SOCRAR: Semantic OCR through Augmented Reality

Jannis Strecker
jannisrene.strecker@unisg.ch
University of St. Gallen
St. Gallen, Switzerland

Kimberly García
kimberley.garcia@unisg.ch
University of St. Gallen
St. Gallen, Switzerland

Kenan Bektaş
kenan.bektas@unisg.ch
University of St. Gallen
St. Gallen, Switzerland

Simon Mayer
simon.mayer@unisg.ch
University of St. Gallen
St. Gallen, Switzerland

Ganesh Ramanathan
ganesh.ramanathan@student.unisg.ch
University of St. Gallen
St. Gallen, Switzerland

ABSTRACT

To enable people to interact more efficiently with virtual and physical services in their surroundings, it would be beneficial if information could more fluently be passed across digital and non-digital spaces. To this end, we propose to combine semantic technologies with Optical Character Recognition on an Augmented Reality (AR) interface to enable the semantic integration of (written) information located in our everyday environments with Internet of Things devices. We hence present SOCRAR, a system that is able to detect written information from a user's physical environment while contextualizing this data through a semantic backend. The SOCRAR system enables in-band semantic translation on an AR interface, permits semantic filtering and selection of appropriate device interfaces, and provides cognitive offloading by enabling users to store information for later use. We demonstrate the feasibility of SOCRAR through the implementation of three concrete scenarios.

CCS CONCEPTS

• **Human-centered computing → Mixed / augmented reality**; **Ubiquitous and mobile computing systems and tools**; • **Computing methodologies → Knowledge representation and reasoning**; **Computer vision tasks**.

KEYWORDS

Web of Things, Optical Character Recognition, Augmented Reality, Knowledge Graph, Ubiquitous Computing

ACM Reference Format:
Jannis Strecker, Kimberly García, Kenan Bektaş, Simon Mayer, and Ganesh Ramanathan. 2022. SOCRAR: Semantic OCR through Augmented Reality. In *Proceedings of the 12th International Conference on the Internet of Things (IoT '22), November 7–10, 2022, Delft, Netherlands.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3567445.3567453

1 INTRODUCTION

In our professional and private lives – ranging from smart home control to industrial automation – we have access to a growing number of services: More and more sensors – even low-cost batteryless sensors that may run for decades [10] – provide us with information on our context and we may utilize virtual and physical functionality through the interfaces these services provide to us. It is then often desired to integrate data processing flows across services (i.e., service mashups) that fulfill a specific user goal.

In this context, we ask how humans might be enabled to interact more efficiently with (virtual and physical) sensors and actuators in their surroundings, and specifically how we might create more fluid transitions between these services and information that is available to humans in non-digital form. To accomplish this fluidity, we propose a system that combines semantic technologies with Optical Character Recognition (OCR) on an Augmented Reality (AR) interface: Our system thereby is enabled to *semantically lift* non-digital information in the field of view of the user and thereby to associate it with available services in the user's surroundings.

The adoption of Knowledge Graphs (KGs) [12] and the expansion of semantic technologies research has gained traction in academia and industry in recent years. This development is driven by the vast (and heterogeneous) data that is being produced by humans and physical objects (e.g., sensor and actuators), and is leading to a proliferation of controlled vocabularies and standards within and across domains. The objective of semantic technologies is to create common machine-readable and machine-understandable descriptions of the data that heterogeneous systems produce and consume, thereby enabling them to interoperate. Building on top of the Internet of Things (IoT), the Web of Things (WoT) takes semantic descriptions as one of its pillars to bring sensors and actuators to the Web. Through semantic descriptions, specifically descriptions that follow the World Wide Web Consortium's (W3C) WoT Thing Description (TD)[1] standard, the WoT achieves uniform machine-understandable descriptions of device interfaces and supports binding to a variety of protocols (e.g., CoAP, MQTT, OPC-UA). TDs can be enriched using ontologies and may utilize well-known vocabularies to further contextualize the *Thing* that is being described, as well as its inputs and outputs.

These semantic descriptions are readily used by automation systems, e.g. during the discovery of a required functionality to accomplish a specific task. Thus, we propose that *human interfaces* should also make use of such semantic descriptions to better support users. This can be done, for instance, by adapting the value read by a sensor to the system of units (e.g., imperial or metric) that best fits the user's context, or by contextualizing data located in the

IoT '22, November 7–10, 2022, Delft, Netherlands
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9665-3/22/11.
https://doi.org/10.1145/3567445.3567453

[1] https://www.w3.org/TR/wot-thing-description

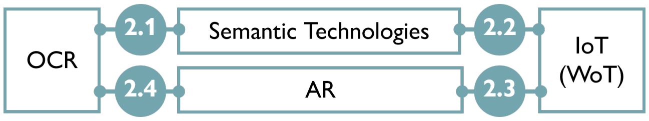


Figure 1: Outline of the SOCRAR related work.

user's physical environment which requires further information to be understandable: A specific reading of a CO2 sensor measured in parts per million (ppm) might be shown to expert users in this very form, while it is automatically converted and presented as *low*, *medium*, or *high* to other users, thereby providing more meaningful information to them. In this context, we propose that combining AR – as *human interface* – and OCR – for extracting characters from digital images that are captured by an AR headset's camera [23] – with semantic technologies and W3C WoT TDs can provide people with useful contextual support in pervasive computing environments.

Our main contribution is the integration of semantic technologies, W3C WoT TDs, OCR, and AR, to create a system capable of detecting data from a user's physical environment, contextualizing it on a semantic backend and communicating useful information to a user through an Augmented Reality interface.

In the following, we present an overview of research that is related to our contribution; we then provide several motivating scenarios in Section 3 – these relate to the potential impact of the SOCRAR system on end users, from helping them better understand information located in their physical environment that today lacks context, to decreasing a user's cognitive load by helping her remember data located in the physical world that might function as an input to another device or service. We subsequently present the system's architecture and implementation in Section 4 and discuss the demonstrated features, future applications, and limitations of our approach and system in Section 5.

2 RELATED WORK

SOCRAR integrates contributions from several fields, namely the IoT, the WoT, semantic technologies, OCR, and AR. Thus, we discuss relevant works in Sections 2.1–2.4 (see Fig. 1).

