trial dataset for classification, obtaining the best results once again when modalities are fused.

Despite the variety of features extracted from different modalities in the mentioned works, they all share a common characteristic: a simple fusion approach is used for final classification, that is, features are combined into a single vector before or after training a classifier without caring about the complementarity between features. As far as we are concerned, there are not works exploring alternative multimodal fusion methods (either early or late approaches) for deception detection in videos.

3. Feature Extraction

3.1. Datasets

3.1.1 Real-life Deception Detection Database

Nowadays, this database [14] works as a baseline for deception detection in "real-life" videos: unlike other databases, these videos were not recorded in a controlled ambient developed for the deception detection task (to the best of our knowledge, it is the only public real-life datasate available). It is composed of 121 trial videos, 61 deceptive and 60 truthful, extracted from the web; labeling was done manually based mainly on the court verdicts, posterior exoneration, verification of police reports against declarations, etc. There is a total of 58 different identities, from which each identity has an unbalanced set of deceptive or truthful videos -usually, a person's videos are uniformly from a single class.

3.1.2 Novel Spanish Abortion/Best Friend Database

Not only deception detection video datasets are scarce, but they are usually composed from American people; even if not, they use English as spoken language. Furthermore, most of them are not publicly available due to IRB restrictions. Motivated by this fact, we are working in the development of a public novel dataset composed of Mexican people speaking Spanish. Inspired by the protocol used by Abouelenien et al. [2], the participants are asked to give a 2-3 minutes description about their genuine position towards abortion, followed by another description of a fake posture on the same topic. Similarly, they are asked to talk 2-3 minutes about their best friend, as well as talking 2-3 minutes about a person they can't stand as if that person was their best friend. The database so far is composed of 42 videos, 21 deceptive and 21 truthful, with 11 different identities.

3.2. Data Extraction

This section is devoted to explain the different features extracted from videos, as well as the tools used for it. First, we introduce the two level hierarchy that will be used from now on:

- Modality: a video is a multimedia type of file, composed basically of two types of media (namely images and audio). With this idea, we call modalities to each stream of data that can be extracted independently from a video
- View: inspired by the "view" concept used by Barbu et al. [4], we call view to all different perspectives or feature sets that can be extracted independently from a modality sharing a semantic relationship

As above-mentioned, from videos we can extract straight forward two modalities. However, we can also extract transcriptions from the audio modality to create a third modality: text. In the next subsections, we discuss the way modalities and views were extracted for analysis.

3.2.1 Visual

The videos from the database are fed to OpenFace [3] 2.1.0, a facial behavior analysis toolkit. This toolkit analyses videos at frame level, returning different type of features: facial landmark detection, head pose estimation, facial action unit recognition (both binary presence and intensity) and eye-gaze estimation. These sets of features are used separately as the views for the visual modality. Facial analysis is of particular interest because not only face is usually the most visible body part when talking with someone but it reveals a vast amount of information about the internal state of speakers, including behavior that can be distinguished between lying and truthfullness as stated by Paul Ekman [8]. Particularly, action units (AU) are useful to identify emotions that want to be kept in secret by a person; however, identifying the existence of such hidden emotions can be misleading of deception as stated by Ekman too [6]. Therefore, analyzing AU further than for emotion recognition can be useful for the deception detection task. Additionally, facial landmarks are not only useful for the automatic detection of AU, but to describe faces and facial behavior. Head pose estimation can be used as an additional descriptor of body language, giving an insight of involuntary movements beyond face.

3.2.2 Acoustical

For each video, FFmpeg is used to extract a WAV audio file. Each of these files is fed to a MATLAB script from COVAREP [5], an open-source repository of advanced speech processing algorithms. Unlike the visual modality, analysis can't be done at frame level and must be done among time-windows. For the experiments done in this work, this windows (sample rate) were of 0.01 seconds (default value). For each time-window, the next views were extracted: glottal flow (NAQ, QOQ, H1-H2, HRF, PSP, MDQ, Peak Slope, Rd, Rd confidence, Creky Voice), voice (F0, V/UV), MCEP (MCEP 0-24), HMPDM (HMPDM 0-24) and HMPDD (HMPDD 0-12) [5].

