February 2024

Indoor Positioning and Navigation by Semantic Localization Based on Visual Context

D Shin

Indoor Positioning and Navigation by Semantic Localization Based on Visual Context

ABSTRACT

Conventional indoor localization techniques rely on high-precision indoor 3/6 degrees-of-freedom (DOF) positioning of the user device which may be infeasible if the device lacks positioning sensors such as GPS or IMU, if such sensors are turned off, or if the sensors have insufficient accuracy. This disclosure describes techniques the use of language modeling techniques for providing indoor navigation capabilities in the absence of such sensor data based on the local visual context obtained with a camera. Text captions describing frames of the user's visual context in an indoor space are generated. A collection of captions for the current and recently captured, timestamped frames of the visual context, and a suitable prompt and metadata are input to a large language model to determine the current location of the user within the indoor space. The techniques can be incorporated within any indoor digital mapping and navigation application via any device capable of capturing the visual context via a camera.

KEYWORDS

- Navigational model
- Visual language model
- Indoor positioning
- Indoor localization
- Semantic localization
- Movement trajectory
- Digital map
- Visual context
- Generalized Pre-trained Transformer (GPT)
- Visual encoder
- Location interpolation
- Text caption

BACKGROUND

Users often need to navigate within indoor spaces. Applications that provide indoor navigation capabilities typically rely on high-precision indoor 3/6 degrees-of-freedom (DOF) positioning of the user device. Conventional device positioning techniques use motion and/or positioning data from various device sensors, such as Global Positioning System (GPS), Inertial Measurement Unit (IMU), etc. Appropriate heuristic filters are applied to the raw sensor data to obtain smooth indoor localization.

The conventional approach described above suffers from two main limitations. First, the sensor data is not always reliable. For instance, multipath propagation for the GPS signal can result in inaccurate positioning even with a custom filter. Second, user devices sometimes lack GPS sensors because of tradeoffs with other factors, such as dimensions, weight, etc. or the GPS sensors may be deactivated.

DESCRIPTION

This disclosure describes techniques for providing indoor navigation capabilities without the need to use data from motion and positioning sensors within a device. Instead, the positioning information is derived from the local visual context, obtained with user permission.

A suitable visual-language model is employed to generate a text caption describing each frame within the user’s visual context. Subsequently, similarity scoring and indoor location interpolation can be performed in the language domain using the localization semantics within the captions. In addition, relevant temporal information, such as navigation time, can be considered to improve the accuracy of the positioning. The operation can predict a new indoor location from few-shot examples of various parts of the indoor space.

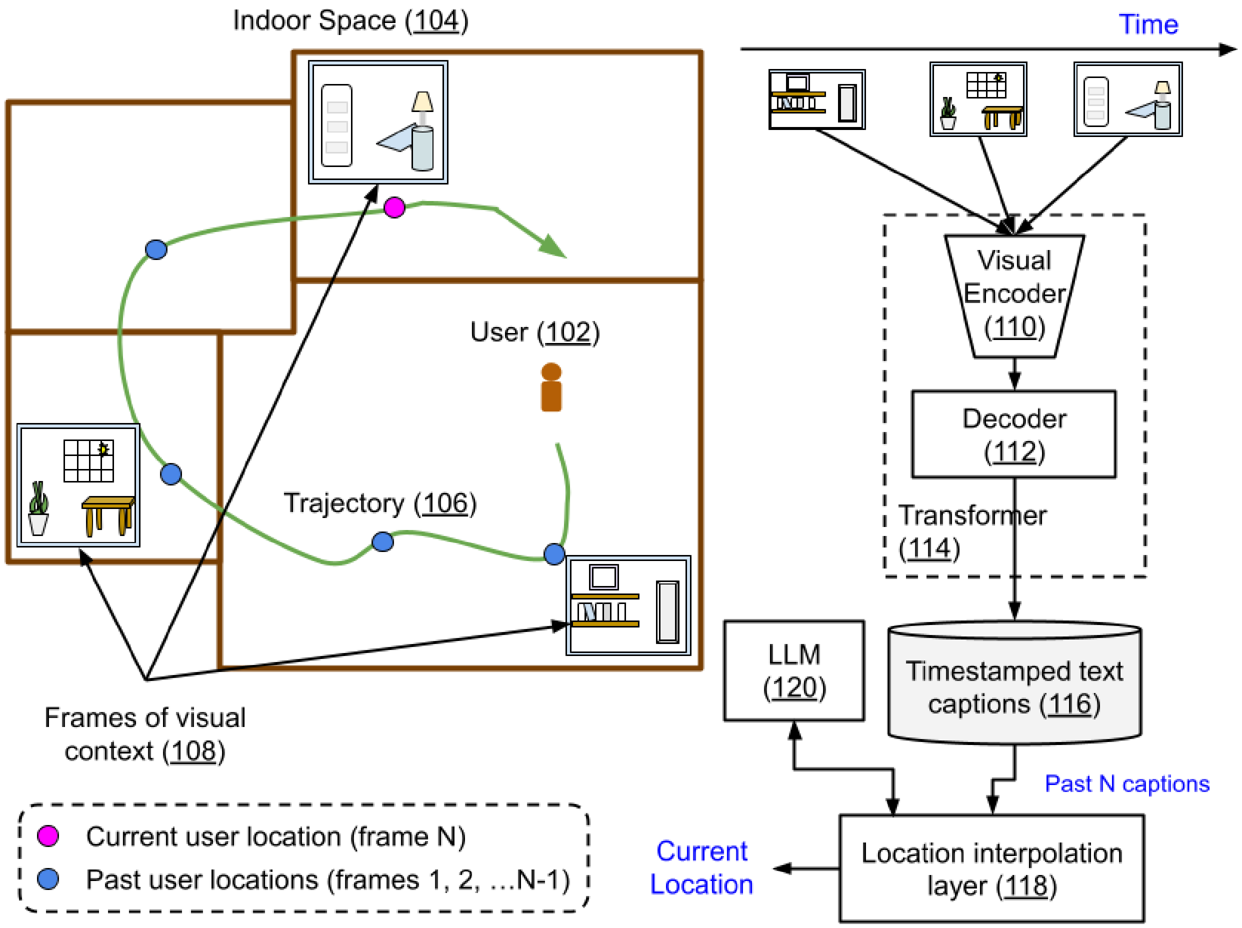
Indoor Space (104)
Time
User (102)
Trajectory (106)
Visual Encoder (110)
Decoder (112)
Transformer (114)
Frames of visual context (108)
LLM (120)
Timestamped text captions (116)
Past N captions
● Current user location (frame N)
● Past user locations (frames 1, 2, …N-1)
Current Location
Location interpolation layer (118)

