Compromise

Data compression paradigm based on omitting self–evident information

Maribor, 29. 06. 2023

David PODGORELEC

University of Maribor

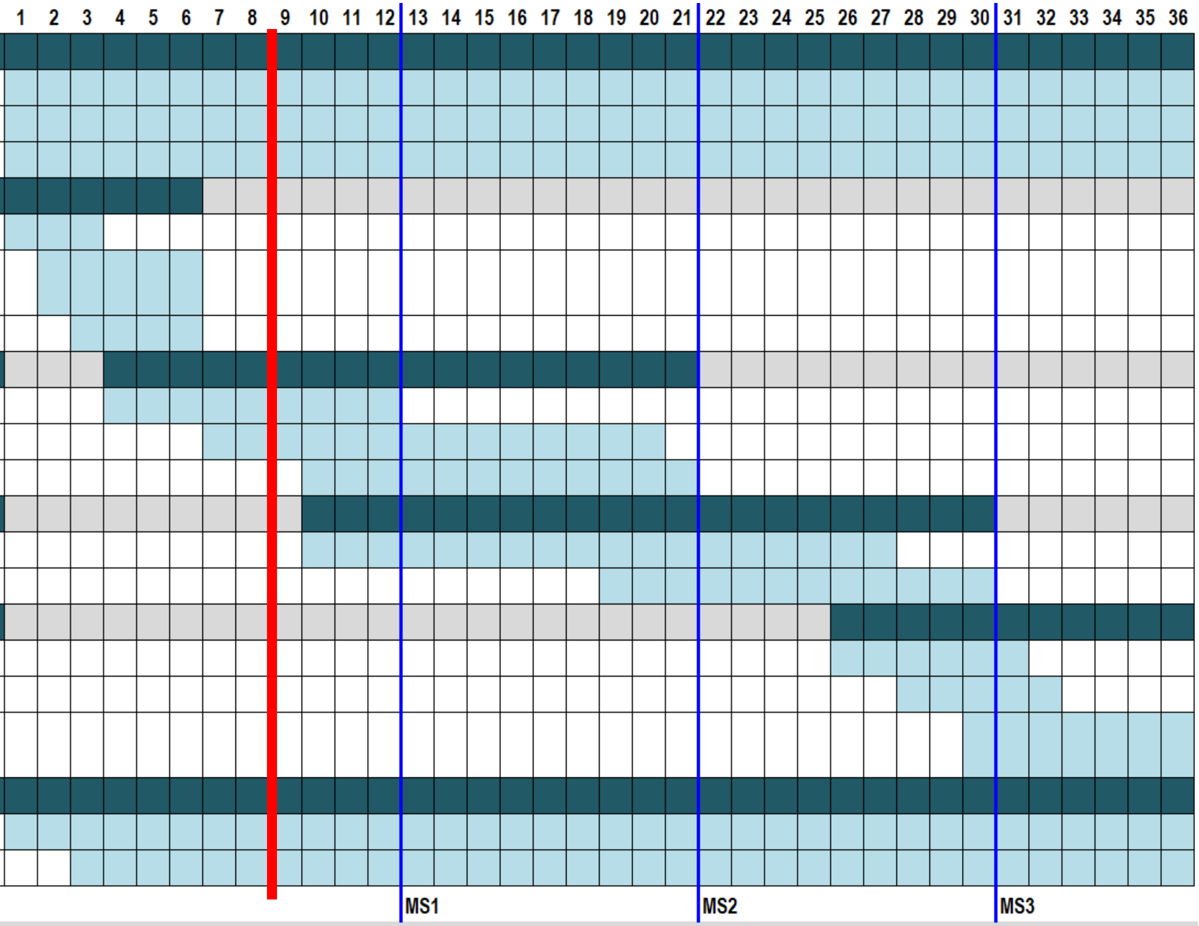
Faculty of Electrical Engineering and Computer Science

Institute of Computer Science

Laboratory for Geospatial Modelling, Multimedia and Artificial Intelligence

Workplan

WP	T	Work package/task title	Start	End
1		**Project management**	**1**	**36**
	1	Administrative and financial project management	1	36
	2	Quality assurance and risk mitigation	1	36
	3	Legal, data and knowledge management	1	36
2		**Definitions and unified taxonomy of features**	**1**	**6**
	1	Generation of domain-dependent feature repertoires	1	3
	2	Definition of feature descriptions and development of methods for their interpretation	2	6
	3	Specification of domain-independent feature taxonomy	3	6
3		**Feature detection, compression, and data restoration**	**4**	**21**
	1	Feature detection	4	12
	2	Data restoration and residual determination	7	20
	3	Lossless compression of features and residuals	10	21
4		**Feature selection and optimised residual determination**	**10**	**30**
	1	Feature selection	10	27
	2	Integration of feature selection and residual determination	19	30
5		**Component integration and hypothesis testing**	**26**	**36**
	1	Adaptation of SOTA methods for comparison	26	31
	2	Component integration	28	32
	3	Analysis of results, iterative improvements of methodology, and hypothesis testing	30	36
6		**Dissemination, exploitation, and communication**	**1**	**36**
	1	Dissemination, exploitation, and communication strategy	1	36
	2	Dissemination activities	3	36



