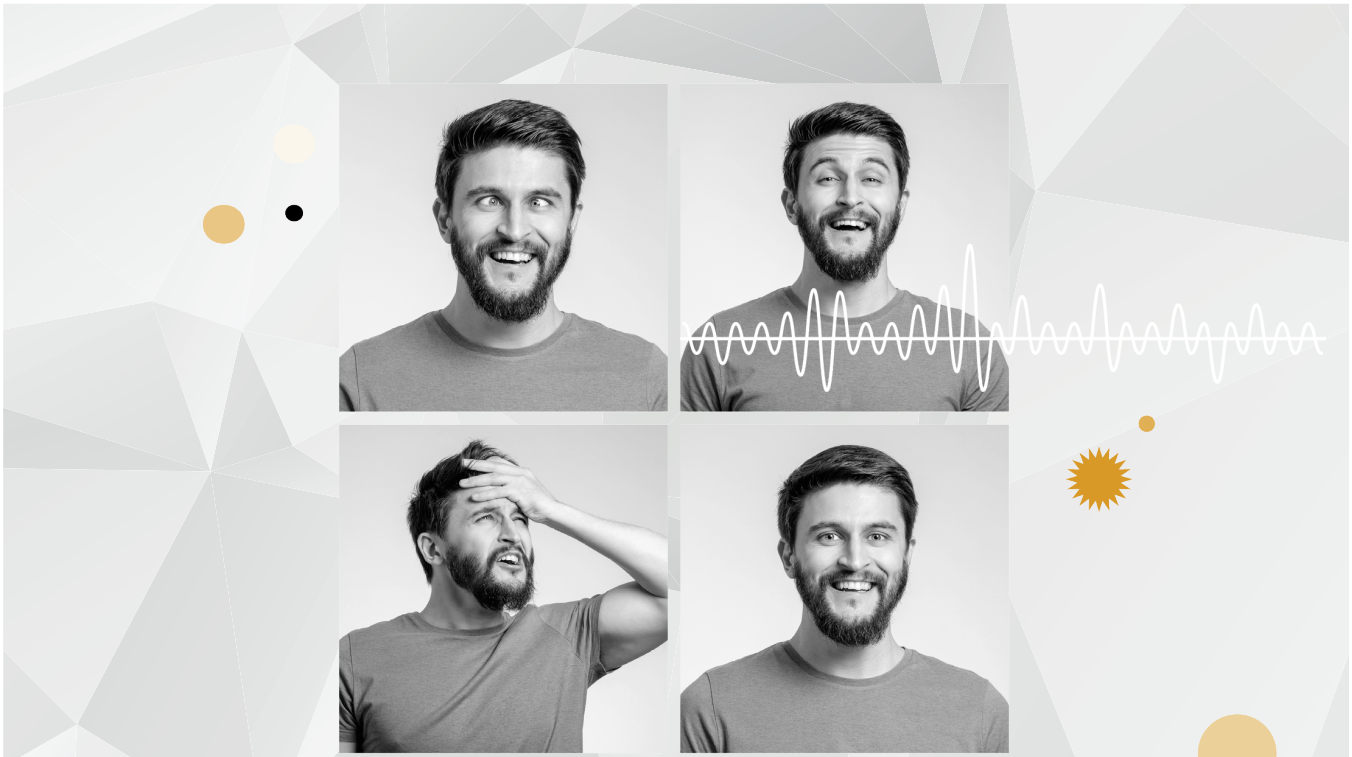
WORKING PAPER SERIES



IMPROVING SIGNAL DETECTION IN SOFTWARE-BASED FACIAL EXPRESSION ANALYSIS

Matthias Unfried, Markus Iwanczok

WORKING PAPER /// NO. 1 / 2016

Improving signal detection in software-based facial expression analysis

Matthias Unfried*† Markus Iwanczok‡

Abstract— *Algorithmic and equipment-based procedures for emotion detection are often afflicted by measurement error or signal noise. In this paper, we analyze the signal-noise-relation of software for automated facial expression analysis used to measure emotional response to marketing stimuli. We isolate the noise and discuss, apply, and evaluate several methods for reducing the noise. Our results show that noise is a challenge in automated analysis of facial movement data, but can be reduced by applying fairly simple methods. Using data from a real market research study we show that noise can be reduced to a negligible level.*

