LINKING MULTIPLE PERSPECTIVES WITH OBJECT-BASED VISUAL CUES FOR SPATIAL VIDEO ANALYSIS

D. Hollenstein[1]*, S. Bleisch[1]

[1] Institute of Geomatics, FHNW University of Applied Sciences and Arts Northwestern Switzerland, Muttenz, Switzerland
(daria.hollenstein, susanne.bleisch)@fhnw.ch

Commission IV, WG IV/9

KEY WORDS: geovisualization, video, map, multiple perspectives, visual localization, uncertainty

ABSTRACT:

The visual analysis of videos in context with mapped information requires support in the challenges of linking different spatial perspectives (e.g., street level and survey perspective), bridging different levels of detail, and relating objects in different visual representation. Uncertainty in the spatial relation between camera views and map complicates these tasks. We implemented visualizations for the visual analysis of street level videos (i.e., video key frames) embedded in their spatial context. As part of this, we developed a design rationale for visual cues that help link the video key frames and a map in cases where the spatial relation between camera view and map is of uncertain accuracy. We implemented three cue types (simplified viewshed, object-based dot cues, street centre line cues) for an image data set with heterogeneous camera localization accuracy and assessed the resulting cue properties. Based on this, we argue in favour of cue designs that minimize uncertain information required for their display at the expense of cues' spatial explicitness in cases of potentially low camera localization accuracy. When localization accuracy is expected to be at least moderate, particularly, dot cues that refer to unambiguous points of reference within easily recognizable objects of ample size present a promising option to support view co-registration.

1. INTRODUCTION

Embedded in their spatial context, video recordings of spatio-temporal events, such as floodings or accidents, are a valuable source of information for in-depth event analysis. However, the analysis of images or videos in context with mapped information entails the visual challenges of linking different spatial perspectives (e.g., street level and survey perspective), bridging different levels of detail and objects in different visual representation (Plumlee and Ware, 2003b, Wang et al., 2007, Wang, 2010). Without visual cues that establish links between the street level camera views and the survey map, this process may be challenging, time consuming or literally impossible for users with limited local knowledge.

Particularly, in cases, where a detailed spatial relationship between map and image objects must be established to understand an event, cues in addition to camera position and direction of view are desirable (Plumlee and Ware, 2003a, Tory, 2003, Wang, 2010). However, cues, such as camera coverage or cues that indicate homologous objects in the camera view and the map, require information on 3D camera position and orientation. In application scenarios such as video surveillance (Wang et al., 2007, Girgensohn et al., 2007), virtual environments (Plumlee and Ware, 2003b) or street view applications (Kopf et al., 2010), the spatial relation between views is usually known and adequate for the targeted use and visualization. For the analysis of unstructured, crowd-sourced video collections, visual localization or image-to-model localization may yield estimates of camera position and orientation (Baboud et al., 2011, Tompkin et al., 2012, Karsch et al., 2014, Brejcha et al., 2018, Meyer et al., 2020). However, such information comes at varying, potentially unknown degrees of accuracy. Making use of this information for visualization purposes may result in unreliable visual cues with compromised functionality.

For a project aimed to leverage eyewitness videos for crisis management, we developed visualizations to support the analysis of street level videos together with a map. We developed a design rationale for visual cues that help link camera views (i.e., video key frames) and a map when the spatial relation between camera and map is of unknown, potentially limited accuracy, while the aim of the visual analysis is detailed scene understanding.

Task Description: The design rationale followed the consideration of non-application specific, low-level 'integrative' tasks (Wang, 2010) in the context of spatial scene understanding, in the planview are aimed at relating information from the map and from the videos:

1. Coarse alignment of views: Identification of approximate camera positions and viewing directions on the map and with respect to each other.
2. Visual assessment of camera localization accuracy.
3. Detailed visual registration of camera views and map: Identification of mapped objects in the camera views and localization of objects in camera views on the map.

Nature of Data and Uncertainty: Our design targets video key frames that are recorded at the street level in built-up areas and georeferenced through visual localization (Meyer et al., 2020). However, the images used for the implementation of the visualizations originate from a street view data set recorded with low-cost sensors (Nebiker et al., 2021). The used data set has properties similar to the data targeted with the design. Likewise, it comes with heterogeneous camera localization accuracy which leads to angular, lateral and depth error (Holloway, 1995) in the

* Corresponding author

https://doi.org/10.5194/isprs-archives-XLIII-B4-2022-455-2022

registration of map and image objects. Further, misalignment may also be introduced by generalisation, positional resolution, and accuracy of the mapped data.

2. RELATED WORK

2.1 Providing cues for the linking of camera and map views

A variety of visual cues, view layout and interaction options have been put forward to support the linking of different spatial perspectives (Plumlee and Ware, 2003b, Plumlee and Ware, 2003b, Tory, 2003). In frameworks and interfaces for the exploration and analysis of videos together with a map, commonly, one or a combination of the following approaches is applied:

Display of information about the camera view on the map: In applications where camera position is the only information available, the display of camera positions and video trajectories on the map is combined with linked highlighting and options for spatial video navigation to support video-based space exploration (Mildner et al., 2013, Jamonnak et al., 2020). GeoVisual (Jamonnak et al., 2020), a GIS for geo-tagged, semantically annotated multimedia data, is aimed at visual and quantitative video-based spatial analysis. Videos are browsed along GPS trajectories on a map or by sliding along a street view interface that displays available images at a location together with context from Google Street View.