Glottal flow has an important contribution to the suprasegmental characteristics of speech and is known to significantly vary with changes in phonation type, so its parameterisation can be useful in many areas of speech research [5]. Mel-frequency cepstral coefficients (MFCC) have been widely used for Automatic Speech Recognition; however, COVAREP extracts an alternative set of MFCCs which are extracted from the "True Envelope" spectral representation (MCEP) which showed usefulness for emotion recognition [5]. F0 and V/UV are used to study the pitch an rhythm of a person while lying or telling the truth. Harmonic model and phase distortion mean and deviations (HMPDM and HMPDD) have been used before for depression detection in videos, showing its usefulness for an affective computing task hard for humans.

3.2.3 Textual

This modality was only extracted for the court-trial dataset due to the available tools. Although this dataset is distributed with manual transcriptions, we want to explore a fully automated mechanism for deception detection. As such, video transcriptions are extracted automatically using Watson Speech to Text from IBM; results are fairly convincing for the court videos in English, but video transcription was discarded for the Spanish dataset as the results from Watson (using the Spanish model) were unacceptable. Based on the study presented by Rill-García et al. [16], the next views were extracted at video level using the Natural Language Toolkit from Python: character n-grams (from 1 to 4, a view for each), Part-Of-Speech n-grams (from 1 to 4, a view for each) and LIWC dictionary encoding. The POS tags were extracted using Syntaxnet. Additionally, a typical Bag-Of-Words representation with NLTK was added as a view as suggested by the work of Pérez-Rosas et al. [14]. Finally, a simple syntax analysis as done by OpenMM [12] is used as a view too; these "syntax features" have been used previously for sentiment and deception detection [11].

3.2.4 Dealing with Time

This far, the visual and acoustical modalities are dependent on each video's length. As different size vectors can't be used by typical binary classifiers, we need a way to get fixed-size vectors from each video. As done by OpenMM [12], an open-source multimodal feature extraction tool, the final representation from a variable length sequence of features is computed as 11 statistical functionals for each feature at view label (see Fig. 1).

The reason behind this is that we want a numeric description of the behavior of the features along the video keeping as much information as possible: mean behavior, extreme values, variation along the video, etc., thus embedding a single feature into a vector of statistics aiming to describe its behavior into variable-length videos.

4. Experiments and Results

A 10-fold cross-validation is used to evaluate the different experiments. However, as we are exploring the multimodal analysis of deception cues, we want to avoid the clas-

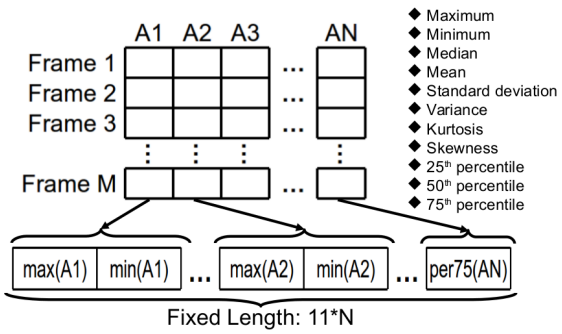

Figure 1. Creation of a fixed size vector from number-of-frames-dependent matrix.

sifiers to degenerate into identity detectors -we don't want to classify a person in the test set as a liar just because all their training examples were deceptive. Therefore, our 10-folds are identity based rather than instance based (no person in the test set was used in the training set as suggested by Wu et al. [17]). As the labels per subject are unbalanced in the court dataset, AUC ROC is used as evaluation metrics. In the case of the Spanish dataset, as labels per subject are balanced, the AUC should be similar to accuracy, so AUC is conserved for convenience.

4.1. Per Modality Study

Before exploring multimodal fusion, we want to explore the effectiveness of each modality independently. Experiments were performed using scikit-learn 0.20.2 with Linear SVC (SVM) as baseline classifier (as it tended to show the best results in preliminary experiments without any hyperparameter tuning). Per modality, studies were performed at two levels: modality and views. That is, we want to evaluate each view separately to gain insight of the performance of "intuitive" features separately, and then evaluate how all this views work together when concatenated as a single feature set. For fair comparison, no hyperparameter optimization is done across experiments.

4.1.1 Visual

Results for the visual modality in the court dataset can be seen in Fig. 2 with blue bars. From the 6 explored views, 3 stand out: binary presence of facial action units, eye landmarks and gaze direction. When combining all 6 views into a single vector, results are still over the 0.5 AUC threshold, but the overall results are worst than picking the best view.

