Move, and I will tell you who you are: Detecting deceptive roles in low-quality data

Nimrod Raiman
University of Amsterdam
Amsterdam, Netherlands
Nimrod@Raiman.nl

Hayley Hung
University of Amsterdam
Amsterdam, Netherlands
H.Hung@uva.nl

Gwenn Englebienne
University of Amsterdam
Amsterdam, Netherlands
G.Englebienne@uva.nl

ABSTRACT

Motion, like speech, provides information about one's emotional state. This work introduces an automated non-verbal audio-visual approach for detecting deceptive roles in multi-party conversations using low resolution video. We show how using simple features extracted from motion and speech improves over speech-only for the detection of deceptive roles. Our results show that deceptive players were recognised with significantly higher precision when video features were used. We improve the classification performance with 22.6% compared to our baseline.

Keywords

multi-party conversation, deception, human behavior

1. INTRODUCTION

Lies and deception are part of everyday life [1]. *"On average we each try to dupe others more then once a day"* Vrij [1, page 11]. Some lies, such as "I'm fine", can be seen as a social lubricant, while others are of course more harmful. *"Lying is such a central characteristic of life that better understanding of it is relevant to almost all human affairs" Ekman [2, page 23].* Despite the quantity and consequences of lies, people are very bad at recognizing them [2]. A possible explanation for this is the ostrich effect [1]: people do not always want to hear the truth.

In this paper we propose a method for detecting deceptive roles in multi–party conversations using audio and low–quality video data. Deception is defined by Vrij [1, page 15] as *"a successful or unsuccessful deliberate attempt, without forewarning, to create in another a belief which the communicator considers to be untrue".* This can be achieved by telling a lie or (partially) concealing the truth. A lie is a specific instance of deception in the form of an untruthful statement. This paper's goal is particularly challenging because a person in a deceptive role is not lying or deceiving all the time. Current approaches for detecting deception are based on either non–verbal audio or video cues. We extend both approaches by combining automatically extracted non–verbal audio and video cues for detecting deceptive roles. Our proposed models are tested on the werewolf corpus [3], a multi-party conversation dataset where players have either a deceptive role (wolf) or a non-deceptive role (villager) (sec.2). The wolves' goal is to deceive the villagers and make them believe they are one of them.

To our best knowledge very little research is done on automated audio-visual analysis of social interaction concerning deception detection [4] and no research has been done on automated detection of deception in multi-party conversations using low-quality video data. This paper introduces a fully automated method for detecting deception with audio and low-quality video data in multi-party conversations. Analysing the content of audio and video recordings of social interaction can be done using three different approaches: the verbal approach [4–6], where the content of a conversation is analysed; the non-verbal audio approach [4–12], where the focus is not on what is being said, but how it is said; and the non-verbal video approach [4, 8–11, 13, 14], where body motions, gestures and facial expressions are analysed.

There is a large range of video features that can help understand social interactions between people. Some are very subtle and detailed (e.g. eye blinking, barely noticeable lip-clenching, eyebrow movement, etc.) [2, 8]. Some are more obvious (e.g. head nod, head shake, arm folding, etc.) [8–11, 13, 14]. In this paper we focus on detecting deceptive roles in audio and low-quality video data by analysing gross body movements. Approaches that work for low video quality and are not based on facial expressions are more widely applicable because of two reasons: first, not all video recording systems (e.g. webcams, mobile phone cameras, surveillance cameras, etc.) have sufficient quality for extracting detailed features such as facial expressions; and second, because in real world situations faces are not always visible for the camera. The maximal size of the faces of the people in the dataset for example is 30 by 30 pixels. Extracting robust detailed facial expression from this is challenging. Also the lighting conditions are not ideal, in many cases the skin colour of the persons in the dataset is very similar to the background. Because of this, detecting skin in the video stream is not trivial.

In this paper we focus on non-verbal cues. Chittaranjan and Hung [7] show that by using automatically extracted non-verbal audio features for detecting deceptive roles in recordings of a multi-party role playing game (RPG), their system's performance is significantly higher than the participants' in the game. Meservy et al. [13] explain an automated

ICMI'11, November 14–18, 2011, Alicante, Spain.