VideoScapes (Tompkin et al., 2012), is an application for causal exploration of city areas through collections of unordered street-level videos. The videos are structured in a graph like manner for this purpose. A survey map shows edges (video sequence trajectories) and nodes (junctions of several video sequences). By hovering over edges, the user browses the corresponding video, while the camera's field of view is displayed and highlighted on the map. At nodes, the user can switch videos. The display of camera viewing directions requires at least information on camera orientation in 2D, as e.g. from a device compass.

The display of cameras' coverage on the map is even more spatially indicative but requires 3D camera position and orientation, the camera field of view and a 3D model of the environment. Wu & Tory (2009) display the spatial coverage of images and gridded coverage heat maps on floorplans of buildings for the use of image collections in construction management (Wu and Tory, 2009). Other applications display camera position, viewing direction and field of view or approximate coverage (Girgensohn et al., 2007, Zhang et al., 2010).

Dynamic view arrangement, alignment, and coupling To help the integration of different perspectives, some applications support the alignment of the map with the camera viewing directions or spatially ordered layouts of views. Wang (2010) presents a testbed interface for 'contextualized video' visualization in a surveillance scenario. The testbed provides the choice of a 2D or a 3D floorplan and associated video views placed at the plan margin, embedded in the plan view or both. Findings from a user study indicate task dependent advantages of different view combinations and layouts (Wang, 2010). For 'integrative tasks', Wang (2010) suggests emphasizing the video and its close spatial context. DOTS (Girgensohn et al., 2007), another video surveillance interface, provides a spatial multi-video player: Video views are arranged according to their spatial layout . This way, a tracked person moves from one view to the next. A survey floorplan is given for context and aligned forward-up with the video view at the centre of the spatial player. When the centre view is switched, the layout is re-arranged around the new view's position.

Apart from the alignment of video and map view, the coupling of zooming and panning functions over several views is applied in similar analysis situations in entirely virtual environments without video views (Plumlee and Ware, 2003b, Pindat et al., 2013). Further, mutually indicating the cursor position in different perspective views is a promising option to support orientation (Kopf et al., 2008, Zhang et al., 2010). Different dynamic transitions between camera views are also explored in this regard (de Haan et al., 2010, Tompkin et al., 2012).

Display of information extracted from the video on the map and in the video view: This cue type requires video information extraction (e.g., objects of interest) and is specifically targeted at supporting spatio-temporal event analysis. Usually, it is coupled with timeline and spatial interaction options. The video surveillance tool DOTS (Girgensohn et al., 2007), displays persons' trajectories, as extracted from the videos, in the video as well as on the floorplan. Videos can be played by dragging a person icon along the floorplan. Stein et al. (2018) present a tool for soccer play analysis. Player tracking and analytics from different camera views are projected onto a normalized pitch and back onto the camera view. Active players are mutually highlighted.

Display of static mapped information in the camera view: In our own approach, we propose the display of map elements in the camera view to provide orientation and connect the street view perspectives with each other and with their spatial context in the survey map. This approach is inspired by similar uses of map data overlays in photographs (Kopf et al., 2008, Karsch et al., 2014) and Augmented Reality (AR) (Veas et al., 2012). Veas et al. (2012) project topographic vector elements into views of a multi-view outdoor AR-application to embed the views in their spatial context. Also, they include vector elements to extend the view of a single camera to show content from variable perspectives. Kopf et al. (2008) augment photographs with geo-data (e.g. roads, landmark and location names) and present it side-by-side to a map. The augmented photograph's view frustum is displayed on the map and while moving the cursor in one view, its position is indicated in the other.

This type of cue seems less explored in application to video analysis. This might be due to the circumstances that for many applications, e.g., video surveillance or sports analytics, the spatial context is restricted and potentially well-known (Girgensohn et al., 2007, de Haan et al., 2010) or perceptually very similar in the video and the map, e.g., a soccer pitch (Stein et al., 2018). However, Zhang et al. (2010) implemented an algorithm for semi-automatic registration of tourist videos to a 3D city model and automated landmark labelling in the videos. They include the annotated videos into a map-based interface, where camera position, field of view and landmark highlighting are displayed dynamically as the video is played. In addition, clicking on building in the video highlights the same on the map. After a user test they conclude that *"deep integration"* of videos with a map can improve navigation (Zhang et al. 2010, p. 268).

