Connecting Everyday Objects with the Metaverse: A Unified Recognition Framework

Liming Xu*
Department of Engineering
University of Cambridge
Cambridge, United Kingdom
lx249@cam.ac.uk

Dave Towey†
School of Computer Science
University of Nottingham Ningbo China
Ningbo, China
dave.towey@nottingham.edu.cn

Andrew P. French
School of Computer Science
University of Nottingham
Nottingham, United Kingdom
andrew.p.french@nottingham.ac.uk

Steve Benford
School of Computer Science
University of Nottingham
Nottingham, United Kingdom
steve.benford@nottingham.ac.uk

***Abstract*—The recent Facebook rebranding to Meta has drawn renewed attention to the metaverse. Technology giants, amongst others, are increasingly embracing the vision and opportunities of a hybrid social experience that mixes physical and virtual interactions. As the metaverse gains in traction, it is expected that everyday objects may soon connect more closely with virtual elements. However, discovering this "hidden" virtual world will be a crucial first step to interacting with it in this new augmented world. In this paper, we address the problem of connecting physical objects with their virtual counterparts, especially through connections built upon visual markers. We propose a unified recognition framework that guides approaches to the metaverse access points. We illustrate the use of the framework through experimental studies under different conditions, in which an interactive and visually attractive decoration pattern, an Artcode, is used as the approach to enable the connection. This paper will be of interest to, amongst others, researchers working in Interaction Design or Augmented Reality who are seeking techniques or guidelines for augmenting physical objects in an unobtrusive, complementary manner.**

***Index Terms*—Artcode, augmented reality, interaction, metaverse, visual marker**

I. Introduction

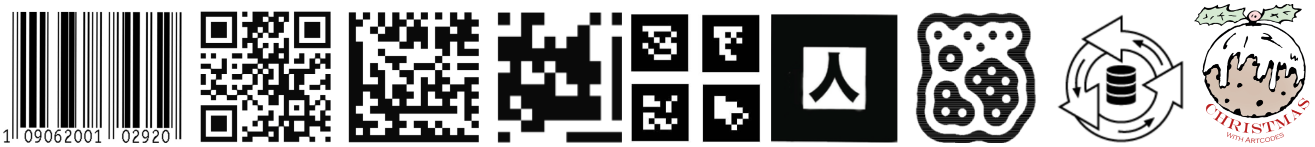
Attending events virtually has become a normalized part of our everyday life, due partly to the COVID-19 pandemic [1]. Increasingly, events are held online, or support attendance through avatars, on platforms such as Zoom, and Gather Town. This form of virtual engagement may well continue beyond COVID-19. Moreover, Facebook's recent rebranding to Meta and Microsoft's announcement of launching into the metaverse strengthen the likelihood of this being part of our new normal [2]. It is therefore reasonable to expect that our future will include a physical world even more augmented by a wide variety of virtual worlds. These virtual worlds may require unobtrusive and easy-to-use access points to a massive integrated network of virtual worlds, or metaverse. Attainment of a fully-realized, immersive metaverse will require efforts and advances in multiple areas, including computer graphics, display hardware, and communication networks [3]. In this paper, we address the issue of connections between the physical and the virtual worlds, proposing a conceptual framework for recognizing access points that may be hidden or camouflaged visual markers.

The term "metaverse" was coined in 1992 by Neal Stephenson in his science-fiction novel *Snow Crash* [4], depicting a 3D virtual world where people can interact with each other, and with intelligent agents, through their avatars [5]. 30 years later, and the development of metaverse is arguably still in its infancy, still with no generally accepted definition [5]–[7]. The development framework of the metaverse, and its characteristics, have been studied in the literature. Benford [5], for example, listed five metaverse properties: a virtual world; a virtual reality; persistence; connection to the real world; and other people. In contrast to the industrial seven-layer metaverse value chain described by Radoff [8], Duan et al. [7] proposed a three-layer metaverse development architecture, representing the physical world, interaction, and the virtual world. In spite of the lack of consensus on definition, there does appear to be general agreement that three basic metaverse properties are: (i) a physical world; (ii) a virtual world; and (iii) the connection between these two worlds.