2.1 OCR and Semantic Technologies

From early-on, the OCR community has considered the semantics of recognized texts. Some research has focused on improving the accuracy of OCR algorithms through semantic enrichment, while others focus on enabling a range of post-OCR tasks. Given such a range of tasks and applications, the interpretation of what *semantics* encompasses differs across the related work.

Concerning research to improve the accuracy of OCR algorithms, Jobbins et al. [14] propose a method that uses external knowledge from a thesaurus. Given such information, the semantic similarity among a set of recognized candidate words is computed and used to propose other words from the thesaurus that have higher semantic similarity. Broda and Piasecki [6] propose a system to improve a handwriting OCR algorithm for medical records in a similar way, by using semantic similarity among the OCRed words. Here, different semantic similarity measures (e.g., cosine, information radius) are explored. However, this system does not need a thesaurus, instead it introduces heuristics for ad-hoc terms such as abbreviations and names of specific drugs.

Park et al. [26] apply OCR to restaurant and shopping receipts: Following a set of guidelines and a pre-defined taxonomy of the different sections in a receipt (e.g., name of the business, items purchased, subtotal, total), people manually draw bounding boxes around these sections and select the correct label (e.g., *store name*) thereby contextualizing the recognized characters. Jian et al. [13] motivate their research with the hesitancy to apply NLP techniques to books that have been digitized through OCR due to OCR-induced errors that might persist in combination with the lack of transparency of the applied NLP methods. Concretely, that contribution presents an exemplar study that analyses how BERT (*Bidirectional Encoder Representations from Transformers*) embeddings are able to encode semantic information of books at the chapter level on OCRed books vs. baseline books. The early results of this work suggest that NLP techniques, specifically BERT, could be utilized on OCRed texts with some success.

Wang et al. [39], propose an algorithm for OCR-based image captioning that takes advantage of not only the text that is recognized in an image, but also of the size and position (i.e., geometric relationship) of recognized OCR tokens. Through a *Long Short-Term Memory plus Relation-aware* (LSTM-R) pointer network architecture, the authors show that their algorithm performs better than others that do not incorporate such geometric relationships. Regarding the meaning of the tokens, semantic features of the recognized words are considered as the likelihood that they might appear in the current context.

Given this existing promising work, we see great potential in pursuing the integration of OCR approaches with semantic technologies – specifically KGs – and posit that this would enable the creation of applications that take advantage of rich domain knowledge that has been made available in a machine-understandable way. Moreover, it could allow not only for integrating data from a user's environment but also to control services around them.

2.2 Semantic Technologies and the IoT (WoT)

Semantics have also increasingly started to play a role in the IoT and the WoT. Over the past decades, we have witnessed extensive research and development of technologies to enable sensing, computation, actuation in connected everyday things, thus bringing the vision of an IoT ever closer to reality. The availability of low-cost computation and networking capabilities has driven the application of the IoT to use cases ranging from consumer devices to industrial processes [20]. However, out of the several challenges already visible a decade ago [11, 20], the problem of technical (e.g., proprietary communication protocols) and semantic (e.g., proprietary vocabularies and information models) interoperability of heterogeneous devices still presents an obstacle to the world-wide integration of connected devices and services, which would ideally stretch across domains. This integration is, today more than ever, relevant as more and more automated clients – autonomous systems – are starting to populate digital environments alongside humans [8].

The focus on interoperability prompted researchers to be inspired by the Web architecture, which has successfully demonstrated that it can bring together IoT devices in a scalable (including on low-power devices) [15] and open [22] way. Beginning with a vision of integrating *Things* as a part of the Web, a recent standardization effort by the W3C has resulted in the creation of means to describe both the semantic context of things and the interaction possibilities offered by them [7]. Thus, the W3C WoT provides a guiding abstract architectural framework[2] which can be used for the implementation of domain-specific solutions. Moreover, the WoT enables the semantic description of devices and services through the Thing Description (TD)[3] specification. By specifying and linking to its TD, a device can for example convey that it is an instance of the `<http://www.w3.org/ns/sosa/Sensor>` class from the Semantic Sensor Network Ontology (SOSA)[4] and that its outputs are of type `<http://qudt.org/vocab/unit/DEG_C>` which is a concept from the QUDT[5] ontology. The sensor's machine-understandable description may then be linked to other descriptions, such as one corresponding to the room in which the sensor is located, perhaps using the Brick[6] ontology. Given this integration, a building automation program in charge of keeping a specific space at a comfortable temperature can utilize the current sensor value, plus the unit used to measure the temperature and the place in which the sensor is located to trigger an action in another physical device, e.g., turning on the air conditioning unit in the correct space.

The semantic description increases the discoverability and reuse of the sensor, and permits semantic integration that dissolves syntactic tight coupling. Concretely, the semantic linking of the sensor's functionality with the requirements of a consumer of its data and the protocol binding of the sensor's TD allows a sensor of this kind to be replaced or upgraded at run time and without causing any disruption, this could prevent having to re-engineer the building automation program.

The application of Web and Semantic Web principles to IoT systems that were hitherto focused on connectivity, is becoming a reality with the adoption of the WoT. The integration of knowledge-based reasoning into (industrial) devices has enabled not only first semantics-based integrations [17], but also applications such as knowledge-driven automated fault detection [28], as well as multiagent-based systems capable of reasoning to achieve their goals, in application fields such as smart farming [29].

2.3 IoT (WoT) and AR

Our proposed approach, SOCRAR, focuses on the human element and on how we might facilitate the interaction of humans with their smart environments. Three decades ago, Mark Weiser envisioned a seamless integration of networked (micro-) computers and displays into the physical world [41]. Weiser's vision is progressively blending in our daily activities and various academic agendas [1, 30]. Earliest Augmented reality (AR) applications go back to the 1960s [34], however only three decades ago AR could truly become a viable research field [2]. Today we have access to technologies that reduce seams between computers, humans and their environment while making Weiser's vision more tangible. Thus, expectations on the IoT (WoT), AR, and on the combination of these technologies are growing [25, 40].

