Spatial Temporal Reasoning Using QSR, Physics, and Image Processing

Nathan Eloe  
nwe5g8@mst.edu

Jennifer L. Leopold  
leopoldj@mst.edu

Chaman L. Sabharwal  
chaman@mst.edu

Computer Science Department Missouri  
University of Science and Technology Rolla,  
MO 65409, USA

Abstract

Qualitative spatial reasoning (QSR) is a powerful tool in automated computer reasoning, a necessary step forward in fields like computer vision and media analysis. Stereo graphical media has rapidly become a prevalent part of technological culture, and the amount of these kinds of data that exists is staggering. Humans interpret depth information using prior knowledge that a computer lacks. This prior knowledge stems from remembered observance of the basic laws of physics. While the computer lacks the intuitive understanding of these principal physical properties, it is capable of determining more precise information through calculation. Herein the authors explore the information that can be gained from an amalgamation of QSR methods and physics, and present some preliminary results from an implementation based on this powerful combination.

1 Introduction

Human perception of three dimensions is a complex field. A person determining the shape of an object must rely heavily on visual cues in the form of shading, edges, and depth information [1]. Where this information is insufficient to unambiguously identify an object, the observer must make a judgment that may not be consistent with that of all other observers [1]. Research has shown that an individual’s desires can influence the way things are perceived (“Wishful Seeing”) [2].

This raises some important questions in the field of computer vision, including:

- As a binary entity, how can the computer deal with ambiguous visual information?
- With no base of previous experience, what information should the computer use to reduce ambiguity?
- How can the information gain of the system be maximized while the computational cost is minimized? In other words, what calculations should be done to obtain the most information with the least work?

With stereoscopic media becoming pervasive in the form of 3D movies and consumer electronics (e.g. televisions and portable gaming devices), there is a growing, urgent need for computational analysis of these data. Such technology could have impact in physical security (e.g. analysis of images from multiple sources such as security cameras), robotic vision, and defense (e.g. identification of potential dangers and suspicious behavior from stereoscopic information).

Image processing techniques can provide insight into a system recorded stereoscopically, but only about what can be seen by the cameras. Humans use experience and prior knowledge to make assumptions about parts of the scene that are hidden from view. One example of this is a speaker behind a podium; an audience member would know that a human standing behind a podium most likely has legs, and that the podium does not continue back into infinity because the objects in the scene (human, podium), are known quantities that have been encountered before. The computer does not have this same experience; image processing techniques alone would only allow the computer to know about what is directly visible.

In this paper, the authors explore the use of Qualitative Spatial Reasoning (QSR) methods and basic physical properties in addition to visual information from the scene to reduce the amount of incomplete visual information. Spatial information gained from QSR and physics is retroactively applied to the scene to further reduce ambiguity; present knowledge about a system is used to revise past assumptions, which improves the precision of current and future data.
2 Background and Related Work

2.1 Image Processing and Disparity

Image processing is an important field in computer and robotic vision. A large portion of research in this area has been devoted to finding computationally efficient algorithms; images are inherently two dimensional, which implies that most naive algorithms are at best $O(m \times n)$ in their computational complexity for an $m \times n$ image. The persistence of high resolution images (full high definition already being common and 5k resolution beginning to emerge) means that these algorithms will be computationally expensive, especially with many image formats being 4-channel (giving an $m \times n \times 4$ data structure size to hold RGBA or HSV format (Hue Saturation Value Alpha) information, two popular information formats).

Disparity [3, 4, 5] and the parallax effect are two concepts exploited in image processing to mimic human perception of depth. It uses the fact that objects (and distances) appear smaller the farther away from the observer they are. Thus, by determining the distance between occurrences of an object in each of a pair of stereo images, the relative distance from the camera to the object can be determined. Disparity has also been used to estimate the motion of objects [5]. It is an invaluable tool in determining spatial information from multiple observations of the same scene.

2.2 Human Perception in Three Dimensions

Human 3D perception is fascinating: by all reckoning, such a feat should be mathematically impossible with the abstract data the brain receives from the eyes [1]. Regardless, humans are capable of making relatively consistent judgments about shapes and motion in three dimensions using only data from two “cameras” (the eyes) and a base of experience. Learned behavior such as object permanence [6] show that prior knowledge is required to make judgments about three dimensional space. Mimicking human perception with a computer is an important facet of computer and robotic vision.

2.3 Qualitative Spatial Reasoning (QSR)

Qualitative Spatial Reasoning (QSR) has varying applications in Geographic Information Systems (GIS), visual programming language semantics, and digital image analysis [7, 8]. Systems for spatial reasoning over a set of objects have evolved in both expressive power and complexity. The design of each system focuses on certain criteria, including efficiency of computation, ease of human comprehension, and expressive power.