Although various devices have been designed for accessing virtual elements or virtual worlds, a map showing the presence of access points to these virtual worlds would guide the connection (and potentially enhance the experience). If this could be provided in an explicit and straightforward manner, for example, through an annotation indicating the presence of such entrances to virtual worlds, then even better! In contexts requiring aesthetic-awareness, such at art galleries, implicit

*This work was completed while the first author was a Ph.D. student at University of Nottingham Ningbo China.

†Corresponding author.



(a) Barcode (b) QR code (c) Data matrix (d) Rohs' code (e) ARTag (f) ARToolkit (g) ReacTIVision (h) D-touch (i) Artcode

Fig. 1: Visual marker examples.

markers integrated into a part of the environment (such as in the surface pattern of an object) may be more appealing. In other environments, like in a corridor or hallway, both implicit and explicit visual markers may be acceptable. In this paper, we report on the use of such surface visual markers for connecting everyday objects with digital materials — such as digital footprints, a virtual world, or a metaverse. We propose a unified recognition framework (URF) for bridging the physical and virtual worlds through visual decorations.

The main contributions of this paper are threefold, summarised as follows:

- We report on the use of visual markers as clues to prompt interaction with virtual worlds.
- We generalize a URF for identifying the presence of access points in public spaces.
- We report on experimental studies conducted using one type of visual marker (Artcodes [9], [10]), illustrating how the proposed URF works.

The rest of this paper is organized as follows. Section II briefly reviews the related work on visual markers in augmented (AR) and virtual reality (VR). Section III introduces the URF and the preliminaries pertaining to this work. Section IV describes experimental studies evaluating the use of Artcodes as access points to virtual elements. Section V includes discussion of the implications of this study. Finally, Section VI concludes this paper and describes future work.

II. Related work on visual markers

A variety of visual markers (see examples in Figure 1), both human-readable and not, have been proposed [9], [11], with two of the most well-known being barcodes [12] (Figure 1a) and QR codes (Quick Response codes, Figure 1b) [13]. The barcode was among the earliest methods of representing data in a visual, machine-readable form, initially patented in 1952 [9]. While barcodes mainly appear in the retail sector, QR codes have become a ubiquitous feature [9]. Barcodes and QR codes were designed to be reliably read by machines, with no error occurring when they are scanned. However, this reliability comes at a cost of limited aesthetics: Neither are visually meaningful to humans, and it can be difficult to distinguish different codes though visual inspection alone.

Many other visual marker systems have similar characteristics to barcodes and QR codes, often with their information being encoded within a matrix of black and white dots, and usually with some form of error detection and correction mechanisms. Examples of such marker systems include the Data Matrix [14] (Figure 1c) and the Rohs visual code [15] (Figure 1d). While these visual markers are effective for encoding data, they were not intended for camera pose estimation and calibration, and are thus not appropriate for use as fiducials in AR systems — a fiducial is a type of marker mounted within an environment to enable estimation of the relative pose between the camera and object. Some example fiducial systems are: ARTag (Figure 1e) [16]; ARToolkit (Figure 1f) [17]; and reacTIVison (Figure 1g) [18]. ARTag markers employ a square border for marker localization, connectivity and perimeter analysis. They have a large library of patterns inside the border and use edge-detection approaches to achieve reliability [16]. ARToolkit markers consist of a thick square black border with a variety of patterns in the interior — the black outline allows for marker localisation and *homography*[1] calculation. The reacTIVision markers are automatically generated by fiducial recognition engines such as Amoeba and D-touch [19]: They have compact geometry and offer a limited space for users to adjust their aesthetic aspects [18].

