

Modeling How Humans Judge Dot-Label Relations in Point Cloud Visualizations

Martin Reckziegel, Linda Pfeiffer, Christian Heine, and Stefan Jänicke

Abstract—When point clouds are labeled in information visualization applications, sophisticated guidelines as in cartography do not yet exist. Existing naive strategies may mislead as to which points belong to which label. To inform improved strategies, we studied factors influencing this phenomenon. We derived a class of labeled point cloud representations from existing applications and we defined different models predicting how humans interpret such complex representations, focusing on their geometric properties. We conducted an empirical study, in which participants had to relate dots to labels in order to evaluate how well our models predict. Our results indicate that presence of point clusters, label size, and angle to the label have an effect on participants' judgment as well as that the distance measure types considered perform differently discouraging the use of label centers as reference points.

Index Terms—Human judgment model, document visualization, label placement.



1 INTRODUCTION

IN order to guide through complex visual sceneries, label placement has been a matter of research for a long time. The field of cartography first addressed this topic to communicate spatial information effectively by labeling geographical features such as countries, cities, and rivers [4]. While this has been a manual task for a long time, current geographic information systems apply automated labeling algorithms [5] to select and to place appropriate labels according to general design guidelines [6], [7]. Though such sophisticated labeling guidelines rarely exist for applications in visualization, several techniques have been developed to aid navigating through visualizations and to understand visualized contents [8], [9]. While *internal labeling* algorithms align labels within regions they are assigned to, *external labeling* algorithms place labels either next to the corresponding glyphs using a recurring pattern (e.g., always at the top right), or next to the visualization. In the latter case, lines are drawn between labels and associated objects to indicate relationships.

Label placement also plays an important role in various information visualization applications. When mapping document collections in a two-dimensional space, documents are represented as dots arranged according to similarity. On-demand labeling with magic lenses [10] and static internal labels aid at communicating the contents of documents placed underneath or close by [1]. An example is shown in Figure 1a. Labels can be seen as a thematic overlay that superimpose relationships between dots and topics. Considering that each label represents a cluster and each dot belongs to a cluster, what are the visual features that induce the relationship between a dot and a label? Do we judge the blue dot in Figure 1a as belonging to the

mesh-cluster due to proximity, or to the *surface*-cluster due to the arrangement of dots? And, do we judge the green dot as belonging to the *vortex*-cluster due to the large font size of the *vortex* label, or do we group it to the *vector*-cluster due to the proximity to the smaller-sized *field* label? Although aspects of how we perceive similar or closely placed glyphs have been studied in the past (e.g., Gestalt principles of grouping [11]), to the best of our knowledge, no empirical studies have been conducted examining the human interpretation of complex scenarios involving dots and labels. In other applications, dot layers are likewise overlaid with labels. In graph visualizations, individual nodes or entire node clusters might be labeled (see Figure 1b). In tag maps, labels form a thematic layer summarizing textual information about geo-referenced data items. As shown in Figure 1c, the underlying data might be displayed as a dot layer. In both cases, the interplay between dots and labels is further influenced—by links between nodes and by geographical features of the background map, respectively. The document spatialization scenario in Figure 1a can be seen as an abstraction of both.

In this paper, we report results obtained in an online study on how humans judge relationships between labels and dots in document spatializations. Based on previous studies and related work in information visualization, we identified visual features that these representations typically vary and that bear on this judgment task. We defined twelve model classes that combine and weigh these features in different ways to predict human performance for this task and selected stimuli optimized to discriminate between these models. The results of our study identify one model that outperforms the others.

2 RELATED WORK

We give an overview of related perceptual and cognitive studies, and introduce the problem of label placement in cartography. In addition, we list information visualization applications in which label placement plays a notable role.

- M. Reckziegel and C. Heine are with Leipzig University.
E-mails: {reckziegel, heine}@informatik.uni-leipzig.de
- L. Pfeiffer is with German Aerospace Center DLR.
E-mail: linda.pfeiffer@dlr.de.
- S. Jänicke is with University of Southern Denmark.
E-mail: stjaenicke@imada.sdu.dk.

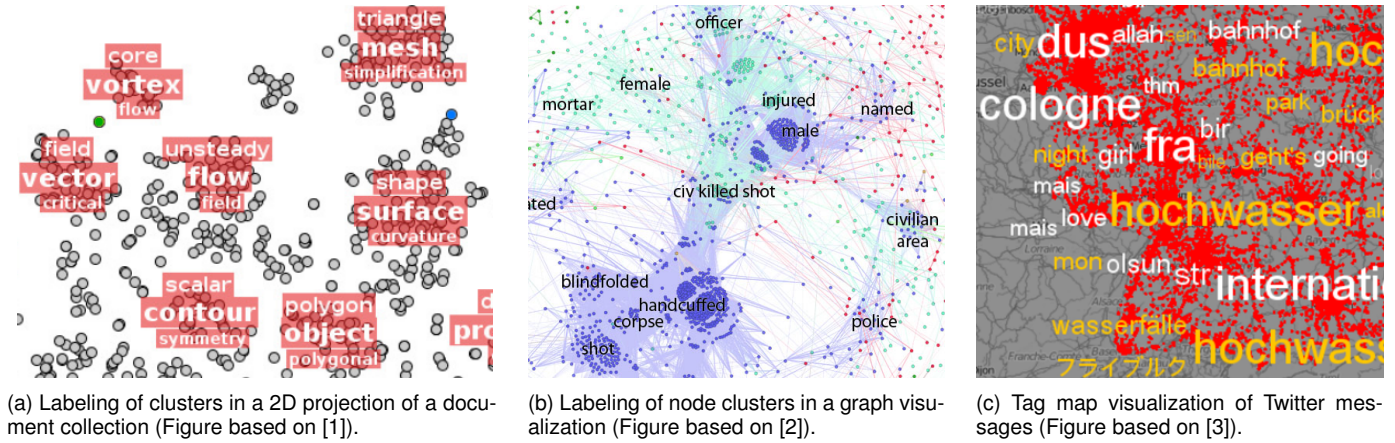


Fig. 1: Related scenarios in which point cloud visualizations are overlaid with labels showing spatially accumulated characteristics.

2.1 Perceptual Studies

In the 1920s, Gestalt theorists studied how humans perceive complex scenarios, how the impression of a *form* emerges, and how the relationship of visual elements that compose a scene is perceived [12]. Gestalt principles were proposed that describe how we see objects by grouping similar elements, how we recognize patterns, and how we simplify complex images [11]. The Gestalt principle of proximity, stating that close objects are perceptually grouped together, is most related to our work. According to the ‘first law of cognitive geography’ [13] people even believe that closer objects are more similar to each other. Therefore, we assume that the distance between a dot and a label has a major impact on how humans judge their relationship.

Perception of distances in spatial layouts is influenced by several effects, e.g., the horizontal-vertical illusion [14], [15]. It relates to the orientation of the imagined connecting line between objects and the overestimation of vertical lines compared to their horizontal counterparts. Furthermore, the perception of distance is affected by the presence of other objects like dots [16]. The more interjacent dots are located between two objects the more distant they will be estimated. The scenario in our study is more complex as we measure the relationship between different object types, a dot and a label, and intervening dots may serve to perceive this relationship much stronger. The Gestalt principles of closure and continuity [11] are related to this phenomenon. Fabrikant and Montello found that emergent point features like the arrangement of dots in the form of lines or shapes have an impact on perceived distances and object similarities, making them appear further apart but more similar [17].

Perception of clusters is not only affected by proximity but also by global spatial concentration, change of object density, and the figure-ground ratio [18], [19]. Similarly to those studies, we developed mathematical models of human judgment, but ours describe the relation between dot clusters and labels depending on different visual features.

2.2 Map Label Placement

The history of creating maps goes back to around 25,000 BC [20] and placing labels on maps to support understanding of geography has been a substantial task ever

since maps were drawn. While many famous cartographers evolved a common style to place labels on maps, Eduard Imhof [6], [21] was the first to publish general principles and requirements to design legible, easily comprehensible maps. He argues that map legibility depends on good label positioning, and each label has only one optimal position on the map. It requires a clear graphic association to the object it belongs to. He distinguishes between three object types: point, line and area. Labels for points should be placed next to, over, or under the object favoring a position to the top-right of the object. Labels for lines should be placed along the line conforming to its curvature. Labels for areas should be stretched, bent, scaled, and spread over the whole space to clarify areal association. While Imhof discussed qualitative aspects of label placement, Pinhas Yoeli [22] was the first to propose an automated system to achieve a good label placement. Later, many strategies for automated label placement were proposed, e.g., [10], [23], [24], [25], [26], [27]. For a comprehensive overview of labeling techniques we refer the reader to *The Map-Labeling Bibliography* [5].

