Published paper:

Sridhar, M, Cohn, AG and Hogg, DC (2011) *From Video to RCC8: Exploiting a Distance Based Semantics to Stabilise the Interpretation of Mereotopological Relations.* In: Egenhofer, MJ, Giudice, NA, Moratz, R and Worboys, MF, (eds.) Spatial Information Theory - 10th International Conference, COSIT 2011, Belfast, ME, USA, September 12-16, 2011. Proceedings. COSIT 2011, Belfast, ME, USA. Springer , 110 - 125 .

http://dx.doi.org/10.1007/978-3-642-23196-4_7

From Video to RCC8: exploiting a Distance Based Semantics to Stabilise the Interpretation of Mereotopological Relations

Muralikrishna Sridhar and Anthony G Cohn and David C Hogg

University Of Leeds, UK,
{krishna,agc,dch}@comp.leeds.ac.uk. [**]

Abstract. Mereotopologies have traditionally been defined in terms of the intersection of point sets representing the regions in question. Whilst these semantic schemes work well for purely topological aspects, they do not give any semantic insight into the degree to which the different mereotopological relations hold. This paper explores this idea of a distance based interpretation for mereotopology. By introducing a distance measure between x and y, and for various Boolean combinations of x and y, we show that all the RCC8 relations can be distinguished. We then introduce a distance measure which combines these individual measures which we show reflect different paths through the RCC8 conceptual neighbourhood – i.e. the measure decreases/increases monotonically given certain monotonic transitions (such as one region expanding). There are several possible applications of this revised semantics; in the second half of the paper we explore one of these in some depth – the problem of abstracting mereotopological relations from noisy video data, such that the sequences of qualitative relations between pairs of objects do not suffer from "jitter". We show how a Hidden Markov Model can exploit this distance based semantics to yield improved interpretation of video data at a qualitative level.

1 Introduction

Mereotopologies have traditionally been defined in terms of the intersection of point sets representing the regions in question. This is true for both RCC8[14], the 4- and 9- intersection calculi[3, 6] and indeed many other mereotopologies covered in [2]. Alternatively, Galton [8] gives a semantics in which the eight RCC relations are distinguished by whether all, some or none of x is inside y or not, and vice versa, and whether there are shared boundary points or not.

Whilst these semantic schemes work well for purely topological aspects, they do not give any semantic insight into the degree to which the different mereotopological relations hold. The authors in [1] and [5] provide a way of describing

[**] The financial support of the EU Framework 7 project Co-friend (FP7-ICT-214975), and the DARPA Mind's Eye program (project VIGIL, W911NF-10-C-0083) is gratefully acknowledged, the latter also supplying "person-ball verb" videos.

relations between regions which have uncertain boundaries. However, neither provides an alternative semantics for mereotopology, nor do they address the issue of learning a robust transformation function from metric to qualitative relations.

For EC [2], and PO [9], more granular calculi have been designed which distinguish the degree to which these relations hold; however, to our knowledge, the other RCC8 relations have not been so refined. In any case these refinements are at the calculus level (rather than the semantic level), and are discretised (a finite number of refinements) and with no metric on the refinements. For the other relations, the degree of holding has not been covered at all in the semantics of the relationship (though there are fuzzy versions of RCC[16]).

This paper explores this idea of a distance[1] based interpretation for mereotopology. By introducing a distance measure between x and y, and for various Boolean combinations of x and y, we show that all the RCC8 relations can be distinguished. We then introduce a distance measure which combines these individual measures which we show reflects different paths through the RCC8 *conceptual neighbourhood graph* (CNG) – i.e. the measure decreases/increases monotonically given certain monotonic transitions (such as one region expanding) as previously discussed in [4].

There are several possible applications of this revised semantics. First, we note that [11] present a way of approximating constraint satisfaction in reduced expressivity constraint networks by using the idea of a relation 'almost' holding – the semantics presented here would fit well with this technique.