The visual appearance of marker systems that rely on geometrical features for localization and encoding is strongly constrained. In the majority of cases, the shape (the geometry) of the markers is automatically generated, allowing little freedom of design. In contrast, another type of visual markers, such as D-touch and its variant Artcodes [9], [20], offer much more flexibility in geometrical form, both for the outline shape and the interior elements. D-touch encodes information through the topological structure of the markers — the adjacency information of connected components, represented in a region adjacency tree [21]. This supports users' creation of their own readable markers that are both aesthetic and meaningful [11]. Artcode implements and extends the D-touch approach, refining their drawing rules, and introducing human-meaningful (but machine-irrelevant) embellishments and aesthetic style guidelines. The Artcode approach provides the creative freedom to produce visually appealing *and* machine-readable markers (patterns) that are meaningful to humans, and that resemble free-form images.

In addition to these visual marker technologies based on geometry or topology, conventional image recognition technologies have also been employed to relate information to a much wider variety of images. Blippar [22] and Google Lens [23], for example, make use of image recognition techniques to embed data into images. However, because these techniques

[1] An isomorphism in projective spaces that is used to calibrate camera pose.

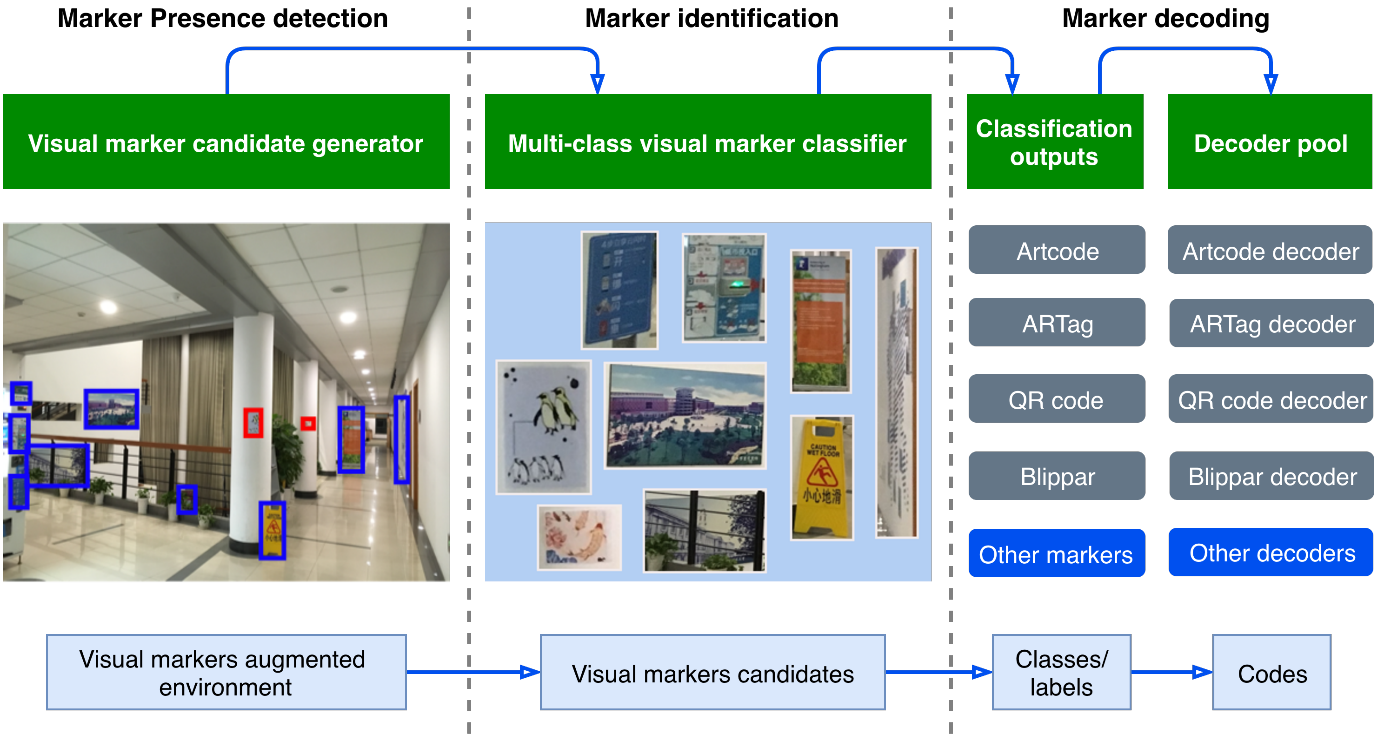


Fig. 2: A unified recognition framework (URF) for visual markers.

