Anger recognition in audio-visual data

Frank Borgonjen
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
f.d.borgonjen@student.utwente.nl

ABSTRACT

Although there exist quite a few techniques to recognize emotions in audio-visual data, there is still a major gap between the capabilities of humans and computers in this field. This paper is a search to the visible and audible features that could be used to automatically detect a person's emotional state from audio-visual data. While detecting emotions in audio-visual data is a broad field, we narrowed the research by picking one emotion, namely anger, that acts as a guideline through the research. Six different features were found that indicate whether a person is angry: facial expressions, gestures and postures, prosody, speech, respiration rate and pupil size. The current status of the automatic detection of these features are discussed and, where possible, improvements are proposed. The research shows that it is not desirable to determine a person's emotional state from one feature and that is why fusion of the features is important and of great value.

Keywords
Emotion recognition, Anger recognition, Audio-visual data

1. INTRODUCTION

The increasing amount of audio-visual data on the internet has resulted in the need for methods to store, index, query and retrieve these amounts of data. In order to let search engines understand the contents of audio-visual data an extensive description of the content has to be available. While manually annotating audio-visual is an expensive and time consuming activity, this has led to the development of techniques that automatically generate semantics from raw data.

Currently many techniques exist that are capable of automatic annotation of audio-visual data. These techniques provide a wide range of annotations that vary from location information [1] to facial expressions. But do these techniques cover all the characteristics that we want? Could we for example get satisfactory results if we search for videos that shows a romance, or an argument between two people? Are the current techniques capable of recognizing human emotions like love, anger and sadness?

Recognizing emotions in (real time) audio-visual data is a valuable addition. More precise queries could be made into audio-visual databases. A query like “Show me all videos of two people that have an argument” or “Show me all videos of two people that have an intimate conversation”. Enhancing query results is not the only advantage. Emotion recognition could also be used in surveillance cameras that are installed in city centers. When it detects an aggressive person or a group of persons it can send a message to the operating room. Another example where emotion recognition could be used is in televisions. Aggressive or too intimate scenes can be detected and appropriate measures like a black screen can occur to protect young viewers. The benefits of emotion recognition requires no additional explanation.

Automatic emotion recognition is quite a challenge because it is a complex task. A lot of visible and audible features like facial expressions and body postures change when someone expresses an emotion. Even people often misinterpret the features that show what emotion a person expresses [13]. In an online psychology test [26], for example, users are asked to swiftly look at pictures of a persons face that expresses a specific emotion. After they looked at the picture the users were asked to select the emotion they think is expressed by the person on the picture. The average score on the test is 5 out of 30 correct answers. This shows that just from looking at someone's facial expression you cannot say with certainty what that person is feeling. More features should be taken into account if you want to be sure what emotion someone feels. What makes it even harder is the fact that people sometimes try to cover their emotions, for example when they are ashamed or when they want to keep up appearances.

Automatic emotion recognition is a very broad research area and we therefore narrowed the research to the automatic recognition of one specific emotion, namely anger. The reason to pick anger as emotion is because the visible and audible features that a person expresses are the most striking for this emotion. This makes it easier for the computer to distinguish these features.

In this paper we show to what extent computers are capable of automatic recognition of anger in audio-visual data and which improvements could be made to obtain a better accuracy.

2. RESEARCH QUESTIONS

There is still a big gap between the capabilities of humans and computers in the field of recognizing emotions. To fill the gap we will give an answer to the following research question:
• How can we improve the automatic recognition of anger in audio-visual data?

To be able to answer this question the following subquestions will be answered first.

• What features are noticeable when people express their anger?
• What techniques are currently available that recognize these features in audio-visual data?
• Which techniques could be useful to recognize features for which no recognition techniques are available?
• What are the weaknesses of the techniques used to recognize these features?
• What can we do to bridge these weaknesses?
• Does fusion of the different features provide a better accuracy?