Keywords— *facial coding; signal-noise-ration; noise reduction*

1 Software-based emotion detection

Passive data collection methods have been growing over the last few years and decades. Especially in market research, these methods are often applied alongside traditional questionnaires in order to augment the analysis of consumer response with direct measures of experience. It is particularly difficult to reliably ascertain emotional reactions (e.g., for advertising tests and usability studies) through direct questionnaires. For this reason, a number of methods have been developed in recent years for directly capturing emotional reactions. One area in this field addresses the inference of emotional reactions by analyzing facial expressions.

Most of the methods for automatic detection of emotional reactions from facial movements are based on the same principle. Algorithms are used to extract particular facial features from images and video recordings of respondents. These are used to either assign the facial expression directly to a specific emotion (e.g., anger) or to an "action unit" (c.f. Ekman and Friesen, 1978), which is in turn used to infer particular emotional reactions. Different algorithms exist both for extraction of features and for classification. However, a common feature of both types of algorithms is that large databases comprising annotated images are necessary to train them in order to reliably classify the recorded facial expressions (c.f., e.g., Pantic and Rothkrantz, 2003; Zeng et al., 2009).

The software GfK EMO Scan was developed specifically for use in market research. It determines the valence of facial expressions from webcam images or video recordings as a measure of emotional experience (Garbas et al., 2013). The software is based on a combination of the SHORE facial expression recognition analyzer (Küblbeck and Ernst, 2006; Küblbeck et al., 2009; Ruf et al., 2011) and a valence detector which was trained with a database that includes several thousand images of different emotional facial expressions.

Analysis of video recordings with the software entails splitting them into individual images (frames) and determining a valence value for each frame. Depending on the frame rate of the recording, this can generate up to 30 valence values per second. The video recordings are then calibrated to the respondent's neutral facial expression.

However, as is the case for most equipment-based and algorithmic methods, factors that degrade the image quality (e.g., image compression, poor lighting, etc.) can result in noise. Under the term noise we subsume measurements which are triggered by one or more quality-degrading factors but not by the actual facial response. In order to investigate and quantify more precisely, a robustness test of the analysis software presented above was conducted.

The aim of this paper is to quantify the magnitude of noise more precisely, investigate the influence of noise

*Corresponding author, matthias.unfried@gfk-verein.org

†GfK Verein, Fundamental Research

‡GfK SE, Computational Statistics

Citation: Unfried, M. and Iwanczok, M. (2016), Improving signal detection in software-based facial expression analysis, GfK Verein Working Paper Series, No. 1 / 2016

on the results and find suitable methods for reducing or even eliminating it.

To this end, we examine the determinants of noise more closely and isolate the noise from the real signal in data that were collected for this purpose in a test scenario. Subsequently, a number of methods will be presented which are able to reduce noise significantly such that it statistically disappears. The results of the test scenario are compared with data from a real study to investigate the relevance of noise for software application in market research. The paper closes with some recommendations of methods which should be used for smoothing the data and factors which should be considered when interpreting the data.

2 Determinants of noise

A wide range of factors can cause and influence noise. In principle, a distinction can be made between software and hardware factors of influence. However, all factors presented below can potentially impact the image quality and detection results.

2.1 Software-related factors

Video codec

A video codec is software that encodes and decodes digital videos. To manage the amount of data to be transferred, online videos (streaming/live streaming) are often coded in such a way that a certain loss in quality occurs. The *Sorenson Spark (H.263 model)*, the *H.264* and the *VP6 on2* codec for online *Adobe Flash* applications are the most common codecs for this purpose.

The frequency of so called key frames is a decisive factor in the creation of noise. Key frames are frames which are transmitted unmodified; all frames between the key frames are only interpolated. This interpolation then produces artifacts, which can vary for each interpolated frame. Reducing the time between key frames reduces noise, while increasing the key frame distance amplifies the noise. However, it should be taken into account that reducing the distance increases the quantity of data that needs to be transferred (c.f., e.g., Slepian and Wolf, 1973; Wyner and Ziv, 1976).