Figure 1: Screenshot of the video recordings

non-verbal video approach for inferring deception in dyadic conversations from a set of features extracted from head and hands movements in a video. Our hypothesis is that since both the voice and the body are a source of information on the current state of mind of a person [1, 2] and because audio and video features contain different information, looking at both will give a more complete picture of the social interactions and the current state of mind of a person.

2. DATA

The "Werewolf" dataset (made by Idiap[1]) used for training and testing the models in this paper is the same as used by Chittaranjan and Hung [7]. The dataset contains 3 hours of conversational data of 4 games played by 2 groups. The groups consisted of 10 and 8 different participants playing the RPG "are you a werewolf?". All participants in the game sat on a chair while they were alive in the game. On the video recordings the upper bodies are fully visible, but the legs are often occluded. A screen shot of the video data is shown in figure 1.

In the simplest setting of this game there is one mayor that leads the game, two of the players have the role of werewolves and the other players are villagers. The game has two phases, night and day. During the night phase the werewolves choose one player which they kill overnight and thus will not wake up in the morning. During the day phase all villagers that are alive and the werewolves in disguise discuss who they want to kill. The villagers try to find out who the werewolves are and it is the werewolves' task to deceive the group to avoid being killed and losing the game. This setting provides a good motivation for the players to lie and conceal their identity. During the game the participants identified with their role, resulting in real emotions. From the dataset only the day phases are used for training and testing the models.

The audio data were recorded using headset microphones, the video data with three frontal cameras. The layout of the recording room is shown in figure 2. The lighting conditions in the recording room were not ideal, the lighting is not equal throughout the room and in many cases the skin colour of the players is very similar to the background colour due to highlights.

3. CUE EXTRACTION

This paper proposes an automated approach for detecting deceptive roles using audio and visual cues in low-quality video data. Since this approach is developed for low-quality data, no attempt is made at detecting detailed features such as micro expressions. Instead, the focus is on gross body motion. Before one's behavior was analysed, speaking activity and motion features were extracted.

[1] http://www.idiap.ch/scientific-research/resources/wolf-corpus

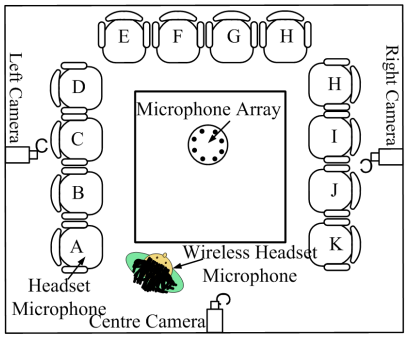


Figure 2: The layout of the recording room

3.1 Speaking activity

We used the method presented by Chittaranjan and Hung [7] to extract the speaker turn segmentation. Output of a voice activity detector (VAD) was synchronised to the corresponding frames of the video. A VAD classifies a segment as speech when the energy in that segment exceeds a certain threshold. The result was a binary vector for each player with ones for the frames where speech was measured and zeros for the frames where this person was silent.

3.2 Motion cues

Each individual's motion was extracted by first locating their body, using the frontal face Haar cascade classifier implementation from OpenCV. Once the faces were located and the number of individuals in the video were known the areas of their bodies was determined. The average width of the extracted body area was empirically set to 140 pixels symmetrically aligned around the face area and the height was set to 250 pixels. The body areas were resized according to their distance to the camera, measured by the size of the chair. Meservy et al. [13] explain in their approach that different areas of and around the body can be informative (e.g. to discriminate between an open and closed posture) concerning motion and location of the hands. Inspired by this approach the extracted area of the body from the video stream is divided in six regions, see figure 3. By analysing each region separately, different clues from the behavior can be captured. The head (**H**) region captures head motion (an increase in head motion can indicate one is being more alert). The region above the shoulders (**AS**) captures motion from self touching in the neck and head area (self touching can indicate anxiety). The left body region (**BL**) and the right body region (**BR**) capture motion involved with a more open posture. Finally the left (**BML**) and right (**BMR**) side of the body middle region capture motion involved with a more closed posture.

