EG UK Computer Graphics & Visual Computing (2015), pp. 1–4
Rita Borgo, Cagatay Turkay (Editors)

Quasi-Hamming Distances: An Overarching Concept for Measuring Glyph Similarity

Philip A. Legg[†1], Eamonn Maguire[2], Simon Walton[3], and Min Chen[3]

[1]University of the West of England, UK, [2] CERN European Laboratory for Particle Physics, and [3]University of Oxford, UK

Abstract

In many applications of spatial or temporal visualization, glyphs provide an effective means for encoding multivariate data objects. However, because glyphs are typically small, they are vulnerable to various perceptual errors. In data communication, the concept of Hamming distance *underpins the study of codes that support error detection and correction by the receiver without the need for corroboration from the sender.*

In this ***extended abstract****, we outline a novel concept of* quasi-Hamming distance *in the context of glyph design. We discuss the feasibility of estimating quasi-Hamming distance between a pair of glyphs, and the minimal Hamming distance for a glyph set. This measurement enables glyph designers to determine the differentiability between glyphs, facilitating design optimization by maximizing distances between glyphs under various constraints (e.g., the available number of visual channels and their encoding bandwidth).*

1. Introduction

Glyph-based visualization [War02, BKC*13] is a common form of visual design where some data records are depicted by pre-defined visual objects, which are called *glyphs*. Glyph-based visualizations are ubiquitous in modern life since they make excellent use of the human ability to learn abstract and metaphoric representations in order to facilitate instantaneous recognition and understanding. Glyphs can be used to encode variables of different data types, categorical (e.g., [LCP*12, MRSS*12]) as well as numerical (e.g., [KW06, DTW*14]). However, glyphs are typically small, and are often designed with a high-degree of similarity in order to facilitate mapping consistency, semantic interpretation, learning and memorization. In many applications of spatial or temporal visualization, such as geo-spatial visualization and event visualization, there can be a large number of small glyphs, which are sometimes overlaid on some background graphics or imagery. Hence we are particularly concerned about the *differentiability* of glyphs and potential perceptual errors in observation and exploration.

Figure 1 shows example cases that may render some glyphs indistinguishable. Zooming-out actions in data exploration can reduce glyph size significantly. For example, they could make some shapes (e.g., circle and hexagon) and textures appear similarly, while confusing the categorization of sizes (e.g., large, medium, small). Meanwhile, environmental lighting conditions and printing or photocopying facilities can cause color and greyscale degeneration. Not only would such changes make some glyphs indistinguishable, but they can also confuse the association of different colors or grayscales. While having a dynamic legend will help alleviate the confusion about various mappings, it demands users to view the legend on a regular basis, incurring additional cognitive load in terms of the effort for the bothersome visual search and memorization of the unstable mapping keys. Other causes could also include color- or change-blindness, and short- or long-sightedness.

In the visualization literature, there are many useful guidelines that could be adopted for glyph designs [BKC*13]. For example, Bertin [Ber83] advised that size is not associative, hence unsuitable for encoding categorical attributes. Tools such as ColorBrewer [HB11] can be used to generate qualitative colormaps for effective separability of attribute values. There were also empirical studies on similarity of simple glyphs (e.g., [DBH14]). Many glyph designers apply their creative intuition to ensure the diversity and legibility of different glyphs. This poses some challenging research questions for effective glyph design, including, '*Is there a theoretical framework to encompass various design guide-*