We find joint efforts to develop immersive displays as seamless interfaces to IoT devices that allow human users to observe, control, and interact with systems in smart environments such as vehicles, homes, farms, industrial shopfloors, and cities [24, 35]. For example, Garcia-Macias et al. developed *UbiVisor*, a prototype browser for IoT devices [18]. The mobile client of the UbiVisor captures the context (e.g., humidity and temperature for a plant) using QR and RFID tags. Then its server makes inferences with the help of semantic models. Finally, it informs the user on an AR display. Mayer et al. present a system that records interactions between IoT devices in a central logging backend [21] for enabling users to observe the causes and effects of interactions among those devices either on a screen or an AR frontend. More recently Zheng et al. proposed STARE, a semantic AR decision support framework [42]. In a smart home context, the results of a user experiment show that STARE reduces information overload and improves the interpretation of IoT data with decision support explanations and information visualizations. While AR headsets are mainly used as an enabler of information visualization in human-computer interaction, we note that they can be employed as well for the detection of biometric signals from users (e.g., eye movements) [25], detection of physical objects (through visual object classification) [33], and recognition of other visual information that is available in the surroundings of users.

2.4 AR and OCR

Since most AR headsets also include forward-looking cameras, the combination of AR applications and OCR is compelling. On AR headsets, OCR has in the past been utilized to detect texts on physical documents, digital displays, or handwritten documents in various domains, such as traffic surveillance or to support the digitization of historical documents [23]. Other research has included OCR in AR applications for translation purposes [36]. For instance, Toyama et al. developed a system in which a user wears a see-through head-mounted display (HMD) that overlays Japanese texts with English translations [37]. HoloDoc displays additional digital information alongside physical documents with the help of AR [16]. Beyond mere visualization for its human user, this work also uses OCR to search for additional information about recognized words using an online search engine. Similarly, Rahman et al. proposed a system to support data collection processes in aquaculture farming [27]. There, OCR was employed on AR glasses to read data from a sensing unit; this data is then sent to a cloud server for processing. However, to overcome errors in the recognition of units in the displayed area, the team implemented a post-processing algorithm that corrected the OCR results.

3 MOTIVATION

To our knowledge and given the state of the art across these four research domains, there has been no research or system yet that leverages the WoT together with semantic technologies, AR, and OCR in order contextualize the users interactions with available connected devices in their environment. In the following, we present three scenarios that motivate our work and suggest the benefits it might bring to users.

[2] https://www.w3.org/TR/wot-architecture
[3] https://www.w3.org/TR/wot-thing-description
[4] https://www.w3.org/TR
[5] https://qudt.org
[6] in https://brickschema.org

Ben, the exchange student. Ben is an exchange student from the USA, who is spending a semester in France. Ben is excited about his time abroad but there are frequently encountered differences that inconvenience him. For example, the temperature display in his room and other displays in different classrooms show the temperature in degrees Celsius. Such details that delay Ben's immediate understanding of his surroundings are everywhere, e.g., on the street signs showing the speed limit, on the food and beverages he buys, and on the recipes he wants to cook. Ben of course knows how to convert most metric units to imperial ones, which are naturally more meaningful to him. However, he needs time and effort to compute such conversions every time he wants to make a decision. Ben is lucky to have a modern dorm room equipped with devices and appliances that can also be controlled at a fine-grain granularity, for example he can control the color of the lights and their intensity and can select an ambient mode. He just needs to learn a few french words. What if Ben could simply use his AR-enabled device capable of contextualizing his field of view and showing data that is meaningful to him?

Carol, the facility manager. Imagine a common IoT scenario, such as Building Automation (BA) systems, which ensure that living and working spaces are comfortable and secure for their occupants. The requirements for space comfort vary widely according to its usage, e.g., office buildings, hotels, manufacturing plants, and laboratories. BA systems are often programmed to operate according to specific purposes. However, the purpose of a specific space might change over time according to the current users' needs (e.g., a sport hall can be used as a conference venue). During the operation of a building, Carol's job is to audit the operation of the BA system, where she must be informed of the intended usage of the different spaces in the building, so she can make sure that the corresponding comfort requirements are met. Currently, Carol's audits are done manually: She browses her physical files and selects the checklists that detail the standard comfort requirements of each space she needs to audit considering the specific usage of the space. Then, Carol physically visits each space, takes measurements of the current conditions, and compares them with the requirements specified on her checklists. Such manual auditing tasks are cumbersome and time-consuming, resulting in increased costs. Could there be a system that makes Carol's work more efficient and allows her to take advantage of the connected sensor on the environment (which log the data but do not have a physical user interface) so she does not have to take redundant measurements in every space she checks?

Diana, the machine operator. People tend to delegate parts of cognitive processes to external media in their environment [9] ranging from paper notes to digital repositories. Such *cognitive offloading* is defined as "the use of physical action to alter the information processing requirements of a task so as to reduce cognitive demand" [31]. Especially in cognitively demanding situations (e.g., an operator working on the shop floor) people need to note down relevant details on physical or digital media (post-its, shopping or to-do lists, etc.) to memorize and recall them later. In her job as a machine operator, Diana is responsible for collecting parts of a machine tool from a stockyard, assembling the tool from these parts, and measuring the tool's offset values. The measured values then need to be entered in a milling machine that uses them to process raw material. Even though the entering of (even slightly) wrong offset values can incur large cost – specifically, damaging the raw material, the machine tool, or even the machine itself – Diana often does not write down the measured values, since this could delay her operations. Her daily work would be significantly simplified, and significant errors could be avoided, if she was equipped with a system that would remember the measured offset values for her, for simple recall when configuring the milling machine.

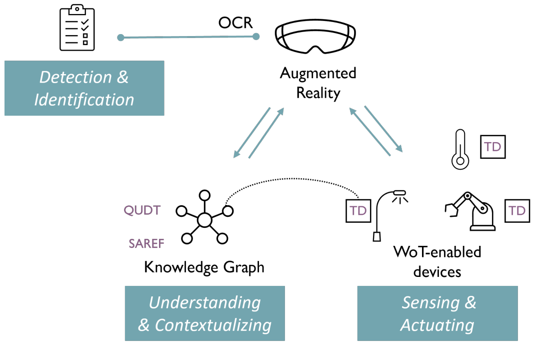


Figure 2: SOCRAR's three modules and their interactions with an AR-enabled device. With SOCRAR, we integrate OCR, AR, semantic technologies, and W3C WoT environments.

