

Spatio-temporal Modelling using Video Input

Peter Whigham

Spatial Information Research Centre
University of Otago, Dunedin, New Zealand
Phone: +64 3 479-8301 Fax: +64 3 479-8311
Email: pwhigham@infoscience.otago.ac.nz

**Presented at SIRC 2000 – The 12th Annual Colloquium of the Spatial Information Research Centre
University of Otago, Dunedin, New Zealand
December 10-13th 2000**

ABSTRACT

Many disciplines study the variation in patterns (events) that occur across space and time. Comparison between these events is often facilitated by the ability to measure the similarity between events based on both spatial and temporal context. This paper describes the initial problems and complexities involved with using spatiotemporal similarity techniques, based upon patterns observed using video input. A template and continuous model of space, incorporating vector representations for movement (time), are both discussed as possible representations for sporting events described by video. The paper includes representation techniques for spatiotemporal events, their comparison using similarity measures and an indexing system for extracting video segments from a database.

Keywords and phrases: spatio-temporal modelling, similarity measures, video input, sporting patterns

1.0 INTRODUCTION

Many disciplines study the variation in patterns (events) that occur across space and time. Comparison between these events is often facilitated by the ability to measure the similarity between events based on both spatial and temporal context. This paper will describe the issues surrounding the design of a system for recording spatio-temporal patterns, based on video input, and the ability to index video segments based on generalised pattern descriptions. Spatiotemporal similarity techniques have been studied in geostatistics and artificial intelligence (Delis, Papadias et al. 1998), and in the context of multimedia (Faloutsos and Lin 1995; White and Jain 1996) and database research (Erwig, Gutting et al. 1997). Previous work using video has focused on single object motion in a highly controlled environment (Chang and Lee 1997), or by automatically extracting features from the video signal (Chang and Lee 1997; Chang, Chen et al. 1998). This paper describes the issues involved with building a system to allow generalisations for a number of objects in a scene defined by a complex video signal, and their extension to spatiotemporal similarity measures for indexing and video retrieval.

Video sources have been selected for data capture because they are readily available, contain a large amount of information and are a visually powerful media for presenting the retrieved concepts. The use of video as the data source for studying spatiotemporal patterns implies several requirements: the recorded events must be spatially referenced; the elements described within the events must be distinguishable and generalisable; the patterns described must be both novel and interesting; and there must be sufficient data available to produce many different (but similar) patterns. These requirements suggest that a sporting domain is an appropriate field for study. In particular, the domain of rugby union has been selected for the following reasons: there is an abundance of video data for rugby union, the playing field gives a spatial reference for any event; the time when any event occurs is normally shown from television broadcasts; there are complex patterns involving both the players and referee that

are of interest to researchers in rugby union, including coaches (of players and referees), analysts, and the general public; and there are many years of historical data, allowing the opportunity for studies to be performed that explore how patterns of play have changed over time.

There are currently commercial systems used by rugby coaches for training, such as the KeyToAnalysis system, built by Community Communications of Christchurch, New Zealand. This system allows the rapid coding of propositional descriptions of events during a rugby game, for example “line break”, “player x tackled”, etc. Although the system can also record spatial location for an event, it is not possible to use similarity measures based on spatial context, nor can a series of spatio-temporal events be used to classify and index video segments (personal communication, Kelvin Duncan, Community Communications). The proposed system described in this paper will extend these concepts to allow indexing of video segments based on spatial and temporal similarity measures.

2.0 THE ATTRIBUTES OF VIDEO INPUT

Most studies involving the use of video come from the field of multimedia. The main focus is techniques to extract features automatically from video that allow indexing into a video stream possible, or the development of techniques for describing changes in spatial patterns. These techniques are applied to very specific and well controlled forms of video, normally either by using a stationary, single view camera with positional references (Chang and Lee 1997), or by having a user input a description of the scene and its changes over time (Vazirgiannis, Kostalas et al. 1999).

2.1 Sporting Video Descriptions

The automatic extraction of features from televised sporting events is currently not possible. Televised sporting events are normally shown with a variety of camera angles, incorporate the use of zoom and scrolling, and swap between views in an unpredictable manner. Based on this input it is not possible to automate extraction of features, since there is no frame of reference. Additionally, the features of interest are neither static nor predictable in form or behaviour. Given these constraints the construction of meaningful descriptions (i.e. the semantics) of the video sequence requires human interaction.