† email: Phil.Legg@uwe.ac.uk

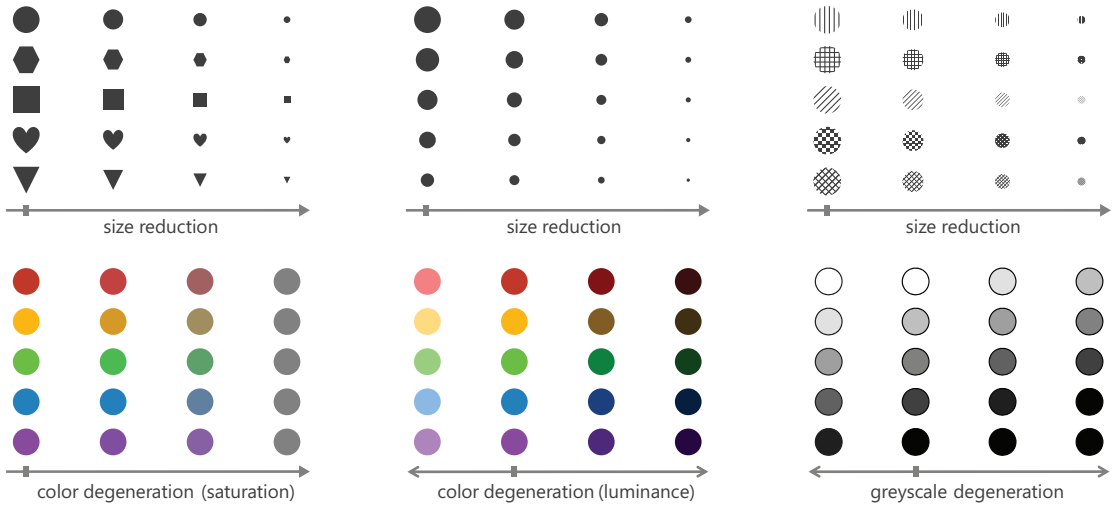

Figure 1: *Here, it can be seen how size reduction and color degeneration can dramatically impact on the effectiveness of the different visual cues, such as shape, size, texture and color. At full resolution, different shape glyphs are well-distinguished. As the size is reduced, shapes that initially were distinct become much more visually similar. The same can be observed for color and greyscale degeneration.*

lines?' and '*Is there a systematic approach to design a fail-safe glyph set?*'

In our recent work (not yet published), we have proposed a conceptual framework for glyph-design based on the Hamming distance (Section 2). Because of the perceptual nature of many design aspects, we have introduced the notion of *quasi-Hamming distance* (QHD) (Section 3). Using this notion, we are able to translate qualitative assessment of perceptual distances in a design to QHDs. When the minimal QHD for a glyph set is 1, the glyph set is vulnerable to the 'noise' during observation and exploration. When the minimal QHD is 2, the glyph set facilitates some error detection, with which the viewer can use interaction (e.g., zooming-in, or looking at the legend) to investigate the error. When the minimal QHD is 3 or more, the glyph set facilitates some error correction at the receiving end. This enables us to adjust the design to ensure a minimal QHD among a set of glyphs. In other words, this provides a systematic approach to optimize the design of a set of glyphs, providing fail-safe glyph encoding. We have outlined several methods for estimating QHD, and conducted two proof-of-concept experiments for estimating QHD based on the grading by human participants and using image-comparison metrics respectively. Using this novel concept, we have also implemented a glyph-based visualization tool in two application case studies.

This **extended abstract** presents a brief outline of the aforementioned work, focusing the concept of QHD.

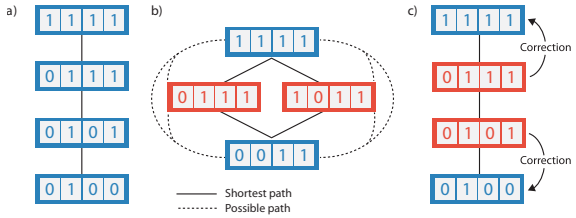

Figure 2: *Computing the minimum Hamming distance between binary representations. Blue values illustrate valid values in a known dictionary, red values are invalid in the dictionary. In (a), all values are valid in the dictionary and any error due to bit changes cannot be detected. In (b), only values 1111 and 0011 are valid. The minimum Hamming distance between these values is 2, as shown by the shortest path. A distance of 2 supports error detection, i.e., if either of these intermediate values (0111 or 1011) are received. In (c), only values 1111 and 0100 are valid in the dictionary. The minimum Hamming distance between these values is 3, shown by the shortest path. A distance of 3 supports error correction, since if 0111 is received then this is closest to 1111, and if 0101 is received then this is closest to 0100.*

2. Hamming Distance

Proposed by Richard Hamming in 1950, the Hamming distance provides a similarity measure between two codewords by calculating the difference on a per-position basis [Ham50]. It is a simple, yet effective approach for error

detection and correction routines, for instance, checking for the successful transmission of data, be it from reading physical media such as a CD, or when receiving information over a network connection. In tele- and data communication, redundancy is widely used for detecting whether the original message was received correctly, and in many situations, for automated correction. In computer vision, the concept has been applied to barcode design (e.g., [Ols11, Bys12]).