For the Spanish dataset, visual modality results can be seen in Fig. 3. Again, eye landmarks stand out among views; but now, instead of binary presence, intensity of AU stands out, as well as facial landmarks.

The change in best-performing view, as well as the increased relative performance of AU intensity could be explained by the camera distance/angle with respect to the speaker in both datasets, as in the Spanish dataset the camera is close to the person in all videos unlike the court dataset, therefore simplifying the facial analysis task.

However, again, the best single-view result outperforms the concatenated results, which are still above the 0.5 AUC threshold (and above the concatenated results from the court dataset).

4.1.2 Acoustical

For the acoustical modality, results obtained from the court dataset are presented in Fig. 2 using orange bars. MCEP achieve the best overall single-view result, matching the best view from the visual modality (AU presence). However, this is the only view from the acoustical modality to surpass the 0.5 AUC threshold. Furthermore, when concatenating all the acoustical views the result is still above the 0.5 threshold but below the obtained with MCEP.

With respect to the Spanish dataset, results using the acoustical modality are shown in Fig. 3. Again, MCEP outstand among views (suggesting the envelope strategy on MFCC can be useful indeed for deception detection), but a particular view performs better this time: voice, composed solely of fundamental frequencies (F0) and voiced/unvoiced segment binary identification. Unlike the court dataset, in the Spanish one there are near-to-zero utterances from other speakers rather than the analyzed person in all videos; also, the speeches are longer. This suggests deception can be detected simply based on the analysis through time of F0 and pauses in the speech of a single speaker.

However, even if single-view results are better in the Spanish dataset, the acoustical modality as a single vector have a better performance in the court-trial dataset.

4.1.3 Textual

As aforementioned, the textual modality was extracted from the court dataset only. Those results can be found in Fig. 2 using green bars. Particularly, the best results are achieved with Bag-Of-Terms representations, namely: bag of words, bag of char 2-grams, bag of char 3-grams and bag of char 4-grams. Also, LIWC encoding and syntax features achieve similar results. A brief parenthesis is used here to provide a possible explanation on the low performance of POS tags: POS tagging is done based on context (for example, orange can be either adjective or substantive); the automatic transcription done by Watson comprise every utterance in the audio stream, retrieving therefore a mixed text containing sentences/words spoken by all the people in the audio file. With such a text, automatic POS tagging it's a harder task, in a text that is hard to understand itself without hearing the original conversation.

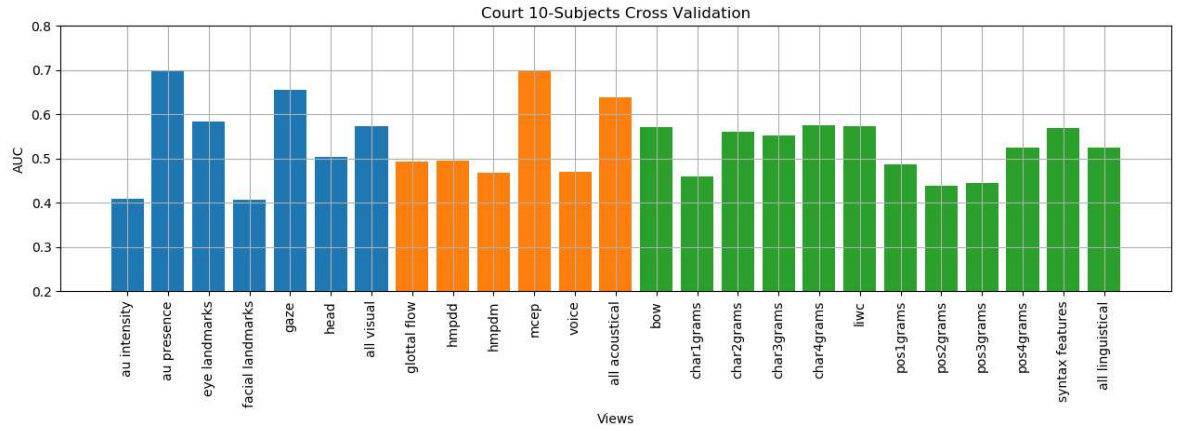

Figure 2. AUC achieved by the different views in the court-trial dataset, including their concatenation (rightmost column).

