Design Study for a Geospatial-Video Data Analysis Query Language

Chanwut Kittivorawong
UC Berkeley, USA
chanwutk@berkeley.edu

Shadaj Laddad
UC Berkeley, USA
shadaj@berkeley.edu

Andrew Lenz
UC Berkeley, USA
andrew.lenz@berkeley.edu

Amy Lu
UC Berkeley, USA
amyxlu@berkeley.edu

ABSTRACT

Geospatial-video databases help data scientists store their video data efficiently, but many of them do not provide intuitive query languages for the data scientists to retrieve the data back. Databases with intuitive query languages provide data abstractions that hide their internal data representations. These data abstractions limit the number of operations that data scientists can perform on their data. In this paper, we address the failure for existing languages to easily enable queries to video frames, as identified through need-finding interviews with users of geospatial-video query languages. We focus on its data abstraction and query language design. Our new data abstraction additionally includes a video frame data-type. Our new query language provides operations for users to reason about video frames in their data.

Author Keywords

User Studies; Geospatial-Video Data; Data Model; Query Language

CCS Concepts

•**Human-centered computing** → **User studies;** •**Information systems** → **Query languages;** *Geographic information systems;*

INTRODUCTION

Capturing moments with video cameras is easy, and storage costs for storing video data is cheap. Video cameras have thus become a widely used tool to document moments through visual information. However, visual information of video data is unstructured, making it difficult to store and look-up. Processing and analyzing is often labor- and time-intensive for the data scientist. A common approach is to extract spatial and temporal information from the video data, using computer vision techniques [8, 12]. The extracted information makes the data more structured and easier to organize. We can them map the spatial information to the real-world location, becoming geospatial information. Geospatial information helps data scientists to make sense of the video data as they correlate with the real world. However, we still lack easy-to-use and expressive language to retrieve and explore the data.

Our goal for this project is to design a language that makes it easier for data scientists to search and explore massive geospatial-video datasets. We explore two geospatial-video data analysis use cases, and use them to shape the design of our tool:

- **Autonomous Driving**: Scenic [2] is a language for generating visual scenes, specifically for autonomous vehicle training data. by comparing the scenes with a real-world autonomous vehicle video data [1]. A Ph.D. student in Computer Science is working on validating visual scenes generated [5], by comparing the scenes with a real-world autonomous vehicle video data [1].

- **Evident of Misconduct:** Data journalists often have large banks of video data, but not nearly enough time to look through these data. This may hinder the ability to accurate retrieve evidence involved in things like police misconducts or false court statements.

In order to uncover what are the essential user necessities of a geospatial-video analysis tool, introduce new features to expediate workflows, and decrease barrier-to-entry for working on these important problems, we conduct user interviews related to the two use cases above.

We have explored prior attempts on geospatial-video data analysis tools. VisualWorldDB [4] presents a new data model to help users make sense of their video data after they are stored in the database. However, the work focuses heavily on the data storage system and optimizations. The proposed SQL-like query language is difficult to use as it exposes its internal data tables from which users can query. Apperception [3] is built on top of the VisualWorldDB idea and aims to improve its usability. Apperception has a much simpler query language compared to VisualWorldDB. In the attempt, Apperception hides all the the internal data table and present a new data abstraction to users. As a result, the query language cannot

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

express some data exploration questions. CLIP [7] is a large language model capable of generating video frames from a natural language description as the prompt. However, the results are not deterministic, and it is difficult to query geospatial or temporal information.

In this project we made following contributions:

1. We interviewed a data journalist and a computer science undergraduate, who are working on geospatial-video data analysis tasks.
2. We designed a Python [9] programming tool based on prior works and participants' need from the interviews. Our tool presents an easy-to-understand data abstraction that will allow programmers of any experience level to easily explore large geospatial-video databases. Our tool also provide an expressive domain-specific language for users to explore their geospatial-video datasets through the data abstraction we provided.
3. Our tool is designed to be extensible. Users will be able to extend our language with their custom functions. These custom functions help the users to express the queries that our provided query language cannot.
4. We design a Graphical User Interface (GUI) for non-programmer to apply the same principle that our python tool provides but without programming experience.

