Waldemar Bartyna[1]
ORCID: 0000-0001-5218-9579

Marcin Stępniak[2]
ORCID: 0000-0002-5113-272X

University of Siedlce
Faculty of Exact and Natural Sciences
Institute of Computer Science
ul. 3 Maja 54, 08-110 Siedlce, Poland

{[1]waldemar.bartyna,[2]marcin.stepniak}@uws.edu.pl

Data Collection and Analysis in Social Anxiety VR Therapies

DOI: 10.34739/si.2024.31.05

Abstract. Virtual reality (VR) therapy has emerged as a promising approach for treating social anxiety disorder by allowing controlled exposure to anxiety-provoking social situations in a safe virtual environment. Collecting and analyzing physiological data from participants during VR therapy sessions is crucial for monitoring their anxiety levels, tailoring the therapy, and assessing treatment outcomes. This article discusses the challenges and solutions for collecting and analyzing rapidly generated physiological data, such as galvanic skin response, heart rate, and respiration rate, in the context of social anxiety VR therapies. It explores techniques for real-time data streaming, storage in efficient file formats like Apache Parquet, and subsequent analysis of the collected data. The article also examines the role of data analysis in improving therapy protocols, personalizing treatment, and advancing our understanding of the physiological responses associated with social anxiety. By leveraging modern data engineering practices, VR therapy platforms can unlock valuable insights and enhance the effectiveness of exposure-based interventions for social anxiety disorder.

Keywords: Sensor data, Data storage, Social anxiety, Virtual reality.

1 Introduction

In our social anxiety diagnosis and therapy platform [8], we use virtual reality to simulate an environment for the therapy participant that triggers certain reactions in them. In the course of conducting therapy sessions, data is collected from physiological sensors, which can be used to monitor the current state of the therapy participant, as well as for later analysis. There are many publications addressing the topic of physiological monitoring during virtual reality therapy [18, 16]. Our solutions focus on the problems of schoolchildren, but the results developed could find wider applications.

At different stages of diagnosis and treatment, the requirements for data recording change. In the data collection phase, the most important consideration is the speed of data recording so that all information is recorded in time without data loss. Faster recording entails leaving more processing power for other components of the system, making it possible, for example, to present live data during the session. After diagnostic or therapeutic sessions, the collected data is used for analysis. In this case, the most important thing is the reading speed, but not just the sequential reading of the full data set. More important is the ability to lookup information directly in the file. This is especially important for huge data sets, on the order of gigabytes. After the analysis stage, the data is archived so that it is also available in the future. The size of the data is most important at this stage, to minimize the cost of the media used for storage. Reading speed plays a secondary role in this case, if we assume that the archived data can be read and converted back to an efficient form.

The assumptions are that the data from the physiological sensors are to be recorded continuously or at short intervals. Thus, in the event of some software or hardware failure, the stored data will be readable (we ignore the failure of the disk, data storage). Therefore, the storage format must allow the addition of new data to previously stored data. In the case of a fixed object format (data chunk, measurement), the best option is tabular formats that allow incremental addition of more rows. Other popular data formats will also be examined to confirm these assumptions.

One option for recording is to use a database. Unfortunately, due to transaction handling and integrity mechanisms, the performance of this solution will be lower than when writing directly to a file. If integrity mechanisms are disabled, the integrity of the database may be lost.

2 Physiological sensors for anxiety therapy

From the many types of physiological sensors, there are a number of choices that are particularly useful for analyzing the condition of the therapy participant. It is worth bearing in mind that a sensor that gives more precise information will not always be the best choice. For example, an EEG sensor provides such information, but its use is cumbersome due to the need to connect electrodes on the participant's head.

ECG sensors [10, 25] measure the electrical activity of the heart. ECG sensors can monitor heart rate and heart rate variability (HRV). Increased heart rate (tachycardia) occurs in stressful

situations. Decreased heart rate variability (HRV) is an indicator of sympathetic system arousal [13].

EDA sensors [9, 11] measure skin conductance, which increases in response to emotional and stressful stimuli, social stress. It can determine the overall level of emotional arousal.