Consider patterns that occur during a game – there are set plays, set actions involving individual players and distinctive collections of players, and then more fluid group and individual patterns. All of these patterns have both a spatial and temporal context and, depending on the purpose of the video user, all may be meaningful patterns to record and be able to index for similarity. Certainly a system for describing and storing these patterns will not be able to predict all forms that are important prior to the use of the system. Hence the system must be constructed in a generic fashion, allowing the user to construct their own description of actions, objects, spatial relationships and time.

It is informative to list the different levels of detail that exist when considering video sporting descriptions. For the following description an Event is defined as a sequence of video frames that are perceived by the user as appropriately described with a single label.

- Event with no spatial or temporal context, which is simply defined by a duration (number of frames). For example, *player x involved in tackle y*.
- Event with spatial context but no temporal reference. For example, *player x involved in tackle y in the centre of the field*.
- Event with spatial and temporal context. For example, *player x involved in tackle y in the centre of the field after 20 mins of play*.

An Event from the previous description is atomic, in the sense that the individuals in the video scene (such as players, the ball, the referee) are not explicitly represented. To extend the description it is necessary to consider that a single stream of video contains objects, where an object is a user-defined entity that has meaning for some part of the duration of the video stream. Note that an object may not exist for the entire duration of a video sequence, and that objects may appear and disappear during a sequence. For example, a rugby scrum (an object) may alter into a series of distinct players, or form into several distinct objects that are coherent within a sequence. To extend the previous description:

- Event with spatial context, composed of a number of distinct objects. The objects have a well-defined topology. For example, *the position of a scrum and back-line.*
- Event with spatial context, temporal reference and a number of distinct objects. The objects have a well-defined topology. For example, *the position of a lineout and backline, 5 meters from the tryline, with 10 minutes to go in the game.*
- Event with spatial and temporal context, with a number of distinct objects that may change during the sequence. For example, *a lineout that forms into a ruck, where finally the ball moves out to the backline, with 10 minutes to go and 20 metres from the tryline.*

3.0 REPRESENTING SPATIO-TEMPORAL PATTERNS

There are two main options regarding the representation of spatio-temporal patterns derived from video: sketch-based and icon-based. The following criterion will be used to assess the appropriateness of each technique: speed and ease of input, required accuracy of spatial location for input and query and their relationship to the implementation of similarity measures. Note that due to the size of individual players, and the video presentation, an event based on the analysis of a video cannot be spatially located with an accuracy of less than one metre, except for several specific fixed starting events (such as a kick-off). A number of other issues limit the spatial resolution for positioning events, including:

- irregular camera angles,
- that many single events move (for example, a scrum is set at a particular position, but may move during what would be considered a single, atomic event), and
- the relative importance of position when describing patterns.

Given these issues, for the domain of rugby union, it is not necessary to have events referenced with accuracy below approximately 1 metre, which is essentially the local neighbourhood of an individual.

Other issues include the symmetrical representation of space and the concept of attack and defense. Although the entire playing field is necessary for location, the defensive and attacking areas of the field are symmetrical given teams typically swap ends during a game. Hence it will be necessary to be able to indicate who is attacking, and their direction of attack, so that this context can be properly used when accessing stored descriptions.

3.1 Sketch-Based Descriptions

Sketch-based descriptions (Egenhofer 1996; Egenhofer 1997) have been previously developed for both the representation of spatial patterns and as a query mechanism. *Sketcho* (Blaser 1999), the system developed using these concepts, allows a user to sketch vector-based objects using a mouse, associate attributes to these objects, and to use a form of case-based reasoning to search for similar patterns in a database. The use of weightings for pattern similarity allows matching to occur based on topological as well as attribute relationships. This approach has the merit of flexibility of representation, and is similar to techniques often used by coaches when describing moves and patterns of play. However, sketching requires certain skills, including the ability to be consistent between different descriptions, and knowledge of the pattern descriptions used in the system. Additionally, sketching with a mouse or pen takes more time than a single “drag-and-drop” operation, and connecting attributes (such as a player name) to a sketch is not a consistent operation with the overall feel of sketching. However, an interface can be constructed that allows a reasonably fast association of attributes to sketches by having a list of attributes, selected using the pen or mouse, to connect with a sketch object. In fact, attaching attributes to a sketch or any other form of representation will still have to deal with this step, and therefore it is not a determining factor. The location of an event, such as a scrum, will require the calculation of either a centre-point of the polygon used to represent the scrum, or allow an area/extent description to be used. Since drawing these objects will be fuzzy and will be done rapidly, the additional flexibility allowed for by sketching does not seem appropriate to this circumstance.