2.2 Dealing with inaccuracy and uncertainty

In surveillance applications or virtual environments, usually, information on the spatial relation between different views is suf-

ficiently accurate for the display of visual cues, such as camera viewpoint, direction, and coverage (the latter also requires an adequate 3D model of the environment). In applications that are more like ours (Kopf et al., 2008, Tompkin et al., 2012, Zhang et al., 2010, Karsch et al., 2014), camera position and orientation from low-cost sensors or from image registration processes varies in accuracy and may compromise the quality and functionality of the resulting visualizations. However, visualization issues related to uncertain input information and unsatisfactory visualizations as a result of poor registration are not discussed broadly in the above-mentioned work. Zhang et al. (2010) discuss limitations of their registration process and problems with the resulting stability of annotations. They conclude that despite these problems, users gain information through the camera-map registration. Issues with GPS position accuracy (Mildner et al., 2013, Tompkin et al., 2012) or varying results of image registration are mentioned (Karsch et al., 2014) elsewhere. However, their effects on the visual results are not discussed in detail.
More commonly, issues of visual coherence that relate to limited registration accuracy, are discussed (Azuma, 1997, Azuma et al., 2001, Zollmann et al., 2021) and addressed in the context of AR-applications: Holloway (1995) developed a model to analyse the consequences of registration error. He decomposes misalignment of virtual and real points into an angular, a lateral and a depth error component. MacIntyre & Coelho (2000) propose level of error filtering (LOE), to dynamically adapt visualizations to the estimated current error of the tracking system. With LOE, an object highlighting frame is sized up and eventually replaced by a textual description as the tracking error increases to prevent ill-placed or misleading attentional cues. This is extended to the use of error ellipses of object vertices to calculate expanded and shrunk convex hulls of objects for error robust highlighting and labelling (MacIntyre et al., 2002). Möller et al. (2012) present an indoor navigational aide that switches visualization modes (from AR to VR) when location estimation becomes unreliable. Pankratz et al. (2013) address the problem of tracking errors in AR-navigational aids with visualizations that indicate the degree of tracking error, e.g., through a change in symbol shape and colour.

Following postulates from AR research (MacIntyre and Machado Coelho, 2000, Möller et al., 2012, Pankratz et al., 2013), we believe that registration errors from visual localization are a problem that should be considered in visualizations building on such data. Otherwise, automated workflows will yield confusing or misleading visualizations when localization accuracy is limited. We draw from studies on AR-visualization (Veas et al., 2012), augmented photographs (Kopf et al., 2008), and 2D-3D view registration (Tory, 2003) to develop a design rationale for visual cues that help link different perspective views when their spatial relation is uncertain. The objective of the cue design is integration with an interface that provides interactive arrangement, alignment and browsing of views to support the outlined tasks (c.f. 1) associated with the analysis of videos in their spatial context (Wang et al., 2007, Wang, 2010).

3. DESIGN RATIONALE

We composed an interface prototype with D3.js (Bostock, 2021) to support the outlined view registration tasks (c.f. 1) in a scene of interest. The interface includes a close-up view of video key frames and their nearby spatial context (Wang, 2010). On-demand, the orientation of the close-up map can be switched from north-up to align with the viewing direction of a key frame (Darken and Cevik, 1999, Plumlee and Ware, 2003a). This interface presents a starting point for the design and integration of visual cues that support the linking of camera and map views.

3.1 Requirements

With the described tasks and data characteristics at hand, we defined design requirements for visual cues that help link street level camera and survey perspective map views under varying conditions of spatial accuracy.

1. **Cues should indicate a camera's viewpoint, viewing direction and depth.** Indicating the viewpoint and viewing direction of one view in another, is a crucial aid for linking information from those two views (Plumlee and Ware, 2003a). Providing depth cues for images may further support the visual registration of far-field image objects with the map (Livingston et al., 2009).
2. **Cues ideally support pre-attentive pattern matching between views.** According to insights from an empirical study (Tory, 2003), this could strongly support the linking of different perspective views.
3. **Referents of object-based cues should be recognizable in the map and in the camera view.** For object-based cues to function, the object of reference should be visually identifiable in both views. E.g., individual parking lots or streetlamps may be visible in a video, but not available in the map, while large trees may be mapped, but not easily identified among other vegetation in street level videos. Beyond, the referent position on the object should be visible and recognizable in the image. Referent position visibility facilitates the mental matching of cue and object (Furmanski et al., 2002), recognizability may help perceive misalignment.
4. **Cue design should consider and minimize the issues of unknown and varying data accuracy.** Spatially explicit cues (e.g., indication of analogues objects) promise to be powerful. However, with increasing uncertainty of the input data (camera localization, map data), such cues are expected to lose their functionality and may become misleading (Azuma et al., 2001, Pankratz et al., 2013). Hence, while trying to conform to requirement III, cue design should minimize the uncertain information required for display.
5. **The density of cues per image and per scene should be high enough to be informative with respect to requirement I and II, but not as high as to cause clutter and confusion.** Too many cues will lead to view occlusion problems and may be difficult to process, particularly for users with limited local knowledge (Wang, 2010).

3.2 Cue Design

Based on the outlined requirements, we designed visual cues that establish links between street level camera and survey map views. Viewshed computation and cue selection was done using GRASS GIS (GRASS Development Team, 2022), image processing relied on the Python library matplotlib (Matplotlib Development Team, 2021).