Bit rate

The bit rate describes the data throughput within a given period of time (e.g., bits per second). Video material can be created and sent with a dynamic or static bit rate.

With a static bit rate, the data volume transferred always remains constant. This can result in limitations for signal transfer depending on the internet connection.

For the variable bit rate (VBR, dynamic streaming), the video material is analyzed and a higher bit rate is assigned to areas in the stream where the image changes more than those where the image does not change as much. This allows image and transmission quality to be higher when internet connection speed is comparatively low.

Video resolution

The size of the face in the video also impacts signal quality and essentially depends on the video resolution. If the face is smaller, fewer details can be captured. If resolution is very low, contours of the face, for example, can be extremely out of focus although the level of blurring can vary across the image. Generally, this results in highly grainy images, although the area where the image is grainy can vary over time, often in conjunction with image compression and definition of key frames.

In order to create satisfactory results, the resolution of the entire video should sufficiently high. This ensures that the area of the face is large enough.

Influence of software-related factors on image quality

Depending on the chosen bit rate and compression method, individual images in videos are subject to varying degrees of artifact formation. The greater the degree of artifact formation, the more strongly the detection results will be influenced. Figure 1 shows the image quality in relation to compression and bit rate.

(a) high compression and low bit rate

(b) low compression and high bit rate

Figure 1: Example of variations in quality for different compression levels

While weak compression and relatively high bit rates result in high image definition (Figure 1(b)), the contours

are rather blurred for strong compression and low bit rate (Figure 1(a)).

Influence of atypical facial features

Detection and classification algorithms are generally trained through large annotated databases. Consequently, the detection result is also dependent on conformity of the recognized face with the database. If the recognized face has any unusual features, this can result in fluctuations of the results values. For example, if someone wears glasses with reflective lenses, this can create problems in eye recognition and therefore influence the detection and values.

Quality check and dynamic adjustment

In applied settings, it is essential that the data volume is kept as low as possible, especially when the video has to be transmitted online. However, this can impact the quality of video material and therefore also detection results. For this reason, it is recommended that the quality of recorded material is already assessed as part of a quality check prior to recording and, if necessary, to adjust video resolution and compression. This makes it possible to respond to inadequate facial recognition through poor lighting, for example. The software for emotion detection used in this study incorporates a quality check by calculating a quality indicator. The indicator states how well the face can be detected. Compression and resolution can therefore be adapted if the indicator falls below a set threshold. In principle, the quality check can be repeated as often as required until the image quality reaches the desired level.

2.2 Hardware-related factors

Camera

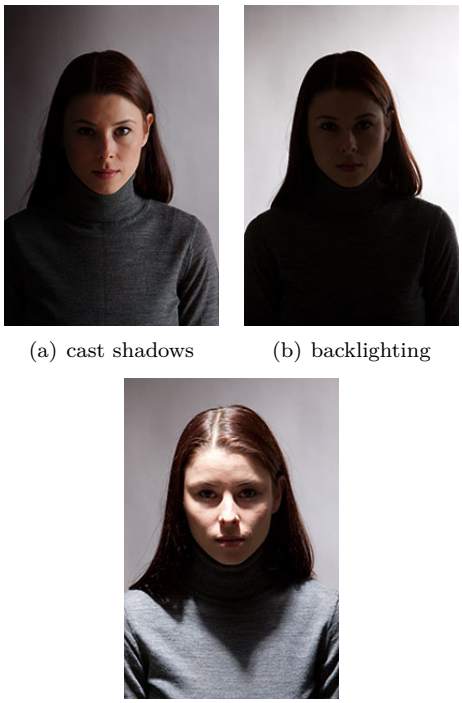
Webcams are generally used for automated emotion recognition. The quality of these cameras influences the image and therefore also the detection result. As many webcams have relatively large wide-angle lenses, the distance between respondents and the camera is particularly important as well as the recording angle. In addition, the direction in which respondents are looking impacts the detection result and if the angle is very wide, for example, the eyes cannot be detected. The algorithm of the GfK EMO Scan is able to correct horizontal deviations in the line of sight by +/- 30 degree.