The first step in the research is to gather all features that are visible or audible when a person is angry by reading research papers and books about emotions. These features form the basis for the research. The next step is to search for known techniques that are capable of automatic recognition of every single feature. We list the techniques and name their advantages and disadvantages. According to these disadvantages we point out gaps in the process. Where possible, a solution to fill the gaps will be proposed and substantiated. For the features where no known recognition techniques exist we discuss techniques from related domains and substantiate why they could be useful. At last we discuss whether fusion of the different features provides better accuracy.

3. FEATURES OF ANGER

There are two different forms of anger, hot or aggressive versus cold or passive anger. Passive anger is more subtle and much harder to recognize than aggressive anger and therefore we only focus on the recognition of aggressive anger. From here on when we talk about anger we mean aggressive anger.

Anger can be recognized by a lot of different features. Unfortunately several of these features such as heart rate and galvanic skin response can only be perceived by using sensors. For this research we are looking for as many features that can be perceived just by analyzing audio-visual data.

A literature study resulted in six useful features that can be perceived from audio-visual data; facial expressions, postures and gestures, vocal expressions, speech, respiration rate and pupil size.

3.1 Facial expressions

The Facial Action Coding System (FACS) [14] by Ekman, Friesen and Hager is a detailed technical guide that explains how facial behavior follows from muscular actions. With this guide it is possible to describe facial expressions by a series of so called action units (AUs) such as ‘inner brow raise’ and ‘upper lip raise’. This work has been a basis for many research in facial expression analysis [21, 34, 23]. Recognizing facial expressions can basically be translated to recognizing a set of action units that belong to that specific emotion.

Cross cultural differences make it difficult to define which emotions are universal and thus it is hard to classify facial expressions into emotions. To what extent emotions are universal has been thoroughly discussed by Ekman (1989, 1994) [10, 11], Izard (1994) [20] and Russel (1994) [29]. Fortunately for this research anger is one of the six basic emotions that is universally recognized.

According to Ekman [12, 13] and Gunes [17] people will show one or more of these facial changes when they are angry:

• Squeezed and lowered eyebrows pulled inward
• Lines between eyebrows
• Vertical, sometimes curved, forehead wrinkles above the eyes
• Tensed upper eyelids and possibly raised, lowered or squared
• Lower eyelids together and possibly raised
• Lips pressed together or open, squared mouth with raised lips
• Dilated nostrils
• Less sclera is shown
• Jaw pushed forward

Research to emotion recognition through facial expressions is mostly performed in the field of human computer interaction where one of the goals is that the computer ‘senses’ the persons emotion. This means that the subjects are often filmed with a webcam in controlled environments (good lighting, only one subject, subject facing the camera, etc.). A universal facial expression recognition system will have to be able to detect emotions in all sorts of audio-visual data. Researchers thus have to take into account that there might be multiple subjects that appear in the video, that the subjects face will often be tilted at some degree and that the background will not be a static picture.

Saxena et al (2009) [31] used motion estimation to track the most prominent subject in video data. Their method could be applied in combination with a facial expression recognition system to avoid mixed emotion detection because of multiple subjects that appear in one shot. An advantage of the system by Saxena et al is that it is not necessary that subjects are filmed in a controlled environment with a static background. One drawback of the system is that it processes images at 10 fps instead of 24 that is used in regular video data. This means that you either can only track subjects that move at moderate speed, or you can track fast moving subjects at a slower speed.

Kumano et al (2009) [21] acknowledged the problem that facial expressions should be recognized independent of the facial pose. Their research has lead to a method that is capable of recognizing anger with an accuracy of 97.5% for horizontal, vertical and sideway facial orientations in the range of ± 40, ± 20 and ± 40 degrees respectively. The direction that Kumano and his colleagues took is an important step towards facial expression recognition in regular audio-visual data.