BVP (Blood Volume Pulse) sensors measure changes in blood volume in blood vessels. They typically use photoplethysmography (PPG) [7], which measures changes in light absorption by tissues, which is correlated with blood flow. BVP sensors can measure pulse and provide data on heart rate and heart rate variability (HRV). Increased heart rate and heart rate variability (HRV) can be used as indicators of stress and anxiety [17].

Respiratory sensors monitor parameters related to breathing [22], such as respiratory rate, volume and airflow. They can be used to diagnose and monitor respiratory disorders such as sleep apnea. There are various ways to measure breathing parameters. Respiratory belt sensors [7] record the movement of the chest or abdomen during breathing. A spirometer [20] measures the volume and flow of air during breathing and is used to assess lung function. Airflow sensors measure airflow through the airways, often using thermistors or other measurement technologies. Monitoring respiratory rate and depth can be useful for assessing a patient's response to social stressors [26]. During the analysis phase, breathing patterns specific to the participant during anxiety situations can be determined.

EEG sensors [21] measure the electrical activity of the brain. They are used in the diagnosis and monitoring of neurological disorders such as epilepsy. They can also be used, for example, to determine a person's level of concentration.

EMG (Electromyogram) sensors [19] measure the electrical activity of muscles, which is useful in diagnosing muscle and nerve diseases. Changes in muscle tension can be observed in stressful situations.

In our platform, we use the Neurobit Optima+ 4 USB device to acquire physiological data. As the name of the device suggests, communication with the computer is via a USB port. The basic set of sensors we use measures BVP, GSR and temperature under the nose. The last parameter is used to determine respiratory rate.

Regardless of the type of device, these return, at a certain frequency (usually high), the results of measurements that our system must effectively record.

3 Available file formats

There are several file formats that allow storing different data structures. Some require that the object structure be kept the same throughout the file, and there are others that allow complete freedom in this regard. Text formats can be stored in a direct human-readable form, without the need for dedicated applications to read the file. Text formats, however, take up much more disk space. This article will examine the most popular file formats, most of which can be described as the de facto standard for exchanging information between different systems.

3.1 XML

XML (Extensible Markup Language) is a textual format used to represent various data in a structured way. XML is platform-independent, allowing documents to be easily exchanged

between heterogeneous systems. With Unicode support, it can store text in various "human" languages. XML is a standard specified by the W3C organization [6].

3.2 CSV and TSV

CSV (comma-separated values) is a format for storing data in text files, where individual records are separated by end-of-line characters and field values, according to the format's name, are separated by commas. The CSV format is described in the RFC 4180 specification [23], but there are many implementations that do not adhere to the rules there. Characters such as a semicolon or tab are often used as field separators (in such cases, the file format is sometimes called TSV - tab-separated values). The first line may contain an optional header in the same format as the regular lines of the record. This header will contain the names corresponding to the fields in the file.

3.3 YAML

YAML [5] (YAML Ain't Markup Language) is a human-readable data serialization format. It is mainly used to store configuration and exchange data between applications written in different programming languages. It uses indentation to denote data structure, which makes it intuitive, but prone to errors due to improper formatting. YAML supports complex data types such as lists, maps and scalar values. It is popular in DevOps and for application configuration.

3.4 BSON

BSON [3] (Binary JSON) is a binary data serialization format. It is more compact than JSON and allows faster data processing, making it ideal for high-performance databases. BSON supports additional data types, such as dates and binary streams, which are not natively supported by JSON.

3.5 JSON Lines

JSON Lines [4] is a file format in which each line is an independent JSON object. It is ideal for stream processing and large data sets, as it allows for easy addition and deletion of records. JSON Lines is simple and compatible with most tools for working with JSON, making it easy to process.

3.6 MessagePack

MessagePack [12] is a binary serialization format that is more efficient in size and processing speed compared to JSON. It is designed to be as compact as possible, reducing network overhead and processing time. MessagePack is ideal for applications that require high performance, such as games or mobile applications.