Imhof points out that a legible label placement does not only depend on the label style and position [6]. The arrangement of other labels on the map, and thus the interplay between all visual features on the map needs to be optimal. Explicitly, he demands that ‘names should disturb other map contents as little as possible’ and asks to avoid occlusion and concealment. In some label placement scenarios in information visualization applications (see subsection 2.3) it is hard to comply with this rule. Objects to be labeled may already occlude, and white space between objects may be too small. Therefore, labels are often placed naively (e.g., always to the right of the dot), and occlusion and concealment among labels and dots are not considered.

2.3 InfoVis Applications

Although label placement is a crucial task in many application areas of visualization, e.g., medical visualization [28], we focus on related applications in information visualization where labels and dots are placed in a plane.

2.3.1 Document Spatializations

A typical approach for the exploration of large document collections is to represent documents as high-dimensional

vectors, determine pairwise distances between documents, and project these vectors to a two-dimensional space using techniques like principal component analysis (PCA) [29] or t-Distributed Stochastic Neighbor Embedding (t-SNE) [30]. Many applications offer magic lenses to support interactive exploration of the resulting point clouds [31]. In such scenarios, a subset of dots is selected and corresponding labels are placed next to the lens. As this procedure automatically communicates a relationship between dots and labels, in the following, we focus on representations having labels that are permanently shown. Wise et al. [32] was one of the first who put *gisting terms* to describe corpus clusters in ‘galaxies’ where each star represents a document. Endert et al. [33] also draws document collections as galaxies, and labels describing the contents of galaxies originate from the galaxy centroids. Choo et al. [34] attach topic identifiers oriented at the top of dot clusters in a two-dimensional scatter plot. Da Silva et al. [35] manually label such clusters that they call neighborhoods to ease identifying least-varying dimensions of a multidimensional data set. In the latter two applications, color is further used to communicate cluster relationships. Kandogan [36] automatically annotates clusters with descriptive labels placed at the clusters’ centroids. In addition, labels for outliers are placed next to the dots. Fortuna et al. [37] draw a contour landscape underneath the dots, and representative keywords are determined for random positions. A rather traditional approach is proposed by Luboschik et al. [38] who put labels on the screen directly next to dots or at a distance using lines to indicate relationships. Han et al. [1] apply Luboschik’s algorithm to label clusters in a point cloud. To improve the assignment of the labels to the clusters, a second layout algorithm draws major tags in the centers of clusters using a semi-transparent background to discriminate dots from labels (see Figure 1a). All listed applications place labels at plausibly chosen positions, but how humans judge relationships between dots and labels and what features influence this judgment has not been studied before.

2.3.2 Graph Visualizations

In analogy to document spatialization, graphs can be laid out in a two-dimensional area, and nodes represent entities [39]. In addition, links between nodes are drawn to highlight their relationships. An overview of multi-faceted graph visualization techniques is provided by Hadlak et al. [40]. Labels for individual nodes or node clusters are drawn to ease navigation in the graph, especially when the graph consists of hundreds or thousands of nodes. Usually, labels are placed pragmatically. For co-citation networks, Chen et al. [41] place non-occluding labels to the right of a few nodes and labels for clusters originate from the cluster centers. Similarly, labels are placed at predefined positions in most graph visualization applications [42], [43], [44]. To avoid occlusions among labels, Wong et al. [45] propose a design that places labels circularly around nodes, and edges are represented by strings instead of lines. When graphs illustrate biological processes, node shapes are enlarged so that labels can be placed inside them [46], [47], [48], [49]. Accompanied with labels, portrait icons are used as nodes in social network visualizations [50], [51]. Many of the listed applications do not only use labels to inform

about nodes. Edges illustrate relations, nodes’ clusters may receive the same color, and node shapes can take different forms. Abstracting from those additional visual features, graph visualizations can be seen as a special case of the basic document spatialization scenario.

2.3.3 Tag Maps

Tag maps are maps containing a thematic layer in the form of a tag cloud [52]. They are related to our work, because they superimpose relations between tags and geographical entities (e.g., cities, rivers, landmarks). The document spatialization scenario is emulated when individual data items are arranged on a dot layer [3]; an example is shown in Figure 1c. In tag maps, tags are placed dependent on the geographical position of their associated data items. Large geo-referenced data sets often contain different tags for the same location, and tag map layout algorithms decide on aggregations and omissions. Tag-cloud-driven tag maps [53] collect information for entire polygonal regions and place the tags regardless of actual geographical locations within that polygon applying a tag placement strategy adopted from tag cloud algorithms [54], [55], [56]. Location-driven tag maps place tags dependent on the actual geographical distribution of corresponding data items. Ahern et al.’s approach [57] applies a geographical clustering and draws one representative tag in each cluster’s barycenter without considering potential occlusions. Thom et al. [58] use a displacement strategy with an Archimedean spiral originating from the preferred location. While keeping the topical diversity on a global scope, this strategy leads to loosening the linkage between themes and geography. As a consequence, geographical entities like cities may be perceived as belonging to unrelated, but closely positioned tags. Reckziegel et al. [53] avoid displacing tags. The tag map algorithm ensures that a tag always represents the relative majority of data items positioned underneath its bounding box. Though tags are less diverse, geographical entities are likely to be related to closely positioned tags. In all tag map approaches, the font size of a tag in the final layout reflects the number of aggregated data items. Being aware that the geographical map that lies beyond labels and dots influences human judgment, we regard tag maps, likewise to graph visualizations, as abstractions of the document spatialization scenario.

3 VISUAL ABSTRACTION

Though neglecting additional visual features as prevalent in graph visualizations and tag maps, the generalized document spatialization scenario entails numerous conditions influencing arrangement and appearance of dots and labels on the screen in the first place and the human judgment of relationships among them in the second place. The design space reflecting those conditions had to be bound in our experimental setup as well as in the development of models describing human judgment in order to process a manageable number of images by the study participants.

Element Distribution Document spatializations contain dots and labels that can be arbitrarily distributed across space. Labels and dots may or may not overlap. As we neither wish to analyze the influence of occlusions among

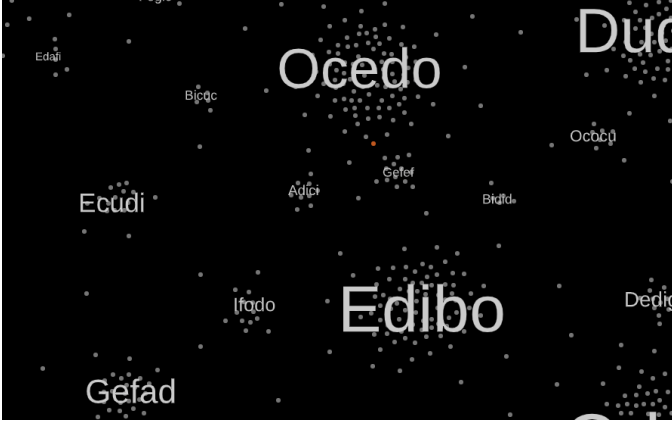


Fig. 2: Example of a stimulus shown to the participants. The query dot the participants had to relate to one of the labels is highlighted in red.

labels nor those among dots, we only allow occlusions among labels and dots. As the presence of dot accumulations and clusters has an influence on human perception (see subsection 2.1), we allow dots either to be distributed with equal density or to form clusters nearby label centers.

Labels Labels can vary in font face, orientation, and font size determining the height of a label. As typical for related scenarios, we use a uniform font face and keep all labels oriented horizontally as rotated tags are perceived as ‘unstructured, unattractive, and hardly readable’ [59]. We regard varying font size—the main visual attribute in tag cloud representations [60]—as equally important for the considered scenario as dot clusters. Further, labels in related applications convey meaning, but we use random character sequences and do not model semantic influence.

Dots Dots can vary in radius, thus, the area they cover, as well as in the appearance of the corresponding glyph. In our setup, we use circular solid filled glyphs of equal radius without strokes to represent dots.

In Figure 2 the visual features considered are varied to exemplify the abstract document spatialization scenario. Although color is often used in related applications to highlight clusters, we disregard color in our study as the number of clusters to be discriminated using color is limited [35]. Also, it frees us from the need to calibrate displays and enables us to conduct the study online.