Facial expressions are affected when a person is speaking and could lead to false classification of emotions. As a solution to this problem, the system of Datcu and Rothkrantz

• Jaw pushed forward

Vertical, sometimes curved, forehead wrinkles above the eyes
• Tensed upper eyelids and possibly raised, lowered or squared
• Lower eyelids together and possibly raised
• Lips pressed together or open, squared mouth with raised lips
• Dilated nostrils
• Less sclera is shown
• Jaw pushed forward

Research to emotion recognition through facial expressions is mostly performed in the field of human computer interaction where one of the goals is that the computer ‘senses’ the persons emotion. This means that the subjects are often filmed with a webcam in controlled environments (good lighting, only one subject, subject facing the camera, etc.). A universal facial expression recognition system will have to be able to detect emotions in all sorts of audio-visual data. Researchers thus have to take into account that there might be multiple subjects that appear in the video, that the subjects face will often be tilted at some degree and that the background will not be a static picture.

Saxena et al (2009) [31] used motion estimation to track the most prominent subject in video data. Their method could be applied in combination with a facial expression recognition system to avoid mixed emotion detection because of multiple subjects that appear in one shot. An advantage of the system by Saxena et al is that it is not necessary that subjects are filmed in a controlled environment with a static background. One drawback of the system is that it processes images at 10 fps instead of 24 that is used in regular video data. This means that you either can only track subjects that move at moderate speed, or you can track fast moving subjects at a slower speed.

Kumano et al (2009) [21] acknowledged the problem that facial expressions should be recognized independent of the facial pose. Their research has lead to a method that is capable of recognizing anger with an accuracy of 97.5% for horizontal, vertical and sideway facial orientations in the range of ± 40, ± 20 and ± 40 degrees respectively. The direction that Kumano and his colleagues took is an important step towards facial expression recognition in regular audio-visual data.

Facial expressions are affected when a person is speaking and could lead to false classification of emotions. As a solution to this problem, the system of Datcu and Rothkrantz...
are related to anger are listed below [25, 36]:

- High movement activity
- Expansive movements
- High movement dynamics

Unfortunately the movement activity, expansiveness and dynamics are also high for elated people; this makes emotion recognition from these characteristics a difficult task. The fact that it is also difficult to distinguish gestures for happy and angry people means that classifying emotions based on gestures and postures is a complex task. To solve this problem there is the need to exclude one of the emotions (anger or elated joy). This can be achieved by fusion of multiple features discussed in section 4.

### 3.2 Postures and gestures

Body posture and gestures are another important feature that indicates the emotional state of a person. These are the most obvious posture and gesture characteristics that are expressed by an angry person [36, 24, 17]:

1. Upper body erect and possibly bent forward
2. Shoulders up
3. Hands stretched out frontal
4. Arms crossed
5. Hands close to face (emphasize spoken words)
6. Hands opening/closing
7. Lateralized movements with hands
8. Pointing
9. Hands clenched
10. Hands on hips/belt

The system proposed by Gunes and Piccardi (2006) [17] is capable of recognizing gestures 1, 3, 5, 7. One of the downsides of their approach is the controlled environment in the videos that is necessary for the system to work properly. Alon et al (2009) [2] provide a more robust system that is focussed on tracking the hands (necessary for gestures 3, 5, 6, 7, 8 and 10) of the subject. The test sets they used contained complex background where people were moving and walking by. For this reason we can assume that gestures can be detected with a large accuracy.

The average use of the different gestures and postures to express emotions summarized by Wallbott (1998) [36] indicate that several gestures that describe anger are also shown when a person is elated. This makes it difficult to distinguish these emotions just by analyzing gestures and postures.

Not only the different gestures and postures are important but also the amount of movement determine what emotion a person experiences. The movement characteristics that are related to anger are listed below [25, 36]:

- High movement activity
- Expansive movements
- High movement dynamics

### 3.3 Prosody (vocal features)

Prosody is the rhythm, stress, and intonation of speech which could reflect the emotional state of the speaker or whether the utterance is, for example, a question or a statement. The ability of infants to detect negative emotional information develops earlier in the vocal domain than in the facial domain [16]. This shows that prosody is of great value to determine whether a persons is angry.