A second application, which we explore more extensively, is an application to abstracting qualitative spatial relations from video data. In video data, objects are frequently represented by shape abstractions such as minimum bounding rectangles (MBR). However, visual noise and other errors introduced by video processing frequently result in instability in mereotopological relations over time, when these are defined point-set theoretically – i.e. 'jitter' can result as relations change frequently depending on the exact position and size of the MBR. We show how using a distance based metric can result in a much more stable qualitative spatial abstraction; the distance based semantics can be used to decide when to transition between relations. We show that good transition points can be learnt automatically by training an HMM [13]. The impact of this improved technique for abstracting qualitative spatial relations from noisy video data can be demonstrated in a procedure to learn event classes from video data.

The rest of the paper is structured as follows. Section 2 introduces the proposed distance based semantics for RCC. Section 3 proposes a way of using this semantics within a HMM based framework, in order to handle noise in video. Section 4 describes experiments on real video data for validating the effectiveness of the proposed approach in handling noise, in relation to the traditional

[1] In this paper, we refer to *distance* as a numerical description of how proximal regions are, without necessarily assuming it is a metric. What is important, as described below, is that it captures certain monotonicity properties over ther conceptual neighbourhood graph.

approaches. In sections 5 and 6 we summarize the work, and point out certain limitations that provide insights into interesting directions for future research .

2 A Distance Based Semantics for RCC

In [14], the binary primitive $\mathsf{C}(x, y)$, x is connected to y was introduced with the semantics that the closure of the region x shares a point with the closure of region y. From this primitive $\mathsf{C}(x, y)$, the eight jointly exhaustive and pairwise disjoint relations of RCC8 can be defined:

$$\Re = \{\mathsf{DC}, \mathsf{EC}, \mathsf{PO}, \mathsf{TPP}, \mathsf{NTPP}, \mathsf{EQ}, \mathsf{TPPi}, \mathsf{NTPPi}\}$$

The 4-intersection model of [6] from which an essentially equivalent set of eight relations can be derived is defined in terms of examining the patterns of intersection between the interior and boundary point sets of a pair of regions x and y: each relation is characterised by a particular combination of $\emptyset$ and $\neg$ symbols, denoting empty and non-empty intersections respectively.

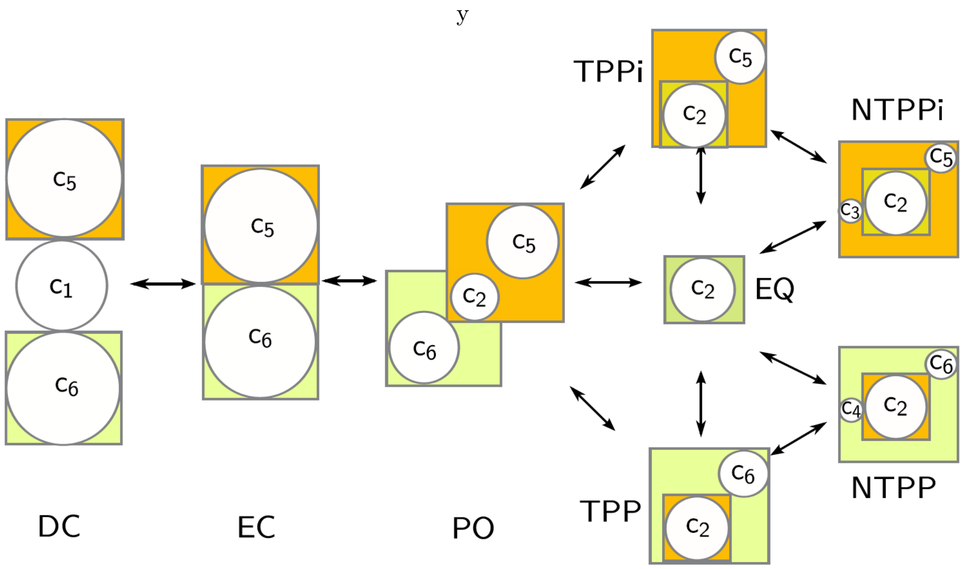

Fig. 1. The RCC8 relations x (green) and y (orange) along with their conceptual neighbourhood. The six circles relating the various Boolean combinations of x and y are also depicted when they have non zero diameter.