A higher quality autofocus can improve the video data. Poor cable connections or dirty lenses can also have a negative effect on the recording quality.

Lighting

Inadequate lighting has a similar effect on video quality. Figure 2 shows different examples of poor lighting and its impact on image quality.

In addition, poor light conditions such as back-lighting or cast shadows can cause contours to be blurred or particular facial features such as the eyes or the mouth to barely be recognizable, which again impacts the detection result. Given that lighting can vary during a recording, image quality can also change and consequently result in noise.



(a) cast shadows (b) backlighting

(c) light from above

Figure 2: Influence of lighting on image quality

3 Reduction of noise

3.1 Test scenario for measuring noise

A test was conducted to isolate noise. For the test, 24 still images showing various emotional expressions of 12 different individuals were extracted and each was developed into a 30 second video file. The resulting 24 video files were analyzed with GfK EMO Scan and calibrated separately. For each video file we used the averaged valence of the video file itself for calibration and subtracted this calibration from all measured valence values (15 values per second). Given that there are no changes of emotion in still frames, this allows pure noise to be observed. As this procedure used video material comprised of still frames, variation in lighting and webcam quality can be excluded as sources of noise. Atypical facial features were also excluded in the selection of pictures as far as possible. Consequently, the only noise sources that remained were compression (video codec), bit rate and key frame setting. In this regard, the settings chosen were those that are also applied in real study applications.

When examining noise individually for each analyzed video, the individual average noise ranges between -33.6 and 14. The lowest value of standard deviation is 5.7 and the maximum deviation is 20. Overall, the maximum negative deflection for individual frames was in the region of -63.9 and the maximum positive deflection was 53.7.

Noise	Minimum	Maximum
Mean	-33.6	14
SD	5.7	20
max negative signal	-63.9	-14.9
max positive signal	9.1	53.7

Table 1: Mean and dispersion of noise with 15 Hz data

3.2 Methods for reducing noise

A range of different mathematical methods are available for smoothing data and thus reducing noise. This includes methods such as the Hodrick-Prescott filter (Hodrick and Prescott, 1997) and the Kalman filter (Kalman, 1960). But each of these has shortcomings: whereas the Hodrick-Prescott filter is good for removing seasonal effects from trend data the Kalman filter requires that the distribution of noise must be known. Data can also be smoothed through approximation with spline curves or simply through temporal aggregation. These two methods have fewer shortcomings, so will be explored in detail.

The idea behind the computation of spline curves is to achieve a smoothed approximation of signal through a piece-wise defined, continuous and differentiable function. This method involves segmenting the time series into intervals and approximating it piece-wise with a polynomial of degree n. The continuous and differentiable nth-degree function is derived from individual polynomials, which are defined for each section, and used to describe the entire time series. The function then is derived from the parameterization of the single polynomials. By fitting the polynomials, the data is smoothed due to the interpolation between the different data points (cf., e.g., de Boor, 1978).

Far easier to implement and simultaneously generating similar results to approximation through spline curves is temporal compression of the data, which means averaging within particular time intervals. Table 2 and Figure 3 show examples of data with a temporal resolution of 10 Hz, 1 Hz, and spline approximated data.

	Mean	SD	Minimum	Maximum
10 Hz	-1.3	2.15	-8.33	5.1
1 Hz	-1.3	1.26	-3.1	1.29
Spline	-1.31	1.3	-4.68	1.57

Table 2: Mean, SD and variance at 10 Hz, 1Hz, and spline approximation

Table 2 shows some moments of the noise distribution when temporal aggregation and spline approximation have been applied. It shows that the standard deviation of noise is reduced by about half. The range falls from 13.43 to 4.39 with temporal aggregation and to 6.25 with spline approximation.