The experiments by Barra et al (2006) [5] show that emotions can be identified by analyzing segmental and prosodic features. In their experiments the segmental features are represented by the mel-frequency cepstrum coefficients (MFCC) which are a representation of the short term power spectrum of a sound [22]. The prosodic features are represented by fundamental frequency (F0) which is one of the most important features that represent the prosodic and syntactic representation of speech [30]. The experiments show that anger can be recognized by the combination of the MCFF and F0 analysis with an accuracy of 87%.

The MCFF feature is the most important in the recognition of anger from prosodic signals, but this feature is not very robust when the audio feature contains noise. To improve the robustness of the MCFF feature, the findings of Tyagi et al (2005) [33] could be applied. They propose a modification to the MCFF algorithm to increase the power of the log Mel-amplitudes. This results in the reduction of low-energy spectral noise which improves the robustness.

### 3.4 Speech

Speech recognition (SR) systems are designed to convert speech signals into text. This perceived text does not only contain informative content but also cues that indicate the emotional state of the speaker. Unfortunately not much research has been performed in detecting emotions from text and the performance is far from satisfactory [7]. The main reason is that for most of the sentences, prosody determines whether it is meant to be angry or happy, not the sentence itself.

The accuracy of speech recognition systems also plays a role in the detection of emotions from spoken words. According to [32] untrained speech recognition systems have a maximum accuracy of ± 80%. A requirement to perceive this accuracy is an audio signal that contains as little noise as possible. This accuracy will probably not be reached because audio-visual data generally contains noise.

When it appears possible to derive information about the emotional state of a speaker from spoken words, there is an upside of using audio signals from audio-visual data. An automatic lip reading system such as the one of Wiggers and Rothkrantz (2010) [37] can be used to perceive better speech to text accuracy when the audio signal contains noise.

All in all it seems that speech recognition will not provide useful information about the emotional state of the
speaker(s) as long as there is no valuable mapping between text and emotion.

3.5 Respiration rate

Psychological studies like the ones from Rainille (2006) [27] and Wagner (2005) [35] have shown that emotions can be recognized by several physiological features. These features include amongst others heart rate, skin conductivity and respiration. Unfortunately only respiration is a feature that might be of some value for the recognition of emotions in audio-visual data while other features can only be measured by sensors.

Respiration can be measured by changes in rate, amplitude, and inspiration/respiration ratio [3]. In case of audio-visual data, the only characteristic about respiration that you can perceive is the respiration rate by investigating the audio signal. According to Clynes and Menuhin (1977) [8] respiration rate accelerates during anger or hate. This unfortunately is also true when someone experiences fear. Nevertheless we already stated that the more features that might indicate that a person is angry, the better.

No technique exists that uses respiration rate to determine what emotion a person experiences in audio-visual data. But there are researches related to this problem. Ruinskiy and Lavner (2007) [28] propose an algorithm that very accurately detects breathing sound in speech and song signals. Just like the experiments discussed in the section about prosody, this algorithm also uses the mel-frequency cepstral coefficients (MCFF) as feature to detect breath sounds. The same improvements to the MCFF algorithm can be made to improve the robustness of this feature.

The challenge that is left after the detection of breathing sounds is determining the respiration rate that gives us information about a persons emotional state. It seems fair to assume that it is not difficult to determine the respiration rate when it is possible to detect breath sounds.

The usefulness of this feature depends on the audio-visual data that is analyzed. Most of the time breath sounds will not, or only partially, be audible in audio-visual data. But when a clear audio signal is available, this feature might be of some help in determining a persons emotional state.

3.6 Pupil size

According to Harrison et al (2007) [18] and Pease (1997, p.86) [24] anger will result in the contraction of the pupils. The fact that the pupil size ranges from $\pm 3$ to $\pm 10$ mm indicates that automatic detection of pupil contraction is a very exact matter. This might also be the reason why changes in a persons pupil size has not been used in the automatic detection of emotional state.

To get to the point that it is possible to measure pupil size changes we need to overcome two challenges:

- A real time eye tracking system
- An accurate measurement of the pupil size

For the former challenge several systems exist like the one of Heinzmann and Zelinsky (2002) [19]. Heinzmann and Zelinsky propose a system that is capable of robust real time eye tracking and gaze point estimation. One of the steps in the eye gaze tracking is the determination of the corners and the position of the iris as shown in figure 1.