4 ARCHITECTURE AND IMPLEMENTATION

To create SOCRAR, a system capable of contextualizing data in a user's field of view, we propose to combine OCR, AR, semantic technologies, and W3C WoT-enabled devices. These technologies interact through three main modules (see Fig. 2): a detection and identification module, an understanding and contextualization module and a sensing and actuation module. The functionality from these modules is integrated through an AR HMD.

Detection and Identification. This module runs on a Microsoft HoloLens 2 (HL2) and uses the Windows Runtime OCR API[7]. An image from the HL2's camera serves as an input for the OCR engine which outputs the text it recognized on the image, splits it into lines and then into individual words. The OCR engine runs locally on the HL2, this could be favourable for use cases in which keeping data in premises is important (e.g., handling confidential data).

Understanding and Contextualizing. Once a string of characters has been recognized through OCR, this module is in charge of giving meaning to such characters. To this end, we take advantage of well-known, well-documented ontologies such as QUDT[8] for units and dimensions, the Smart Applications REFerence ontology (SAREF)[9] for describing IoT devices, and the Building Topology Ontology (BOT)[10] for describing spaces. Thus, when a character is recognized, the SOCRAR system queries the Knowledge Graph (KG), for example to verify if the recognized characters correspond to a unit described in QUDT. In case it is found that they indeed represent a unit, the semantic representation of this unit is retrieved, including the conversion factor to other compatible systems of units.

[7] https://docs.microsoft.com/en-us/uwp/api/windows.media.ocr
[8] https://qudt.org
[9] https://saref.etsi.org
[10] https://w3c-lbd-cg.github.io/bot

Additionally, the KG hosts the Thing Description (TD) ontology[11] and its instances, which describe WoT enabled devices that are related to the description of the physical space they are located in.

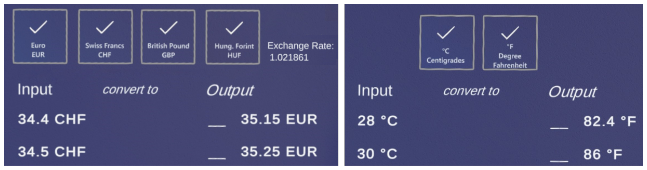
Sensing and Actuating. Given an environment in which devices' programming interfaces are described using TDs that are accessible through a KG, the SOCRAR system is capable of interacting with these devices by making the appropriate requests described in their TDs. Thus, when a user is about to enter a room and looks at the room's sign, the OCR engine recognizes a set of characters, which are then used to query the KG. After finding out that they correspond to a room (described through the BOT ontology), a query can be made to find the available WoT devices in that room. From the resulting device's TDs, the system can make, for instance, an HTTP request to obtain the current value of a humidity sensor, or a request to make a robotic arm grab an object.

Augmented Reality. The user interface is visualized on the HL2 using building blocks from the Mixed Reality Toolkit [12]. Today, hand-held devices (e.g., smartphones or tablets) enable many end-users to have access to AR-based solutions. We used HL2 because we developed our system for the near future when, as envisioned by others [25], head-mounted or wearable AR displays will be as pervasive as smartphones today. Moreover, compared to a hand-held device the HL2 allows a user to use her hands to interact with devices in her environment. In our implementation, the HL2 is able to communicate with the KG and the W3C WoT devices through HTTP requests, however the system can be extended to any (preferably TD-bound) communication protocol.

In the following, we discuss the operation and features of the SOCRAR system across the three scenarios introduced in Section 3.

4.1 Ben: Contextualizing Data in the Wild

To demonstrate the SOCRAR system's capabilities of contextualizing data, we have implemented three exemplary input conversions of currencies, units, and words (i.e., colors). Recalling Ben's acclimatization process in a new country, the SOCRAR system is capable of converting the values of a temperature display to the unit that he is more familiar with (e.g., degrees Celsius to Fahrenheit), in this case imperial (see Fig. 3). To further assist Ben, the price tags of products in shops and prices of menus in restaurants can be converted to a currency that he understands intuitively. Thus, when Ben goes abroad, the SOCRAR system can assist him in automatically converting the prices he sees (e.g., from Swiss Francs to Euros). Concretely, in Ben's scenario, the SOCRAR system captures an image of Ben's field of view. This image is then evaluated by the *Detection and Identification module*, which detects all the text contained in the picture. Each OCRed word is transmitted to the *Understanding and Contextualizing* module, which is in charge of verifying whether it might correspond to a unit, a currency, or a color by triggering a query on the KG to find if there exists a unit with a value on the `qudt:udunitsCode` or `qudt:ucumCode` property that matches the OCRed text received. In case a unit is found, the KG responds with the unit's URI and its description. This is useful for further queries, since the information for automatically performing unit conversion can be easily retrieved from the KG later in this way. In the same way, the QUDT ontology is used to determine if an OCRed text could be referring to a currency. In this case, the property associated to a unit of type `qudt:CurrencyUnit` that the text is checked against is `qudt:expression`, which is a three-letter code that represents each currency (e.g., EUR - Euro; or CHF - Swiss Franc). To verify colors, we used the SAREF ontology extension[13] and added language tags to the different color instances, which allow for the translation to other languages.



Input	convert to	Output
34.4 CHF	__	35.15 EUR
34.5 CHF	__	35.25 EUR

Input	convert to	Output
28 °C	__	82.4 °F
30 °C	__	86 °F

Figure 3: Two screenshots from SOCRAR's AR application. The image of the left shows an exemplary currency conversion in the SOCRAR system, e.g., for prices in a restaurant menu. The image on the right shows an exemplary temperature unit conversion.