RELATED WORK

VisualWorldDb

In this paper [4], the author introduced us to the concept of Visual World Applications (VWAs). VWAs are applications that rely on spatial and temporal information from visual data. The authors presented VisualWorldDB a video database management system optimized for VWAs. With VisualWorldDB, users can create a *world* that represents an actual world. Then, users can ingest multiple geospatial-videos into the *world*. VisualWorldDB then places all the *visual objects* present in the videos into the *world*. Each *visual object* has properties that include its location in the real world. After the users have all the *visual objects* added to the *world*, they can query for these *visual objects* instead of their videos. In the end VisualWorldDB retrieves the videos associated with the *visual objects* queried.

Despite the simplified data model, VisualWorldDB focuses on its storage systems and optimization. The paper does not go into much discussion about VisualWorldDB's query language. The paper presents an example program in VisualWorldDB's language to ingest videos into a *world*. The language has a SQL-like syntax that expose internal data tables for users to manually input rows of data themselves. This approach is not intuitive to the users. First, the program users write is error prone. All the internal data table should have an invariant that keep the *world* representation valid. However, users have access directly to the internal data table, so they can unintentionally break some of the invariant. Second, the program that the user writes is verbose. To add a video camera into a created *world*, users need to:

1. Add the camera as an object to the world into an Objects table.
2. Add the position of the camera as a point in to the world into a Point table.
3. Add the orientation of the camera as another point into the world into the same Point table.
4. Add a video file associated with the camera into the created world.

Adding a camera to the *world* could be done as one operation. Instead, we need four operations to perform this task. In this project, we take inspiration from this *visual world* concept.

Apperception

Apperception [3] has a similar *world* concept as VisualWorldDB but with a more usable Application Programming Interface (API). Apperception stores all the geospatial data in MobilityDB [13]. However, it provides a Python API for users to interact with its geospatial-video data storage. Apperception provides a Python *World* object that represent a *world*. There are multiple operations that users can do with a *world*.

- `add_camera` to add a camera and video to the *world*. Compared to VisualWorldDB, users only need one operation to add a camera to a *world*.
- `recognize` to extract objects from an existing video and place them into the *world*.
- `filter` to filter only objects that the user is interested in.

After all the operations done, users will have a *World* that contains only the objects that they are interested in. Then, they can use a `get_*` method to observer the state of the world.

- `get_trajectory` gives users the trajectory of each of the objects in the world. The users can then use an `overlay_trajectory` method to overlay the trajectories into the original video.
- `get_videos` gives the videos of all the objects in the world.

Apperception's interface is easy to use and not verbose. It wraps around the internal geospatial-video data storage, so the users do not have direct access to the data table. As a result, it is less likely for them to accidentally violate the data store invariant.

A *World* is immutable, which means that it cannot be modified. Any operation done to a *World* create a new *World* that represents the original one with the operation applied to. This approach for making a *World* immutable helps users to easily reason about the program they have written. In addition, having a immutable DataFrame-style syntax is common in tools [11, 6, 10] for data exploration tasks. Users who are already familiar with data exploration tools could learn how to use Apperception quickly.

However, Apperception's language is designed for users to find objects in a *World*. It is hard for users to describe a query to find a specific moment of an object using only Apperception language. For example, if a user has an hour-long video that

capture the same car for the whole video. The car crashes at 45th minute of the video. The user can use Apperception language to query for a car that crashes. Apperception will return a *World* that only contains the car. But, Apperception cannot provide the information about the when the incident occurs.

[Mick: what is the problem in apperception?: cannot explain more complicated query] [Todo: What techniques did you adopt from others?]

Rekall

[Mick: todo: query by frames] [Todo: What techniques did you adopt from others?]