3.2.1 Camera position and simplified viewshed on the map: Camera position and a simplified viewshed is symbolized with a circle and an outline on the map (Fig. 2A). A colour-code links the symbols with the respective camera frame. These cues indicate the street level view's recording position and direction on the map, but not in the camera view. Also, they do not provide direct links between camera views or indicate image-objects' geolocation, except where the (near-end) lateral side of the viewshed intersects with mapped objects that are recognizable in the image. The computation of the viewshed requires information on camera position, viewing direction, and the camera field of view. The viewshed was limited to 75 m from the camera and the vertical angle of the camera field of view was not considered in the computation. The viewshed was calculated using a simplified model of the environment. It contained information on terrain elevation (Amt für Geoinformation Kanton Basel-Landschaft, 2018) as well as buildings from a 2D data set (Amt für Geoinformation Kanton Basel-Landschaft, 2021) that best matched the visible building outlines as seen from the street level view. A uniform height was set for all buildings. A fixed elevation above ground was assumed for the camera. This was possible due to the nature of the image data set and eliminated problems arising from large errors in camera z-coordinates (Nebiker et al., 2021). The viewshed raster was resampled at 2 meters for vectorization, isolated areas ($<$100 m^2) were discarded and the viewshed outline buffered (2 m), smoothened and clipped (5 m) around the camera position. This generalization results in an outline that is easy to read but introduces inaccuracies itself (Fig. 1). The simplified viewshed was used also in the selection of visible cue objects in the key frames (see below).

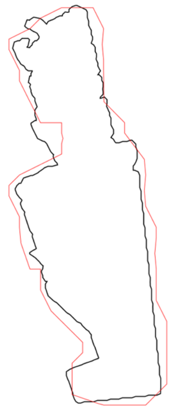

Figure 1. Outline of a viewshed computed at 0.25 m [black] and at 2 m [red] resolution.

3.3 Object-based line cues

We display colour-coded road centre lines (Bundesamt für Landestopografie swisstopo, 2021) in the camera view and the map to support view linking (Fig. 2C). Line cues that reference unambiguous linear features have the potential to indicate the approximate camera viewpoint, viewing direction and image depth by indicating the location of image objects relative to cues, and to support pattern matching between views, all given that, the linear features are sufficiently salient. In the case of road centre lines, this implicates, they include distinct segments, e.g., crossroads. Line cues with a recognizable referent may also serve as a visual estimate of registration accuracy. The display of linear objects requires the registration of map points in x, y and z with the camera view.

3.4 Object-based dot cues

We designed dot cues using a street number data set with individual points at building fronts (Amt für Geoinformation Kanton Basel-Landschaft, 2021). Points that are inside the viewsheds of at least two cameras were selected for the display of dot cues. The dots are equally colour-coded through all views (Fig. 2B). In the camera view, the cues are displayed 2m above ground and scaled and sorted with respect to viewing distance. The image cues are displayed with a uniform blur and transparency. In combination with the distance-dependant scaling and sorting, this may support the interpretation of depth in the camera view (Drascic and Milgram, 1996). Similar to line cues, dots with recognizable referents (buildings) in the camera view might indicate approximate camera position, direction of view and the geolocation of image objects through their relation to cues. Dot cues may also provide patterns to match between different views. Dot cues still require the registration of map points in x, y, with the camera view. However, choosing an object of ample size (building) and within the object a referent position that is unspecific with respect to elevation (thus neglecting the dimension which is the least informative regarding orientation in our case), makes the registration more flexible with respect to error and thus may help preserve cue functionality when camera localization accuracy is limited. Through their referent position at the building front, near the entrance, the dots may provide some visual indication of registration accuracy. Compared to the line cues, however, this function is expected to be weaker in favour of the improved error robustness.

3.5 EVALUATON AND DISCUSSION OF VISUAL CUE QUALITY

We implemented the three cue types on a selection of 110 street level images from a data set with heterogeneous georeferencing accuracy (Nebiker et al., 2021). The images were selected to yield scenes (15) of 7-8 images. All selected images covered the street centre in some part. The images of a scene were displayed in a static layout with a map at the centre (Fig. 2). For every scene, the localization accuracy varies between images. A formal evaluation of the designs' functionality would have been desirable at this point but was not attempted, due to the lack of a data set with more homogeneous localization accuracy.

We conducted a visual assessment of the resulting cue quality in the scenes with the outlined tasks (c.f. 1) in mind: We assessed the resulting visualizations with respect to readability and expected support in coarse and detailed image-to-map registration and error indication. For dot and line cues, this included the consideration of feature information value. Further, we identified cases where misplaced cues potentially become misleading. In relation to this, we quantified misaligned dot cues (dot centre appears on wrong building) and incorrect viewshed-building intersection for the near-end lateral side of the camera views. For isosceles-like viewsheds, building intersections of both lateral sides were considered. Based on this visual assessment, we infer and discuss the potential and limitations of the different designs with respect to the design rationale. The findings from the evaluation procedures are reported per cue type. The sections contain concluding findings in italics, followed by the observations that led to these findings and some suggestions for design improvements.