In this section, we introduce an alternative semantics for RCC8 (and thus effectively also for Egenhofer's relations). For the sake of simplicity, we restrict our analysis to rectangular one-piece regions with no holes. We discuss the limitations of the proposed framework for other shapes in section 6 pointing to possible ways of generalizing our current approach. We confine our attention here to rectangles aligned to two orthogonal axes, which naturally correspond to rectangular bounding boxes, that are obtained using low video analysis. It is worth noting that addressing the noise arising from low level video analysis has partly inspired the proposed approach.

Our point of departure from previous work is to note that the standard semantics says nothing about how far apart two regions are when they are disconnected (DC). In much earlier work, a CanConnect(x, y, z) relation was introduced [10] which holds when the rigid body x is sufficiently large so as to be able to connect regions y and z if translated into a suitable position. This gives rise to the idea of measuring the degree of disconnection between x and y by a third region. We wish to choose a canonical shape region for this, and the obvious choice is the n-sphere for n-dimensional space. Although RCC can be interpreted in arbitrary dimensions, for the sake of simplicity we restrict our attention in this paper to 2D, so we thus measure the degree of disconnectedness between a pair of regions x and y by the smallest circle which can connect them. We will call this circle $c_1(x, y)$. As x and y approach each other, the diameter of $c_1(x, y)$ will decrease until they become EC, and the diameter of $c_1(x, y)$ is zero[2].

Inspired by this, we now introduce further circles to provide a measure for the other RCC8 relations. Considering the RCC conceptual neighbourhood, the next relation to hold after EC is PO. A circle $c_2(x, y)$ being the largest circle in the intersection of x and y neatly captures the degree to which x and y partially overlap – as x and y transform/translate towards TPP/TPPi/EQ, so the diameter of $c_2(x, y)$ will increase[3].

If we now consider the value of the expression $|c_1(x, y)| - |c_2(x, y)|$, (where $|...|$ denotes the diameter of the circle) then it can easily be seen that it will start off positive, reduce to 0 when $\mathsf{EC}(x, y)$ holds, and then become negative for all the other relations. To distinguish all eight relations, we need to introduce further measurements. We can do this by considering the other Boolean combinations of x and y. Thus $c_3(x, y)$ denotes the smallest circle which can connect y to the complement of x; dually $c_4(x, y)$ denotes the smallest circle which can connect x to the complement of y; $c_5(x, y)$ denotes the largest circle in the region $x - y$ and $c_6(x, y)$ denotes the largest circle in $y - x$. These circles are all depicted for the RCC8 relations in figure 1.

To show that these six circles are sufficient to distinguish all the RCC8 relations, consider Figure 2. By inspection of the columns labelled by $c_1(x, y) - c_6(x, y)$, it can be seen that each row is unique, and thus the RCC8 relation which holds can be determined by inspection of these six circles and whether their di-

[2] Technically a circle has to have a non-zero diameter. But for convenience, here we refer to a point as a circle of diameter 0.