[Mick: todo: CLIP/LiveReel?] [Todo: What alternative techniques have people used to work on this problem?] [Todo: What techniques did you adopt from others?]

FRUCTURE FOR RACKET

When th

INTERVIEWS

[Todo: What research questions drove the design of the study?]
In our user interviews, we

[Todo: How did you collect data about your target population?]

[Todo: How did you find participants?]

[Todo: How did you select participants?]

[Todo: What procedure did you use for collecting data?]

[Todo: What materials, if any, did you show participants?]

[Todo: What data did you collect?]

[Todo: What did you learn from the user data?]

[Todo: How did you learn it? What led you to these conclusions? Be specific.]

[Todo: What quotes, stories, numbers, or patterns back up these conclusions or offer extra detail?]

[Todo: Which conclusions surprised you?]

[Todo: How did your conclusions from the user data affect what you did next?]

[Todo: How should your conclusions from the user data affect what future researchers do?]

[Todo: andrew?]

In order to identify short-comings for current tools and guide the design of our work, we conducted contextual inquiry and observed users performing tasks involving geospatial video analysis using existing tools. An area of the pipeline which we were especially interested in is how researchers specify queries that identify situations with multiple objects, potentially with some specific behavior over time. The following section discusses the findings from our two interviews in greater detail.

Data Journalist

User L is a data journalist at a public radio station who must look at video evidence on the order of thousandsto collect and report on evidence, specifically surrounding police misconduct. With permission, we recorded L performing the video queries required for this task. In L's workflow, she goes through several videos and makes quick notes on their utility for her aim; for relevant videos, she "clicks around a little" through frames to roughly orient herself with what is happening in the scene. Once key frames are identified, L takes

[Todo: focus on why we need to filter by frame instead of just filtering by objects like in apperception] [Todo: to scope down the part of each video that are as of useful to them]

Programmer

[Todo: focus on our syntax and data model that it is understandable] [Todo: focus on the part where he suggests us to implement a syntax where users can filter by frame and then switch to filter by objects]

DATASET

For our prototype, we use the nuScenes dataset [1], a collection of camera, lidar, and radar images to facilitate autonomous driving research. Data is collected from 15 hours of driving data in Boston and Singapore, and key frames are sampled at a rate of 2Hz for labeling by human annotators, resulting in a rich collection of information suitable for various downstream needs in autonomous driving research. The structure of the decomposition of video frames into data and annotation structure was a good representative of geospatial settings for our language design.

DATA MODEL

To interface with the data, the user uses the *Annotation*, *Instance*, *Frame*, and *Video* abstractions. Figure 6.0.1 illustrates this data model in greater detail. Each video is denoted by an unique scene ID, and frames are indexed by the order in which they appear. All instances have an unique instance token which can be used to query specific properties for that instance specifically. By passing the scene ID, the user can load the video with corresponding annotations and properties needed for downstream tasks. The user may also directly instantiate any Frame and Instance directly from the nuScenes database, if provided with the frame order and instance token in addition to the scene ID. All metadata are stored in Annotations, with the exception of "Category" (e.g.), which is an attrithe frame order

[Mick: how are we different from apperception] [Todo: to focus: the added idea of exploring video data by frames instead if by instances like in apperception]

DOMAIN-SPECIFIC LANGUAGE

[Todo: to focus:
- by-frames operations: as a response to Lisa's use case where we need to scope the output video to the only interesting parts of each videos.
- by-frames <–> by-instances: as a response to Yousef's comments to allow the language to express more queries]