4 MODELS

Implementing the conditions outlined above, we hypothesized several mathematical models to predict how humans associate dots with labels in two-dimensional images. An image I consists of a finite set of horizontal labels $L = \{l_1, \dots, l_{|L|}\}$ and a finite set of dots. Each label l is a sequence of characters, each character’s glyph covering a certain area of the image. Dots are located at their respective centers $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_{|X|}\}$.

Our models aim to predict which label $l \in L$ a human would associate with a query dot q . Consider a simple image with labels A and B , no dots except one query dot q , either positioned at \mathbf{x}_1 , \mathbf{x}_2 , or \mathbf{x}_3 as shown in Figure 3. If q would be positioned at \mathbf{x}_1 , it will most likely be associated with A ,

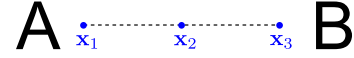


Fig. 3: A simple stimulus with one query dot at one of three possible positions \mathbf{x}_1 , \mathbf{x}_2 and \mathbf{x}_3

if positioned at \mathbf{x}_3 , it will most likely be associated with B . Somewhere along the line between \mathbf{x}_1 and \mathbf{x}_3 there will be a position \mathbf{x}_2 where the judged association switches from A to B . Extending this thought to the plane containing multiple labels and additional dots, we could, in theory, assign each label an *association region*. The result would look like a weighted Voronoi diagram with weights depending on label properties and dots positioned in close vicinity. In practice, however, due to limited visual acuity and perceptual choice being a probabilistic process [61], humans might not give the same answer if asked again to relate the dot to a label for the same configuration. Therefore, our models will predict probabilities for each label $l \in L$ to be assigned to a query dot q at position \mathbf{x} . For simplicity, we will describe our models in terms of a scoring function; the probabilities are then just the normalized scores:

$$p_I^M(l, \mathbf{x}) = \frac{s_I^M(l, \mathbf{x})}{\sum_{l' \in L} s_I^M(l', \mathbf{x})} \quad (1)$$

The score s_I^M of model M in image I is a function that gives higher values the more likely a label l is assigned to \mathbf{x} . The different models M will score according to the influencing factors discussed below. All our models’ scores will be positive, hence no label is assigned the probability 0, although it can be arbitrarily close to 0; this ensures that our models’ log-likelihoods (see subsection 5.4) remain finite.

4.1 Distance

Distance is the first influencing factor we took into account. Since labels are complex irregularly shaped areal objects, formalizing the notion of distance is not straightforward. We define the distance between label and dot as the Euclidean distance between the dot’s center and the ‘nearest’ label point (illustrated in Table 1a) that can be:

label center $\mathbf{n}^{LC}(l, \mathbf{x})$ is the center of the bounding box of l .

label bounding box $\mathbf{n}^{LB}(l, \mathbf{x})$ is a point inside the bounding box of l which is nearest to \mathbf{x} .

character bounding box $\mathbf{n}^{CB}(l, \mathbf{x})$ is a point inside the union of the bounding boxes of each character of l which is nearest to \mathbf{x} . It reflects the irregular shape of the label better.

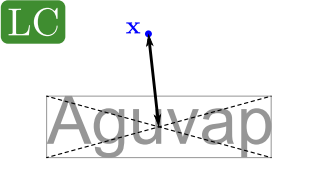
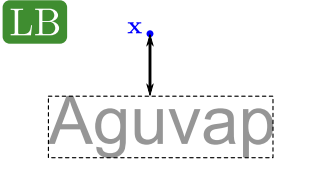
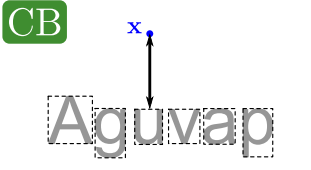
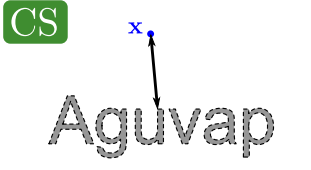
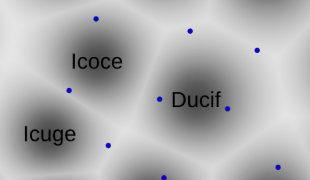
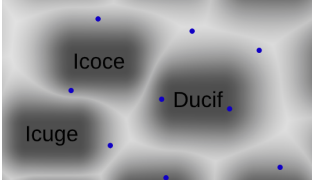
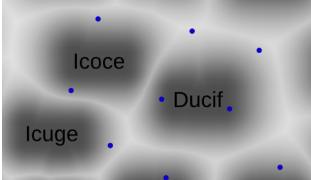
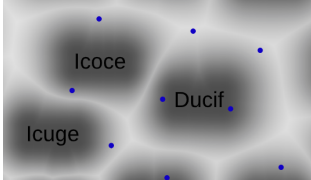





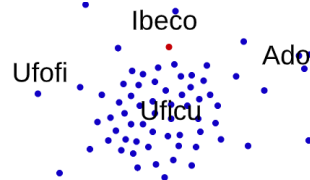
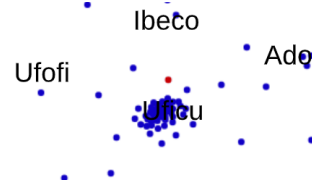
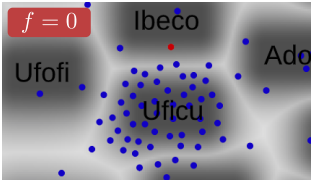
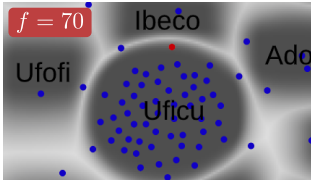
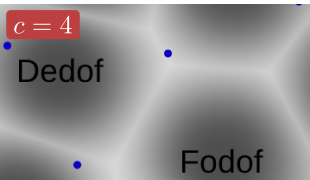
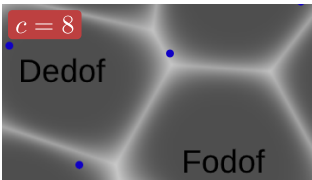
character shape $\mathbf{n}^{CS}(l, \mathbf{x})$ is a point inside the area covered by the glyphs of l which is nearest to \mathbf{x} .

4.2 Label Size

As currently no studies about the influence of label size on distance perception exist, but many information visualization applications use font size to communicate importance, we decided to study their influence in our experiment. We hypothesized that with increasing label size the region of association increases beyond what distances predicted, and we aimed to assess how much larger. This effect seems to depend on whether the query dot was above, below, or to

TABLE 1

Visualization of model scores for the different parameters. The darker the pixels in the grey scale images the higher their maximal score over all labels. All combinations of the four distance types (**distT**) and the three label size weights (**weightT**) result in 12 model classes for which all respective real valued parameters, displayed with a red background, will be fitted.