3.6 4.1 Viewshed cues

The viewshed provides limited support in the fast coarse alignment of images and map: The viewshed does provide the information needed for the coarse alignment of views. However,

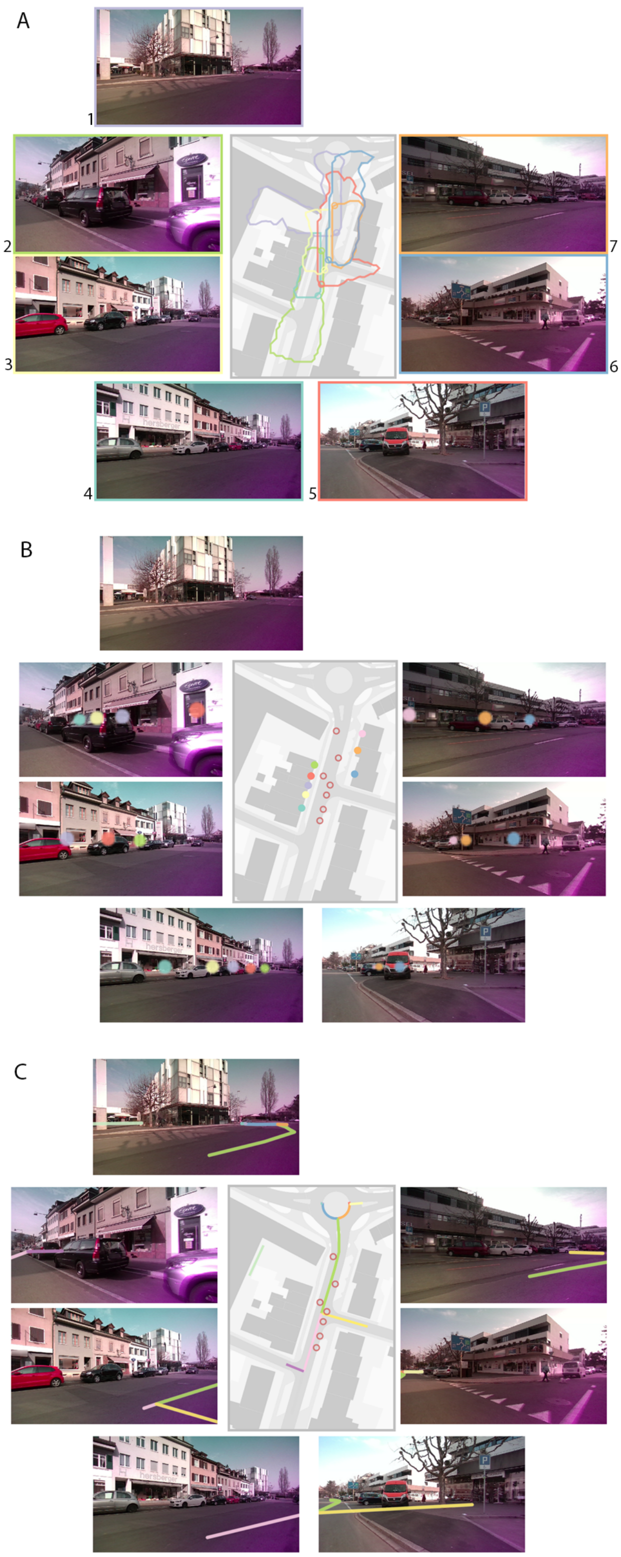

Figure 2. Identical scene with different types of colour-coded visual cues that link camera views and map. From top to bottom: simplified viewshed, dot cues, street centre line cues. (images: (Nebiker et al., 2021), geo-data: (Amt für Geoinformation Kanton Basel-Landschaft, 2021, Bundesamt für Landestopografie swisstopo, 2021)). Image 2 illustrates a potentially misleading case for both, viewshed and dot cues. In image 4, the near-end lateral viewshed outline intersects with the correct building on the map and might support correct object-wise alignment, while red and green dot cues misalign in the image and may be confusing

the 7 to 8 outlines clutter on the map and it takes some time to visually disentangle them and identify a specific outline and its corresponding camera position (Fig. 2A). Adding overlay blending, linked highlighting and image scrolling, when integrating this cue type in an interactive visualization, may mitigate this problem.

The display of the simplified viewshed outline provides the most robust support for the detailed linking of image and map objects: From the visual assessment of the scenes, we hypothesize that the intersection of the near-end lateral outline of the viewshed with a building close to the camera is probably the strongest cue for establishing object correspondence. Thus, misleading cues may result in cases where the near-end lateral viewshed outline intersects with the wrong building (e.g., Fig. 2A, image 2). Incorrect intersections occur in 21.8% of the images. Some of these errors may be due to viewshed generalisation.

The viewshed provides some indication of camera localization accuracy: Errors are recognizable, when a viewshed suggests the visibility of the entire road width or of buildings on both street sides when, in fact, only one side is visible in the image (Fig.2A, image 2) or vice versa. The far end of the viewshed is most indicative, as angular errors are most pronounced at distance. The mentioned intersection errors at the near end, however, will be more difficult to identify and have the potential to mislead the user (Fig.2A, image 2). Blurring viewshed outlines may help alert the user about potential inaccuracies (Pankratz et al. 2013).

3.7 Dot Cues

The visual assessment confirms the potential of dot cues to support coarse scene understanding by fast identification of approximate camera positions and viewing directions: The visual assessment revealed that overall, dot cue visualizations provide pattern like cues, that are easy to read (Fig. 2B). However, 5.5% of the images provide no cue and another 24.5% of the images show only one dot, thus, cue information value is limited. Only in parts, cue shortage is due to the restriction of cues to referent points that appear in more than one image. In some cases, a building is well visible in the image, but the building's referent point is not (Fig. 2, images 1& 3). In some cases, the restriction of cue visibility to 75 m, may discard cues on prominent buildings that are recognizable from larger distance and could help in the linking of camera views (Fig.2, image 4). A more adaptive cue design that considers building size might improve dot cue functionality.