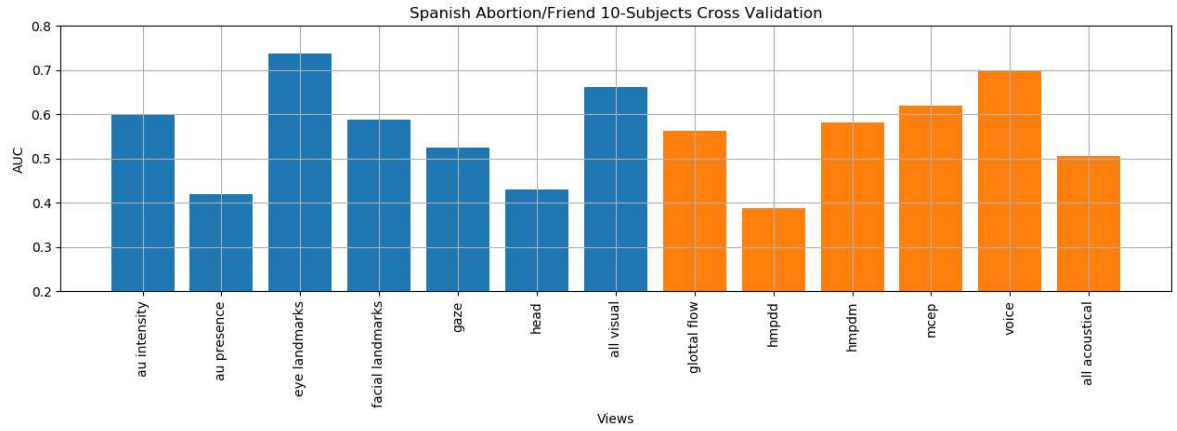

Figure 3. AUC achieved by the different views in the Spanish dataset, including their concatenation (rightmost column).

4.2. Complementarity

Once each modality has been analyzed, we want to find out if it is potentially useful to combine them in order to achieve better results. In order to do this, we analyze the results obtained at instance level to see how complementary they are at both views and modalities levels: that is, even if each type of features has many mistakes, we want them to be wrong at different instances -so that, if we combine them in a proper way, we get better results.

To know the best possible result after fusion, we use the Maximum Possible Accuracy (MPA) metric: at instance level, if any of the views/modalities classified the instance correctly, the instance is considered as correctly classified. This is then an optimistic measure of a perfect fusion.

Also, we want a numeric measure to evaluate how diverse are the errors between views/modalities. For this purpose, we use the Coincident Falure Diversity (CFD) metric, which ranges from 0 (when all views/modalities always make the same label predictions) to 1 (when misclassifications are unique to one view/modality).

As it can be seen (see Fig. 4, 5), not only the CFD is far from 0 both at views and modality levels, but the MPA is greater at views level rather than at modalities level. This suggests there is, in fact, complementarity both at views and modalities level; also, it seems like there are complementarity reasons to split the different modalities into views.

4.3. Baseline Fusion Methods

In order to do multimedia fusion, we took an approach based on classifier ensembles. With this paradigm, we ensemble views/modalities rather than classifiers. As baseline methods, we used two traditional ensemble methods both at views and modalities levels: hard majority votes and stacking. These methods are in the category of late fusion (fusion is done after independent classifications).

Additionally, two early fusion (fusion is done before any classification) approaches were used as baseline too. The

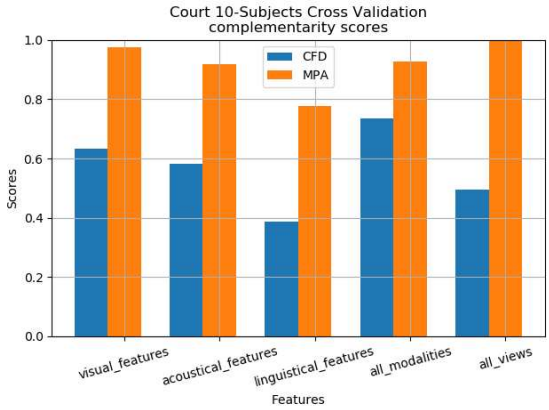

Figure 4. CFD between views and modalities from the court dataset, as well as their MPA.