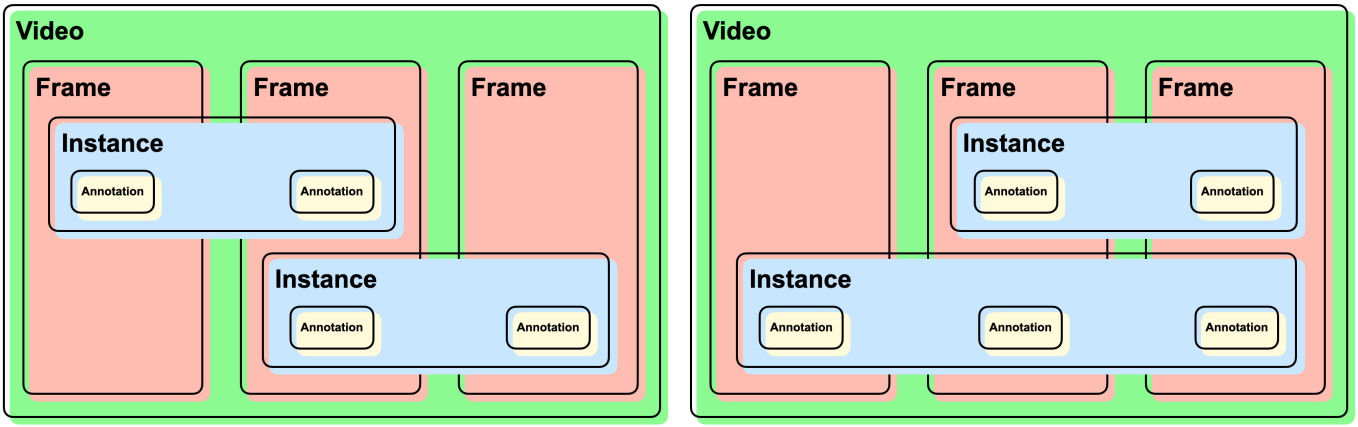

Figure 6.0.1: The wealth of image, annotation, spatial, and temporal information in geospatial data is simplified into in our language into Video, Frame, Instance, and Annotation. A video consists of several consecutive frame, which point to a specific image, and instances (e.g., cars, pedestrians) may persist through many frames. Each instance contains a category attribute, and an annotations attribute for accessing more detailed metadata of interest.

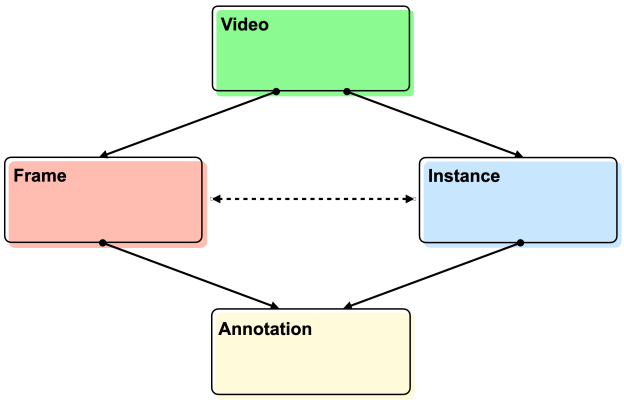

Figure 6.0.2: [Mick: todo]

To address the pain points we noticed in our user interviews, we chose to design a new language for describing queries over geospatial video data as a domain-specific language embedded in Python. Instead of building a new language from scratch, which would require designing a new syntax and editor integrations that users would have to learn about, we are able to inherit the large ecosystem of existing tooling around Python. Furthermore, some potential users of our system are already familiar with Python, which makes it further easier to integrate into existing pipelines.

Design Guidelines

There are three key factors constraining the design of our language: incremental query creation, support for using the language through a graphical user interface, and supporting queries that combine predicates over time and space. Before we dive into how we design the language around these constraints, let us walk through how we derived the factors based on our user-studies and their consequences on the space of language designs.

Incremental Queries

Across the interviews, a key pattern that stuck out was the process by which the participants would identify portions of video that were of interest—either by manual scrubbing through video files or the creation of queries. Rather than define the end-to-end query, the participants would start with a coarse grained search, looking for portions of the video that are likely to have the information they are looking for. Across these sections, the participants would then refine their search/query by playing back the video in real-time or querying individual frames in more detail.