In order to analyze the impact of noise on the recorded valence values for emotional reactions, still images of one person with a variety of positive and negative facial expressions were strung together to create a 30 second

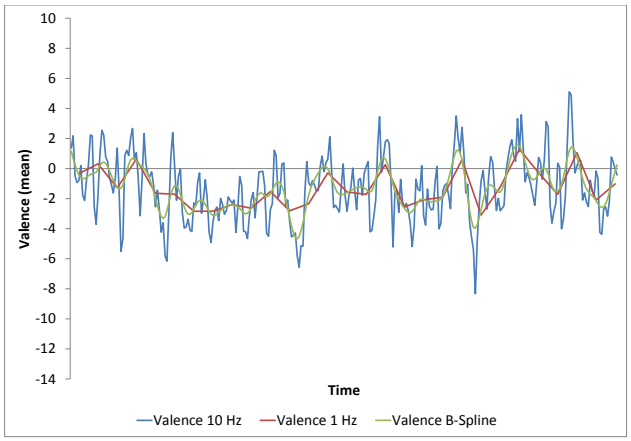


Figure 3: Reduction of noise following temporal compression and B-splines

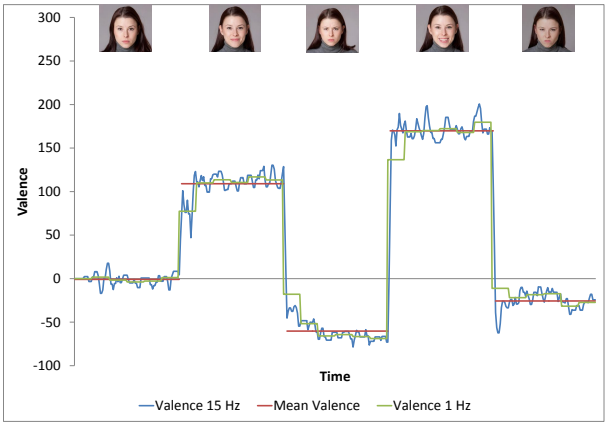
video. The still frame for each emotional state was displayed for 6 seconds. Two tapes were developed using two different actors. Each recording was calibrated to neutral facial expressions for that actor.

Figure 4 shows the noise for typical emotional reactions. Significant deflections were certainly evident between the individual fixed images, but the valence fluctuates around the average within each image. These valence fluctuations are greater around the large deflections at the phase shift. This is due to the interpolation of frames between key frames. As the video is made by using still frames, there is no continuous transition between different emotional states and thus, the interpolation by the video codec produces these distorted values. However, the interpolation distortion would be much smaller for real recordings. For example, the transition from a neutral face into a smile would be more smooth. Similarly, the measured average valence of the respective emotional expression clearly exceeds noise, which is particularly apparent for compression to 1 Hz.

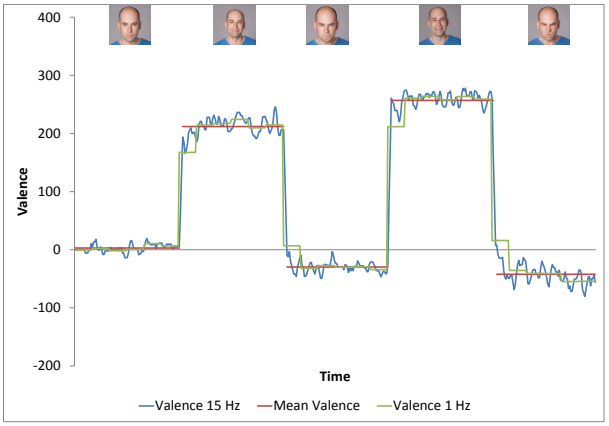
Noise over time can therefore be significantly reduced through temporal aggregation. If this method is applied to the data of each respondent, a further reduction of noise can be achieved by aggregating the data across respondents. To this end, averages across individuals are computed at each point in time. The following figures show the impact of cross-sectional aggregation for different frequencies.

In Figure 5(a), data was aggregated from 15 Hz to 10 Hz and an average was taken for all respondents. In addition, the confidence intervals ($\alpha = 0.1$) were calculated and a t-test (two-sided, two samples with heteroscedasticity, $\alpha = 0.1$) was applied to determine whether the averages significantly deviate from zero. It shows that only a few frames remain where average noise significantly deviates from zero.