Hamming Distance defined upon two equal-length binary codewords is the number of bits at which they differ. When considering a code consisting of k codewords, it is defined as the minimal Hamming distance between all pairs of codewords within in the code.

Figure 2 illustrates the concept of the Hamming distance. The concept is supported by having a dictionary of known terms than a receiver would expect to receive. In (a), all possible values are valid in the dictionary, giving the minimum Hamming distance of 1. This code is unable to support error detection and correction since any erroneous change of bits will result in a valid value. In (b), there are two valid values, 1111 and 0011. The minimum Hamming distance for this code is thus $H = 2$. Should a received value be either of these intermediate hops, an error would be detected since neither of these values is valid. In (c), suppose we now have the values 1111 and 0100. Here, the minimum Hamming distance is increased to $H = 3$. If one of the intermediate values is received in transmission, then an error is detection as before. However, we can also calculate which valid value that the received erroneous value is closest to. This allows us to correct the erroneous value. By increasing the separability between the two valid values, we can compensate for potential errors that may occur in the transmission.

In many ways, visualization is akin to the concept of transmitting data across a network [CJ10]. Therefore, in addition to creating a visualization that conveys information, it is vital that this visualization contains this concept of error checking, to ensure that the message being transmitted is not misunderstood.

3. Quasi-Hamming Distance for Glyph Design

The concept of Hamming distance for a code can be adapted to provide a strong foundation for measuring similarity of glyphs. The design of effective glyphs is not a trivial task. In order to provide an effective encoding for a large multivariate dataset using a glyph representation, the designer needs to consider how each data attribute will map to a particular visual channel that is available to them. Some applications may involve as many as 20 variables that are to be encoded into a relatively small glyph [DTW*14]. The application domain may introduce further semantics that need to be represented in the glyph encoding, such as domain-specific metaphors [LCP*12]. Where the application requires a large number of different glyph representations (e.g., when depicting different activities), the collection of glyphs can be considered to be the construction of a visual language for communicating the data. Glyphs within the same visual language should share commonality, much in the same way that certain letter combinations in a vocabulary (e.g., 'pre-', 'un-' and '-tion' in English) are shared by many words with similar semantic meanings. This helps to aid familiarity and memorization when learning the language being communicated. However, there is a delicate balance to achieve between this commonality, and the separability that helps to support error detection and correction in the visualization communication.

By using the minimum Hamming distance as a basis for our design principles, we can systematically aim to achieve this balance. However, measuring the distances and errors in visual perception is clearly not as simple as measuring those represented by binary codewords. We thereby propose an approximated conceptual framework based on the principle of Hamming distance, and we call it *Quasi-Hamming Distance* (QHD). The term 'quasi' implies that the distance measure is approximated, as is the quantitative measure of perceptual error. The main research questions are thereby (i) whether we can establish a measurement unit common to both measures, and (ii) how we can obtain such measurements.

Here, let us consider the challenges of measuring visual separability. Different visual channels are used within visualization, including color, shape, size, orientation and texture. Supposing we have a series of glyphs (Figure 3) the task would be to quantify the minimal QHD of this set of glyphs. In order to achieve this, one may choose a variety of methods, including:

- Estimation by expert designers,
- Crush tests,
- Task-based evaluation,
- User-centric estimation,
- Computer-based similarity measures.

Our experience of user-centric estimation and computer-based similarity measures has shown that QHD can be estimated. We hope to report this experience soon in a formal publication.