Our insight from this observation was that users of video query languages will want to develop the query incrementally, where they can start with a query that identifies sections of the video with general patterns related to their specific goal (such as the presence of multiple cars in the same frame) and iteratively refine it to identify the specific scenario (such as cars moving towards each other) while receiving feedback from the query engine.

Supporting incremental query creation places significant constraints on the design of the language since the language must offer modular APIs to support users adjusting their query pipelines over time and must also have a performant runtime that can execute updated queries efficiently. Furthermore, it requires us to ascribe meaning to intermediate steps of a query, including complex concepts such as windows over time or groups of objects, so that users can see the results of each component of their pipeline.

Graphical User Interface

Based on our interview with Lisa Pickoff-White, who works on data journalism but primarily uses manual scrubbing methods to query videos rather than writing code, we identified that a key feature to make our language useful is offering a graphical interface for constructing programs for users who are not comfortable with text languages. At the same time, however, we want to support a lower-level Python API for constructing queries that expert users such as Yousef would be more comfortable using.

To support this graphical interface, we decided to design it as a layer on top of the Python API with mappings from each concept in the low-level to features in the interface. This architecture means that the low-level API should be strongly typed, since types in the API can be used to constrain the types

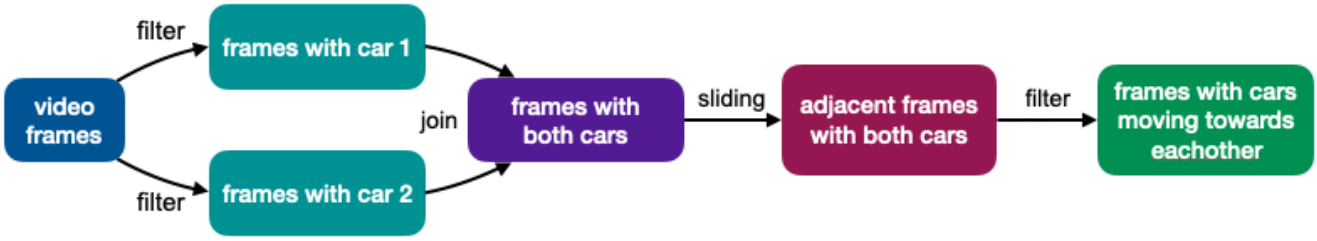

Figure 7.0.1: An example of a dataflow users can construct for a complex query.

of programs in the graphical interface to those that are correct by construction. Furthermore, it requires us to include several built-in APIs for common scenarios to reduce the overhead of describing various queries.

Queries over Time and Space

Finally, we observed in both interviews that queries over video data often require repeatedly moving between the axis of time and detected objects, something that existing query systems do not support. For a concrete example, consider an query over autonomous vehicle data that is looking for collisions. We want to find chunks of the video where two cars are moving to each other, but this requires predicates that both reason about multiple frames together and multiple objects together. After a user writes a query to identify cars, they will need to identify *frames* with multiple cares, and then process windows of frames to check if the detected objects are moving towards each other.

Queries like this are difficult or impossible to describe with existing query languages, which either focus exclusively on time or objects as the primary axis being queried over. We aim to address this in our design by providing APIs that mix these two concepts together, enabling queries such as filtering over the properties of objects in windows of time. This requires us to extend the data model, as we discussed in the previous section, and also provide APIs for switching between the frame and instance views of video data.

Language Design

With these design goals, we decided to develop a DSL based on declarative and functional programming principles. Rather than writing out queries as imperative statements that mutate data, have unclear dependencies, and use hard-to-optimize control constructs, this approach allows us to optimize and execute the queries users write from a global perspective. By modeling queries as dataflows, such as the example in Figure 7.0.1, our language can also provide high-quality intermediate feedback since every intermediate node can be mapped to output samples. Finally, the declarative approach makes it simpler to build a graphical interface, since the dataflow model corresponds to a natural visual model of connecting primitive operations with dependencies.