Distance Type (distT)	<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p>LC</p>  </div> <div style="text-align: center;"> <p>LB</p>  </div> <div style="text-align: center;"> <p>CB</p>  </div> <div style="text-align: center;"> <p>CS</p>  </div> </div> <div style="display: flex; justify-content: space-around; margin-top: 10px;">     </div> <p>(a) The top row illustrates how the distance between a dot x and a label for the different distance types label center (LC), label bounding box (LB), character bounding box (CB) and character shape (CS) is determined. The bottom row visualizes the resulting model scores. Darker pixels have lower minimum distance over all labels. There is a clear difference in shape of the regions of association between LC and the other types. There are only subtle differences between CB and CS. Furthermore, the scores for LC fall off more quickly.</p>
Label Size Weight (weightT)	<div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 30%;"> <p>V</p> $\omega^V(l) = H(l)^{-h}$ </div> <div style="width: 30%;"> <p>$h = 0.3$</p>  </div> <div style="width: 30%;"> <p>(b) The vertically-weighted type (V) weights the distance based on the height $H(l)$ of the label l's bounding box and the height influence parameter $h \geq 0$. The visualization to the left shows the influence of this weighting for $h = 0.3$. The label 'Fofef's region of association increases compared to labels with a smaller height, respectively font size.</p> </div> </div> <hr/> <div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 30%;"> <p>A</p> $\omega^A(l) = (W(l) \cdot H(l))^{-a}$ </div> <div style="width: 30%;"> <p>$a = 0.3$</p>  </div> <div style="width: 30%;"> <p>(c) The area-weighted type (A) weights the distance based on the area $W(l) \cdot H(l)$ of the labels l's bounding box and the areal influence parameter $a \geq 0$. Similarly to the vertically-weighted type, the label 'Fofef's region of association in the visualization to the left is increased, here even larger for an equal value of the area influence parameter compared to the height influence parameter.</p> </div> </div> <hr/> <div style="display: flex; justify-content: space-between; align-items: flex-start;"> <div style="width: 30%;"> <p>E</p> $\omega^E(l) = \sqrt{\frac{\delta}{\ \delta\ } \begin{pmatrix} W(L)^{-w} & 0 \\ 0 & H(l)^{-h} \end{pmatrix} \frac{\delta}{\ \delta\ }}$ </div> <div style="width: 30%;"> <p>$w = 0.3$ $h = 0.3$</p>  </div> <div style="width: 30%;"> <p>$w = 0.3$ $h = 0$</p>  </div> <div style="width: 30%;"> <p>$w = 0$ $h = 0.3$</p>  </div> </div> <p>(d) The ellipse-weighted type (E) weights the distance according to a scaling of space different in x- and y-direction based on the width $W(l)$ and height $H(l)$ of the label's bounding box using the abbreviation $\delta = \mathbf{x} - \mathbf{n}^{\text{distT}}(l, \mathbf{x})$. For both $w = 0.3$ and $h = 0.3$ shown in the left most visualization of the model scores, the resulting image is different than for the isotropic cases, because label width and height are not equal resulting in anisotropic weighting of distances. The increase of region of association of larger labels is not as strong for $w = h$ as in the isotropic models. The effects of different influence parameters for width and height are shown in the middle and right most visualization.</p>
Force (f)	<div style="display: flex; justify-content: space-around; align-items: center;">   </div> <div style="display: flex; justify-content: space-around; align-items: center; margin-top: 10px;">   </div> <p>(e) Four illustrations showing an example of the force-directed approach to account for dot accumulations. The two illustrations on the left side show the positions of the dots before and after the displacement for $f = 70$. Before the displacement, the red dot is closer to the label 'Ibeco' while after the displacement it is closer to 'Ufocu'. The two visualization on the right side show the resulting model scores for $f = 0$ (no deformation) and $f = 70$. The force-directed deformation gives the label 'Ufocu' a larger region of association.</p>
Certainty (c)	<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;"> <p>$c = 4$</p>  </div> <div style="text-align: center;"> <p>$c = 8$</p>  </div> </div> <p>(f) Visualization of the model behaviour for different values of the certainty parameter c for the LC distance type. With increasing c more pixels get dark, meaning that the probability of the likeliest label increases, hence the model is more 'confident' of its predictions.</p>

the side of the label. We were not sure whether humans can perceptually discount the width of a label knowing that it depends on its textual content and that font size is used to indicate importance. We try to capture this influence by different methods of ‘weighting’ distances. Two isotropic methods weight the distances according to the height (see Table 1b) and the area (see Table 1c) of labels. The influence of different heights, respectively areas, on the weighting is defined by the parameters $h \geq 0$ and $a \geq 0$. An anisotropic method weights the distances according to a scaling of space differently in x - and y -direction based on the width and height of the labels’ bounding boxes (see Table 1d). This method controls the amount of weighting by the parameters $w \geq 0$ and $h \geq 0$.

4.3 Dot Clusters

The presence of dots can have an influence on distance perception (see subsection 2.1). We hypothesized that, if dots form clusters around a label, they might increase the label’s region of association, but if the dots are distributed uniformly they do not. We modeled this effect by deforming the plane, so that regions around clusters contract, resulting in larger regions of association for labels with clusters.

The deformation method performs five iterations of a force-directed algorithm. In each iteration labels attract dots and dots attract each other. Let δ denote the vector from a dot to another dot or to the center of a label’s bounding box, and let $z = \min\{1, \frac{\|\delta\|}{f}\}$. The force is defined as

$$\mathbf{F} = (2z^6 - 3z^4 + 1)\delta.$$

$f \geq 0$ is the parameter regulating the intensity by which dot clusters get contracted. There is no deformation of space if $f = 0$ and the resulting vector \mathbf{F} has zero length, for any distance $\|\delta\|$ greater than f . We move each dot in the direction $\epsilon\mathbf{F}$ for each force acting on it. Labels are not moved. We determined $\epsilon = 0.01$ empirically to give good results. Table 1e shows an example deformation of space and how this affects the region of association.

4.4 Scores

We combined the three influencing factors into twelve model classes. Their scoring function has the form

$$s_I^M(l, \mathbf{x}) = \left(\omega^{\text{weightT}}(l) \|\chi - \mathbf{n}^{\text{distT}}(\chi)\| \right)^{-c}$$

where χ is the position \mathbf{x} after force-directed deformation, weightT and distT name one of the three weighting and four distance types, respectively, and $c \geq 0$ is a parameter capturing a notion of certainty. Higher values for c will increase the relative differences in scores and respectively probabilities, illustrated in Table 1f. It can thus be used to fit human accuracy for repeated trials.

Combining all possibilities for weightT and distT generates twelve model classes, each of which has additional parameters. All model classes have a certainty parameter c and a force parameter f . The vertically-weighted models have an additional height influencing parameter h , the area-weighted models an additional area influencing parameter a , and the ellipse-weighted models two more parameters: height influence h and width influence w .

5 METHODS

We conducted an empirical experiment in order to fit the model parameters described in section 4 and to evaluate our twelve models. Proven to deliver valuable results for perceptual studies in visualization [62], we performed a crowdsourcing experiment in the form of an online study to reach a large number of participants in a short study duration of four days. In order to mitigate the disadvantages compared to controlled laboratory studies [63], we (1) requested participants to participate only once, (2) removed outliers introduced by random responding (see subsection 5.4), and (3) ensured a uniform display of the same stimuli on different screens (see subsection 5.2). We conducted two pilot studies to improve our study design.

5.1 Participants

We invited students and scientific staff of different universities and research institutes via mailing lists to participate in the experiment. Furthermore, we announced the experiment on social media channels. To motivate taking part in the study, participants had the chance to win one out of ten gift certificates. We had 93 participants, from which we excluded four from the analysis (see section 6 for further details). Of the remaining 89 participants 34 stated that they were female and 53 male (two gave no answer). The average age of the participants was 33.5 years (SD = 11.3). 32 participants stated to have good sight, 51 to have corrected sight and five to have reduced sight (one gave no answer).

5.2 Procedure

The study was offered in English and German. After choosing the language, participants were presented a consent form, which served to inform potential participants on the study’s goal, data privacy, and legal considerations. After accepting the consent form, we asked participants to adjust a colored square to the size of a credit card to ensure uniform stimulus sizes across displays. We required a display of at least 13 inches diagonal. Participants were requested to sit about 70 cm away from their screen to ensure stimuli appeared under uniform visual angles.

We asked participants to choose a quiet place, to switch off their mobile phones, and to switch the browser to full screen mode in order to reduce distractions. Then, participants were briefly informed on the upcoming stimuli and their task: ‘Quickly click on the word, to which, in your opinion, the dot highlighted in red most likely belongs to.’ In a short training session consisting of a few trials with increasing difficulty, participants could familiarize with the task. The experiment consisted of 214 stimuli presented in four blocks, every block containing 51-54 stimuli. At the beginning of each block we reminded participants of their task, and after each block we asked participants to take a short break to minimize any fatigue effects. Stimuli were presented in a randomized order, and we recorded answers (the clicked labels) and response times. At the end, we asked participants to fill out a questionnaire. Following questions about demographic information, we asked to rank on a five point Likert-scale (ranging from ‘0 = no influence’ to ‘4 = very strong influence’) how strong, in their opinion,

the following features influenced their decision to choose a label: distance between the highlighted dot and the label, presence of other dots, size of the labels. Next, open-ended questions upon further significant decision criteria and general comments were asked. We added the Likert-scales in order to get more detailed feedback compared to only asking open-ended questions as in the pilot studies.

5.3 Stimuli

A stimulus in our experimental design is an image comprising dots and textual labels – an abstraction of images in our target application as described in section 3. Dots can vary in their position, and labels can vary in their position, content, and font size. We procedurally generated 51 images, from which we selected portions as our stimuli. We describe these generators in the remainder of this subsection. For each image, we selected four query dots using an optimization algorithm that aims to improve model discrimination per stimulus. The details and evaluation for this inessential step are explained in the supplemental material, since it requires a substantial mathematical background and would distract the reading flow. With four query dots for each generated image we have 204 stimuli, to which we added 10 ‘tracer bullet’ stimuli, i.e. stimuli with known correct answer.