For the latter challenge there is no research that shows that such a system exists. Fortunately there are researches that are related to this measuring problem. The system by Heinzmann and Zelinsky that is discussed above is able to calculate the distances between eye corners and the iris to estimate the gaze point. From the fact that accurate measurements can be derived from fairly low quality video data we can assume that it is also possible to derive measurements of the pupil size.

One other challenge that has to be taken into account is the determination of the edge between the iris and the pupil. For people with a very dark iris color this is a more difficult task than for people with bright iris colors. The computational canny edge detection approach [4] is one of the techniques that could be used to optimize the edge estimation between the iris and the pupil.

By combining the techniques discussed above we can assume that it is possible to measure the pupil size, or at least the ratio in comparison to the size of the iris. Examining the ratio between pupil and iris size will make it possible to determine the moments that a person contracts or dilate his/her pupils which might indicate a change of emotional state.

4. FUSION

Most of the research done in emotion recognition is focused on one single feature such as facial expression. As mentioned in the previous chapter, results of emotion detection from single features will often not be enough to determine whether a person is angry. To bridge this deficiency it seems obvious that emotion recognition should be determined by fusion of the features from the previous chapter.

Fusing the results after all features have been processed will probably not give the best accuracy. Facial expressions for example are influenced by the movements of the lips when someone is speaking. And gestures mostly occur simultaneously when someone emphasizes certain words. These dependencies should be taken into account when fusing information from the different features. Another challenge is the absence of features what is very likely to happen. Gestures for example are not visible when you see a close up of someones face. A system that fuses information from multiple features should be able to deal with these absences.

Probabilistic graphical models such as Bayesian networks and hidden Markov models [15] have been proven to work very well for this kind of fusion problems. These models can handle noisy features, missing feature information and hypotheses. Zajdel et al (2007) [38] for example use a dynamic Bayesian network to fuse information about gestures and audio features to detect aggressive human behavior. Results show that fusion of these features provide a detection accuracy of about 78% compared to 45% and 67% for just audio or video respectively.

The system presented by Castellano et al (2008) [6] uses three different features (face, body gestures and speech) to recognize emotions. They show results of fusion with Bayesian classifiers at two different levels in the classifi-
cognition process. The first level is at the so called feature level which means that features are independently classified, and then fused. The second level is at the so called decision level where all features are fused in one Bayesian classifier. Results show that anger can be better detected after decision level fusion and joy can be better detected after feature level fusion. This system recognizes anger with an accuracy of 96.67% after fusion while recognition from facial expressions, gestures and speech alone provide an accuracy of 56.67%, 80% and 93.33% respectively.

From these outcomes it seems fair to assume that fusion of the different features will provide better detection results.

5. CONCLUSIONS
This paper shows what the state of the art is capable of in the field of emotion recognition and especially the recognition of anger in audio-visual data.

The research has pointed out that there are six different visible and audible features that can be used to determine whether a person is angry namely facial expressions, gestures and postures, prosody, speech, respiration rate and pupil size.

Facial expressions, gestures and postures and vocal expressions are features that have proven to be valuable input for anger recognition. Respiration rate and pupil size are features that could be of some value in anger recognition, but these features depend on specific visual (pupil size) and audible (respiration) data that is not present most of the time. Anger recognition from speech is a feature that will have little accuracy. This is due to the difficulty of mapping words to emotions and the fact that many words or sentences have different emotional meanings when pronounced differently.

The performance of these features apart from each other will not give satisfactory results to determine whether a person is angry. Fusion of the features seem to be the outcome to improve anger detection in audio-visual data.

6. ACKNOWLEDGMENTS
I would like to thank my mentor Maarten Fokkinga for his valuable input in choosing my research topic and providing me with useful related literature. I would also thank my fellow-students of the ‘Data Application and Integration’ track for their input and reviews.

7. REFERENCES
[22] Mel-frequency cepstrum.