[3] It may be noted that in some cases, the diameter of c_2 and (other circles defined below) may not change for prolonged periods (e.g. if we were to take two rectangles of the same height and translate one of them horizontally inside the other). However, since the principal purpose of the work is give more information near the relation boundaries, it is sufficient that the diameter of these circles changes significantly near these boundaries.)

	c1(x,y)	c2(x,y)	c3(x,y)	c4(x,y)	c5(x,y)	c6(x,y)
	minC(x,y)	maxC(x∩y)	minC(-x,y)	minC(x,-y)	maxC(x-y)	maxC(y-x)
DC	+	0	0	0	+	+
EC	0	0	0	0	+	+
PO	0	+	0	0	+	+
TPP	0	+	0	0	0	+
NTPP	0	+	0	+	0	+
EQ	0	+	0	0	0	0
TPPi	0	+	0	0	+	0
NTPPi	0	+	+	0	+	0

Fig. 2. A table showing whether the diameter of the six circles are 0 or non zero for the RCC8 relations. minC(x, y) denotes a minimum sized circle which can connect x and y and maxC(x) denotes the maximal sized circle which can fit into x.

ameter is zero or non-zero, in just the same way as the 4-intersection model allows the eight Egenhofer relations to be distinguished. Here we require eight values to characterise each relation rather than the four in the 4-intersection model (though less than the nine of the 9-intersection model).

This thus gives an alternative way of defining the standard set of eight mereotopological relations, which differs from the RCC8 definitions based on $\mathsf{C}(x, y)$, or the 4-/9-intersection model, or indeed the modal semantics found e.g. in [15]. This may have some theoretical interest, but our purpose in defining this semantics was to address a problem arising in abstracting RCC8 relations from video (in fact, the problem also occurs even in the simpler RCC-5 calculus – see our earlier work in which we briefly outline a much simpler version of the approach here for RCC-5[19, 18]). Typically in video interpretation, objects/blobs are identified and then tracked, such that a unique identifier can be associated with the different positions of the object at different times. However, it is not just the position of the object which can change, the shape can change too, either because it actually changes shape (as in the case of a person changing their posture), or because, in the image plane, an object appears to get larger as it moves closer to the camera, or because of visual noise which results in the the object detection software assigning a different shape to the object in different frames. It is this final problem which is of particular concern to us, since the changes are not "real changes" but rather artefacts of the software system. Often the size/position of the object will change rapidly from frame to frame, resulting in "jitter". Such problems are well known and endemic in computer vision. The use of shape abstraction primitives, such as bounding boxes, or the convex hull of an object can help alleviate these problems but the issue still remains. The problem of interest to us here is when this jitter causes undesirable changes of spatial relation – for example when two objects approach each other, the RCC8 relation between the bounding boxes may not simply transition from

DC to EC and then to PO following the arcs in the RCC8 conceptual neighbourhood. Rather, there is likely to be a jittering of relations, such as DC, EC, PO, EC, DC, PO, EC, PO, DC, EC, PO (where each of these relations indicates that it holds over some maximal interval with no intervening relations). There are a variety of computer vision smoothing techniques (such as a Kalman filter [20]) which can be applied to the tracks and shape abstractions which can help reduce such jitter; however we have not found these to be satisfactory, possibly because they are not specifically aimed at the discretisations of a qualitative calculus, but generally simply aim to smooth in continuous spaces such as are generally found in low level computer vision representations.

Our approach, detailed in the second half of this paper, is to try to learn when to transition from one relation to another. We do this by training a Hidden Markov Model (HMM) with one state for each relation. In order to build such an HMM, it is convenient to have a variable which can be used to assess which state holds. The idea of using the distance based semantics for RCC8 is attractive in this regard since it allows the possibility of combining the different circle measures to produce an overall measure which can be used to decide when to transition from each relation; a vector of measures could also be used, e.g. taking each circle diameter individually as in input, but this makes the learning task more difficult.

So we turn to the question of how to combine the circular measures to provide an appropriate value for the HMM. Consider the path through the conceptual neighbourhood DC, EC, PO, TPP, NTPP – this can be seen in the first five rows of the table in figure 2. Initially c_1, c_5, c_6 all have positive diameter whilst the others have zero diameter. Then c_1 becomes zero, followed by c_2 becoming non zero, then c_5 becomes zero, then c_3 becomes non zero, then finally c_4 becomes non zero. Thus c_1 and c_5 have become zero, whilst c_2 and c_4 have become non-zero; c_3 and c_6 remain unchanged. Note that all changes are qualitatively monotonic along this path, i.e. there is no change back to a previous qualitative value.