often use neural networks and vector matching for encoding and decoding information, it is challenging (or impossible) to explain and interpret how the system works to non-technical designers or users. More recently, new systems that use deep-generative networks to automatically generate markers have been proposed, including learnable visual markers [24], E2ETag [25] and DeepFormableTag [26].

III. Unified Recognition Framework (URF)

As AR and metaverse applications become more pervasive, we will live in a world with dispersed access points to connect with virtual elements. There will be an increasing number of entrances to these elements within our surrounding environment, through a variety of virtual markers, both visible and "hidden". Identifying the probable existence of these entrances will be the first step to triggering the follow-up interaction. Considering the many types of entrance that may co-exist, a unified recognition framework (URF) will be needed. In this section, we present such a conceptual URF for general visual marker presence recognition and identification.

Given the number of extant visual markers, both in academia and in industry, and the high likelihood of many more systems emerging in the future, attempting to explicitly include all in this URF would be unrealistic. We therefore only include a selection of some typical markers to show the basic URF components. The left part of Figure 2 shows a common scene, an indoor area of a building with various visual markers (highlighted in the picture). Not all of the annotated objects are readable — some are explicitly-placed readable Artcodes (in red boxes), while others (in blue boxes) are commonplace objects that could be enhanced as visual markers.

As shown in Figure 2, the URF involves three stages: marker presence detection; marker identification; and marker decoding. The *detection* stage involves detecting visual markers in the surrounding environment. Given the scenario in the left part of Figure 2), for example, this stage would detect the possible presence of visual markers using image processing and computer vision techniques, and would output a set of localized candidate visual markers. This output set is then passed to the *identification* stage (the middle of Figure 2) to determine if they *are* markers, and, if so, what class of markers they belong to (Artcodes, QR codes, Blippar images, etc.). A key component of the identification stage is a *multi-label classifier* that accepts the candidate markers, and outputs their corresponding classes or labels. The final stage is the *decoding*, which includes a *decoder pool* from within which the corresponding decoder identifies and decodes the embedded message in the visual marker.

Once the data (codes) carried by the visual marker are identified, the connected visual information (labelled by the visual marker) can be triggered. In this URF, visual marker detection and identification are two independent stages, but in reality, these two things are often done together. Although the URF is a conceptual framework, describing the essential components and a feasible pipeline to bridge the physical and virtual worlds, the concrete implementation may differ from one scenario to another. A possible URF implementation may be an *all-in-one* brokering system that recognizes the presence of all (or most) of the visible or hidden visual markers, then calls the corresponding decoders or identifiers, and then steps into the embedded virtual worlds.

The next section presents experimental studies examining discovery of the presence of visual markers using a concrete marker system, Artcode [9], [27].

IV. Experimental studies

The URF proposed in the last section includes the two primary elements: visual marker discovery and identification, with discovery of the markers being a *prerequisite* to the follow-up identification. Moreover, providing hints and clues to the location of (camouflaged) access points to virtual worlds may encourage people to explore those connections, thus creating new interaction opportunities. Given the importance of visual marker discovery in the URF pipeline, we conducted two case studies into how digital clues can be provided to guide users with devices (such as AR headsets) to approach the object and enter the metaverse. Artcodes, which are both meaningful to humans, and readable by scanners, were selected as the marker system.

A. The Artcode approach

Artcodes[2] are human-designable topological visual markers, developed based on the D-touch system [11]. By incorporating additional drawing constraints and aesthetic embellishments, Artcodes enable more visually pleasing and interactive patterns than d-touch [9]. Figures 1h and 1i show examples of d-touch and Artcode markers. A valid Artcode consists of two parts: a recognizable foreground (the food image in Figure 1i); and some image-based background (the text in Figure 1i). The foreground is intended for reading by machines, but the background can be designed for human consumption. Artcodes can be beautiful, interactive motifs that can decorate the surface of everyday objects without impacting the aesthetics of the object in the way that QR codes would.