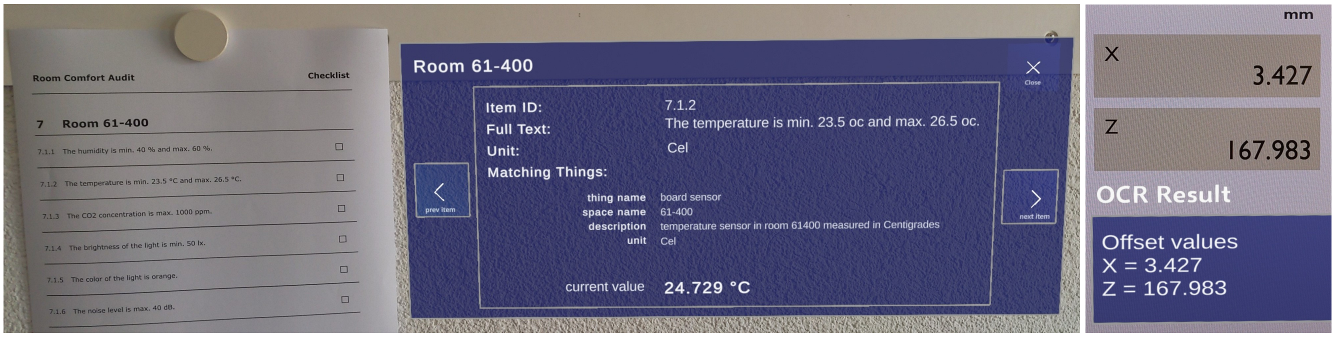
Once the text is contextualized with the information retrieved from the KG by the *Understanding and Contextualizing* module, the SOCRAR system displays the recognized and validated unit, currency, or color. In our current implementation, the user may then click buttons to convert the detected values to another unit, do a currency conversion, or translate a color to another language. For up-to-date exchange rates, the system uses a free API[14], which is called with the previously contextualized parameters.

4.2 Carol: Auditing the Operation of a BA System through a Checklist

Carol, the facility manager, may use the SOCRAR system to verify the correct operation of a BA system. Carol needs to check a laboratory on the lowest floor of the building. When she is about to enter the laboratory, she looks at the label placed at the entrance. The SOCRAR system then takes a picture of the current field of view and uses the *Detection and Identification* module to analyze it; through the *Understanding and Contextualizing module*, the recognized text is contextualized and, using the semantic representation of the building specified through the BOT ontology, the correct machine-understandable representation of the laboratory is found.

Carol brings a printed checklist with her which lists required environmental values for workplaces (e.g., from labor law) and next uses the SOCRAR system to verify that these settings hold for the laboratory. When Carol looks at the physical checklist (Fig. 4-a), the SOCRAR system's *Detection and Identification* module extracts every checklist item and saves them to create a digital representation of the checklist – a checklist item in this scenario contains one or two value-unit pairs (e.g., "The CO2 concentration is max. 1000 ppm." or "The temperature is min. 23.5 °C and max. 26.5 °C."), or a color (e.g., "The color of the light is white"). Then, the *Understanding and Contextualizing* module queries the KG to find whether the recognized characters represent a unit or a color. To this end, the used

[11] https://www.w3.org/TR/wot-thing-description
[12] https://github.com/microsoft/MixedRealityToolkit-Unity
[13] https://saref.etsi.org/saref4envi/v1.1.2
[14] https://api.exchangerate.host



Room Comfort Audit Checklist

7 Room 61-400

7.1.1 The humidity is min. 40 % and max. 60 %. □

7.1.2 The temperature is min. 23.5 °C and max. 26.5 °C. □

7.1.3 The CO2 concentration is max. 1000 ppm. □

7.1.4 The brightness of the light is min. 50 lx. □

7.1.5 The color of the light is orange. □

7.1.6 The noise level is max. 40 dB. □

Room 61-400

Item ID: 7.1.2
Full Text: The temperature is min. 23.5 oc and max. 26.5 oc.
Unit: Cel
Matching Things:

thing name	board sensor
space name	61-400
description	temperature sensor in room 61400 measured in Centigrades
unit	Cel

current value 24.729 °C

X 3.427

Z 167.983

Offset values
X = 3.427
Z = 167.983

(a) (b)

Figure 4: Two screenshots from SOCRAR's AR application. (a) The input checklist for BA auditing on the left is transferred to a digital representation using OCR. The blue panel on the right shows the second item from the checklist in AR. (b) An example of cognitive offloading with SOCRAR showing offset values in AR recognized from a tooling machine's display.

queries evaluate if there is a match on a unit's `qudt:udunitsCode` or `qudt:ucumCode` property, or if there is a color match of type *s4envi:Color*[15] (in any of the supported languages, currently: English, French, German, and Spanish).

After a unit is validated, another request is sent to the KG to find the device(s) in the current room whose measurements are expressed in the discovered unit. This is possible since the devices in the environment expose their W3C WoT TDs which describe their APIs and hold semantic information on the input/output types (i.e., the units). Since the device's TD is related to the semantic representation of the room it is located in, it is possible to filter the relevant device to support Carol. Once found, the SOCRAR system utilizes the information on the device's TD to construct and send a request to obtain the current measurements. Given the contextualized data retrieved from the KG, the SOCRAR system presents Carol with a holographic panel containing each item on her checklist along with the recently retrieved device readings. Carol browses this panel to check off all items on her checklist.

4.3 Diana: SOCRAR for Cognitive Offloading

By extending our solution with the ability to store and later recall information from the *Detection and Identification* module, our system is furthermore able to support Diana by automatically taking note of the measured offset values of a newly assembled tool and feeding these values as input to a milling machine on her request. Specifically, after Diana has assembled the tool and measured its offset, the SOCRAR system recognizes the offset values in the right format from a digital display. The recognized values are then displayed in an AR headset (Fig. 4-b), Diana accepts the values as input and walks to the milling machine. Diana now clicks a button to send the values to the milling machine's API (described on the machine's TD) to configure it appropriately.

5 DISCUSSION

Our SOCRAR approach and system combines AR, semantics, WoT, and OCR into a solution that can support humans in their everyday interactions with connected devices. It provides in-band mediation of information sources for users and output-oriented in-band value retrieval and display. Below we discuss the implemented features, potential extensions, and shortcomings of SOCRAR.