3.7 Apache Avro

Apache Avro [1] is a row-based data serialization format that allows both binary and textual representation of data. It is optimized for storing large amounts of data and integrating with Apache Hadoop [24]. Avro uses schemas that are stored along with the data, making it easier for different systems to read and interpret the data.

3.8 HDF5

HDF5 (Hierarchical Data Format version 5) [14] can be used to store and organize large amounts of data. It is ideal for scientific and engineering data because it allows for storing complex data structures and metadata. HDF5 is very efficient in terms of input/output operations and allows data to be stored hierarchically.

3.9 Parquet

Apache Parquet [2] is a columnar file format that is optimized for storing and processing large data sets. It is particularly popular in the Hadoop ecosystem and analytics systems because it enables efficient compression and fast access to data. Parquet supports various compression schemes and data types, which makes it flexible and efficient.

4 Comparison of file formats

The results of data writing tests using the various formats will depend on the test platform, mainly the performance of the CPU, the amount of memory and the speed of the disk. In addition, the format of the data we will write is important. More unique data will directly affect the size of the resulting file. The use of a complex hierarchical structure can affect both the writing speed and file size.

The object structure for each measurement was adapted to the types of sensors used in the system, i.e. BVP, GSR, EMG and temperature. Individual measurements will contain the following data:

Timestamp - information about the time when the measurement was taken. It is necessary to ensure that the precision is sufficient to distinguish and sort individual measurements. For example, at a frequency of 2000 samples per second, we need to record with an accuracy of 0.0001 seconds (4 decimal places). The assumption was made that in binary form the date and time would occupy 64 bits, and in text form it would be in accordance with the ISO 8601 standard [15] with a fixed UTC time zone and an additional fractional part of seconds with an accuracy of 100 nanoseconds.

Signal ID - refers to a specific record in the database, containing information about the type of sensor and additional parameters specific to the sensor during the therapy session. There will be 4 bytes for the ID values.

Value - a double-precision floating-point number. The numeric format with the specified precision will be sufficient to store readings from the selected sensors.

Signal status - can take values: valid, no signal, data loss, out-of-range. One byte will be sufficient to store all possible states.

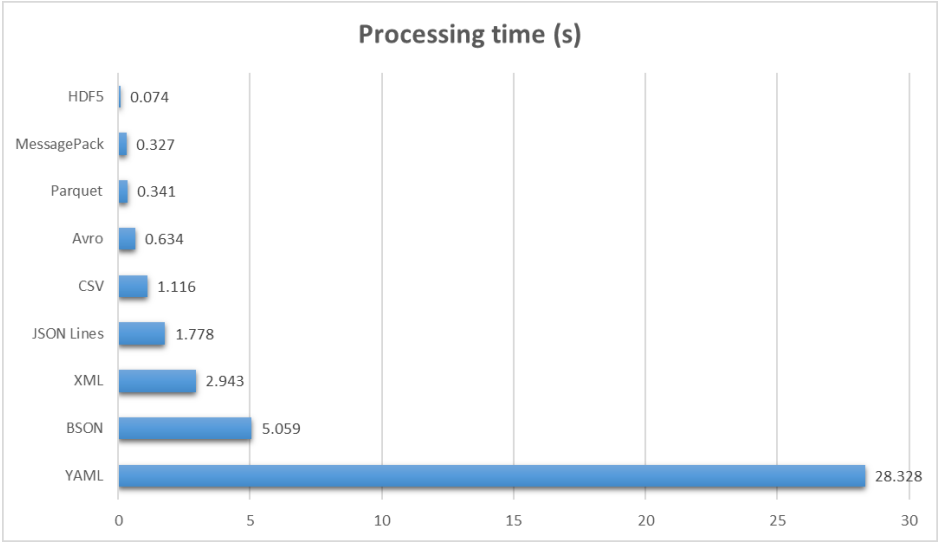

Figure 1. Comparison of the execution time of write operations

The components of the test platform were as follows:

- Operating System: Windows 11 Pro 64-bit
- CPU: Intel Core i7 @ 2.90GHz Comet Lake 14nm Technology
- RAM: 16 GB Dual-Channel DDR4 @ 1329MHz (19-19-19-43)
- Motherboard: Dell Inc. 0VG93V (U3E1)
- Storage: SSD M.2 NVMe GOODRAM PX500 512 GB

A console application written in C# on the .NET 8 platform was used to test different storage formats. In addition to the platform's standard libraries, the following libraries were used: Apache.Avro, CsvHelper, HDF5-CSharp, MessagePack, MongoDB.Bson, Newtonsoft.Json, Parquet.Net, YamlDotNet.