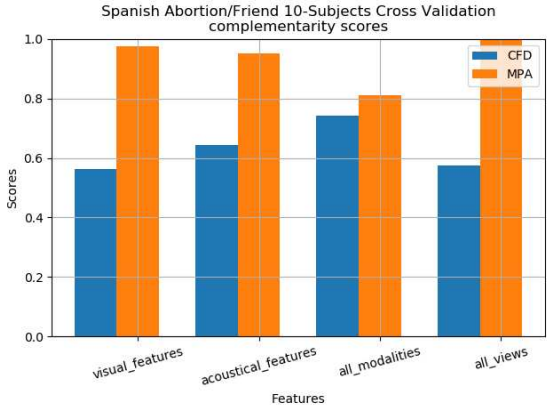

Figure 5. CFD between views and modalities from the Spanish dataset, as well as their MPA.

first one was simply concatenating all the views/modalities into a single vector for classification before performing a classification task. The second one is a temporal "informed" fusion between modalities presented by Morales [11] and used originally for multimodal depression detection in videos.

4.4. Multimodal Boosting Methods

For the boosting strategies, we parted from the Boosting With Shared Sampling Distributions (BSSD) first presented by Barbu et al. [4] for multiple representation fusion based on Adaboost, where weak learners are built at "view" level (our "view" definition was inspired by them) at each iteration. The weak learner with the lower error rate is chosen at each step, and its errors are used to calculate a new probability distribution of the training instances for the next iteration, giving greater weights to the wrongly classified instances. All views share the same sampling distribution, so each weak learner at each iteration gives greater importance to those examples that were "harder" to predict in the previous iterations.

BSSD was originally used with views for visual modality tasks only. We use it in our experiments as well as some variances proposed by us and described next for multimodal deception detection in videos using 50 iterations.

4.4.1 Hierarchical BSSD

Our first approach was extending BSSD with a hierarchical strategy, by using BSSD per modality and then using the label calculated for each modality as a new feature for late fusion. For consistency, the classifier used for late modality fusion is the same used to build weak classifiers. This approach improves the results obtained by BSSD using all the views separately or using modality vectors as views in the court dataset (Fig. 6). In the Spanish database (7), however, it is not the case; there is not a clear advantage from any of the three approaches mentioned in this paragraph.

4.4.2 Nested Cross-validation BSSD

Our hierarchical proposal ensures all the modalities are taken into account for the final decision; however, there is a special case where BSSD is not capable of doing so. BSSD's nature implies most consistent data types (views) dominate over time, as at each iteration it chooses a weak learner built over a single view (the weak learner with the minimum error rate is chosen). However, this error rate is calculated over the training data. Validating within the training data is not a good pointer of the overall performance of a classifier (weak learner, in this case), as it is highly susceptible to over-fitting. In the case where any of the views used for BSSD achieves a perfect score in the training data (probably due to over-fitting), the algorithm will ignore any other view at all iterations.

To explore the benefits of using other validation metrics rather than accuracy in the training set for selecting the best view at each iteration, we modified the BSSD algorithm: for each view, an expected performance is calculated per iteration using a nested weighted 10-fold cross validation over the training data; the weights are normalized per fold from the sample distribution used for the training data at a given iteration. The view with the lowest expected error is used then to build the weak learner corresponding to that iteration. Furthermore, this approach was implemented too doing identity based nested cross-validation as explained in Section 4. Both approaches are reported as "bssd x cv" and "bssd x cv subject", respectively, in Fig. 6, 7. For the court dataset, this validation tends to improve the performance of BSSD when using modalities, but it actually decreases it when using views. In the Spanish dataset, this validation doesn't seem to improve classification at all.

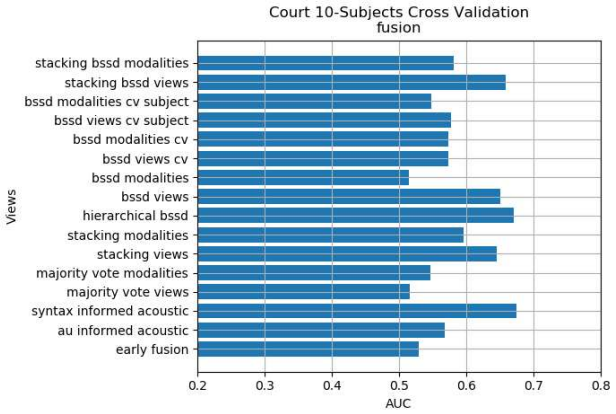

Figure 6. AUC achieved by different fusion methods in the court dataset.