5.1 Demonstrated Features

A SOCRAR-enabled user interface helps its users to understand their environment better. Our approach represents a generalization of in-band translation services [37] and extends that approach with semantically grounded information. We demonstrated this functionality in Section 4.1 (Ben), in which the SOCRAR system is able to recognize inputs from the surroundings of a user and store this information in a semantically contextualized form. Hence, our system does not simply store the string "25 °C" but a semantically lifted version of this, i.e., the string and the additional information that this string describes (in this case temperature) in a machine-readable and machine-understandable form. This means that a user can discover values, save them, take them with her, and use them later (e.g., to input them into a physical machine, an API or a virtual process) unmodified (as shown in Section 4.3) or in modified form (i.e., as "°F"). This is immediately possible for all conversions that are already supported by a linked knowledge base (e.g., QUDT).

As demonstrated in Section 4.2, SOCRAR is able to retrieve the machine-readable and understandable representation of a device's programming interface (i.e., through its W3C WoT TD) by evaluating the recognized units and the user's context. This means that the user will no longer have to manually look for mappings. Instead, the system will automatically match the information in-band and without disturbing her. In case there is only one candidate mapping, the system confidence on having found the correct device for a user's needs will be high. If multiple matching are available, the user's input will be needed.

In the same way as the SOCRAR system today is able to discover device APIs with matching *output* types, it can easily be extended to enable matching based on *input* types, which would create a powerful engineering support tool for creating service mashups with physical devices. Thus, instead of falling back to the manual modeling of data flows (e.g., through program code, blocks-based

[15] PREFIX s4envi: <https://saref.etsi.org/saref4envi/>

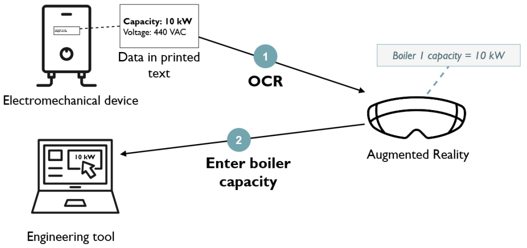


Figure 5: SOCRAR for configuring tools in a BA system. From the user's field of view, a parameter is recognized through OCR (1) and stored to later be used as an input of a digital engineering tool (2).

systems, or data-flow-based systems such as Node-RED[16]), users could use the SOCRAR approach to integrate virtual and physical services while taking advantage of its semantic filtering based on input/output types. In this way, a human engineer will get fewer alternatives or, ideally, will only need to check and verify system-proposed mappings.

5.2 Potential Future Applications of SOCRAR

Configuring Tools and Devices: In Section 4.3 we describe how the *Detection and Identification* module of SOCRAR is used for cognitive offloading in a machine tooling scenario. Consider a similar example where an engineer has to provide a boiler's capacity as an input parameter for a control program (Fig. 5). With the SOCRAR system, she would be able to detect the heating capacity from the boiler's data sheet and copy it. Then, the *Understanding and Contextualizing* module could semantically discover an empty value in the engineering tool that has the same unit as the one of the boiler heating capacity. After a user's authorization, the SOCRAR system could paste that value in the appropriate entry, relieving the engineer from having to remember the long string of characters.

Gaze-Enabled SOCRAR: Previous research on multi-modal HCI (e.g., using hands, gaze, gestures etc.) has focused on transferring content among various displays (e.g., mobile, desktop, wall-size, or HMDs). Turner et al. show how eye and hand movements can be effectively coordinated in transferring content (i.e., cut, paste, drag, drop, summon, and cast operations) between a hand-held and a wall-mounted display [38]. By means of field-of-view tracking, Gluey can execute similar operations (i.e., copy, paste, selection, color picking) among distributed and spatially registered displays [32]. Mäkelä et al. use a combination of gaze and mid-air gestures for transferring content between situated displays and a mobile display [19]. However, these approaches do not leverage semantic technologies as we propose with SOCRAR. Furthermore, they detect user's instant point of interest through eye trackers, which is a default feature in current AR headsets (e.g., the HL2 we use in the SOCRAR system). We believe that it is relevant and feasible to integrate gaze-based features in SOCRAR. Building on previous literature (e.g., [3, 4, 33]), we plan to explore several research paths in this direction, such as attention-aware contextualization of IoT and WoT streams, assessment of users' awareness and attention to objects in their environment, and gaze-contingent user assistance. While eye tracking today still has several shortcomings that are specifically relevant outside of laboratory settings [5], we believe that gaze-enabled SOCRAR would be useful in various industrial operations such as maintenance, remote support, and training.

5.3 Limitations

The current SOCRAR system can be improved at several ends. While our OCR engine performed well in our laboratory environment, minor changes in the lighting conditions often caused poor results. In more realistic environments, other solutions such as Tesseract OCR [17] might be more robust. We further observed that the OCR engine sometimes does not identify all units correctly. Therefore, we implemented a post-processing step that ensures the correctness of the units (e.g., "oC" is corrected to "°C"). To extract specific parts from the OCR's result, e.g. checklist items, these had to be written in a predefined format. In a future version of the SOCRAR system, we plan to make the extraction of specific texts more flexible.

Regarding the KG, we are currently using the QUDT, BOT, TD, and SAREF ontologies, where we assume that devices expose a W3C WoT Thing Description. To expand SOCRAR to more complex use cases that require information of other domains, manual work on selecting the right ontology and instantiating selected concepts is necessary. This could be resource-consuming, since ontologists often need to collaborate with domain experts.

6 CONCLUSIONS

In this paper, we presented SOCRAR, an approach and system that combines AR, OCR, semantic technologies, and W3C WoT-enabled environments. SOCRAR is able to detect sets of characters in a user's physical environment and contextualize them through a semantic backend. This is useful in scenarios in which the contextualized data can be immediately communicated to a user through an AR interface to improve their understanding of the environment, as well as, in scenarios in which interacting with connected devices or even virtual tools is necessary. We demonstrated the feasibility of SOCRAR through the implementation of three scenarios. As a next step, we plan to extend SOCRAR to the scenarios described in Section 5.2 and conduct a user experiment to evaluate the usability of the proposed system.