If the data aggregated for respondents is further compressed to a frequency of 1 Hz (one value per second), the deflections fall even further. Statistical tests show that average values per second are no longer different from

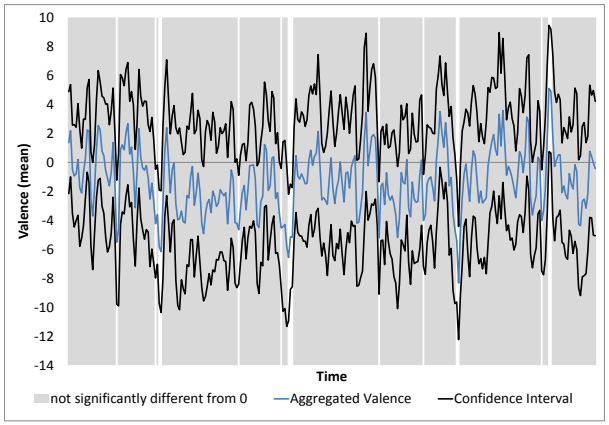


(a)

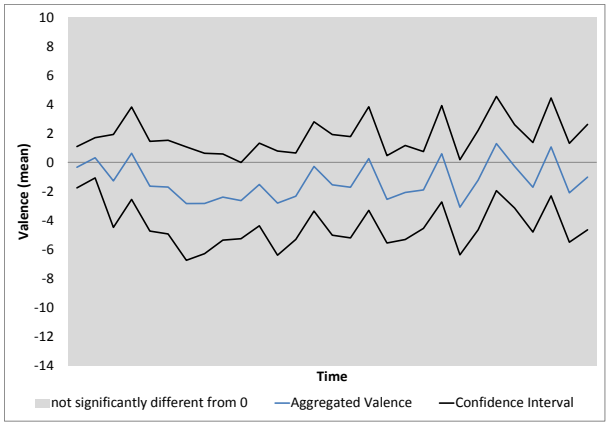


(b)

Figure 4: Noise for different emotional facial expressions



(a)



(b)

Figure 5: Reduction of noise with temporal aggregation of data to 10 Hz and 1 Hz

zero. Figure 5(b) depicts data with frequency 1 Hz data aggregated across all respondents.

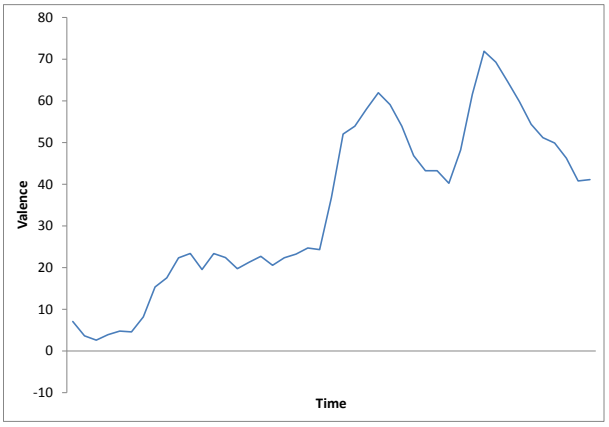
3.3 Relevance of noise in field studies

From consideration of the variance of valence in the context of real application of the software, for example an advertising test, it becomes apparent that noise can be significantly reduced in real-life scenarios. A survey in which the software was used to test different commercials will be used for comparison purposes. The study was conducted in a test studio. Respondents were recorded on a webcam while they watched TV commercials. The recordings were analyzed using GfK EMO Scan (cf. Garbas et al., 2013).