Dot cues are less suitable to support detailed image-to-map registration, when camera localization accuracy is low. Dot cues poorly convey registration inaccuracies: The two findings are related. The visual assessment revealed that in 42.7% of the images, at minimum one dot appears on the wrong building (c.f. Fig. 2B, images 2 & 4). Further, not only the referent position within the building, but also, mapped building units are not always easily recognizable in the images. Incorrect placement (dot on wrong building) and difficult to identify referent objects, result in cues that may be misleading in the task of detailed object-to-object registration (c.f. Fig. 2B). However, where dot cues appear on the correct building, they have the potential to also support detailed image-to-image and image-to-map registration.

3.8 Line Cues

Street centre line cues provide information for coarse view alignment but are not necessarily easy to read (Fig. 3): In 28.2% of the images, the selected section of the street network does not provide features that are salient enough to indicate the viewing direction of the camera. For some scenes, on the other hand, the network results in visually complex cues that probably prevent fast pattern matching. Other features are difficult to read because of registration errors.

Support for detailed mental image to map and image to image registration is limited: For street centre line cues, branching roads and paths technically help object-wise view registration. However, they are not necessarily (well) recognizable in the street view (c.f. Fig. 2C, image 3 & 7). As a result, it is difficult to judge cue accuracy and misplaced cues become misleading. Therefore, the cues provide limited support in detailed view registration and in the localization of image objects on the map.

Street centre line cues provides some indication of registration error (Fig. 2C, Fig. 3): Line cues are indicative of lateral and angular errors because deviations from the street centre are perceived (Azuma, 1997). Also, positive deviations from the ground plane (floating lines) are readable.

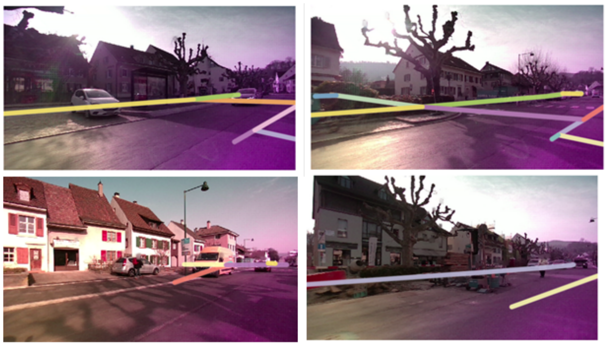

Figure 3. Examples of street centre line cues that are difficult to read because of a complex network situations (top), because of registration error (bottom left) or lack of directional information (bottom right) (images: (Nebiker et al., 2021); geo-data: (Bundesamt für Landestopografie swisstopo, 2021)).

4. DISCUSSION

In our work on visualizations to support the visual analysis of geolocated videos together with a map, we focused on the design of visual cues that link the street level camera views and the survey perspective map. Based on a design rationale, we implemented three cue types (simplified camera viewshed, coloured street centre lines and dot cues that reference points at buildings fronts) on a data set of 110 street view images with heterogeneous localization accuracy (Nebiker et al., 2021). We visually assessed cue properties in the resulting visualization to gain insight on the resulting cue functionality. The visual assessment led to the following insights: While by design, the display of key frames' simplified viewshed on the map is not as informative as the mutual display of object-based cues, the viewshed cues provide the most reliable information for detailed visual image to map registration for the used data set. Identifying image objects on the map with the help of line and dot cues will be difficult or impossible, depending on the scene, when camera pose accuracy is low. Line cues provide some information for coarse view alignment but are not necessarily easy to read. Dot cues, however, have the potential to support fast coarse alignment of scenes even when registration accuracy is limited. None of the designed cue types reliably communicates all types of error.

Test data set and evaluation methodology: The used data set has properties similar to the data that was targeted with the design. However, the findings regarding cue salience are dependent on the specific nature of this data set and likely deviate from the results in an unstructured collection of street view videos. Also, the robustness of the viewshed in the used data set is partly owed to a stable camera field of view and camera elevation above ground.
The chosen evaluation methodology has limitations: While the resulting visualizations can be visually assessed with respect to each requirement of the design rationale, the assessment of actual cue functionality with respect to different tasks (e.g. coarse alignment, object-wise registration) in a multi-view scene requires a user study. Some of the identified draw backs might not affect cue functionality strongly: E.g., sparse dot cues might not be a major issue with respect to the coarse alignment of views, because of salient image features that help bridging between views with few cues. A user evaluation of cue functionality based on a test data set with controlled variation in localization accuracy would be desirable, as this could provide insights on differential cue functionality in relation to different levels of localization accuracy.

Integration in interactive, dynamic visualizations: Further, the integration of color-coded cues in an interactive multi-view display that covers more than one scene raises the question of scalability with respect to the used color-coding. Also, work is needed to evaluate cue functionality in more dynamic, interactive visualizations with many views.