Because of their unobtrusive and non-obvious properties, the presence of an Artcode is not usually obvious: Close inspection may be needed to discover an Artcode when there are no visual clues. Detection of Artcodes through their general visual features, identifying their probable locations by means of a *heat map*, is therefore a meaningful approach. Given the space limitations of this article, interested readers are referred to the literature for more information about Artcodes, including their design, detection, and identification [9], [19], [20], [27], [28].

B. Experimental setting

We conducted experiments to explore Artcode detection in an environment, and deliver clues to guide the subsequent interaction. We assumed a realistic interaction scenario, in which users may wear or carry devices in a physical space, standing far away from the Artcodes: When they discover the presence of an Artcode, they can follow clues to approach the target for further interaction. Rather than fully simulating this scenario, we simplified it while maintaining its core characteristics: Users gain increasing amounts of details as they approach the target.

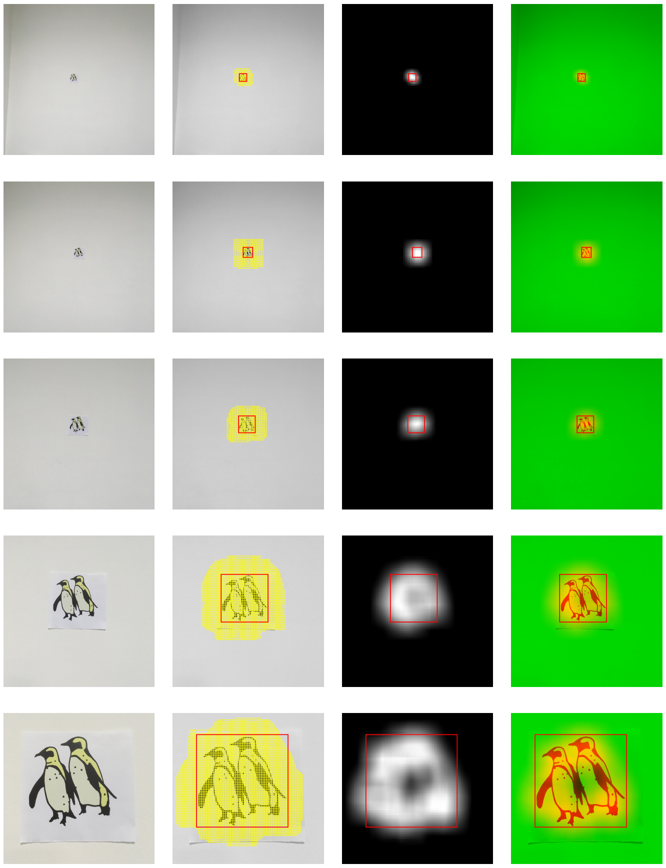
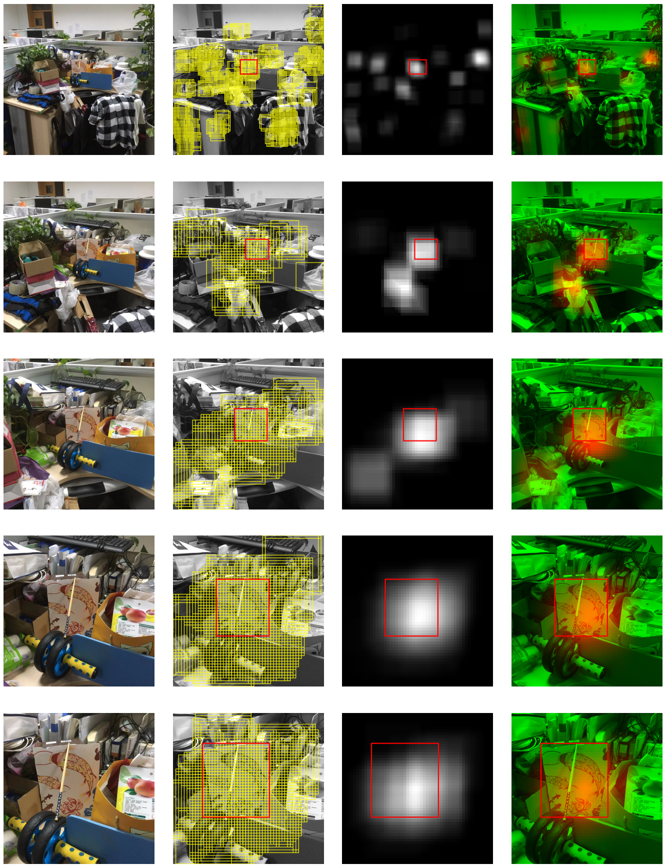