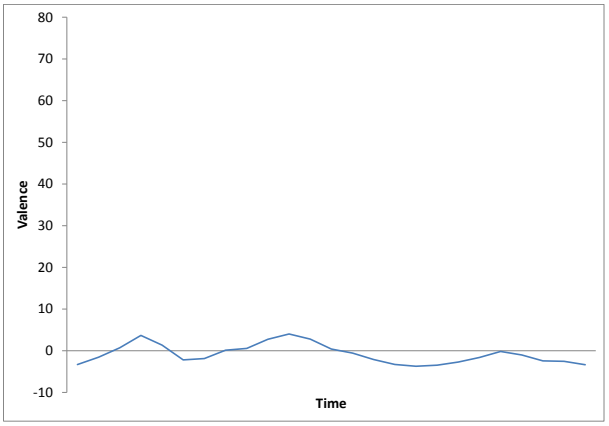
Figure 6 shows the results for two different commercials at a frequency of 1 Hz. According to our test scenarios, the individual standard deviations (at 1 Hz), and therefore the individual noise levels, are between 2.6 and 10.5 and averages at around 4.6.

Considerable deflections can be seen for the automotive commercial. The aggregated valence ranges from approximately 3 to around 72 and is about 35 over time. The standard deviation for aggregated data is around 20. It is apparent that the result clearly differs from random fluctuations. Emotional reactions that go far beyond the measure of noise are also evident at an individual level. The individual standard deviation range is between 7.5 and around 110.5. The average standard deviation across all individuals is 33.5. Of the overall sample, the share of respondents with a standard deviation of less than 10, which is only noise according to the test, is only 2.2%. The bandwidth of calculated valence ranges from around 70 as a minimum to more than 300 as a maximum. Compared with individual noise values from the test, significant emotional reactions are evident for almost all respondents in this respect.

A different picture emerges for the dish-washing liquid commercial. Here, the aggregated valence only ranged from around -4 to 4 and the average is approximately -1 with a standard deviation of around 2. Thus, there are no significant deflections and respondents do not display any emotional reactions on average when they view the commercial. On an individual basis, the share of respondents for whom the standard deviation is in the noise range is considerably higher, at 17%. Additionally, even respondents with a high standard deviation, i.e. with signals outside the noise range, show less intense emotional reactions than respondents in the automotive commercial.

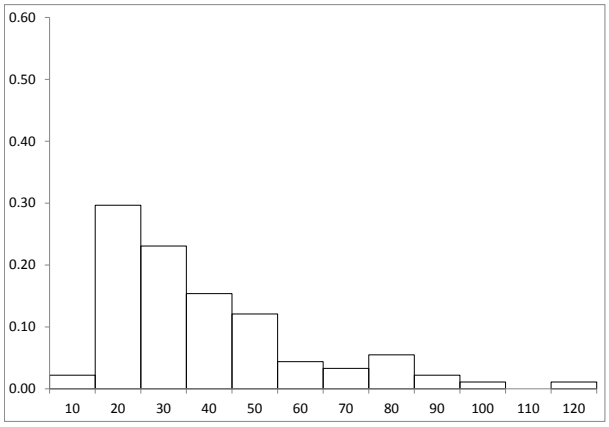


(a) results for commercial of automotive manufacturer (N=91)

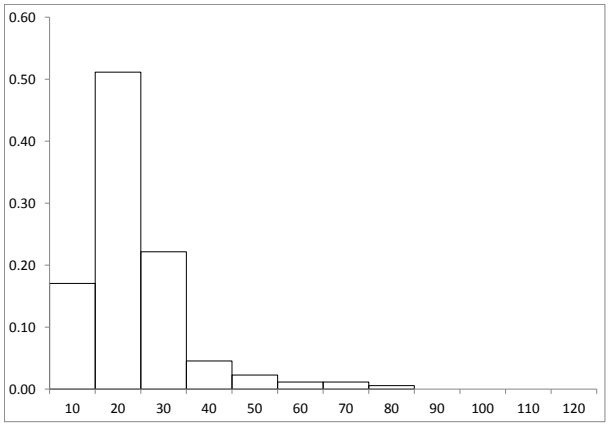


(b) results for commercial dishwashing liquid (N=176)

Figure 6: Aggregated valence for commercials; frequency: 1 Hz



(a) individual standard deviation; automotive manufacturer



(b) individual standard deviation; dishwashing liquid

Figure 7: Distribution of standard deviation for both commercials; frequency: 1 Hz

Overall, following aggregation of the data, no significant emotional reactions were detected. Figure 7 shows the distribution of individual standard deviations for both commercials.