We used two different techniques to detect body motion:
Temporal difference. This technique is based on subtracting consecutive frames. It is computationally cheaper but less accurate than full optical flow (described next). The result of subtracting frames from each other is assumed to be the motion between the two recorded frames. The outputs from this method were thresholded. Finally for each region the binary values are summed to give the visual activity (M_r).

$$M_r = \sum_{p \in P_r} f(p), \text{ where } f(p) = \begin{cases} 0 & \text{if } p \geq \theta \\ 1 & \text{if } p < \theta \end{cases},$$

P_r is the set of pixels in region r and θ is the threshold parameter. θ was empirically set to remove noise from areas of uniform texture.

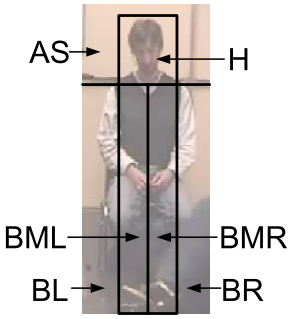


Figure 3: Six regions the body area is divided in

Optical flow. We used the method of Liu [15] to extract the optical flow. This technique allows us to achieve a more accurate mapping of motion between two consecutive frames. The outputs of this algorithm are two flow fields, in horizontal and vertical directions. These flow fields are transformed in two different features; the *motion magnitude* and the *orientation of the motion*. A histogram of the motion magnitudes and orientation is stored per region per frame.

4. DETECTING DECEPTIVE ROLES

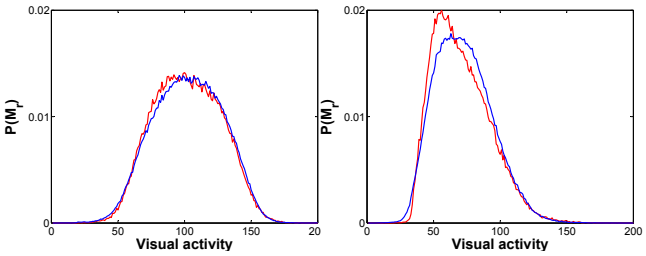
People in a deceptive role talk and move differently from people who are not deceiving [1, 2, 7, 13]. In order to discriminate between deceivers and non-deceivers we apply supervised learning with two learning algorithms: RVM and SVM. Classification is done after each day phase of the game. First the z-scores of the features are calculated over all features of each player. We use z-scores in order to take out user–specific behavior (e.g. visual activity, size of body) by subtracting the mean from each measurement and dividing it by the standard deviation of its distribution. This enables us to compare features from different individuals. For each game day, for each feature, we obtain a large set of z-scores (one per frame). From these, we extract statistics (e.g. entropy, mode, median, range, skew and kurtosis) that are used to learn the models.

5. EXPERIMENTS

For all players in the game, the frames where they speak are processed separately from the frames where there was no voice activity measured through their microphones. This allows experiments to be done on motion while speaking and while not speaking.

Because each game has only two wolves and many more villagers, the number of datapoints (all day games for all players) for villagers and wolves is not balanced. To balance both datasets, random subsampling is applied to the villagers datapoints. This results in 36 data points for wolves and 36 data points for villagers. To obtain comparable results with [7] 18-fold cross-validation was done. To ensure a representative variation for the experiments this process is repeated 1000 times.

The purpose of our experiments is to evaluate how much information about deception can be extracted from gross body motion, and to identify whether different parts of the body provide different information. Our focus is not on the specifics of the classification algorithm, and we report results with two different classifiers so as to make comparison with earlier work possible. These classifiers are: the Relevance Vector Machine (RVM), which was also used in [7] and the Support Vector Machine (SVM), because it is considered one of the best off-the-shelf general-purpose classifiers.



(a) H region (b) BML region

Figure 4: visual activity plots for regions H and BML. Red line shows wolves' and blue the villagers' distribution

Experiment 1: temporal difference. Features that are extracted from this data's distribution are entropy, mode, median, range, skew and kurtosis. This experiment is done on speech frames, non-speech frames and all frames.