5.3.1 Labels

To avoid influence of label meaning, we randomly pick letters alternating between vowels and consonants. All letters have equal probability of being picked. The first letter is always capitalized. Apart from this, we systematically vary the density of labels in an image, their font size, and number of characters. For varying densities we use Poisson-disk sampling [64] to generate the center points of labels, keeping the minimal distance between any two centers at three different values. Font sizes are randomly chosen between $[f_{min}, f_{max}]$ in a non-uniform manner using an exponential curve. This makes larger font sizes less likely than smaller ones. In the baseline setting, we set $f_{min} = f_{max}$, so that all labels have the same font size. There are two other settings with $f_{min} < f_{max}$. We also vary the number of characters of the labels, used to approximately regulate their aspect ratio. In the baseline setting, all labels have 5 characters, while in the second setting the length varies between 3 and 8 characters and in the third between 3 and 12 in the same non-uniform manner as described above.

The label generation is implemented by rejection sampling. The probability of accepting a new label increases with larger differences in font size or aspect ratio to neighboring labels, except in their baseline setting. In any case, a label is rejected if label bounding boxes overlap or the minimal distance criterion is violated. The sampling process is repeated until no more candidates are accepted after a fixed number of iterations. We systematically vary all three variables, density, font size, and aspect ratio, separately. While cycling through the possible settings for one of them, we keep the others at baseline. As such we obtain seven different label generators.

5.3.2 Dots

We keep the radius and the color of dots constant and we vary only dot density. We use Poisson-disk sampling

and vary the minimum distance parameter for to generate dot positions. We get three settings for uniform density: sparse, medium, and dense. In the fourth setting, we imitate document visualizations, where labels are usually placed in cluster centers. Therefore, we place a 2D multivariate normal distribution centered very close to each label’s center with random covariances. The Poisson-disk samples are then rejection-sampled according to this density function. In the fifth and last setting, we set the covariances such that dot clusters protrude into regions of association of foreign labels. To avoid filling the space with too many clusters, we depend the likelihood of large clusters on the absence of nearby large clusters, similar to the acceptance criteria for large font sizes and aspect ratios above. In total, we have five different dot generators.

5.3.3 Final Stimulus Set

Given the label and dot generators, we produced 51 images as follows: We combined each of the dot generator settings, duplicating the two cluster settings to obtain more of such stimuli and resulting in a total of seven different dot settings, with each of the seven label generators to obtain 49 images. As each of the label generators only varies one variable per image, we added two images where the label font size as well as the aspect ratios vary together. For each of these 51 images, we extracted four query dots. We clipped the resulting 204 images so that the query dot is at a random location within a 1.6:1 rectangle.

Furthermore, we generated ten images using the generators above and manually chose the query dot very close to a label, ensuring a very high probability of choosing this label. These stimuli were presented in regular intervals to identify if study participants were clicking randomly on any label rather than performing the experimental task (see also section 6).

The coloring of the stimuli was influenced by the results of the pilot study. It seemed that the visual salience of labels and dots had a major impact on participants answers. Thus, we tried to choose their colors in order to make them similarly salient. Furthermore, we made the highlighted dot most salient in order to reduce visual search times. The final coloring is shown in Figure 2.

5.4 Analysis

By analyzing the experimental data, we try to answer the following main questions:

- Which of our model classes predicts human judgment in the presented scenarios best and how do they compare against each other?
- Do model parameters affect prediction quality and can the participants’ subjective feedback on what influenced their decision be linked to model parameters?
- Are there any factors not yet considered in our models?

To quantitatively compare our twelve model classes we first tried to find the best real valued parameters fitting the experimental data using a maximum likelihood estimation [65]. The likelihood of observing the sequence of labels $l^{(i)}$ for the stimuli sequence $I^{(i)}, \mathbf{x}^{(i)}$ under model M is:

$$\mathcal{L} = \prod_{i=1}^n p_{I^{(i)}}^M(l^{(i)}, \mathbf{x}^{(i)}),$$

where p is given as in Equation 1 and n is the number of participants multiplied by the number of stimuli shown per participant. The log-likelihood depends on three to four parameters (depending on the model) for which the derivatives are difficult to compute because of the force-directed deformation. To find the maximum, we used the Nelder-Mead method [66], which is derivative-free. We used the default values for its parameters $\alpha = 1$, $\gamma = 2$, $\rho = 0.5$, $\sigma = 0.5$, and 15 runs to increase the chance of the found local maximum being a global maximum. Nelder-Mead does not enforce bounds on parameters, but our models require all parameters to be non-negative. Therefore, we transformed them to their absolute value. The first initial value was an educated guess at $a = 0.5$, $w = 0.5$, $h = 0.5$, $f = 54.7$ and $c = 6.5$, while the other 14 were random positions within the bounds of the model class’s parameter space. We ranked models using the Akaike information criterion (AIC) [67], which corrects log-likelihood scores based on the number of model parameters k . AIC can be computed as $AIC = -2 \ln \mathcal{L} + 2k$. We also computed Bayesian information criterion (BIC) scores [65] $BIC = -2 \ln \mathcal{L} + k \ln n$, where n is the number of observations. BIC penalizes the influence of the number of parameters differently. We performed two-sided pair-wise sign tests for the differences in these scores between any two of the twelve model classes. The tests were performed for a significance level of $\alpha = 0.05$, using Bonferroni correction.

To judge the effectiveness of the model parameters we fitted parameters of models reduced by one parameter each and compared these to the complete models by a likelihood ratio test. Additionally, we used bootstrapping to estimate the standard error of model parameters [65]. Furthermore, we analyzed participants’ Likert-scale ratings on the importance of the decision criteria by looking at their descriptive statistics.

Answers to the open-ended questions were coded and categorized following a bottom-up approach and their counts reveal details about the decision criteria in our models as well as further potential decision criteria. Kendall’s rank correlation test was then used to identify how the subjective ratings correlate with the model parameters.

Beforehand, as first step, we analyzed the label answers and response times to identify outliers using two dissimilarity scores. One was computed from the pairwise dissimilarities between participants based on the number of times they disagree in their answers. The other was computed from the differences in response time profile. We looked for outliers in scatter plots created using classical multidimensional scaling (MDS), minimizing the squared error between the Euclidean distances of the coordinates and the dissimilarities.

6 RESULTS AND DISCUSSION

The MDS plots for similarity of participants’ response times showed two outliers—participants who on average took significantly longer for each task than other participants. This suggests that they may have been distracted during the experiment, or failed to observe our instructions to answer quickly. The MDS plot for answer similarity indicated that two other participants differed strongly in their answers

from the majority. We suspected these participants had selected labels randomly. One of these participants failed two of the ten tracer stimuli confirming our suspicion. Thus, we overall removed four participants from the analysis.

6.1 Model Comparison

The twelve model classes, summarized in Table 1, result from combining each of the three types of weighted label size (V, A, E) with each of the four distance types (LC, LB, CB, CS). While V and A weight distances isotropically to a label with respect to its height or area, E weights distances differently depending on the angle to a label with respect to its width and height. Table 2 summarizes the results of the parameter fitting.

6.1.1 Fitted Model Quality

As the AIC and BIC scores rank model classes equally, only AIC scores are used for further comparison. Recall that due to the link between AIC and log-likelihood scores, these can be interpreted easily. Two AIC scores s_1 and s_2 with their difference $\Delta = s_2 - s_1$ mean that the model with AIC s_1 is $e^{\Delta/2}$ times as likely to be the Kullback-Leibler best model, given the observed data [68].

To put these values into perspective, we give some bounds. The model using the least amount of information of a stimulus just predicts each label with the same probability. Because the probabilities are equal, its AIC score only depends on the number of observations and labels, and is 112878 in our case. A less naive model that considers the data is one predicting each label that was actually observed in the study with equal probability. In our case such a model has an AIC score of 33530. On the other hand, the best possible model for our data would be one that predicts exactly the observed frequencies of each label; it would have an AIC score of 22948. The scores of our fitted models lie within these bounds, predicting the observations not perfectly, but vastly surpassing the naive models.

For each value of the label size weighting type, the AICs are such that $LC > LB > CB > CS$. Similarly, for each value of the distance types, the AICs are such that $V, A > E$. Given the poor performance of VLC and ALC, we found ELC to perform comparatively well. To investigate further on the relative performance of model classes, we looked at the pair tests comparing all twelve model classes against each other (see supplemental material). All pairs showed significant differences except for the pair VLB vs. ALB. We ordered model classes by the number of times they performed significantly better against each other. The classes in Table 2 are ordered top to bottom accordingly indicating that the LC distance type performs the poorest. Overall, the ECS model class achieved the best AIC scores and also performs significantly better compared to all other classes. Furthermore, the CB and LB distance types combined with E are good alternative choices as they also perform significantly better than the rest of the classes.