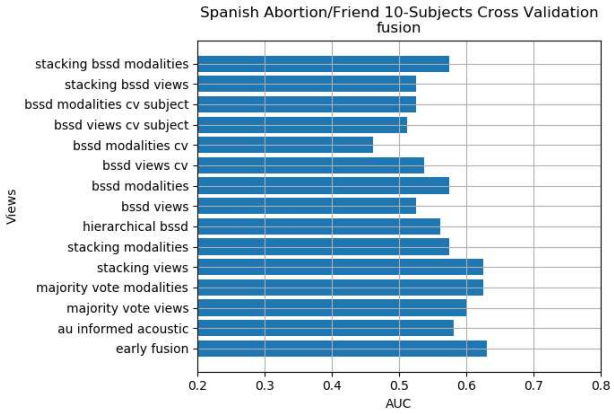

Figure 7. AUC achieved by different fusion methods in the Spanish dataset.

4.4.3 Stacking BSSD

Lastly, Adaboost (and therefore BSSD) classifies an instance with a linear function of the labels predicted for such instance from each weak learner trained. The weights from this linear function are learned by the boosting algorithm as a function of the error rate of each weak learner; however, there might be a benefit from learning these weights outside the boosting algorithm. Hatami and Ebrahimpour [9] try a similar approach, using the weak learners obtained by a boosting algorithm as base classifiers for a staking method, achieving better results than using boosting alone. We use the same approach, using the weak learners generated by BSSD as base classifiers for stacking.

For the court dataset (6), this stacking approach shows a clear improvement at either view or modality level with respect to the original BSSD method. On the other hand, for the Spanish dataset (7), performance of stacking and standard BSSD is very similar.

4.5. Independent Modalities vs Fusion

Among fusion methods, for the court dataset (Fig. 6) stacking stands out among the traditional methods, particularly when stacking views (reaching up to 0.645 AUC). This method, however, is improved with the boosting methods, namely BSSD, stacking BSSD and hierarchical BSSD (0.651, 0.659 and 0.671, respectively). Furthermore, these results were surpassed with the "syntax informed" method proposed by Morales [11] (0.675). These results, however, are slightly below the best single-view ones, corresponding to a tie between MCEP and binary presence of AU (0.699).

Although the fused results are below the best single-view ones, we can see that the boosting fusion methods are robust to the inclusion of features with low performance (many of the included views were below the 0.5 threshold when evaluated solely). This is due to automatic feature set selection during training, which can be easily analyzed by looking at the views selected for each weak learner: the best feature sets are used automatically, using lower performing ones only for specific cases which are hard for the best feature sets. Furthermore, when the hierarchical approach is used, we ensure that all the different modalities extracted from the videos are used.

When it comes to the Spanish dataset (Fig. 7), however, we have a different scenario. Again, stacking views is among the best classic fusion methods (0.625), along with majority vote of modalities and outperformed by a simple early fusion (0.631); again, an "informed" method has a good result (0.600). But this time, traditional methods were not surpassed by the boosting ones, achieving a best result of 0.575 using either normal or stacking BSSD with modalities. For this dataset, the best fused result is considerably below the best single-view one, namely gaze direction (0.769). Given the small number of training instances, this could be explained by the curse of dimensionality.

4.6. LSTM study case

As the best results for fusion were acquired with the court dataset, we wanted to preliminary explore the use of Deep Laarning as an additional baseline for this database. This was motivated by 3 reasons: 1) stacking methods used here have a layer-nature, similar to that of neural networks (NN); 2) the performance of "informed" methods showed good results using a fewer amount of views, 3) DL is a state-of-art technique for many video-related tasks. Putting together those reasons, Long Short-Term Memory (LSTM) networks were used as a way to use NN including the temporal-sequence nature of videos, analyzing features at frame level.