5. CONCLUSION

We addressed visualizations to support the visual analysis of street level videos in relation to their spatial context. We developed a design rationale for visual cues that help link street level camera views and a survey map in situations where the aim of the analysis is detailed scene understanding, but the spatial relation between camera views and map is uncertain. We designed cues that transfer information from the camera view to the map (viewshed cues) and cues that transfer information on mapped objects to the camera views (dot and line cues). We implemented these cues in an image data set with heterogeneous camera localization accuracy. Based on a visual assessment of the resulting cue properties, we suggest using cue designs that minimize uncertain information required for their display when localization accuracy is expected to be low. For the used data set, we suggest using viewshed-based cues when localization accuracy is expected to be low to support detailed camera-map co-registration. Including object-based dot cues when accuracy is expected to be moderate or high has the potential to support fast coarse alignment of different perspective views. Street centre line cues are not necessarily easy to read which limits their functionality. None of the designed cue types reliably indicates all types of error in camera view-map view registration. A combination of cue types maybe helpful to support all relevant visual tasks. However, cue efficiency with respect to specific tasks and different levels of registration accuracy warrants a user evaluation with accuracy-controlled data.

ACKNOWLEDGEMENTS

We thank Jonas Meyer, Ursula Kälin, and Stephan Nebiker for the provision of the data and instructions. This research is part of the EVAC Project of the National Research Programme ”Big Data” (NRP 75, www.nrp75.ch) of the Swiss National Science Foundation (SNSF).

REFERENCES

Amt für Geoinformation Kanton Basel-Landschaft, 2018. Digitales terrain modell (dtm) 2018. baselland.ch/politik-und-behorden/direktionen/volkswirtschafts-und-gesundheitsdirektion/amt-fur-geoinformation/geoportal/geodaten/geodatenprodukte/hoehenmodelle/digitales-terrain-modell-dtm (22 March 2022).

Amt für Geoinformation Kanton Basel-Landschaft, 2021. Av-daten. baselland.ch/politik-und-behorden/direktionen/volkswirtschafts-und-gesundheitsdirektion/amt-fur-geoinformation/geoportal/geodaten/geodatenprodukte/amtliche-vermessung (22 March 2022).

Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., MacIntyre, B., 2001. Recent advances in augmented reality. *IEEE Computer Graphics and Applications*, 21(6), 34-47.

Azuma, R. T., 1997. A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6(4), 355–385.

Baboud, L., Čadík, M., Eisemann, E., Seidel, H.-P., 2011. Automatic photo-to-terrain alignment for the annotation of mountain pictures. *CVPR 2011*, 41–48.

Bostock, M., 2021. D3 data-driven documents. d3js.org (22. March 2022).

Brejcha, J., Lukác, M., Chen, Z., DiVerdi, S., Cadík, M., 2018. Immersive trip reports. UIST ’18, Association for Computing Machinery, New York, NY, USA, 389–401.

Bundesamt für Landestopografie swisstopo, 2021. swisstlm3d. swisstopo.admin.ch/de/geodata/landscape/tlm3d.html (22 March 2022).

Darken, R., Cevik, H., 1999. Map usage in virtual environments: orientation issues. *Proceedings IEEE Virtual Reality (Cat. No. 99CB36316)*, 133–140.

de Haan, G., Piguillet, H., Post, F., 2010. Spatial Navigation for Context-Aware Video Surveillance. *IEEE Computer Graphics and Applications*, 30(5), 20-31.

Drascic, D., Milgram, P., 1996. Perceptual issues in augmented reality. M. T. Bolas, S. S. Fisher, M. T. Bolas, S. S. Fisher, J. O. Merritt (eds), *Stereoscopic Displays and Virtual Reality Systems III*, 2653, International Society for Optics and Photonics, SPIE, 123 – 134.

Furmanski, C., Azuma, R., Daily, M., 2002. Augmented-reality visualizations guided by cognition: perceptual heuristics for combining visible and obscured information. *Proceedings. International Symposium on Mixed and Augmented Reality*, 215–320.

Girgensohn, A., Kimber, D., Vaughan, J., Yang, T., Shipman, F., Turner, T., Rieffel, E., Wilcox, L., Chen, F., Dunnigan, T., 2007. Dots: Support for effective video surveillance. *Proceedings of the 15th ACM International Conference on Multimedia*, MM ’07, Association for Computing Machinery, New York, NY, USA, 423–432.

GRASS Development Team, 2022. Geographic Resources Analysis Support System (GRASS) Software. Open Source Geospatial Foundation. grass.osgeo.org (20 March 2022).

Holloway, R. L., 1995. Registration errors in augmented reality systems. PhD thesis, University of North Carolina at Chapel Hill.

Jamonnak, S., Zhao, Y., Curtis, A., Al-Dohuki, S., Ye, X., Kamw, F., Yang, J., 2020. GeoVisuals: a visual analytics approach to leverage the potential of spatial videos and associated geonarratives. *International Journal of Geographical Information Science*, 34(11), 2115-2135. https://doi.org/10.1080/13658816.2020.1737700.

Karsch, K., Golparvar-Fard, M., Forsyth, D., 2014. ConstructAide: Analyzing and Visualizing Construction Sites through Photographs and Building Models. *ACM Trans. Graph.*, 33(6). https://doi.org/10.1145/2661229.2661256.