Fig. 1: Analyzing descriptions of visual context to determine indoor position

Fig. 1 shows an example operational implementation of the techniques described in this disclosure. A user (102) is moving within an indoor space (104) along a trajectory (106). While moving through the space, with user permission, the surrounding visual context (108) is captured at a high rate by a camera on a user device. The goal is to determine the current indoor location at any time. In Fig. 1, the blue dots illustrate past locations of the user device while the pink dot illustrates the current location.

Frames of the captured visual context are input to a visual encoder (110) that is mixed into the attention layers of a decoder (112) within a transformer model architecture (114) (e.g., [1] [2[]). The output of the decoder includes a text caption describing the input frame. Over time, this operation results in a timestamped collection of text captions (116) descriptive of the user's visual context being accumulated.

To predict a current user location, the last N captions (i.e., the caption for the current frame of the visual context and the previous N-1 captions) are processed via a location interpolation layer (118). Concatenation of the N captions along with a suitable prompt is input to a suitable large language model (LLM) (120). Engineering the prompt to the LLM for obtaining the current location may include prepending or appending extra metadata in the tokenized layers. The output of the location interpolation layer provides the likely current indoor location of the user.

In the example of Fig. 1, the user moves from the room at the bottom right via rooms on the bottom left and the top left to the room on the top right. The camera captures frames (3 frames illustrated in Fig. 1) through user’s trajectory. The current location of the user is determined based on the captured frames and associated captions, depicted with a purple dot in Fig. 1.

The operation can be implemented with any suitable Generalized Pre-trained Transformer (GPT) type model. The visual context can be captured in any suitable form, such as video and/or images. The number of captions and the rate of capture for the visual context used to determine the user’s current location can be set by the developers and/or determined dynamically at runtime.

The described techniques can be incorporated within any indoor digital mapping and navigation application to determine user location, to provide directions, etc. The techniques can be used via any device capable of capturing the visual context via a camera. The operation is similar to the familiar experience of navigating in a new indoor environment by interpolating between a few visually diverse prominent semantic anchors. Implementation of the techniques enables users to navigate accurately in indoor spaces even if location and positioning sensors in the device are inaccurate, turned off, or absent, thus enhancing the utility and user experience (UX) of applications that provide indoor navigation capabilities.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information about a user’s context, a user’s device camera feed, a user’s preferences, or a user’s current location), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user’s identity may be treated so that no personally identifiable information can be determined for the user, or a user’s geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes techniques the use of language modeling techniques for providing indoor navigation capabilities in the absence of such sensor data based on the local visual context obtained with a camera. Text captions describing frames of the user’s visual context in an indoor space are generated. A collection of captions for the current and recently captured, timestamped frames of the visual context, and a suitable prompt and metadata are input to a large language model to determine the current location of the user within the indoor space.

The techniques can be incorporated within any indoor digital mapping and navigation application via any device capable of capturing the visual context via a camera.

REFERENCES

1. https://deepmind.google/discover/blog/tackling-multiple-tasks-with-a-single-visual-language-model/
2. Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc et al. "Flamingo: a visual language model for few-shot learning." *Advances in Neural Information Processing Systems* 35 (2022): 23716-23736.