For this experiments, a LSTM layer with 200 hidden units is fed with sequences of frame analysis, the output

is fed to a fully-connected layer with 100 hu and its output is finally fed to a fully-connected layer with a single output. This architecture was tested with the visual and acoustical modalities only, achieving 0.560 and 0.730 AUC, respectively (SVM results were 0.574 and 0.638, respectively). Two fusion approaches were tested then: feeding the network with both modalities concatenated (early fusion) and feeding two different LSTM with each modality and concatenating both outputs to fed the first fully-connected layer (late fusion). For this fusion strategies, 0.665 and 0.610 AUC were reached, respectively. Both results are below the one achieved by hierarchical BSSD with SVM (0.671).

The same experimental setup was used with the Spanish dataset. However, the results were considerably low, reaching an AUC of 0.384 and 0.294 for the visual and acoustical modalities respectively. When attempting early fusion, the network reached 0.475 AUC only, improving the single-modality results but being still below the 0.5 threshold.

5. Conclusions

In this paper we explored high-level features (at "view" and "modality" level) extracted from videos, using open tools, for the deception detection task. Those features where evaluated in two datasets for deception detection (one constructed from real-life court trials from the web, and other constructed from Mexican people speaking in Spanish about a sensible topic and a personal topic).

Despite the cultural, language, context and topic-related differences, there were indeed shared views that showed a tendency as good discriminators of deception independently of the analyzed person, namely: AU, eye landmarks, gaze direction (visual modality), and MCEP (acoustical modality).

For the court dataset, BoW, bags of char n-grams, and the extraction of syntax features seem to be able to detect deception from automatically transcripted texts (even for multispeaker audios). For the Spanish dataset, there's evidence suggesting that F0 and voiced/unvoiced periods are good features to detect deception from a person talking uninterruptedly (at least for 2-3 minute periods).

Additionally, a study of complementarity between the studied features was performed on both datasets, showing evidence on the convenience of approaching the deception detection in videos as a multimodal problem. Given such evidence, we presented preliminary work on multimodal classification inspired by classification ensemble methods, including a novel pair of boosting based methods competitive with a Deep Learning approach such as LSTM.

Future work involves analysis of fusion methods using the most predictive features studied in this paper, as well as the tuning of hyperparameters for classifiers that can exploit such features (including DL architectures). Moreover, we are working in the expansion of Spanish data for deception detection in videos.

References

[1] Mohamed Abouelenien, Verónica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. Detecting deceptive behavior via integration of discriminative features from multiple modalities. *IEEE Transactions on Information Forensics and Security*, 12(5):1042–1055, 2017.

[2] Mohamed Abouelenien, Verónica Pérez-Rosas, Bohan Zhao, Rada Mihalcea, and Mihai Burzo. Gender-based multimodal deception detection. In *Proceedings of the Symposium on Applied Computing*, pages 137–144. ACM, 2017.

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[4] Costin Barbu, Jing Peng, and Guna Seetharaman. Boosting information fusion. In *2010 13th International Conference on Information Fusion*, pages 1–8. IEEE, 2010.

[5] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarepa collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE, 2014.

[6] Paul Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000(1):205–221, 2003.

[7] Paul Ekman. *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition)*. WW Norton & Company, 2009.

[8] Rosenberg Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[9] Nima Hatami and Reza Ebrahimpour. Combining multiple classifiers: diversify with boosting and combining by stacking. *International Journal of Computer Science and Network Security*, 7(1):127–131, 2007.

[10] Hamid Karimi, Jiliang Tang, and Yanen Li. Toward end-to-end deception detection in videos. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1278–1283. IEEE, 2018.

[11] Michelle Renee Morales. Multimodal depression detection: An investigation of features and fusion techniques for automated systems. 2018.

[12] Michelle Renee Morales, Stefan Scherer, and Rivka Levitan. Openmm: An open-source multimodal feature extraction tool. In *INTERSPEECH*, pages 3354–3358, 2017.

[13] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675, 2003.

[14] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International*

Conference on Multimodal Interaction, pages 59–66. ACM, 2015.

[15] Verónica Pérez-Rosas and Rada Mihalcea. Cross-cultural deception detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 440–445, 2014.

[16] Rodrigo Rill-García, Luis Villaseñor-Pineda, Verónica Reyes-Meza, and Hugo Jair Escalante. From text to speech: A multimodal cross-domain approach for deception detection. In *International Conference on Pattern Recognition*, pages 164–177. Springer, 2018.

[17] Zhe Wu, Bharat Singh, Larry S Davis, and VS Subrahmanian. Deception detection in videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.