It may also be remarked that, assuming the change in size/position of x and y is monotonic (i.e. x/y change smoothly and without "reversal" between the relations DC and NTPP), then the actual metric value of the $c_i(x, y)$ will in general[4] be monotonic too; e.g. consider $|c_5(x, y)|$: its value will be constant whilst DC and EC hold, but once PO starts to hold, then the diameter of $c_5(x, y)$ will steadily reduce until it becomes zero on the transition to TPP.

For the dual path, DC, EC, PO, TPPi, NTPPi, a similar analysis applies, but with the roles of c_3/c_4 flipped and similarly for c_5/c_6[5]. This leads to a formulation of a measure $d(x, y)$ in which c_1, c_5 and c_6 contribute positively to the measure and c_2, c_3, c_4 negatively:

$$d(x, y) = |c_1(x, y)| + |c_5(x, y)| + |c_6(x, y)| - |c_2(x, y)| - |c_3(x, y)| - |c_4(x, y)|$$

[4] Some limitations on this monotonicity are discussed in section 6.

[5] We could also analyse the other paths through the conceptual neighbourhood of RCC8, and indeed the various processes which engender these, following the analysis of [4] .

However, this measure is dependent on the absolute sizes of x and y; e.g. if x and y are both scaled to twice their size, then $|c_5(x, y)|$ will double too. Thus it is appropriate to normalise the measure. Rather than normalise the entire expression, it is better to normalise each component; this also allows for the possibility of x or y changing their size over time (such as when the entire transition from DC to NTPP is caused by an expansion of the region y [4].

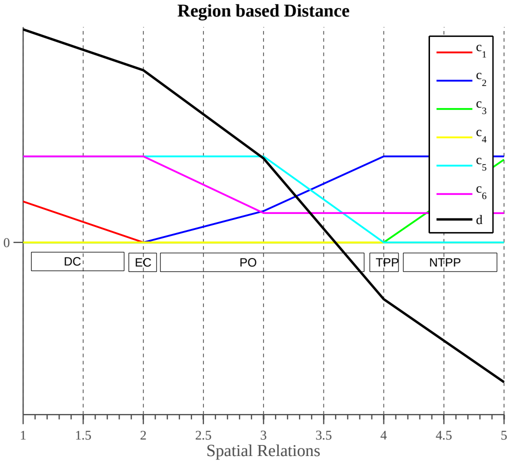

Fig. 3. Illustration of circles c_1 to c_6 as the relationships changes from DC to NTPP. The thick line represents the normalized region based distance obtained using c_1 to c_6 (Original in colour).

We thus consider the normalisation to be applied to each of the $c_i(x, y)$ expressions. The term $c_1(x, y)$ is independent of the size of x and y but is dependent on the size of the universe (we assume here a bounded universe, such as the field of view in a camera image); thus we divide $|c_1(x, y)|$ by $\delta(x \cup -x)$, where $\delta(x)$ denotes the diameter of largest circle inside the bounding box of x, as a measure of the size of x. For $c_2(x, y)$, the maximum value will be when the smaller region is part of the other, so normalising by dividing by the minimum of x and y is appropriate. For $c_3(x, y)$, it only has non zero diameter when NTPP(x, y) holds. The maximum value for $|c_3(x, y)|$ is $0.5 * (\delta(y) - d(x))$, ie when circles on which $\delta(x)$ and $\delta(y)$ are based are concentric. So this suggests normalising $|c_3(x, y)|$ by dividing by $0.5 * (\delta(y) - \delta(x))$. $c_4(x, y)$ is dual to $c_3(x, y)$, so the term should be divided by $0.5 * (\delta(x) - \delta(y))$. For the $|c_5(x, y)|$ term, the maximum value is $\delta(x)$ so this is the normalising factor, and dually for the $|c_6(x, y)|$ term, it should be divided by $\delta(y)$. This results in a revised, normalised distance measure:

$$d(x,y) = \frac{|c_1(x,y)|}{\delta(x \cup -x))} + \frac{|c_5(x,y)|}{\delta(x)} + \frac{|c_6(x,y)|}{\delta(y)} - \frac{|c_2(x,y)|}{min(\delta(x),\delta(y))} - \frac{|c_3(x,y)|}{0.5 * (\delta(x) - \delta(y))} - \frac{|c_4(x,y)|}{0.5 * (\delta(y) - \delta(x))} \tag{1}$$

The normalized measures for each of the circles c_1 to c_6, together with the normalized region based distance $d(x,y)$ is shown in figure 3. Below we refer to $d(x,y)$ as the Region Based Distance (RBD) between x and y.

3 Application to Handling Noise in Video

Recent work [19] [17] has shown the benefits of qualitative spatio-temporal relationships in representing and learning about human activities. In [19], activities are regarded as being composed of events, which are modelled as interactions between a set of objects in space and time. Events such as *unloading*, representing interactions between trolleys, planes and loaders, were learned from videos of aircraft activities.

Interactions were represented in terms of a graph structure that captures the temporal evolution of qualitative spatial relationships between the respective pairs of objects. These pairwise spatial relationships were computed from their tracks, where a track is simply a temporal sequence of minimum bounding rectangles (MBR) covering each object. However, visual noise and other errors introduced by video processing frequently result in instability in mereotopological relations over time, when these are defined point-set theoretically – i.e. 'jitter' can result as relations change frequently depending on the exact position and size of the MBR.

The following describes a solution to this problem using a HMM that overlays a temporal model in order to regularizing these rapidly flipping spatial relationships. The states of the HMM are labelled by the RCC8 relationships. The observations are a sequence of region based distances between the respective pairs of object MBRs. The probability distribution between the states and the observations are modelled by an *observation model* for each state. With a trained HMM, it is possible to predict the most likely sequence of spatial relationships given a sequence of observed RBDs.

The regularizing effect of the HMM is achieved by defining transition probabilities on RCC8 in such a way that it encourages objects to remain in the same state, while allowing transitions that are constrained by the connections in the RCC8 CNG. In other words, the HMM prevents rapidly flipping transitions by encouraging transitions to take place only when there is sufficient evidence from the observations, that is compelling enough to proceed to the next state. We show that good transition probabilities and observation model can be learnt automatically by training the HMM on manually annotated training videos. In this manner, the HMM learns a temporal model that can be regarded as an approximation of the way humans perceive these spatial transitions. The following describes the proposed approach more formally.

Optimal Spatial Sequence for a Pair of Tracks

1. Let $\tau = (..., o_t, ...), \tau' = (..., o'_t, ...)$ be a pair of tracks. It is assumed that all the corresponding MBRs $o_t \in \tau$ and $o'_t \in \tau'$ are observed together[6].
2. Let $D(\tau, \tau') = (..., d(o_t, o'_t), ...)$ be an observed sequence of RBDs between (τ, τ'). Here $d(o_t, o'_t)$ is the RBD (equation 1) between corresponding MBRs (o_t, o'_t) at time t.
3. Let $S(\tau, \tau') = (..., s(o_t, o'_t), ...)$ be a *hypothesized* sequence of qualitative spatial relationships between (τ, τ'). Here $s(o_t, o'_t) \in \Re$ is the hypothesized spatial relationship between the corresponding MBRs (o_t, o'_t) at time t.