(a) Input (b) Proposal (c) Gray (d) Fused

Fig. 3: Simple Artcode detection study in clean background, good lighting.

Two studies were conducted, both involving five images sequences (Figures 3a and 4a) captured with a smartphone moving from far away to close proximity to an Artcode. The size of the Artcode gradually increases as the smartphone moves towards to the target, from top to bottom in the left-most column of the figures (Figures 3a and 4a). Recognition is more challenging from further away. Apart from this, the two studies other settings differed as follows: The first study, Figure 3, used a simple Artcode design, in good lighting, with an uncluttered scene, and an unoccluded Artcode. The second study, Figure 4, involved a more difficult scenario, using a complex Artcode design, shaded lighting, a cluttered scene, and a partially occluded Artcode.

Considering space limitations, and the focus of this paper, the technical details for building the Artcodes-detection machine-learning model are omitted. Similarly, the details underlying the various elements in Figures 3 and 4 (including generation of the proposals and presence maps) are also omitted. Interested readers are again referred to the literature for more information [9], [20], [27].

[2] https://www.artcodes.co.uk/



(a) Input (b) Proposal (c) Gray (d) Fused

Fig. 4: Complex Artcode detection study in cluttered background, poor lighting.

TABLE I: Decoding results for the images in Figures 3a and 4a.

Image / Decoded	1st (top)	2nd	3rd	4th	5th (bottom)
1st study (Figure 3)	×	×	√	√	√
2nd study (Figure 4)	×	×	×	×	×

C. Results

Figures 3 and 4 contain the content and results of the two studies. The four columns in each figure, from left to right, are: (a) the input images (b) the Artcode proposals, annotated with yellow rectangles; (c) the gray Artcode presence heat map; and (d) the fused image (created by combining the input image (a) with the heat map (c)). The red boxes indicate the ground-truth Artcodes.

In addition to the presence detection results in Figures 3 and 4, Table I presents the decoding results (generated according to Artcode decoding procedures [9]). Ticks and crosses in the table indicate whether the given image was successfully decoded or not, with ticks ("√") indicating success; and crosses ("×") indicating failure.

It is clear that the detection proposals in both studies cover the actual marker areas — the penguins in Figure 3, and the fish in Figure 4 — in all image sequences, with dense accumulation of the proposal rectangles centering around the target markers. This is further evidenced in the presence maps (gray and fused), where the marker areas are distinctly visible as heat spots (the bright areas in the 3rd and 4th columns of Figures 3 and 4). The Artcode proposals in all five of the first study images center around the true Artcode areas, identified by the red boxes: In the second study, in contrast, although the Artcode proposals cover the true Artcode areas, there are multiple proposals that are not around the actual target, especially for the images that were captured from a greater distance (in the top three rows of Figure 4a).

The cluttered scene in the second study affects the detection, increasing the number of false positives: Many non-Artcode objects in this scene may look like Artcodes, with their generic visual features potentially causing the classifier to label them as Artcodes. However, although redundant heat spots were generated, the actual target Artcodes are also identified: Figures 4c and 4d show multiple detections (indicated by heat spots), but one of them does contain the actual target Artcode. Heat spots in the presence maps can alert the user to the possible existence of access points to the metaverse, encouraging the user to come close for follow-up examination and identification.