Kopf, J., Chen, B., Szeliski, R., Cohen, M., 2010. Street Slide: Browsing Street Level Imagery. *ACM Trans. Graph.*, 29(4). https://doi.org/10.1145/1778765.1778833.

Kopf, J., Neubert, B., Chen, B., Cohen, M., Cohen-Or, D., Deussen, O., Uyttendaele, M., Lischinski, D., 2008. Deep Photo: Model-Based Photograph Enhancement and Viewing. 27(5). https://doi.org/10.1145/1409060.1409069.

Livingston, M. A., Ai, Z., Swan, J. E., Smallman, H. S., 2009. Indoor vs. outdoor depth perception for mobile augmented reality. *2009 IEEE Virtual Reality Conference*, 55–62.

MacIntyre, B., Coelho, E., Julier, S., 2002. Estimating and adapting to registration errors in augmented reality systems. *Proceedings IEEE Virtual Reality 2002*, 73–80.

MacIntyre, B., Machado Coelho, E., 2000. Adapting to dynamic registration errors using level of error (loe) filtering. *Proceedings IEEE and ACM International Symposium on Augmented Reality (ISAR 2000)*, 85–88.

Matplotlib Development Team, 2021. Matplotlib. matplotlib.org (22 March 2022).

Meyer, J., Rettenmund, D., Nebiker, S., 2020. Long-Term Visual Localization in Large Scale Urban Environments Exploiting Street Level Imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2, 57–63.

Mildner, P., Claus, F., Kopf, S., Effelsberg, W., 2013. Navigating videos by location. *Proceedings of the 5th Workshop on Mobile Video*, MoVid ’13, Association for Computing Machinery, New York, NY, USA, 43–48.

Möller, A., Kranz, M., Huitl, R., Diewald, S., Roalter, L., 2012. A mobile indoor navigation system interface adapted to vision-based localization. *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia*, MUM '12, Association for Computing Machinery, New York, NY, USA.

Nebiker, S., Meyer, J., Blaser, S., Ammann, M., Rhyner, S., 2021. Outdoor Mobile Mapping and AI-Based 3D Object Detection with Low-Cost RGB-D Cameras: The Use Case of On-Street Parking Statistics. *Remote Sensing*, 13(16). https://www.mdpi.com/2072-4292/13/16/3099.

Pankratz, F., Dippon, A., Coskun, T., Klinker, G., 2013. User awareness of tracking uncertainties in ar navigation scenarios. *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 285–286.

Pindat, C., Pietriga, E., Chapuis, O., Puech, C., 2013. Drilling into complex 3d models with gimlenses. *Proceedings of the 19th ACM Symposium on Virtual Reality Software and Technology*, VRST '13, Association for Computing Machinery, New York, NY, USA, 223–230.

Plumlee, M., Ware, C., 2003a. An evaluation of methods for linking 3d views. *Proceedings of the 2003 Symposium on Interactive 3D Graphics*, I3D '03, Association for Computing Machinery, New York, NY, USA, 193–201.

Plumlee, M., Ware, C., 2003b. Integrating multiple 3d views through frame-of-reference interaction. *Proceedings International Conference on Coordinated and Multiple Views in Exploratory Visualization - CMV 2003 -*, 34–43.

Stein, M., Janetzko, H., Lamprecht, A., Breitkreutz, T., Zimmermann, P., Goldlücke, B., Schreck, T., Andrienko, G., Grossniklaus, M., Keim, D. A., 2018. Bring It to the Pitch: Combining Video and Movement Data to Enhance Team Sport Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 24(1), 13-22.

Tompkin, J., Kim, K. I., Kautz, J., Theobalt, C., 2012. Videoscapes: Exploring Sparse, Unstructured Video Collections. *ACM Trans. Graph.*, 31(4). https://doi.org/10.1145/2185520.2185564.

Tory, M., 2003. Mental registration of 2d and 3d visualizations (an empirical study). *IEEE Visualization, 2003. VIS 2003.*, 371–378.

Veas, E., Grasset, R., Kruijff, E., Schmalstieg, D., 2012. Extended Overview Techniques for Outdoor Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics*, 18(4), 565-572.

Wang, Y., 2010. Design and evaluation of contextualized video interfaces. PhD thesis, Virginia Tech.

Wang, Y., Krum, D. M., Coelho, E. M., Bowman, D. A., 2007. Contextualized Videos: Combining Videos with Environment Models to Support Situational Understanding. *IEEE Transactions on Visualization and Computer Graphics*, 13(6), 1568-1575.

Wu, F., Tory, M., 2009. Photoscope: Visualizing spatiotemporal coverage of photos for construction management. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, Association for Computing Machinery, New York, NY, USA, 1103–1112.

Zhang, B., Li, Q., Chao, H., Chen, B., Ofek, E., Xu, Y.-Q., 2010. Annotating and navigating tourist videos. *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, Association for Computing Machinery, New York, NY, USA, 260–269.

Zollmann, S., Langlotz, T., Grasset, R., Lo, W. H., Mori, S., Regenbrecht, H., 2021. Visualization Techniques in Augmented Reality: A Taxonomy, Methods and Patterns. *IEEE Transactions on Visualization and Computer Graphics*, 27(9), 3808-3825.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B4-2022-455-2022 | © Author(s) 2022. CC BY 4.0 License.