The spatial reasoning system chosen for this investigation is VRCC-3D+ [9], an expansion and implementation of the RCC-3D [10] system designed by Albath et al. As opposed to other RCC systems (many of which have no implementation), the relations in VRCC-3D+ express both connectivity (in 3D) and obscuration. Obscuration will change from viewpoint to viewpoint, but connectivity is a global property that can be used to discern new information at every perspective in the system.

For this work, the authors focus on the obscuration element of the VRCC-3D+ relation. The connectivity portion of the relation will become important as the system is expanded to handle an arbitrary number of cameras and camera vantage points. VRCC-3D+ identifies four basic kinds of obscuration: no obscuration (nObs), partial obscuration (pObs), complete obscuration (cObs), and equal obscuration (eObs). The system further breaks each base obscuration into three different classes: regular obscuration (object A obscures object B), converse obscuration (object A is obscured by object B), and equal/mutual obscuration (object A and object B obscure each other). At this point in the investigation, this further classification is unimportant; it only matters if obscuration is present between two objects, not which object is being obscured.

2.4 Inertia and Conservation of Mass and Energy

There are a multitude of physical properties that can be used to discern information about spatial relationships. Every property used to derive spatial information introduces a new computational cost and has an upper limit to the amount of information it can deduce. The ideal property would be one that would give insight into the system without requiring any new calculation. When this is impossible, the goal should be to maximize the ratio of information gain to computational cost. One of the goals of this research is to discover a combination of physical properties that maximizes this ratio.

As a starting point, two physical properties will be examined: inertia and conservation of mass and energy. Inertia is best described by its colloquial definition: an object at rest tends to stay at rest, an object in motion tends to stay in motion. More formally, inertia is the resistance a physical object has against a change in its state of motion or rest. This can provide useful insight into the physical relationship of two objects. Given two objects, if one passes behind another, it can be used to determine whether or not the objects collided at any point. In terms of spatial connectivity, this collision will correspond to a change from a disconnected (DC) state to an externally connected (EC) state. This in turn gives useful information because it defines a known point on the (possibly hidden) boundary of one or both objects.
Conservation of mass and energy will also be used in conjunction with inertia to gain additional information. If an object becomes obscured by another object, its trajectory can be estimated. If the actual trajectory is different than the calculated trajectory, then something must have changed the state of motion or rest of one or both objects. Using the difference in expected and actual position at a given time to revise earlier calculations results in a corrected physical model that yields additional information about the entire system.

2.5 Current Work in Qualitative Spatial and Temporal Reasoning

Qualitative spatial and temporal reasoning has been an active field in recent years. Takahashi [11] explored using a new expansion to RCC-8 in which he uses two specific vantage points at right angles to a scene. Connectivity and obscuration were determined from each location to give a more precise determination about the objects in the scene. Takahashi’s work differs from this work in that he uses a front and bird’s-eye (“side” and “upper”) view to obtain information. In contrast, this work focuses on emulating human sight using stereo images, which will be expanded to include information from additional visual sources.

Renz [12] proposed efficient algorithms for determining tractable subsets of RCC-8 and the Interval Algebra by phrasing the problem as a consistency satisfaction problem CSPSAT(S) and refining the set when necessary. Directly determining the relations between objects in space and time is not a direct consequence of these tractable subsets, but any reduction in the size of the subset of possible relations can increase the efficiency of determining actual relations between objects [13]. These tractable subsets can be used to aid in the disambiguation of information from multiple sources and will be exploited in this research.

Renz and Ligozat [14] performed a theoretical analysis of spatial temporal reasoning systems and showed that if a system exists such that weak composition does not result in actual composition, path consistency no longer applies. In these cases, algebraic closures of the system must be used to determine composition. They examine the effects of weak composition on spatial temporal reasoning systems and provide a methodology to analyze spatial and temporal calculi. While purely theoretical, this work benefits qualitative spatial and temporal reasoning. Path consistency and composition are two important attributes of a QSR system that have been exploited to aid in automated reasoning; analysis of this work to show these facets of spatio-temporal reasoning are not violated will be important to the continued usefulness of the system.

Ye and Hua [15] explored using depth cameras to determine three dimensional spatial relations. They did not apply their work to a series of images over time, and use specialized depth finding cameras to determine depth (the Xbox Kinect). As the research presented in this paper is expanded to include additional visual sources, Ye and Hua’s work may be investigated as another kind of information source.

In 2007, Santos [16] investigated a framework in which the depth and motion of an object may be reasoned with while accounting for the observer’s viewpoint. He presented a formal logic based approach to reasoning about depth and motion that he used in a robotic vision application called the Depth Profile Calculus (DPCC). DPCC uses depth maps obtained through disparity calculations to determine information about three dimensional space, but ignores many other visual cues available (such as color, lighting, and other physical properties). In this work, the authors use similar methods, but incorporate additional information to get a more correct view of the world.