According to the decoding results (Table I), the top two images in the first study (those captured from the furthest distance) could not be decoded, due to the low resolution and loss of details. The closer three input images in the first study, however, ware successfully identified and decoded, opening up the "hidden" virtual worlds. This represents a simplified realistic interaction, where the users often come closer to a target after first getting the general impression (the hint or clue).

The more complicated environment in the second study, including a more sophisticated Artcode, poorer lighting, clutter, and occlusion (with a chopstick in the way), resulted in none of the five images being successfully decoded. This also represents a common, real-world situation, where the target image may be obscured from certain angles. In this case, the presence maps should motivate the user to get nearer, and to remove the obstruction, or to explore new viewing angles for better identification. The explorative interaction process allowed by the proposed URF would enable various designs (e.g., design for serendipity [29], [30]), and open up new interaction opportunities for connecting to the metaverse.

V. Discussion and Implications

The two studies present a simplified and concrete implementation of the proposed framework, illustrating the key steps of detecting and identifying visual markers before decoding them, and accessing the metaverse. Currently, implementing the proposed URF for all known visual markers may not be feasible — partly due to the ever-expanding set of such markers, and the regular emergence of new interaction devices. However, this investigation using Artcodes as a representative marker provides evidence for the URF's applicability. This paper, and the URF generally, can also serve as guidance for metaverse access point design, using visual markers (especially in an unobtrusive but explorative manner). The proposed framework also includes a mixed interaction manner, combining physical movement and digital engagement in an augmented physical world with ubiquitous connection access points.

VI. Conclusion and future work

In this paper, we have explored the problem of connecting with virtual worlds (or the metaverse) in an augmented physical world. We have presented a unified recognition framework (URF) consisting of three components for designing and implementing an explorative access point. A concrete implementation of this URF using Artcodes as access points was used to illustrate the process. An example of visual markers, Artcodes are both machine-readable and human-meaningful decorative patterns that represent the kind of access tool that will become increasingly commonplace in the future. The initial discovery of the presence of markers (indicated by a heat map) and the follow-up, closer inspection and detection were demonstrated by the two studies in the paper. The URF would enable the design of a kind of brokering system that can invoke appropriate recognition algorithms to deal with different types of access points, and may inspire interaction design in the metaverse age. While this study used smartphones and Artcodes, our future work will include the investigation of other AR devices and other visual markers.

Acknowledgments

This work is supported by the Natural Science Foundation of China (project no. 61872167). The authors acknowledge the financial support from the Artificial Intelligence and Optimisation Research Group (AIOP), the Faculty of Science and Engineering (FoSE), the International Doctoral Innovation Centre, Ningbo Education Bureau, Ningbo Science and Technology Bureau, and the University of Nottingham.

References

[1] T. Johnson, "Virtual events: planning the new normal." Namizi Research, July 2020. Accessed on 08 January 2022.

[2] R. Waters, "Microsoft takes on facebook by launching metaverse on teams." Financial Times, November 2021. Accessed on 08 January 2022.

[3] J. D. N. Dionisio, W. G Burns III, and R. Gilbert, "3D virtual worlds and the metaverse: Current status and future possibilities," *ACM Computing Surveys (CSUR)*, vol. 45, no. 3, pp. 1–38, 2013.

[4] N. Stephenson, *Snow Crash: A Novel.* Spectra, 2003.

[5] S. Benford, "Metaverse: five things to know — and what it could mean for you." The Conversation, November 2021. Accessed on 20 December 2021.

[6] K. J. Nevelsteen, "Virtual world, defined from a technological perspective and applied to video games, mixed reality, and the metaverse," *Computer Animation and Virtual Worlds*, vol. 29, no. 1, p. e1752, 2018.

[7] H. Duan, J. Li, S. Fan, Z. Lin, X. Wu, and W. Cai, "Metaverse for social good: A university campus prototype," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 153–161, 2021.