6.1.2 Parameters Values

The parameter c reflecting the certainty of the influential area around the labels is fitted between 4.6 and 6.3 for the different models—between the two examples shown

TABLE 2
The results of the best parameter fitting

model class	scores			parameter log-likelihood estimation					
	$\ln \mathcal{L}$	AIC	BIC	c	f	A	V	E	
						a	h	w	h
VLC	-16 690	33 386	33 410	4.84±0.07	55.14±1.85		0.365±0.016		
ALC	-16 176	32 358	32 381	5.10±0.08	51.69±2.13	0.183±0.007			
ELC	-14 567	29 142	29 174	6.29±0.22	53.87±1.64			0.702±0.015	0.721±0.024
ALB	-14 766	29 538	29 561	4.56±0.12	56.20±1.59	0.014±0.008			
VLB	-14 761	29 528	29 552	4.56±0.13	56.23±1.51		0.040±0.017		
ACB	-14 619	29 244	29 268	4.67±0.12	56.73±1.43	0.041±0.008			
VCB	-14 598	29 201	29 225	4.68±0.12	56.77±1.49		0.103±0.017		
ACS	-14 499	29 003	29 027	4.77±0.13	56.27±1.29	0.043±0.008			
VCS	-14 476	28 958	28 982	4.77±0.13	56.57±1.46		0.108±0.016		
ELB	-14 471	28 949	28 980	4.94±0.19	56.57±1.51			0.125±0.023	0.058±0.036
ECB	-14 407	28 821	28 852	4.99±0.16	56.89±1.56			0.208±0.024	0.187±0.036
ECS	-14 259	28 526	28 558	5.11±0.19	56.85±1.56			0.217±0.021	0.194±0.033

The lower the AIC and BIC scores the better, the higher the log likelihood score ($\ln \mathcal{L}$) the better. The a , w , h and f parameters are not scale-invariant and need to be adjusted for a different number of pixels per degree of visual angle than used in our study (45.29px per degree). Next to the parameter values the standard error, empirically calculated from the bootstrap samples (see supplemental material for details), is given.

in Table 1f. As one can see, these values generate areas where the certainty to choose between two or more labels is near guessing level which confirms the initial assumptions about the probabilistic aspects of the decision process and supports our choice for probabilistic models.

The respective font size parameters, reflecting how much differences in label size influence distances and regions of association, are fitted much higher for LC compared to the other distance types. This is expected, as the extend of the labels is not considered before weighting the distances, which a higher font size weight is able to compensate. Further, the certainty parameter c is also fitted higher in LC as the areas of uncertainty are larger compared to the others given the same value for c (see Table 1a, where we used the same setting for c for all subfigures) caused by the same effect. The force parameter f is fitted lower for LC compared to the other types. In contrast to LC, the font size parameters of LB always fit the lowest. This is due to LC over-estimating distances while LB is under-estimating them given the fact that the AIC scores for those types are worse than for the other two.

Comparing the size-weighted parameters of the model classes amongst each other, the area parameter a obtains the lowest values followed by the height parameter h of the vertically-weighted model class followed by the two parameters w and h of the elliptical class. Again, this is expected concluding from the model score visualizations in Table 1b-d, where for the same values of the respective parameters, the effects on the regions of association are ordered equally. We further found that the width parameter w is always fitted a bit higher than h for the elliptical model E. As such, a difference in width between two labels has a slightly higher effect on the label choice than a difference in height when judging query dots in the respective horizontal or vertical direction to those labels. This matches with the horizontal vertical illusion effect described in subsection 2.1.

TABLE 3
Impact by Decision Criteria: Likert-scale Ratings

	0	1	2	3	4	Mean	SD
Distance	0	1	8	22	58	3.54	0.71
Presence of Dots	3	12	17	38	19	2.65	1.07
Font Size	7	24	29	20	9	2.00	1.11

6.2 Decision Variables

The likelihood ratio tests showed significant differences for each submodel with one parameter removed, except for elliptical models, where both font size weighting parameters need to be removed (details in the supplemental material). Hence, each of the parameters chosen is relevant to the performance of the complete model. This can also be read from the standard errors shown in Table 2 for the force and font size parameters, which indicate that these parameters are unlikely to be zero and thus have no influence. Similarly, the certainty parameters is unlikely to have the value one and thus no influence.

The likelihood ratios show that usually the certainty parameter had most influence, followed by the force parameter, then font size parameters. Only in models using distance type LC, the font size parameters have more influence than the force parameter. It seems that fitting the font size parameter compensates for the LC distance type not depending on label geometry, which depends on font size.

The Likert-scale ratings, which are summarized in Table 3, show only few ratings that attest no influence to 'presence of dots' or 'font size', whereas 'distance' was an important decision criterion for every participant. The mean ratings in fact show that distance was the most important decision criterion, which confirms our previous assumption. Regarding the other two decision criteria, the presence of emergent point features has a higher impact than font size. This shows that the subjective ratings align with the results of the likelihood ratio tests (except for the LC mod-

TABLE 4
Decision Criteria mentioned by the Participants

Decision Criterion	Count
Point Clusters	11
Cluster Size	5
Cluster Shape	3
Cluster Density	2
Other Point Features	8
Line Features	3
Relative Position	9
Reading Direction	2
Previous Point of Attention	6
Overlapping Points	2
Visual Presence of the Dot in Question	1
Labels Passed while Looking for the Highlighted Dot	1
Label Shape	1
Label Content	1
Typography	1
Centrality of the Label	1
Number of Labels	1
Overall Aesthetics	1
Intuition	3

TABLE 5
Kendall's rank correlation coefficient between fitted parameters of the best model (ECS) and Likert-scale ratings

	w	h	f	c
Distance	$\tau = -.0364$ $p = 1$	$\tau = -.0756$ $p = 1$	$\tau = -.1206$ $p = 1$	$\tau = .1642$ $p = .6249$
Presence of Dots	$\tau = .2739$ $p = .0079$	$\tau = .3273$ $p = .0006$	$\tau = .4563$ $p = .0000$	$\tau = -.2883$ $p = .0040$
Font Size	$\tau = .2909$ $p = .0031$	$\tau = .2576$ $p = .0145$	$\tau = .2500$ $p = .0202$	$\tau = -.2178$ $p = .0743$

p-values adjusted according to Bonferroni.

els). Within their open-ended answers shown in Table 4, many participants defined the decision criterion ‘presence of points’ further. Eleven participants stated that the presence of point clusters had an impact on their decision and nine participants mentioned other features, such as line features being influential. This aligns with the aspects described in literature (see subsection 2.1). Refining the influence of clusters, their size (5), density (2), and shape (3) were seen by the participants as possible criteria.

We further investigated the relationship between the subjective ratings and the parameters by fitting an individual model for each person and computing correlations between the fitted parameters and the Likert-scale ratings (see Table 5). The ratings for presence of dots and font size correlate positively with parameters modeling font size (w,h) and cluster (f) influence at medium effect sizes ($.27 < \tau < .46$). This indicates that the subjective judgments of the influencing factor importance is also reflected in the fitted parameters of our models.

Within their open-ended answers participants introduced further decision criteria that were important to them. Multiple participants mentioned an influence of the mouse cursor position or previous locus of attention (6), but this is an influence we could not remove and remains a limitation of our study design. Another recurring decision criterion was the relative position of dot and label (9). Some participants reported that they tended to choose a label positioned right, left, above, or below the highlighted dot. Two of them

explicitly mentioned writing direction to have an impact. By manually checking the selected labels, we found that the most frequently clicked label was more often above the query dot than below even though there were cases in which another label was closer. These two findings indicate that the relative position of dot and label as a decision criterion is worth further investigations. Several non-recurring decision criteria mentioned by the participants, concerning label characteristics, label positions, or non-varied aspects like color, may need further investigation as well.