3 Computational Spatial and Temporal Reasoning

As an initial exploration, the authors constrained the system of interest as follows:

- The system is modeled as a single rolling green sphere that passes behind a stationary blue sphere but does not collide (Fig. 1).
- The system is simulated using Blender 2.64 [17] with two separate camera positions to guarantee that frames from the cameras would be showing different perspectives of the same point in time.
- The cameras were aligned such that the direction of views were parallel and the top row of the left camera’s image corresponded to the top row of the right camera’s image. This differs from human vision slightly, as the computer does not need to “focus” on an object by pointing both cameras at it; its visual information is more complete than a human’s over the entire image.
- The floor of the system was transparent.
- The moving sphere’s trajectory was perpendicular to the view direction of the cameras.

The preceding constraints were placed on the system to allow simplifications that are considered to be unimportant in the context of this work:

- Masking the image using HSV (Hue Saturation Value) values was used for image segmentation into objects.
- Disparity was calculated for each object by finding the center of the matching bounding box of objects and determining the difference in the x direction.
Analysis of stereoscopic videos is a three step process: *frame analysis*, *obscuration analysis*, and *object analysis*.

### 3.1 Frame Analysis

In the context of this work, a *frame pair* is a pair of stereo images from a left and right oriented camera that portray the same moment in time from different perspectives. For every frame pair in the videos the following actions are taken:

- The images are converted from the default representation to HSV
- Range filtering is used to determine the locations of both objects in the images.
- The disparity and bounding rectangle are calculated for each object.
- The bounding rectangle and disparity for each object are stored, along with the frame number.

### 3.2 Obscuration Analysis

For this paper, obscuration and object analysis occur with respect to the left camera. The results could be refined by using information from both cameras.

The following pseudocode is used to determine the obscuration from the left camera at every step. The list of bounding rectangles and disparities from the frame analysis is stored in steps. The green sphere corresponds to object A in the pseudocode, and the blue sphere is object B, and bbox refers to an object’s bounding box:

```python
obss = []  # the list of obscurations
for s in steps:
    if object A has a bbox in s:
        xa = A's bbox x location in s
        wa = A's bbox width in s
    if object B has a bbox:
        xb = B's bbox location in s
        wb = B's bbox width in s
    if the bboxes overlap:
        lastO = 'pObs'
    else:
        lastO = 'nObs'

    else:
        lastO = 'cObs'
    obss.append(lastO)
```

The eObs obscuration type is combined with cObs; not enough information exists in this experiment to distinguish between cObs and eObs.

---

Figure 1: Images from analyzed video: as seen from the left camera. The green sphere is further from the cameras than the blue sphere, and as such appears smaller.
Note that there is no distinction as to which object obscures the other, just that some obscuration occurs. It was visually verified that this code correctly identified changes in obscuration with respect to the left camera’s video feed. Fig. 1 shows the frames identified as changes in obscuration occurred from the left camera’s perspective.

3.3 Object Analysis

In the object analysis step, the positions and depths of each object are determined. Position is determined using the right most edge of the bounding box. If no obscuration is detected from either perspective, the depth and position of the object are directly recorded. Otherwise a polynomial is fit to the previous values recorded and used to estimate the current location. Every direction of movement (x,y,z) is handled independently. Due to the simplicity of the nature of this system, a linear fit was used; as the system is generalized, the order of the polynomial can be increased to handle differing kinds of acceleration and forces.

4 Experimental Results

Fig. 2 shows the positions of objects from a birds eye view of the system. Every marker on the graph shows an observed or estimated object location of a particular object. This information can be remarkably helpful in learning about the structure of the system. For example, it may be possible to determine from using only the stereo images that the blue sphere does not extend into infinity due to the perspective nature of the projections. However, depending on the intrinsic properties of the camera, there could be a large area in space that may or may not contain the blue sphere. Using the information gained from projecting the path of the green sphere behind the blue sphere, it can be concluded that the green sphere did not collide with the blue sphere, so an upper bound is placed on how far back the blue object can extend.

This figure illustrates that this line of inquiry shows promise: a relatively accurate extrapolation of the green ball’s location is feasible with a relatively small number of data points. This estimation could be improved further by including the observed position of the green ball in later frames, then using that information to retroactively correct the estimations of the balls location. This will allow information inferred from that estimation to be refined even further.

5 Conclusions and Future Work

Using physical properties in conjunction with QSR and image processing methods is a promising direction in the field of computational vision and spatio–temporal reasoning. This could have applications in physical security (automated CCTV analysis), media analysis, and many other multimedia fields.

In this paper, the authors have initiated an exploration into using these three areas to accomplish automated spatio–temporal reasoning. The results of this initial research are encouraging but leave room for improvement and refinement. This work will be continued to allow analysis of systems with fewer constraints, add additional physical properties that are considered, and eventually be applied to video feed of live events from cameras, not just rendered physical simulations. Different combinations of physical properties and image processing techniques will be investigated to find a high information gain to computational cost ratio.

References


Figure 2: Observed and extrapolated positions. Each data point is a frame. Motion is from left to right.