Experiment 2: motion magnitude: For this experiment one histogram is calculated from all the motion magnitude histograms of all the frames of each player in a game day. The same features as in experiment 1 are extracted from these combined histograms.

Experiment 3: motion orientation: Similarly as in experiment 2, the motion orientation histograms for each region and player per game day are combined. The same features are extracted as mentioned in experiments 1 and 2

Experiment 4: combined video: In this experiments all features from the experiments 1, 2 and 3 are combined. PCA is applied on the features to reduce the dimensionality and get rid of non-informative features.

Experiment 5: combined audio video: For this experiment all video features from experiment 4 were used and fused with the audio features used by Chittaranjan and Hung [7]. PCA is applied on the features to reduce the dimensionality and get rid of non-informative features. Each of these experiments was carried out three times; first, for the features extracted from frames with speech; second, for features extracted from frames without speech; third, for features extracted from all frames.

6. RESULTS

Figure 4 a and b show normalized plots of the z-scores of the distribution of the temporal difference for the H and BML regions. The red lines represent the wolves' and blue lines represent the villagers' visual activity. Plots 4a and b clearly show that wolves move less then villagers. This phenomenon varies per region but occurs in all and is consistent with prior research [1, 2, 13]. Plots 4a and b also show BML region is more informative for deception detection than region H as the distribution of the wolves is clearly more skewed to lower visual activity values.

For all experiments the results are shown in tables 1 and 2. The best performing result of Chittaranjan and Hung [7] was used as baseline for our experiments. Their method's best performance, f-measure of 0.62, was obtained by fusing all the audio features.

Our best performing model is the one trained on all frames and all the video features fused (exp. 4). This model's f-measure is 0.76, which is an improvement of 22.6% over the

Table 1: Mean F-measures of liar (L) and non-liar (NL) classification using RVM.

Ex.	all frames		speech		non-speech	
	Reg.	L vs. NL	Reg.	L vs. NL	Reg.	L vs. NL
1	BML	0.62 ± 0.13	BML	0.61 ± 0.13	BML	0.62 ± 0.12
2	BML	0.66 ± 0.12	BML	0.67 ± 0.13	BML	0.66 ± 0.13
3	BL	0.60 ± 0.13	BL	0.61 ± 0.12	BL	0.60 ± 0.13
4	all	0.71 ± 0.18	all	0.66 ± 0.17	all	0.68 ± 0.18
5	all	0.72 ± 0.18	all	0.67 ± 0.17	all	0.70 ± 0.18

Table 2: Mean F-measures of liar (L) and non-liar (NL) classification using SVM.

Ex.	all frames		speech		non-speech	
	Reg.	L vs. NL	Reg.	L vs. NL	Reg.	L vs. NL
1	BML	0.62 ± 0.13	BR	0.60 ± 0.13	BML	0.62 ± 0.13
2	BML	0.69 ± 0.12	BML	0.68 ± 0.13	BML	0.71 ± 0.13
3	BL	0.61 ± 0.13	BML	0.65 ± 0.12	BL	0.61 ± 0.13
4	all	0.76 ± 0.17	all	0.66 ± 0.16	all	0.72 ± 0.18
5	all	0.75 ± 0.18	all	0.68 ± 0.17	all	0.73 ± 0.18

baseline. For all experiments the best performing regions are the body regions. From the body regions it seems that the area in the middle of the body (**BML** and **BMR**) is most informative for detecting deception in multi-party conversations. The **BML** and **BMR** regions' performance is almost identical.

From the models that are trained on either speech or non-speech frames, the model trained on non-speech frames and all the video features combined with the audio features, as used by Chittaranjan and Hung [7], (exp. 5) performs best with an average f-measure of 0.73. In the experiments where features are combined (exp. 4 and 5), models trained on non-speech frames always outperform models that are trained on speech frames. A possible explanation is that when people talk the focus is on them, therefore they try to control their body motion more. Another explanation is that only 12.16% of all frames are speech frames which might effect the quality of the model.

