

Contents

Resumen	iii
Contents	xv
1 Introduction	1
1.1 Motivation	1
1.2 Scientific goals and research hypotheses	8
1.3 Structure of the thesis	11
2 Towards achieving a high degree of situational awareness and multimodal interaction	13
2.1 Introduction and related work	14
2.2 Architecture overview	17
2.3 Visual interpretation and validation	23
2.4 NLP using chatbots	26
2.5 NLP using transformers with questions and answers	28
2.6 ML for anomaly detection	31
2.7 General multimodal AR approach	32
2.8 Experimental setup and evaluation	35
2.9 Conclusions	41
3 Environment awareness, multimodal interaction, and intelligent assistance	45
3.1 Introduction	46
3.2 Related work	49

3.3	System architecture overview	51
3.4	System implementation	57
3.5	System evaluation	63
3.6	Analysis and results	66
3.7	Discussion	69
3.8	Conclusions	70
3.9	System evaluation	73
4	Large Language Models for in situ knowledge documentation and access with AR	77
4.1	Introduction	78
4.2	Background and context	82
4.3	Proposed system	86
4.4	Evaluation and results	93
4.5	Conclusions	97
4.6	Appendix	100
5	Conclusions	103
5.1	Conclusions	103
5.2	Future work	107
5.3	Scientific contributions	108
	Bibliography	109

List of Figures

1.1	Early AR head-mounted display system, showcasing components for visual display, head tracking, and voice command input. (Caudell and Mizell 1992)	2
1.2	Virtual Reality Continuum. (Milgram et al. 1995)	2
1.3	Projection-based AR. (Pejsa et al. 2016)	3
1.4	Example of and Eye multiplexed AR application. (Google)	3
1.5	The nine pillars of Industry 4.0. (Kadir 2020)	4
1.6	HCI paradigms. (Rekimoto 1995)	5
2.1	General architecture overview	18
2.2	Detailed diagram of the physical and semantic layers	19
2.3	Using a CNN with regression to interpret the values of a pressure gauge	24
2.4	Classification example	25
2.5	Regression example	25
2.6	The operator can ask questions in natural language about this machine. The AR system gives information regarding what element is the operator in front of, so the question is complemented with the required context	27
2.7	Some examples of real questions using a manual of a pressure gauge	31
2.8	The combination of several values can be seen as standard or as an anomaly, and visual cues are possible in AR	33
2.9	When the machine is switched on, the app lets the operator move to the next step	34
2.10	Automatic value extraction from a pressure gauge	35
2.11	Tasks comparison box plot	41
2.12	Classification and regression CNN architectures	43

3.1	System architecture layers.	51
3.2	Reading the pressure value with a CNN in non-sensorized machines.	54
3.3	Users can interact multimodally to obtain AR and NL feedback.	56
3.4	Linking answers in NL to AR cues.	58
3.5	Scanned mesh (Unity).	59
3.6	Link between a physical element and transformer identifier (Unity).	60
3.7	Extruder workflow using the application, setting the nozzle temperatures. Visual validation, AR guiding, and voice interaction.	62
3.8	Comparative of tasks, group A (Semantic AR) Vs. group B (Only AR) Vs. group C (Traditional).	73
3.9	Comparative of answers (Group A Vs. Group C).	74
4.1	SMEs annotate scanned environments.	88
4.2	Shop floor operator retrieves anchored information via NL query or area selection.	89
4.3	Two shop floor roles: SMEs add information, operator retrieves it via NL/AR.	90
4.4	Time comparison between groups A (No app) and B (With app).	96
4.5	Rest API calls.	101

List of Tables

2.1	Using a transformer to ask questions in NL in technical documentation . . .	30
2.2	Task completion times without and with semantic layer	39
2.3	Descriptive and statistical contrasts	40
3.1	Transformer predictions to questions (Intel/bert-large-uncased-squadv1.1-sparse-80-1x4-block-pruneofa).	66
3.2	Scores for the 4 QA transformers tested.	66
3.3	Descriptive execution times of tasks and statistical contrasts.	67
3.4	Two to two comparisons P -values (Bonferroni correction).	67
3.5	Descriptive answer times and statistical contrasts.	68
3.6	Descriptive and comparative scores to the questions raised about the performance of the tasks.	68
3.7	Task time results (Group A Vs. Group B Vs. Group C).	75
3.8	Question time results (Group A Vs. Group C).	75
4.1	GPT-JT tests label information as "new", "redundant", or "contradictory" based on context.	94
4.2	Comparison of time between groups A and B, along with the corresponding p-values	98
4.3	Likert questionnaire	99