Our DSL consists of four core types corresponding to the elements of our data model: videos, frames, instances, and annotations. To develop dataflows over these structures, we then provide collection types wrapping these elements, which represent nodes that will process incoming streams of video data. In the time domain, we also include a collection type that captures groups of frames rather than just individual elements, which is useful when performing tasks like tracking the motion of an object over a window of time.

To create complex queries, users can develop a dataflow pipeline by repeatedly transforming these collection types using functional operations. In addition to the classic transformations such as map, which transforms individual elements independent of each other, and filter, which applies a predicate to each element, we also include several operations specific to the domain of video data. First, we include the `join` operator, which allows a user to combine multiple collections by matching elements with a specific property. For example, users may join

[Shadaj: walk through functional APIs]

[Shadaj: examples of code]

[Shadaj: explain how to alternate between time/space]

EXTENSIONS

While we expect that the language will be sufficiently verbose to construct most types of queries, we also recognize that some of our users may not have the time, desire, or programming expertise to implement the functions that they may want to use to filter the videos. Therefore, we built out some of the possible helper functions that could be useful for filtering. These functions expect to be called within filter and thus all return boolean values. The ones provided below should take in the result of the sliding operation, as they intend to compare information about the instance across adjacent frames in order to determine something about the instance's movement.

Utility functions for filtering:
(Note: `iX` and `fY` (where `X` and `Y` are whole numbers) will refer to Instances and Frames, respectively.)

- `move_away(i1, i2, f1, f2)`:
  This function calculates the distance between Instance `i1` and Instance `i2` in Frame `f1` and Frame `f2`, then compares these distances. If the distance between the instances in Frame `f2` is greater than the difference between the instances in Frame `f2`, then we consider the instances to be moving away from one another.
- `accelerating(i, f1, f2, f3)`:
  This function uses the location of Instance `i` to determine

if it has accelerated between Frame `f1` and Frame `f2`, and Frame `f2` and Frame `f3`. To determine this, it compares the difference in location between both pairs of frames: if the difference between the locations in Frame `f2` and Frame `f3` is greater than the difference between the locations in Frame `f1` and Frame `f2`, we consider Instance `i` to have accelerated.

- `decelerating(i, f1, f2, f3)`:
  This function uses the location of Instance `i` to determine if it has decelerated between Frame `f1` and Frame `f2`, and Frame `f2` and Frame `f3`. To determine this, it compares the difference in location between both pairs of frames: if the difference between the locations in Frame `f2` and Frame `f3` is less than the difference between the locations in Frame `f1` and Frame `f2`, we consider Instance `i` to have decelerated.

- `stopped(i, f1, f2, tol: float)`:
  This function uses the location of Instance `i` to determine if it has stopped from Frame `f1` to Frame `f2`. To determine this, it compares the difference in location between the two frames to the tolerance `tol`: if the difference is greater than `tol`, we consider the instance to have moved; otherwise, it has stopped.

USER INTERFACE

Although we did not implement a user interface for the language, we were able to prototype some designs for such an interface. Based on these explorations, we propose a block-based projectional editor interface to provide the most convenience and accessibility for programmers, particularly those with less experience who we expect to use this tool. To facilitate this, we would allow users to fill the holes in skeleton queries by selecting entries from a drop-down menu which would also have a search bar; this will provide suggestions to users who want to explore the dataset, but also allow users who know what want to see to search for what they are looking for.

As the output of a query in this language would be a series of clips or videos taken from the overall collection of videos, we envision the interface as having a region in which the output of the queries are displayed. However, this creates some problems when we have too many output videos. To amend this, we would split the list of videos into multiple pages to prevent the tool from loading all the videos at once. Then, we show the number of output videos. If the number of output videos is too large, we suggest that the users add more filters to the output to narrow down their search.