Out of all the uncombined video features (exp. 1, 2 and 3), regardless of speech activity, the motion magnitude of region **BML** performed the best. The average f-measure of this model is 0.69, which is an improvement of 11.3% over the baseline. This model's performance increased to 0.71 when it was trained on non-speech frames only. In the experiments with the uncombined features, the models learned on features from all frames never outperform the models that are trained on features extracted from speech or non-speech frames.

We performed a t-test to support our models' performance. The results from experiments 4 and 5 all outperform the baseline significantly at $\alpha = 0.05$. None of the models from experiments 1, 2 and 3 were significantly better than the baseline.

7. CONCLUSIONS

We successfully show that simple non-verbal video features, extracted automatically from low-quality video, can be used as indicators of deception in multi-party conversations. Our results also shows an increase in performance when classifiers trained on video features and on only speech frames. This is a promising result for audio-visual analysis of social interaction for detecting deception.

8. ACKNOWLEDGEMENTS

This research was supported in part by a Marie Curie Research Training Network fellowship in the project "AnaSID" (PIEF-GA-2009-255609) and by the SIA project 'Smart Systems for Smart Services'. We thank Gokul Chittaranjan for providing help and code for extracting audio features and the RVM.

References

[1] Aldert Vrij. *Detecting Lies and Deceit: Pitfalls and Opportunities, 2nd Edition*. Wiley, 2008. ISBN 978-0-470-51624-9.

[2] Paul Ekman. *Telling lies*. Norton, New York, New York, 2001. ISBN 0-393-32188-6.

[3] Hayley Hung and Gokul Chittaranjan. The idiap wolf corpus: exploring group behaviour in a competitive role-playing game. ACM MM, 2010.

[4] Matthew Jensen, Thomas Meservy, Judee Burgoon, and Jay Nunamaker. Automatic, multimodal evaluation of human interaction. *Group Decision and Negotiation*, 19:367–389(23), July 2010.

[5] Sebastian Germesin and Theresa Wilson. Agreement detection in multiparty conversation. ICMI-MLMI, 2009.

[6] Julia Hirschberg, Stefan Benus, Jason M. Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Gir, Martin Graciarena, Andreas Kathol, and Laura Michaelis. Distinguishing deceptive from non-deceptive speech. In *Interspeech*, 2005.

[7] Gokul Chittaranjan and Hayley Hung. Are you a werewolf? detecting deceptive roles and outcomes in a conversational role-playing game. In *ICASSP*, 2010.

[8] Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. ACII, 2009.

[9] Nick Campbell. An audio-visual approach to measuring discourse synchrony in multimodal conversation data. In *Interspeech*, pages 2159–2162. ISCA, 2009.

[10] Hayley Hung, Dinesh Jayagopi, Chuohao Yeo, Gerald Friedland, Sileye Ba, Jean-Marc Odobez, Kannan Ramchandran, Nikki Mirghafori, and Daniel Gatica-Perez. Using audio and video features to classify the most dominant person in a group meeting. ACM MM, 2007.

[11] Dinesh Babu Jayagopi, Hayley Hung, Chuohao Yeo, and Daniel Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Trans. Audio, Speech and Lang. Proc.*, 17:501–513, March 2009.

[12] S Benus, F Enos, J Hirschberg, and E Shriberg. Pauses in deceptive speech. *Speech Prosody*, 18:2–5, 2006.

[13] Thomas O. Meservy, Matthew L. Jensen, John Kruse, Douglas P. Twitchell, Gabriel Tsechpenakis, Judee K. Burgoon, Dimitris N. Metaxas, and Jay F. Nunamaker Jr. Deception detection through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intelligent Systems*, 20:36–43, 2005.

[14] Thomas O. Meservy, Matthew L. Jensen, John Kruse, Judee K. Burgoon, and Jay F. Nunamaker. Automatic extraction of deceptive behavioral cues from video. In *ISI*, pages 198–208, 2005.

[15] Ce Liu. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. PhD thesis, Massachusetts Institute of Technology, May 2009.