3.2 Icon-Based Descriptions

Icon-based, or template-based, descriptions use a set of icons, pictures or building block images, to construct a description. This implies that they are suitable for situations where there are a set number of well-defined events, and that these events do not vary significantly in spatial extent. Note, however, that an icon may be stretched to suit a particular description. An icon description can be associated with a particular grouped object (such as a scrum, lineout or backline) or an individual player or referee. Since a “drag-and-drop” operation is a fast method for associating an event with a location, and icons can be pre-defined for a particular sporting domain, they appear to be an appropriate solution to the representation of events for this problem. Additionally, since absolute location of events is fuzzy under most circumstances, the use of an icon for positioning does not reduce the descriptive power of the patterns. For icons that are static in size (such as a player position) the centre-point of the icon can be used as the reference point. For objects that may vary in size, such as a lineout, a centre-point and bounding box description is adequate. Both of these restrictions are easily adapted to similarity measures for indexing, and therefore seem more appropriate than the sketch-based approach.

3.3 Representation of Space and Time

The representation of space depends on the type of similarity technique that will be used. The main choice is whether to have the space gridded, so that each location has been quantised, or to allow location to be continuous (within the resolution of position specified by the input device). The use of a grid-based description means that there are a set number of horizontal and vertical positions for any object. This allows certain types of similarity measures to be implemented that use binary representations for position and time (see for example (Delis, Papadias et al. 1998)). The use of continuous space would imply that the similarity techniques would use a form of euclidean metric for the distance between two patterns, and similarity would likely be measured by the minimum transformations required to match both patterns. However, there are some implications for maintaining topological relationships that need to be addressed with this approach. Either of these selections also have implications for the form of indexing used to access the database of patterns when searching for similarity of patterns.

Time is based on game time, which depending on the game has certain characteristics. All games are divided up into one or more parts, where normally the time is restarted at the beginning of each part. For example, with Rugby Union there are 2 halves, each of 40 minutes duration, whereas other sports have different divisions of time. This time division is significant, since a typical query may be interested in events that occurred with 5 minutes to go in the game, or 5 minutes to go in the half (or other division). Often turning points in a game occur towards the end of a division of play due to tiredness, anxiety, lack of concentration or some other combination of conditions. Similarity queries must be able to handle the relative query based on time before a break, as well as absolute time during a division.

Handling temporal issues such as relative versus absolute time does not impact on the indexing structure of the temporal label for an event. This is possible since the query, be it relative or absolute, can always be converted to an absolute description before indexing the database for similar events. The more difficult issue with temporal events are how they are best represented by the user (i.e. how time changes are input) and how an indexing structure can be used to incorporate events that change.

4.0 SPATIO-TEMPORAL SIMILARITY TECHNIQUES & INDEXING

Previous authors have studied the design of database systems for storing and retrieving spatio-temporal models (Raafat, Yang et al. 1994; White and Jain 1996; Erwig, Gutting et al. 1997). The descriptions usually start with an initial object shape and position, and then store incremental changes. An alternative is to create a new object each time an object changes, rather than keeping the changes to a single initial object. The implications for this are mainly to do with the data storage and having to deal with connections between objects. However, since it is possible within one scenario to have objects change, move, appear and disappear this is not necessarily an issue. The situation here is similar, since objects may change in extent as well as position. For example, a simple line representing a backline may change in length, break up into smaller components, or disappear during the description. Formal descriptions for the design of database structures of spatio-temporal events usually give a tuple-description, which incorporate either spatial and temporal components separately, or as a combined object. An excellent description of the basic issues surrounding multi-media indexing and similarity measures is presented by Vazirgiannis (Vazirgiannis 1999).

A simple approach to spatio-temporal indexing is to have two separate indexes: a spatial (two-dimensional) index (xmin, ymin, xmax,ymax) , and a temporal index (starttime, duration). Several data structures have been defined for indexing spatial data, mainly based on R-trees (Guttman 1984) or derivatives of this approach. Since R-trees are designed for n-dimensional indexing, it is also possible to combine the spatial and temporal indexes to form a single indexing scheme. This produces a three-dimensional index for the complete spatio-temporal information of each object. These indexing techniques are designed to quickly eliminate those objects that cannot satisfy the query, with the remaining objects having to be studied in more detail. This elimination is achieved by the way the indexing is structured, so that objects having either a spatial or temporal value far from the desired query are physically separated (in the R-tree) by the indexing mechanism. For the purposes of this paper it is sufficient to note that a number of n-dimensional approaches to indexing are available.