ACKNOWLEDGMENTS

This project was funded by the Swiss Innovation Agency Innosuisse (#48342.1 IP-ICT) and the Basic Research Fund of the University of St.Gallen.

REFERENCES

[1] Gregory D. Abowd. 2012. What next, Ubicomp? Celebrating an Intellectual Disappearing Act. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (Pittsburgh, Pennsylvania) *(UbiComp '12)*. ACM, New York, NY, USA, 31–40. https://doi.org/10.1145/2370216.2370222

[2] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre. 2001. Recent advances in augmented reality. *IEEE Computer Graphics and Applications* 21, 6 (2001), 34–47. https://doi.org/10.1109/38.963459

[3] Michael Barz, Sebastian Kapp, Jochen Kuhn, and Daniel Sonntag. 2021. Automatic Recognition and Augmentation of Attended Objects in Real-Time Using Eye Tracking and a Head-Mounted Display. In *ACM Symposium on Eye Tracking*

[16] https://nodered.org/

[17] https://github.com/tesseract-ocr/tesseract

Research and Applications (Virtual Event, Germany) *(ETRA '21 Adjunct)*. ACM, New York, NY, USA, Article 3, 4 pages. https://doi.org/10.1145/3450341.3458766

[4] Kenan Bektas. 2020. Toward A Pervasive Gaze-Contingent Assistance System: Attention and Context-Awareness in Augmented Reality. In *ACM Symposium on Eye Tracking Research and Applications* (Stuttgart, Germany) *(ETRA '20 Adjunct)*. ACM, New York, NY, USA, Article 36, 3 pages. https://doi.org/10.1145/3379157.3391657

[5] Kenan Bektas, Jannis Rene Strecker, Simon Mayer, and Markus Stolze. 2022. EToS-1: Eye Tracking on Shopfloors for User Engagement with Automation. In *AutomationXP22: Engaging with Automation, CHI'22*. CEUR Workshop Proceedings. http://www.alexandria.unisg.ch/266339/

[6] Bartosz Broda and Maciej Piasecki. 2007. Correction of medical handwriting OCR based on semantic similarity. In *International conference on Intelligent data engineering and automated learning*. Springer, 437–446. https://doi.org/10.1007/978-3-540-77226-2_45

[7] Victor Charpenay, Sebastian Käbisch, and Harald Kosch. 2016. Introducing Thing Descriptions and Interactions: An Ontology for the Web of Things.. In *SR+ SWIT@ ISWC*. 55–66.

[8] Andrei Ciortea, Simon Mayer, Fabien Gandon, Olivier Boissier, Alessandro Ricci, and Antoine Zimmermann. 2019. A Decade in Hindsight: The Missing Bridge Between Multi-Agent Systems and the World Wide Web. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. 1659–1663.

[9] Andy Clark and David Chalmers. 1998. The extended mind. *Analysis* 58, 1 (1998), 7–19.

[10] Andres Gomez. 2020. On-Demand Communication with the Batteryless MiroCard: Demo Abstract. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems* (Virtual Event, Japan) *(SenSys '20)*. ACM, New York, NY, USA, 629–630. https://doi.org/10.1145/3384419.3430440

[11] Dominique Guinard, Vlad Mihai Trifa, and Erik Wilde. 2010. Architecting a mashable open world wide web of things. *Technical Report/ETH Zurich, Department of Computer Science* 663 (2010).

[12] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. Knowledge Graphs. *ACM Comput. Surv.* 54, 4, Article 71 (jul 2021), 37 pages. https://doi.org/10.1145/3447772

[13] Ming Jiang, Yuerong Hu, Glen Worthey, Ryan C Dubnicek, Ted Underwood, and J Stephen Downie. 2021. Evaluating BERT's Encoding of Intrinsic Semantic Features of OCR'd Digital Library Collections. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 308–309. https://doi.org/10.1109/JCDL52503.2021.00045

[14] A Jobbins, G Raza, L Evett, and N Sherkat. 1996. Postprocessing for ocr: Correcting errors using semantic relations. In *LEDAR. Language Engineering for Document Analysis and Recognition, AISB 1996 Workshop, Sussex, England*.

[15] Matthias Kovatsch. 2015. *Scalable Web technology for the Internet of Things*. Ph.D. Dissertation. ETH Zurich.

[16] Zhen Li, Michelle Annett, Ken Hinckley, Karan Singh, and Daniel Wigdor. 2019. HoloDoc: Enabling Mixed Reality Workspaces That Harness Physical and Digital Content. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow Scotland UK, 2019-05-02). ACM, 1–14. https://doi.org/10.1145/3290605.3300917

[17] Altti Ilari Maarala, Xiang Su, and Jukka Riekki. 2016. Semantic reasoning for context-aware Internet of Things applications. *IEEE Internet of Things Journal* 4, 2 (2016), 461–473. https://doi.org/10.1109/JIOT.2016.2587060

[18] J Antonio Garcia Macias, Jorge Alvarez-Lozano, Paul Estrada, and Edgardo Aviles Lopez. 2011. Browsing the internet of things with sentient visors. *Computer* 44, 5 (2011), 46–52. https://doi.org/10.1109/MC.2011.128

[19] Ville Mäkelä, Mohamed Khamis, Lukas Mecke, Jobin James, Markku Turunen, and Florian Alt. 2018. Pocket Transfers: Interaction Techniques for Transferring Content from Situated Displays to Mobile Devices. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173709

[20] Friedemann Mattern and Christian Floerkemeier. 2010. From the Internet of Computers to the Internet of Things. In *From active data management to event-based systems and more*. Springer, 242–259.