[8] J. Radoff, "The metaverse value-chain." Medium, April 2021. Accessed on 31 December 2021.

[9] R. Meese, S. Ali, E.-C. Thorne, S. D. Benford, A. Quinn, R. Mortier, B. N. Koleva, T. Pridmore, and S. L. Baurley, "From codes to patterns: designing interactive decoration for tableware," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*, pp. 931–940, 2013.

[10] S. Benford, A. Hazzard, A. Chamberlain, and L. Xu, "Augmenting a guitar with its digital footprint," in *Proceedings of the international conference on New Interfaces for Musical Expression (NIME'15)*, pp. 303–306, 2015.

[11] E. Costanza and J. Huang, "Designable visual markers," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*, pp. 1879–1888, 2009.

[12] N. J. Woodland and S. Bernard, "Classifying apparatus and method." US Patent, October 1952. US2612994A.

[13] ISO/IEC, "Information technology — automatic identification and data capture techniques — QR code bar code symbology specification," February 2015. ISO/IEC 18004:2015.

[14] ISO/IEC, "Information technology — automatic identification and data capture techniques — Data Matrix bar code symbology specification," September 2006. ISO/IEC 16022:2006.

[15] M. Rohs and P. Zweifel, "A conceptual framework for camera phone-based interaction techniques," in *International Conference on Pervasive Computing*, pp. 171–189, Springer, 2005.

[16] M. Fiala, "Artag, a fiducial marker system using digital techniques," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, pp. 590–596, IEEE, 2005.

[17] H. Kato and M. Billinghurst, "Marker tracking and hmd calibration for a video-based augmented reality conferencing system," in *Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality (IWAR'99)*, pp. 85–94, IEEE, 1999.

[18] R. Bencina, M. Kaltenbrunner, and S. Jorda, "Improved topological fiducial tracking in the reactivision system," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, pp. 99–99, IEEE, 2005.

[19] E. Costanza, S. B. Shelley, and J. Robinson, "D-touch: A consumer-grade tangible interface module and musical applications," in *Proceedings of Conference on Human-Computer Interaction (HCI'03)*, (Bath, UK), Springer-Verlag, September 2003.

[20] L. Xu, *Artcode detection in images.* PhD thesis, University of Nottingham, 2019.

[21] E. Costanza and J. Robinson, "A region adjacency tree approach to the detection and design of fiducials.," *Video, Vision and Graphics*, 2003.

[22] "Blippar." https://www.blippar.com/. Accessed on 31 December 2021.

[23] "Google Lens." https://lens.google/. Accessed on 31 December 2021.

[24] O. Grinchuk, V. Lebedev, and V. Lempitsky, "Learnable visual markers," *Advances in Neural Information Processing Systems (NIPS'16)*, vol. 29, pp. 4143–4151, 2016.

[25] J. Brennan Peace, E. Psota, Y. Liu, and L. C. Pérez, "E2etag: An end-to-end trainable method for generating and detecting fiducial markers," *arXiv e-prints*, pp. arXiv–2105, 2021.

[26] M. B. Yaldiz, A. Meuleman, H. Jang, H. Ha, and M. H. Kim, "Deepformabletag: end-to-end generation and recognition of deformable fiducial markers," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–14, 2021.

[27] L. Xu, A. P. French, D. Towey, and S. Benford, "Recognizing the presence of hidden visual markers in digital images," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pp. 210–218, 2017.

[28] L. Xu, D. Towey, A. P. French, S. Benford, Z. Q. Zhou, and T. Y. Chen, "Using metamorphic relations to verify and enhance artcode classification," *Journal of Systems and Software (JSS)*, vol. 182, p. 111060, 2021.

[29] P. André, M. Schraefel, J. Teevan, and S. T. Dumais, "Discovery is never by chance: designing for (un) serendipity," in *Proceedings of the seventh ACM conference on Creativity and cognition (C&C'09)*, pp. 305–314, 2009.

[30] L. Danzico, "The design of serendipity is not by chance," *Interactions*, vol. 17, no. 5, pp. 16–18, 2010.