5.0 SYSTEM DESIGN

The system will be designed in a generic framework, where the definition of space, time and objects will be constructed via a database definition. This will alter the system interface, so that the appropriate venue, teams and players will be presented on the screen in such a way as to allow fast access for constructing events with individuals. The object definitions in the database will associate icon bitmaps with objects, so that the system can be tailored for different sporting or other domains. The video source, digitized with non-playing components removed, will be stored as a single stream in the database. This stream will have time codings associated with it so that the system can automatically give "game-time" when displaying. The interface for recording or querying for events will be the same, since the description of objects to connect to a video stream is the same description as that for similarity querying. However, the use of similarity measures using weights (for example, ignoring or reducing the importance of the temporal component) will require additional information to be included with this activity. Since many video segments may satisfy a query this interface must also allow either fingernail sketches of each segment to be display, or present the object descriptions (i.e. the patterns used to define the event) and a textual description giving the context of the game, time, etc. From a list of these descriptions the user can select the desired video sequences.

6.0 DISCUSSION AND CONCLUSION

The design and construction of a spatio-temporal video indexing system for generalised events will be a challenging task. This paper has presented some of the basic issues surrounding the construction of such a system, in the context of rugby union as the particular domain. Many issues have not been resolved, including the most appropriate interface for defining object movement, how to update the temporal component during a series of events that are considered a single scenario, how to define similarity when time is involved, and many others. For example, should two complex events be considered similar if one has three objects that move over time, and the other has three objects that are identical at the start, but do not have a temporal component? Is it a minimal matching that is required, or should a longer sequence always be considered the minimum length for similarity purposes?

There are many issues involved in the design of this system, with an overriding principle that the system must be fast for input and querying, extensible, generic, and flexible enough to cover many different domains. The definition of space, time and events in a generic framework has not been resolved fully resolved at present and the development of this system will offer the opportunity to study and deal with many of these issues.

ACKNOWLEDGMENTS

The author would like to thank Alec Holt, Ken Hodge, Colin Aldridge and Kelvin Duncan for discussions that led to the definition of this project. This system will be constructed at the University of Otago during 2001. This research is supported by an Otago Research Grant.

REFERENCES

Blaser, A. (1999). Users Guide for Sketcho. Orono, National Center of Geographic Information and Analysis, University of Maine: 50.

- Chang, C. and S. Lee (1997). A Video Information System for Sport Motion Analysis. *Journal of Visual Languages and Computing* **8**: 265-287.
- Chang, C. W. and S. Y. Lee (1997). Video Content Representation, Indexing and Matching in Video Information Systems. *Journal of Visual Languages and Computing* **8**(2): 107-120.
- Chang, S.-F., W. Chen, H. J. Horace, H. Sundaram and D. Zhong (1998). A Fully Automated Content Based Video Search Engine Supporting Spatio-Temporal Queries. *IEEE Trans. CSVT* **8**(5): 602-615.
- Delis, B., D. Papadias and N. Mamoulis (1998). Assessing Multimedia Similarity: A Framework for Structure and Motion. *Proceedings of ACM Conference on Multimedia (SIGMM)*, Brighton, UK, ACM Press.
- Egenhofer, M. (1996). Spatial-Query-by-Sketch. VL '96: IEEE Symposium on Visual Languages, M. Burnett and W. Citrin, Eds., IEEE Computer Society, Boulder, CO.
- Egenhofer, M. (1997). Query Processing in Spatial-Query-by-Sketch. *Journal of Visual Languages and Computing* **8**(4): 403-424.
- Erwig, M., R. H. Gutting, M. Schneider and M. Vazirgiannis (1997). Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases, *Praktische Informatik IV Fernuniversitat Hagen*: 21.
- Faloutsos, C. and K. Lin (1995). *FastMap*: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, M. Carey and D. Schneider, Eds., San Jose, California, USA.
- Guttman, A. (1984). R-trees: A Dynamic Index Structure for Spatial Searching. *Proceedings of ACM SIGMOD International Conference on Management of Data*.
- Raafat, H., Z. Yang and D. Gauthier (1994). Relational Spatial Topologies for Historical Geographical Information. *Int. Journal of Geographical Information Systems* **8**(2): 163-173.
- Vazirgiannis, M. (1999). Interactive Multimedia Documents, G. Goos, J. Hartmanis and J. V. Leeuwen (Eds.) *Lecture Notes in Computer Science*: **1564**, Springer.
- Vazirgiannis, M., I. Kostalas and T. Sellis (1999). Spatial and Temporal Specification & Authoring of Multimedia Scenarios. *IEEE - Multimedia*.
- White, D. A. and R. Jain (1996). Similarity indexing: Algorithms and performance. *SPIE Vol.2670*, San Diego.