4 Conclusion

An analysis of causes of noise and measurement distortions in the automated recognition of emotions from facial expressions was presented. Noise was extracted for a test scenario in which facial recognition software was used to analyze videos comprising still frames. It was possible to isolate noise because the videos did not include any changes in facial expression.

The analyses showed that data from automated emotion detection could be biased due to noise. However, several methods exist which can reduce and statistically eliminate noise. Some of these methods were discussed in more detail and were applied to the data of the test scenario. In particular, data was smoothed through temporal aggregation, spline approximation and cross-sectional aggregation.

It was shown that noise could be significantly reduced by applying these methods to the point that deviations did not statistically exceed zero valence.

Temporal compression to a frequency of 1 Hz is particularly effective and easy to apply. Only the means aggregated over each second have to be computed for each respondent.

If the data is then further aggregated across a sufficiently high number of respondents, noise statistically disappears, both at 10 Hz and 1 Hz.

Summing up, automated recognition of emotions from facial expressions can generate valuable insights and deliver reliable results. However, it is essential that some methodological particularities are taken into account. Noise can occur when examining individual cases, but by considering the characteristics of the data and applying a few simple methods, it can be reduced to statistical insignificance.

To this end, it is recommended to aggregate the data to 1 Hz or at least to 10 Hz, first. Secondly, this temporally aggregated data should be additionally aggregated over a sufficiently high number of observations. Although from a statistical point of view a larger sample size is necessary to obtain statistically robust and asymptotically normal distributed results (internal simulations suggest

at least N=70) cross-sectional aggregation of only 20 respondents should be appropriate to eliminate the pure technical noise from the data.

If these recommendations are considered, only a low level of noise remains. For this reason, it is advisable not to interpret valence values of between -10 and 10 for 10 Hz data and between -5 and 5 for 1 Hz data as emotional reactions but to regard them as neutral. In addition, it is important that the correct settings for elements such as video compression and the video codec are selected. However, it should not be ignored that there is a trade-off between video quality and the volume of data that has to be transferred.

References

de Boor, C. (1978). *A Practical Guide to Splines*. Springer.

Ekman, P. and Friesen, W. V. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press.

Garbas, J.-U., Ruf, T., Unfried, M., and Dieckmann, A. (2013). Towards Robust Real-time Valence Recognition from Facial Expressions for Market Research Applications. *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on. IEEE, 570–575.*

Hodrick, R. J. and Prescott, E. C. (1997). Postwar U.S. Business Cycles: An Empirical Investigation. *Journal of Money, Credit and Banking, 29(1), pp. 1–16.*

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME-Journal of Basic Engineering, 82, pp. 35–45.*

Küblbeck, C. and Ernst, A. (2006). Face Detection and Tracking in Video Sequences Using the Modified Census Transformation. *Image and Vision Computing, 24(6), pp. 564–572.*

Küblbeck, C., Ruf, T., and Ernst, A. (2009). A Modular Framework to Detect and Analyze Faces for Audience Measurement Systems. *GI Jahrestagung, ser. LNI, 154, pp. 3941–3953.*

Pantic, M. and Rothkrantz, L. J. M. (2003). Toward an Affect-sensitive Multimodal Human-computer Interaction. *Proceedings of the IEEE, 91(9), pp. 1370–1390.*

Ruf, T., Ernst, A., and Küblbeck, C. (2011). Face Detection with the Sophisticated High-speed Object Recognition Engine (SHORE). In Heuberger, A., Elst, G., and Hanke, R., editors, *Microelectronic Systems*, pages 243–252. Springer.

Slepian, D. and Wolf, J. K. (1973). oiseless Coding of Correlated Information Sources. *EEE Transactions on Information Theory, 19(4), pp. 471–480.*

Wyner, A. D. and Ziv, J. (1976). The Rate-Distortion Function for Source Coding with Side Information at the Decoder. *IEEE Transactions on Information Theory, 22(1), pp. 1–10.*

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1), pp. 39–58.*