MS1 Proof of concept **MS2** The first operational prototype based on redundant feature set **MS3** Optimized system based on selected features

Deliverables

- At least 3 papers in international open access journals (1)
- at least 6 conference papers (0)
- organisation of 2 dedicated presentation events (ToDo)
- 1 patent application (ToDo)
- Website (OK) and a profile on at least 1 social network (ToDo)
- ongoing results at the end of individual WPs
  - D2.1 – Domain-dependent feature descriptions in all four testing domains [M6]. (in individual applications – to be collected and make domain-independent catalogue)
  - D2.2 – Specification of domain-independent superclasses in a unified feature taxonomy [M6]. (OK)
  - D3.1 – Programming library (API) with sufficient functionality for system operation [M21]. (not in this shape, but OK)
  - D3.2 – Test program for library verification [M21]. (individual applications – OK)

Objectives

Hypothesis: The universal methodology of lossless or near-lossless data compression, which will be based on unified feature taxonomy and restoration methods, will be more efficient than the existing compression procedures for raster images, digital audio, biomedical signals, and sparse voxel grids.

SO1	To develop a universal data compression methodology with a unified taxonomy of features from diverse domains, and a common framework for lossless, near-lossless, and lossy compression.
SO2	To upgrade the prediction of original data by integrating the techniques of feature selection and data restoration.
SO3	To improve the compression ratios in lossless and near-lossless mode in comparison with the existing approaches.
SO4	To improve the accessibility and reusability of features and feature-based restoration.
SO5	To deliver a verification environment for hypothesis testing in four pilot domains: raster images, digital audio, biomedical signals, and sparse voxel grids.
SO6	To disseminate the project results.

KPIs

	Means of achieving objective	KPIs
SO1	By defining domain-specific feature descriptions and classifying them in a domain-independent taxonomy; by developing the procedures for feature detection in source data; by selecting and upgrading domain-independent feature and residual representations before compression; by determining the criteria and developing the procedures for error control in near-lossless and lossy compression.	Specification of parameters and interpretation of all features in the selected test domains; specification of a unified taxonomy, which categorises all types of domain-specific features into generic domain-independent classes; a feature set (upon each test) whose size, after categorisation and before feature selection, does not exceed 10% of original data stream; specification of differences between lossless, near-lossless, and lossy compression.
SO2	By developing the procedures for optimised selection of a detected feature subset; by introducing a domain-independent methodology of restoration from features for predicting data and residual values; by integrating the methodology and feature selection into a joint optimisation procedure.	Representation of residuals with a range that is at least 80% smaller than in the source data; the size of the optimised feature set will not exceed 0.5% of the original data stream.
SO3	By analysing, upgrading, and hybridising the established approaches for lossless compression (e.g., LZW, RLE, arithmetic coding) with the goal of achieving the highest possible compression ratios; by adapting these upgrades and hybrid methods to near-lossless compression.	Lossless compression of raster images to 15-30%, digital audio to 10-50%, biomedical signals, with the focus on EEG/ERP, to 10-30%, and sparse voxel grids to 20-50% of original size. The average results in all data domains will be at least comparable to the existing SOTA methods, considerably better in individual cases, and never significantly worse.

KPIs

	Means of achieving objective	KPIs
SO4	By finding the compromise between SO3 (compression ratio) and time efficiency of feature and residual decoding; by domain–dependent optimisation of restoration methods, based on the domain-independent methodology from SO2.	Feature recall that is 20-times faster than derivation from the original data stream, and comparably fast recall of restored data. In all test domains the decoding of compressed features and the reconstruction of data will be at least 5-times faster than encoding.
SO5	By selecting relevant test datasets; by choosing the compression and reconstruction efficiency metrics (compression ratio, speed, restoration quality); by selecting SOTA methods for comparison; by integrating the solutions of SO1-SO4 into a common framework and developing a test application.	Validation of KPIs for SO1-SO4 in all four pilot domains.
SO6	By publishing in scientific journals and conferences, and by organising public presentations.	At least 3 papers in open access journals, 6 conference presentations, and 1 patent application. Successful demonstration of methodology in at least 2 public events.