7 CONCLUSION AND FUTURE WORK

Label placement guidelines from cartography are not directly applicable to information visualization scenarios. When labeling point clouds – as is typical in document visualizations – rather naive strategies are used instead, that may lead to misinterpretation of the dot label relationship. We developed twelve model classes for predicting how humans judge this relationship, focusing on the geometric properties of labeled point cloud representations. We performed a controlled study to assess which model class predicts human behavior for judgment of dot-label relations most accurately. We identified a single best model, but also investigated the performance difference between the other models to inform guidelines for label placement. Our findings support the following guidelines for designing label placement algorithms:

Measure distances using the actual label shapes. Current naive label placement strategies often just equate the label center with the point cloud center and do not take occlusions and concealment into account. Our results indicate that distance is a crucial criterion for judging the relationship between a dot and a label. However, the label center does not appear to be the point of reference for distance measure. A precise layout algorithm should therefore measure distances between dots and labels using their actual shapes. However, as this is computationally expensive, using reference points on the bounding boxes of the label or its characters instead are reasonable alternatives.

Consider label sizes and visual clusters. Our results also indicate that font size and the visual clustering structure of dots have an impact on the judgment. Hence, label placement algorithms should consider both factors in their layout strategy. Considering clusters should be preferred, as dot clustering does have a higher impact on the judgments than differing label sizes.

Weight label sizes anisotropically. When label sizes are considered by the layout algorithm, a weighting that anisotropically distorts space according to label width and height should be used.

Our results also indicate, that there may be further factors influencing human judgment of the dot-label relationship. Subjective feedback and some manual checking indicated that not only distance, but also the relative position of the label and the dot could be such a factor to be studied in future work. In addition, turning towards other characteristics of visualization design, such as dot size, shape, and color, could deliver valuable insights on

how human perception of relationships between dots and labels is further influenced. Many related applications use color to highlight the relationship among entities, but how colors interact with the perception of distance and cluster density, and how similar colors influence the perception of relationships have yet to be studied and modeled. Also, we are interested in developing simplified models that still have good predicting power.

REFERENCES

- [1] Q. Han, M. John, S. Koch, I. Assenov, and T. Ertl, "LabelTransfer - Integrating Static and Dynamic Label Representation for Focus+Context Text Exploration," *2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)*, pp. 1–8, 2018.
- [2] J. Stray, "A full-text visualization of the Iraq War Logs," 2010, <http://jonathanstray.com/a-full-text-visualization-of-the-iraq-war-logs> (Retrieved 2019-10-03).
- [3] D. Thom, R. Krüger, and T. Ertl, "Can Twitter Save Lives? A Broad-Scale Study on Visual Social Media Analytics for Public Safety," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 7, pp. 1816–1829, July 2016.
- [4] K. Field, *Cartography: A Compendium of Design Thinking for Mapmakers*. Esri Press, 2018.
- [5] A. Wolff, "The Map-Labeling Bibliography," 2019, <https://i11www.iti.kit.edu/~awolff/map-labeling/bibliography/> (Retrieved 2019-03-27).
- [6] E. Imhof, "Positioning Names on Maps," *The American Cartographer*, vol. 2, no. 2, pp. 128–144, 1975.
- [7] T. A. Slocum, R. B. McMaster, F. C. Kessler, and H. H. Howard, *Thematic Cartography and Geovisualization*, 3rd ed., ser. Prentice Hall Series in Geographic Information Science. Prentice Hall, 2009.
- [8] T. Götzelmann, K. Ali, K. Hartmann, and T. Strothotte, "Form Follows Function: Aesthetic Interactive Labels," in *Proceedings of the First Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, ser. Computational Aesthetics'05. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2005, pp. 193–200.
- [9] M. A. Bekos, B. Niedermann, and M. Nöllenburg, "External Labeling Techniques: A Taxonomy and Survey," *Computer Graphics Forum*, vol. 38, no. 3, pp. 833–860, 2019.
- [10] J.-D. Fekete and C. Plaisant, "Excentric Labeling: Dynamic Neighborhood Labeling for Data Visualization," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '99. New York, NY, USA: ACM, 1999, pp. 512–519.
- [11] M. Wertheimer, "Untersuchungen zur Lehre von der Gestalt. II," *Psychological Research*, vol. 4, no. 1, pp. 301–350, 1923.
- [12] C. Ware, *Information Visualization, Second Edition: Perception for Design (Interactive Technologies)*, 2nd ed. Morgan Kaufmann, Apr. 2004.
- [13] D. R. Montello, S. I. Fabrikant, M. Ruocco, and R. S. Middleton, "Testing the first law of cognitive geography on point-display spatializations," in *Spatial Information Theory. Foundations of Geographic Information Science*, W. Kuhn, M. F. Worboys, and S. Timpf, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 316–331.
- [14] G. M. Brosvic and B. D. Cohen, "The horizontal-vertical illusion and knowledge of results," *Perceptual and Motor Skills*, vol. 67, no. 2, pp. 463–469, 1988, pMID: 3217193.
- [15] L. Armstrong and L. E. Marks, "Differential effects of stimulus context on perceived length: Implications for the horizontal-vertical illusion," *Perception & Psychophysics*, vol. 59, no. 8, pp. 1200–1213, Dec 1997.
- [16] P. W. Thorndyke, "Distance estimation from cognitive maps," *Cognitive Psychology*, vol. 13, no. 4, pp. 526 – 550, 1981.
- [17] S. I. Fabrikant and D. R. Montello, "The effect of instructions on distance and similarity judgements in information spatializations," *International Journal of Geographical Information Science*, vol. 22, no. 4, pp. 463–478, 2008.
- [18] T. A. Slocum, "Predicting visual clusters on graduated circle maps," *The American Cartographer*, vol. 10, no. 1, pp. 59–72, 1983.
- [19] Y. Sadahiro, "Cluster perception in the distribution of point objects," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 34, no. 1, pp. 49–62, 1997.
- [20] A. Wolodtschenko and T. Forner, "Prehistoric and early historic maps in Europe: Conception of cd-atlas," *e-Perimetron*, vol. 2, no. 2, pp. 114–116, 2007.
- [21] E. Imhof, "Die Anordnung der Namen in der Karte," *International Yearbook of Cartography*, vol. 2, pp. 93–129, 1962.
- [22] P. Yoeli, "The Logic of Automated Map Lettering," *The Cartographic Journal*, vol. 9, no. 2, pp. 99–108, 1972.
- [23] J. Christensen, J. Marks, and S. Shieber, "An Empirical Study of Algorithms for Point-Feature Label Placement," *ACM Transactions on Graphics (TOG)*, vol. 14, no. 3, pp. 203–232, 1995.
- [24] B. Shneiderman and H. Kang, "Direct Annotation: A Drag-and-Drop Strategy for Labeling Photos," in *2000 IEEE Conference on Information Visualization. An International Conference on Computer Visualization and Graphics*. IEEE, 2000, pp. 88–95.
- [25] M. Formann and F. Wagner, "A packing problem with applications to lettering of maps," in *Proceedings of the seventh annual symposium on Computational geometry*. ACM, 1991, pp. 281–288.
- [26] M. Van Kreveld, T. Strijk, and A. Wolff, "Point Labeling with Sliding Labels," *Computational Geometry*, vol. 13, no. 1, pp. 21–47, 1999.
- [27] M. Nöllenburg and A. Wolff, "Drawing and Labeling High-Quality Metro Maps by Mixed-Integer Programming," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 5, pp. 626–641, 2011.
- [28] S. Oeltze-Jafra and B. Preim, "Survey of Labeling Techniques in Medical Visualizations," in *Proceedings of the 4th Eurographics Workshop on Visual Computing for Biology and Medicine*, ser. VCBM '14. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2014, pp. 199–208.
- [29] H. Abdi and L. J. Williams, "Principal Component Analysis," *WIREs Comput. Stat.*, vol. 2, no. 4, pp. 433–459, Jul. 2010.
- [30] L. v. d. Maaten and G. Hinton, "Visualizing High-Dimensional Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 01 2008.
- [31] C. Tominski, S. Gladisch, U. Kister, R. Dachsel, and H. Schumann, "Interactive Lenses for Visualization: An Extended Survey," *Computer Graphics Forum*, vol. 36, no. 6, pp. 173–200, Sep. 2017.
- [32] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing the non-Visual: Spatial Analysis and Interaction With Information from Text Documents," in *Proceedings of Visualization 1995 Conference*, vol. 51, 01 1995, pp. 51–58.
- [33] A. Ender, P. Fiaux, and C. North, "Semantic Interaction for Visual Text Analytics," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '12. New York, NY, USA: ACM, 2012, pp. 473–482.
- [34] J. Choo, C. Lee, C. K. Reddy, and H. Park, "UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, pp. 1992–2001, 2013.
- [35] R. R. O. da Silva, P. E. Rauber, R. M. Martins, R. Minghim, and A. Telea, "Attribute-based visual explanation of multidimensional projections," in *EuroVA@EuroVis*, 2015.
- [36] E. Kandogan, "Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations," in *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct 2012, pp. 73–82.
- [37] B. Fortuna, M. Grobelnik, and D. Mladenić, "Visualization of Text Document Corpus," *Informatika (Slovenia)*, vol. 29, pp. 497–504, 11 2005.
- [38] H. Schumann, M. Luboschik, and H. Cords, "Particle-based labeling: Fast point-feature labeling without obscuring other visual features," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 06, pp. 1237–1244, nov 2008.
- [39] R. Tamassia, *Handbook of Graph Drawing and Visualization (Discrete Mathematics and Its Applications)*. Chapman & Hall/CRC, 2007.
- [40] S. Hadl, H. Schumann, and H.-J. Schulz, "A Survey of Multi-faceted Graph Visualization," in *Eurographics Conference on Visualization (EuroVis) - STARS*, R. Borgo, F. Ganovelli, and I. Viola, Eds. The Eurographics Association, 2015.
- [41] C. Chen, F. Ibeke-SanJuan, and J. Hou, "The Structure and Dynamics of CoCitation Clusters: A MultiplePerspective CoCitation Analysis," *Journal of the American Society for information Science and Technology*, vol. 61, no. 7, pp. 1386–1409, 2010.
- [42] J. Abello, F. V. Ham, and N. Krishnan, "ASK-GraphView: A Large Scale Graph Visualization System," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 669–676, Sep. 2006.