The tested dataset contained 2 million measurements.

Figure 1 contains average write times from 5 iterations of the test procedure. A significant advantage of binary formats over text formats can be seen. The exception here is the BSON format. What is surprising is the very long time to write a file in YAML format. Its structure is similar to the JSON Lines file, so the write time should be similar. The likely cause is a suboptimal implementation of the used library. Further analysis is needed in this regard.

Analyzing the size of the resulting files for each format (see Fig. 2), it can be seen that binary formats, which further optimize and compress the data, take up the least space. In the middle we have text and binary formats with a concise structure for each portion of the data (a small overhead of additional information). At the end, we have text files that record the names of object properties/components for each record. Therefore, at the very end, we have an XML format that encloses its elements with tags that contain their names.

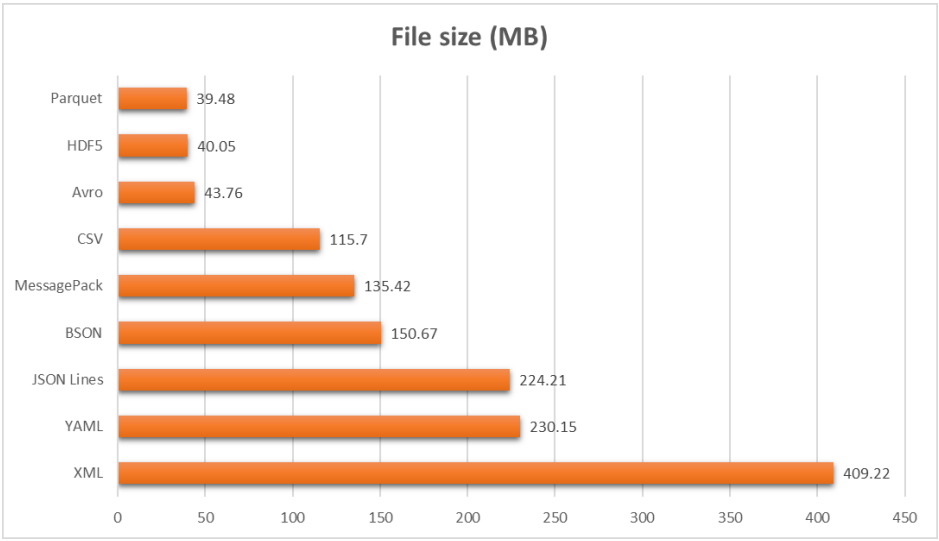

Figure 2. Comparison of the sizes of the resulting files

5 Conclusions

The undisputed winner among the analyzed formats is HDF5. With outstanding writing speed, the obtained file size is close to the best. A good choice might also be to use Avro and Parquet formats, which generate files of similar size, and the time of this operation is not significantly longer than that of the winner. If we want a text format that is human-readable, then the popular CSV format will be the best choice here. The tests conducted are only a prelude to further research. As it was mentioned, further phases of therapy require a different approach to data storage and processing. Some formats can be configured accordingly (such as the degree of optimization or compression), which will also affect the results obtained. There will also be an analysis of different implementations of file handling libraries, including those implemented by the authors.