4. Conclusion

In this extended abstract, we have outlined an overarching concept, QHD, for supporting the systematic design of glyph visualization. The concept is based on the principles of the minimum Hamming distance, which is widely used in communication. At its very essence, visualization is a means of communication through a visual medium, and so just as with traditional communications, it is essential that the message being communicated is transmitted effectively. With our systematic design approach, the visual design can incorporate self-error-correction at the receiver's end based on the separability of the glyph representations. A substantial part of work has been reported in a PhD thesis [Mag15]. We hope

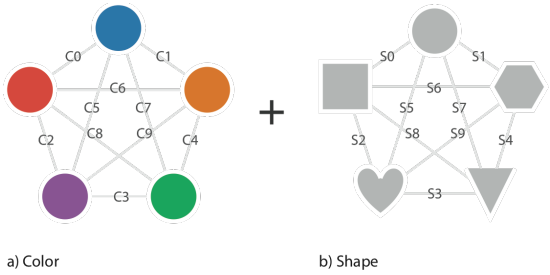

Figure 3: *Graph representation of different glyphs. Each edge in the graph denotes the minimum QHD between the pair of connected glyphs, with (a) using only color and (b) using only shape. The challenge is how to derive the distances that separate each of the connected glyphs, which can be achieved by a number of methods.*

to report full details of this work, including our empirical studies and application case studies, in a formal publication.

References

[Ber83] BERTIN J.: *Semiology of Graphics*. University of Wisconsin Press, 1983. 1

[BKC*13] BORGO R., KEHRER J., CHUNG D. H., MAGUIRE E., LARAMEE R. S., HAUSER H., WARD M., CHEN M.: Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics State of the Art Reports* (May 2013), EG STARs, pp. 39–63. 1

[Bys12] BYSTRYKH L. V.: Generalized dna barcode design based on hamming codes. *PLoS One 7*, 5 (2012), e36852. 3

[CJ10] CHEN M., JAENICKE H.: An information-theoretic framework for visualization. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1206–1215. 3

[DBH14] DEMIRALP C. D., BERNSTEIN M. S., HEER J.: Learning perceptual kernels for visualization design. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1933–1942. 1

[DTW*14] DUFFY B., THIYAGALINGAM J., WALTON S., SMITH D. J., TREFETHEN A., KIRKMAN-BROWN J. C., GAFFNEY E. A., CHEN M.: Glyph-based video visualization for semen analysis. *IEEE Transactions on Visualization and Computer Graphics* (to appear, 2014). 1, 3

[Ham50] HAMMING R. W.: Error detecting and error correcting codes. *Bell System Technical Journal 29*, 2 (1950), 147–160. 2

[HB11] HARROWER M., BREWER C. A.: *ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps*. John Wiley and Sons, Ltd, 2011, pp. 261–268. 1

[KW06] KINDLMANN G., WESTIN C.-F.: Diffusion tensor visualization with glyph packing. *IEEE Transactions on Visualization and Computer Graphics 12*, 5 (2006), 1329–1336. 1

[LCP*12] LEGG P. A., CHUNG D. H. S., PARRY M. L., JONES M. W., LONG R., GRIFFITHS I. W., CHEN M.: MatchPad: Interactive glyph-based visualization for real-time sports performance analysis. *Computer Graphics Forum 31*, 3 (2012), 1255–1264. 1, 3

[Mag15] MAGUIRE E. J.: *Systematising Glyph Design for Visualization*. PhD thesis, Department of Computer Science, University of Oxford, March 2015. 4

[MRSS*12] MAGUIRE E., ROCCA-SERRA P., SANSONE S.-A., DAVIES J., CHEN M.: Taxonomy-based glyph design with a case study on visualizing workflows of biological experiments. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2603–2612. 1

[Ols11] OLSON E.: Apriltag: A robust and flexible visual fiducial system. In *Proc. IEEE International Conference on Robotics and Automation* (2011), pp. 3400–3407. 3

[War02] WARD M. O.: A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization 1*, 3/4 (2002), 194–210. 1