[21] Simon Mayer, Yassin N. Hassan, and Gábor Sörös. 2014. A Magic Lens for Revealing Device Interactions in Smart Environments. In *SIGGRAPH Asia 2014 Mobile Graphics and Interactive Applications* (Shenzhen, China) *(SA '14)*. ACM, New York, NY, USA, Article 9, 6 pages. https://doi.org/10.1145/2669062.2669077

[22] Simon Mayer, Jack Hodges, Dan Yu, Mareike Kritzler, and Florian Michahelles. 2017. An open semantic framework for the industrial Internet of Things. *IEEE Intelligent Systems* 32, 1 (2017), 96–101. https://doi.org/10.1109/MIS.2017.9

[23] Rishabh Mittal and Anchal Garg. 2020. Text extraction using OCR: a systematic review. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 357–362. https://doi.org/10.1109/ICIRCA48905.2020.9183326

[24] Nahal Norouzi, Gerd Bruder, Brandon Belna, Stefanie Mutter, Damla Turgut, and Greg Welch. 2019. A systematic review of the convergence of augmented reality, intelligent virtual agents, and the internet of things. *Artificial intelligence in IoT* (2019), 1–24. https://doi.org/10.1007/978-3-030-04110-6_1

[25] Jason Orlosky, Misha Sra, Kenan Bektaş, Huaishu Peng, Jeeeun Kim, Nataliya Kos'myna, Tobias Höllerer, Anthony Steed, Kiyoshi Kiyokawa, and Kaan Akşit. 2021. Telelife: The Future of Remote Living. *Frontiers in Virtual Reality* 2 (2021), 147. https://doi.org/10.3389/frvir.2021.763340

[26] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. CORD: a consolidated receipt dataset for post-OCR parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.

[27] Ashfaqur Rahman, Mingze Xi, Joel Janek Dabrowski, John McCulloch, Stuart Arnold, Mashud Rana, Andrew George, and Matt Adcock. 2021. An integrated framework of sensing, machine learning, and augmented reality for aquaculture prawn farm management. *Aquacultural Engineering* 95 (2021), 102192. https://doi.org/10.1016/j.aquaeng.2021.102192

[28] Ganesh Ramanathan, Maria Husmann, and Simon Mayer. 2021. Interoperability vs. Tradition: Benefits and Challenges of Web of Things in Building Automation. In *11th International Conference on the Internet of Things* (St.Gallen, Switzerland) *(IoT '21)*. ACM, New York, NY, USA, 57–63. https://doi.org/10.1145/3494322.3494330

[29] Ganesh Ramanathan, Danai Vachtsevanou, Kimberly Garcia, Jérémy Lemee, Samuele Burattini, Kenan Bektas, and Simon Mayer. 2022. Semantic Knowledge for Autonomous Smart Farming. In *Proceedings of the 7th IFAC Conference on Sensing, Control and Automation Technologies for Agriculture*. http://www.alexandria.unisg.ch/266619/

[30] Alessandro Ricci, Michele Piunti, Luca Tummolini, and Cristiano Castelfranchi. 2015. The Mirror World: Preparing for Mixed-Reality Living. *IEEE Pervasive Computing* 14, 2 (2015), 60–63. https://doi.org/10.1109/MPRV.2015.44

[31] Evan F Risko and Sam J Gilbert. 2016. Cognitive offloading. *Trends in cognitive sciences* 20, 9 (2016), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

[32] Marcos Serrano, Barrett Ens, Xing-Dong Yang, and Pourang Irani. 2015. Gluey: Developing a Head-Worn Display Interface to Unify the Interaction Experience in Distributed Display Environments. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Copenhagen, Denmark) *(MobileHCI '15)*. ACM, New York, NY, USA, 161–171. https://doi.org/10.1145/2785830.2785838

[33] Janick Spirig, Kimberly Garcia, and Simon Mayer. 2021. An Expert Digital Companion for Working Environments. In *11th International Conference on the Internet of Things* (St.Gallen, Switzerland) *(IoT '21)*. ACM, New York, NY, USA, 25–32. https://doi.org/10.1145/3494322.3494326

[34] Ivan E. Sutherland. 1968. A Head-Mounted Three Dimensional Display. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I* (San Francisco, California) *(AFIPS '68 (Fall, part I))*. ACM, New York, NY, USA, 757–764. https://doi.org/10.1145/1476589.1476686

[35] Aulia Akhrian Syahidi, Kohei Arai, Herman Tolle, Ahmad Afif Supianto, and Kiyoshi Kiyokawa. 2021. Augmented Reality in the Internet of Things (AR+ IoT): A Review. *The IJICS (International Journal of Informatics and Computer Science)* 5, 3 (2021), 258–265.

[36] Lamma Tatwany and Henda Chorfi Ouertani. 2017. A review on using augmented reality in text translation. In *2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA)*. IEEE, 1–6. https://doi.org/10.1109/ICTA.2017.8336044

[37] Takumi Toyama, Daniel Sonntag, Andreas Dengel, Takahiro Matsuda, Masakazu Iwamura, and Koichi Kise. 2014. A Mixed Reality Head-Mounted Text Translation System Using Eye Gaze Input. In *Proceedings of the 19th International Conference on Intelligent User Interfaces* (Haifa, Israel) *(IUI '14)*. ACM, New York, NY, USA, 329–334. https://doi.org/10.1145/2557500.2557528

[38] Jayson Turner, Jason Alexander, Andreas Bulling, Dominik Schmidt, and Hans Gellersen. 2013. Eye pull, eye push: Moving objects between large screens and personal devices with gaze and touch. In *IFIP Conference on Human-Computer Interaction*. Springer, 170–186. https://doi.org/10.1007/978-3-642-40480-1_11

[39] Jing Wang, Jinhui Tang, Mingkun Yang, Xiang Bai, and Jiebo Luo. 2021. Improving OCR-based image captioning by incorporating geometrical relationship. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1306–1315.

[40] Roy Want, Bill N Schilit, and Scott Jenson. 2015. Enabling the internet of things. *Computer* 48, 1 (2015), 28–35. https://doi.org/10.1109/MC.2015.12

[41] Mark Weiser. 1999. The Computer for the 21st Century. *SIGMOBILE Mob. Comput. Commun. Rev.* 3, 3 (jul 1999), 3–11. https://doi.org/10.1145/329124.329126

[42] Mengya Zheng, Xingyu Pan, Nestor Velasco Bermeo, Rosemary J. Thomas, David Coyle, Gregory M. P. O'hare, and Abraham G. Campbell. 2022. STARE: Augmented Reality Data Visualization for Explainable Decision Support in Smart Environments. *IEEE Access* 10 (2022), 29543–29557. https://doi.org/10.1109/ACCESS.2022.3156697