References

1. Apache avro documentation. `https://avro.apache.org/docs/current/spec.html`, accessed: 2024-07-09
2. Apache parquet documentation. `https://parquet.apache.org/documentation/latest/`, accessed: 2024-07-09
3. Binary json specification. `http://bsonspec.org/spec.html`, accessed: 2024-07-09
4. Json lines. `https://jsonlines.org/`, accessed: 2024-07-09
5. Yaml ain't markup language (yaml). `https://yaml.org/spec/1.2/spec.html`, accessed: 2024-07-09
6. Extensible markup language (xml) 1.1 (second edition). W3c recommendation, W3C - World Wide Web Consortium (September 2006), `http://www.w3.org/TR/2006/REC-xml11-20060816/`

7. Allen, J.: Photoplethysmography and its application in clinical physiological measurement. Physiological Measurement **28**(3), R1–R39 (2007)
8. Bartyna, W., Stępniak, M.: The vr therapy platform for social anxiety treatment. Studia Informatica. System and information technology **29**(2), 27–39 (Dec 2023). https://doi.org/10.34739/si.2023.29.02
9. Boucsein, W.: Electrodermal Activity. Springer Science & Business Media (2012)
10. Clifford, G.D., Azuaje, F., McSharry, P.E.: Advanced Methods and Tools for ECG Data Analysis. Artech House (2006)
11. Dawson, M.E., Schell, A.M., Filion, D.L.: The electrodermal system pp. 159–181 (2007)
12. Furuhashi, S.: Messagepack: It's like json, but fast and small. https://msgpack.org/index.html, accessed: 2024-07-09
13. Gorman, J.M., Sloan, R.P.: Heart rate variability in depressive and anxiety disorders. American Heart Journal **140**(4), S77–S83 (2000)
14. HDF Group: Hdf5 user's guide. https://portal.hdfgroup.org/display/HDF5/HDF5, accessed: 2024-07-09
15. International Organization for Standardization: Iso 8601-1:2019 - date and time — representations for information interchange — part 1: Basic rules. ISO 8601 (2009)
16. Kampmann, I.L., Emmelkamp, P.M., Morina, N.: Meta-analysis of technology-assisted interventions for social anxiety disorder. Journal of Anxiety Disorders **42**, 71–84 (2016)
17. Kim, J., Cheon, E.J., Bai, D.S., Lee, Y.H., Koo, B.H.: Stress and heart rate variability: A meta-analysis and review of the literature. Psychiatry Investigation **15**(3), 235–245 (2018)
18. Kothgassner, O.D., Felnhofer, A., Hlavacs, H., Beutl, L., Palme, R., Kryspin-Exner, I., Kastenhofer, E.: Salivary cortisol and cardiovascular reactivity to a public speaking task in a virtual and real-life environment. Computers in Human Behavior **62**, 124–135 (2016)
19. Merletti, R., Parker, P.A.: Electromyography: Physiology, Engineering, and Non-invasive Applications. John Wiley & Sons (2004)
20. Miller, M.R., Hankinson, J., Brusasco, V., Burgos, F., Casaburi, R., Coates, A., et al.: Standardisation of spirometry. European Respiratory Journal **26**(2), 319–338 (2005)
21. Niedermeyer, E., da Silva, F.L.: Electroencephalography: Basic Principles, Clinical Applications, and Related Fields. Lippincott Williams & Wilkins (2004)
22. Quanjer, P.H., et al.: Lung volumes and forced ventilatory flows. report working party standardization of lung function tests, european community for steel and coal. official statement of the european respiratory society. European Respiratory Journal Supplement **16**, 5–40 (1993)
23. Shafranovich, Y.: Common Format and MIME Type for Comma-Separated Values (CSV) Files. RFC 4180 (Informational) (October 2005), http://www.ietf.org/rfc/rfc4180.txt
24. Shvachko, K., Kuang, H., Radia, S., Chansler, R.: The hadoop distributed file system. In: Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on. pp. 1 –10 (May 2010). https://doi.org/10.1109/MSST.2010.5496972
25. Thayer, J.F., Lane, R.D.: A model of neurovisceral integration in emotion regulation and dysregulation. Journal of Affective Disorders **61**(3), 201–216 (2000)
26. Wilhelm, F.H., Gevirtz, R., Roth, W.T.: Respiratory dysregulation in anxiety, functional cardiac, and pain disorders. Assessment **8**(3), 251–270 (2001)