Mapping the Language to the Interface

First, the tool will store the collection of videos as some variable, which we will refer to as `videos`. A typical query in our language is shown in the Listing 9.1.1.

```
1  people = videos \
2      .flatten_instances() \
3      .filter(lambda i:  \
4          i.property["type"] == "person")
5
6  cars = videos \
7      .flatten_instances() \
8      .filter(lambda i:  \
9          i.property["type"] == "car")
10
11 result = people \
12     .join(cars, on="frame") \
13     .sliding(2) \
14     .filter(lambda i1, i2, f1, f2:
15         move_away(i1, i2, f1, f2)
```

Figure 9.1.1: Example Query in our Programming Language

In Listing 9.1.1, we show an example query where a user is looking for all frames where a person and a car move away from one another. For clarity, this is split into three queries: the first two filter the videos for instances of people or cars, and the third joins these sub-queries to find sequences of frames in which the instances move away from one another; however, these can be combined into a single line of code.

In Figure 9.1.2, we show what the query would look like using our proposed interface. The highlighted sections represent the holes that the user has filled in.

```
Find frames with a person and a car
    moving away ...
```

Figure 9.1.2: Example Query in our Proposed Interface

Frames

When trying to find a certain frame, e.g., frames that occur between 7am and 8am, we will have a query in our interface that looks like Figure 9.1.3. This is equivalent to the code in Figure 9.1.4, which uses the `flatten_frames` method to allow users to filter along the frames.

```
Find frames between 7am and 8am
```

Figure 9.1.3: Instance-based Query in Interface

```
1  videos.flatten_frames() \
2      .filter(lambda i:  \
3          i.property["time"] > 7.00 and
4          i.property["time"] < 8.00)
```

Figure 9.1.4: Instance-based Query in Language

Instances

When trying to find frames that contain a certain instance, e.g., a car, we will have a query in our interface that looks like Figure 9.1.5. This is equivalent to the code in Figure 9.1.6, which uses the `flatten_instance` method to allow users to filter along the instances.

```
Find frames with a car
```

Figure 9.1.5: Instance-based Query in Interface

```
videos.flatten_instances() \
    .filter(lambda i:  \
        i.property["type"] == "car")
```

Figure 9.1.6: Instance-based Query in Language

Multiple Instances in the Same Frame

When looking for frames that have multiple instances, e.g., a person and a car, we will have a query that looks like Figure 9.1.7. To translate this to our language, as seen in Figure 9.1.8, we'll first filter for the frames that contain a person, then filter for the frames that contain a car, and finally join these two lists of using the `join` method, returning only the frames containing both instances. To translate from the structure in the interface to this back-end, the interface should create a new list of frames for each new instance added, and join on matching frames for each pair of instances separated by the "and".

```
Find frames with a person and a car
```

Figure 9.1.7: Example Query in our Proposed Interface

```
people = videos \
    .flatten_instances() \
    .filter(lambda i:  \
        i.property["type"] == "person")

cars = videos \
    .flatten_instances() \
    .filter(lambda i:  \
        i.property["type"] == "car")

result = people \
    .join(cars, on="frame") \
```

Figure 9.1.8: Example Query in our Programming Language

Comparisons with Adjacent Frames

When comparing information between adjacent frames – for instance, when trying to find positional or motion-related information about an instance (in this case a person)– we will have a query that looks like Figure 9.1.9. To translate this to our language, as seen in Figure 9.1.10, we'll first filter for the frames that contain a person, then use the `sliding` operation to create pairs of frames, and finally filter these pairs to find those between which the person does not move.

```
Find frames with a person stopped
```

Figure 9.1.9: Example Query in our Proposed Interface

```
people = videos \
    .flatten_instances() \
    .filter(lambda i:  \
        i.property["type"] == "person")

result = people \
    .sliding(2) \
    .filter(lambda i, f1, f2:  \
        stopped(i, f1, f2) \
```

Figure 9.1.10: Example Query in our Programming Language

Notice that in this case, the `sliding` operation is placed before the second `filter` operation. When any utilify functions are used to compare between frames, we must apply a `sliding` operation first so we can access both frames.