- [43] S. Jänicke, J. Focht, and G. Scheuermann, "Interactive Visual Profiling of Musicians," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 200–209, Jan 2016.
- [44] B. Shneiderman and C. Dunne, "Interactive network exploration to derive insights: Filtering, clustering, grouping, and simplification," in *Graph Drawing*, W. Didimo and M. Patrignani, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 2–18.
- [45] P. C. Wong, P. Mackey, K. Perrine, J. Eagan, H. Foote, and J. Thomas, "Dynamic visualization of graphs with extended labels," in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005*. IEEE, Oct 2005, pp. 73–80.
- [46] M. Streit, M. Kalkusch, K. Kashofer, and D. Schmalstieg, "Navigation and Exploration of Interconnected Pathways," *Computer Graphics Forum*, vol. 27, no. 3, pp. 951–958, 2008.
- [47] F. Schreiber, T. Dwyer, K. Marriott, and M. Wybrow, "A generic algorithm for layout of biological networks," *BMC bioinformatics*, vol. 10, p. 375, 2009.
- [48] M. Rohrschneider, C. Heine, A. Reichenbach, A. Kerren, and G. Scheuermann, "A novel grid-based visualization approach for metabolic networks with advanced focus&context view," in *Proceedings of the 17th International Conference on Graph Drawing*, ser. GD'09. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 268–279.
- [49] R. Jianu, K. Yu, L. Cao, V. Nguyen, A. R. Salomon, and D. H. Laidlaw, "Visual integration of quantitative proteomic data, pathways, and protein interactions," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 4, pp. 609–620, July 2010.
- [50] J. Heer and D. Boyd, "Vizster: visualizing online social networks," in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005*, Oct 2005, pp. 32–39.
- [51] J. Novak, I. Micheel, M. Melenhorst, L. Wieneke, M. Düring, J. G. Morn, C. Pasini, M. Tagliasacchi, and P. Fraternali, "Histograph – a visualization tool for collaborative analysis of networks from historical social multimedia collections," in *2014 18th International Conference on Information Visualisation*, July 2014, pp. 241–250.
- [52] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, "Generating Summaries for Large Collections of Geo-referenced Photographs," in *Proceedings of the 15th International Conference on World Wide Web*, ser. WWW '06. New York, NY, USA: ACM, 2006, pp. 853–854.
- [53] M. Reckziegel, M. F. Cheema, G. Scheuermann, and S. Jänicke, "Predominance Tag Maps," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 6, pp. 1893–1904, June 2018.
- [54] D. Nguyen and H. Schumann, "Taggram: Exploring Geo-data on Maps through a Tag Cloud-Based Visualization," in *2010 14th International Conference Information Visualisation*, July 2010, pp. 322–328.
- [55] K. Buchin, D. Creemers, A. Lazzarotto, B. Speckmann, and J. Wulms, "Geo Word Clouds," in *2016 IEEE Pacific Visualization Symposium (PacificVis)*, April 2016, pp. 144–151.
- [56] N. Ferreira, L. Lins, D. Fink, S. Kelling, C. Wood, J. Freire, and C. Silva, "BirdVis: Visualizing and Understanding Bird Populations," *IEEE transactions on visualization and computer graphics*, vol. 17, pp. 2374–83, 12 2011.
- [57] S. Ahern, M. Naaman, R. Nair, and J. H.-I. Yang, "World Explorer: Visualizing Aggregate Data from Unstructured Text in Geo-referenced Collections," in *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL '07. New York, NY, USA: ACM, 2007, pp. 1–10.
- [58] D. Thom, H. Bosch, S. Koch, M. Wörner, and T. Ertl, "Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages," in *2012 IEEE Pacific Visualization Symposium*, Feb 2012, pp. 41–48.
- [59] M. Waldner, J. Schrammel, M. Klein, K. Kristjánsdóttir, D. Unger, and M. Tscheligi, "FacetClouds: Exploring Tag Clouds for Multi-dimensional Data," in *Proceedings of Graphics Interface 2013*, ser. GI '13. Canadian Information Processing Society, 2013, pp. 17–24.
- [60] F. Viégas, M. Wattenberg, and J. Feinberg, "Participatory Visualization with Wordle," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 15, no. 6, pp. 1137–1144, Nov 2009.
- [61] R. Ratcliff and G. McKoon, "The diffusion decision model: Theory and data for two-choice decision tasks," *Neural Computation*, vol. 20, no. 4, pp. 873–922, 2008, pMID: 18085991.
- [62] J. Heer and M. Bostock, "crowdsourcing graphical perception: Using mechanical turk to assess visualization design."
- [63] A. T. Woods, C. Velasco, C. A. Levitan, X. Wan, and C. Spence, "Conducting perception research over the internet: a tutorial review," *PeerJ*, vol. 3, p. e1058, Jul. 2015.
- [64] R. Bridson, "Fast poisson disk sampling in arbitrary dimensions," in *ACM SIGGRAPH 2007 Sketches*, ser. SIGGRAPH '07. New York, NY, USA: ACM, 2007.
- [65] L. Wasserman, *All of statistics: A concise course in statistical inference*. Springer, 2004.
- [66] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The computer journal*, vol. 7, no. 4, pp. 308–313, 1965.
- [67] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, December 1974.
- [68] K. P. Burnham and D. R. Anderson, *A practical information-theoretic approach: Model selection and multimodel inference*, 2nd ed. Springer, 2002.

Martin Reckziegel received his master degree (diploma) in computer science in 2013. Currently, he is a PhD candidate at the Image and Signal Processing group at Leipzig University. His doctoral studies are co-supervised by Prof. Gerek Scheuermann and Asst. Prof Stefan Jänicke. His research interest includes unsolved problems in geovisualization, driven by research questions in application fields such as digital humanities and environmental sciences.

Linda Pfeiffer received a MSc degree in Intelligent Media and Virtual Reality from Chemnitz University of Technology in 2015. She is currently with the Visual Analytics group at the Institute of Data Science, German Aerospace Center DLR. Her research interests include interactive visualizations and their related social, perceptual and cognitive aspects in application domains such as air traffic control and environmental sciences.

Christian Heine holds a MSc and a PhD degree in computer science from Leipzig University, Germany. He also holds a BA degree in philosophy, neuroscience, and cognition from Magdeburg University. Previously he has worked at ETH Zürich, Switzerland, Heidelberg University, TU Chemnitz, and TU Kaiserslautern, Germany. His research interests include theories of visualization, perception and cognition theory, topology-based visualization, and graph drawing.

Stefan Jänicke is an Assistant Professor at the Department of Mathematics and Computer Science at the University of Southern Denmark, Odense, Denmark. He received his PhD degree in computer science in 2016 from Leipzig University, Germany. His PhD thesis investigates the utility of visualization techniques to support the comparative analysis of digital humanities data. He has gained experience in developing information visualization and visual analytics techniques in numerous interdisciplinary research projects addressing current research questions in (digital) humanities, linguistics, social sciences, biology and sports. His research interests relate to information visualization with a focus on text- and geovisualization.