Using the Interface

We will now demonstrate how to construct the query shown in Figure 9.1.2 using the proposed block-based interface. Note: `(def)` means that this is the default option.

First, we must select the type of query. In this case, since we want the frames where a car and person move away from one another, we will select the query skeleton corresponding to finding frames with a specified instance.

```
Find frames with Instance: all instances (def)
    ...
```

Figure 9.2.1: A

To specify that we want instances a car, we will replace the `Instance: all instances (def)` with `a car`.

```
Find frames with a car ...
```

Figure 9.2.2: B

Then, to add to the query, we will select the `...` and choose the option corresponding to a new instance.

```
Find frames with a car and
     Instance: all instances (def) ...
```

Figure 9.2.3: C

To specify that this second instance should be a person, we will fill in the `all instances (default)` hole with `a person`.

```
Find frames with a car and a person ...
```

Figure 9.2.4: D

Then, to indicate that we want to filter on the frames the contain a car and a person, we will select the `...` and choose the option corresponding to a new filter.

```
Find frames with a car and a person
    Filter: no filter (def) ...
```

Figure 9.2.5: E

Finally, as we want scenes where the car and person move away from one another, we will replace the `Filter: no filter (def)` with `moving away`. This will give us the desired query.

```
Find frames with a car and a person
    moving away ...
```

Figure 9.2.6: F

CONCLUSION

[Mick: todo]

REFERENCES

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2019. nuScenes: A multimodal dataset for autonomous driving. (2019). DOI: http://dx.doi.org/10.48550/ARXIV.1903.11027

[2] Daniel J. Fremont, Edward Kim, Tommaso Dreossi, Shromona Ghosh, Xiangyu Yue, Alberto L. Sangiovanni-Vincentelli, and Sanjit A. Seshia. 2020. Scenic: A Language for Scenario Specification and Data Generation. (2020). DOI: http://dx.doi.org/10.48550/ARXIV.2010.06580

[3] Yongming Ge, Vanessa Lin, Maureen Daum, Brandon Haynes, Alvin Cheung, and Magdalena Balazinska. 2021. Demonstration of Apperception: A Database Management System for Geospatial Video Data. *Proc. VLDB Endow.* 14, 12 (jul 2021), 2767–2770. DOI: http://dx.doi.org/10.14778/3476311.3476340

[4] Brandon Haynes, Maureen Daum, Amrita Mazumdar, Magdalena Balazinska, Alvin Cheung, and Luis Ceze. 2020. VisualWorldDB: A DBMS for the Visual World.. In *CIDR*.

[5] Edward Kim, Jay Shenoy, Sebastian Junges, Daniel Fremont, Alberto Sangiovanni-Vincentelli, and Sanjit Seshia. 2021. Querying Labelled Data with Scenario Programs for Sim-to-Real Validation. (2021). DOI: http://dx.doi.org/10.48550/ARXIV.2112.00206

[6] The pandas development team. 2020. pandas-dev/pandas: Pandas. (Feb. 2020). DOI: http://dx.doi.org/10.5281/zenodo.3509134

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. (2021). DOI: http://dx.doi.org/10.48550/ARXIV.2103.00020

[8] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2015. You Only Look Once: Unified, Real-Time Object Detection. (2015). DOI: http://dx.doi.org/10.48550/ARXIV.1506.02640

[9] Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

[10] Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman (Eds.). 56 – 61. DOI: http://dx.doi.org/10.25080/Majora-92bf1922-00a

[11] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. 2022. *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

[12] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple Online and Realtime Tracking with a Deep Association Metric. (2017). DOI: http://dx.doi.org/10.48550/ARXIV.1703.07402

[13] Esteban Zimányi, Mahmoud Sakr, and Arthur Lesuisse. 2020. MobilityDB: A Mobility Database Based on PostgreSQL and PostGIS. *ACM Trans. Database Syst.* 45, 4, Article 19 (dec 2020), 42 pages. DOI: http://dx.doi.org/10.1145/3406534