The goal is to use a HMM model Θ to predict the most likely sequence of spatial relationships $\hat{S}(\tau, \tau')$, given a sequence of observed distances $D(\tau, \tau')$, between the tracks τ, τ'

$$\hat{S}(\tau, \tau') = \arg \max_{S(\tau,\tau')} P(S(\tau, \tau') | D(\tau, \tau'), \Theta)$$

A HMM to Obtain an Optimal Spatial Sequence

In order to address this problem, we formulate a HMM that models the joint probability distribution of the observed and hidden states as given by the tuple $\Theta = (\Re, A, B, \pi)$, where

1. $\Re$ are the states of the HMM. They correspond to the spatial states in RCC8.
2. $A = (a_{ij})_{ij}$ is the state transition matrix representing the probability $a_{ij} = P(st_{t+1} = s_j | st_t = s_i)$ of transition from state $st_t = s_i \in \Re$ to $st_{t+1} = s_j \in \Re$. Only those transitions that are physically possible, *as given by the CNG of the RCC8* have non-zero transition probabilities.
3. $B = (b_i(\delta_t))_{it}$ is the observation model, where $b_i(\delta_t)$ represents the probability $P(\delta_t | st_t = s_i)$ of observing a RBD δ_t while being in state $st_t = s_i \in \Re$. The observation models for each state $s_i \in \Re$ are modelled as normal distributions[7] $\mathcal{N}(\mu_i, \sigma_i^2)$. Here μ_i represents the mean RBD for the state s_i and σ_i^2 represents the variance of this distance for this state.
4. $\pi = (\pi_i)_i$ is the initial state distribution, where π_i represents the probability of state $st_t = s_i \in \Re$ being the initial state.

The above HMM model Θ is trained with a dataset of sequences of region based distances (between tracks) that are manually annotated with the subjectively correct spatial relations. The Baum-Welch algorithm [13] is used to learn the parameters of the model Θ of the HMM. For a given pair of co-temporal tracks, a Viterbi decoder [13] is used to find the most likely sequence of spatial relationships.

[6] We are interested in computing the spatial relationships only when the objects are observed together [19].

[7] While non-symmetric distributions may be better suited as observation models, the normal distribution offers the simplicity with respect to learning the parameters of the model. We have also experimentally observed that the results with the normal distribution closely approximates those arising from these non-symmetric distributions.

4 Experiments

The following paragraphs describe the two video datasets and the evaluation of the proposed approach on these two datasets.

Datasets

Two real video datasets are used to evaluate the proposed framework. The first consists of activities in an airport apron showing servicing of aircraft between flights. The second video dataset consists of activities representing simple verbs such as throw (a ball), catch etc. We will henceforth refer to these two datasets as the *airport apron dataset* and *person-ball verbs dataset* respectively.

The processing of the real data sets involved two stages: detection and tracking. For the first stage, a multi-class object detector [12] based on HOG features was trained on a separate part of the dataset and applied to each frame of the rest of the dataset. The trained classes for the aircraft apron datasets were (i) plane; (ii) trolley; (iii) loader; (iv) bridge; (v) plane-puller. The trained classes for the *person-ball verbs dataset* were person and ball respectively. The second stage involves applying our implementation of the tracking technique reported in [21] to the detected blobs. We chose this technique since it performs global optimization to obtain the most likely set of tracks.

Evaluation of the HMM based Procedure

The following experiment evaluates the proposed framework and compares it with the point set intersection technique, which is regarded as the baseline. In order to train and test the HMM, the tracked dataset is randomly divided into two parts: The first consisting of two thirds is used for training, and the remaining for testing. Ten such random partitions are created for evaluation. The training data is hand annotated by associating pairs of tracks in the training set with a corresponding sequence of spatial relationships[8]. These annotations are subjectively assigned by the annotators. A part of an annotated sequence is shown in figure 4.

Instead of labelling the entire data of several thousand frames, only those segments where there are changes in spatial relationships are considered for training and testing the HMM. This is because the main purpose of the HMM is to learn a stable transition between the spatial states, rather than parts where there is a high certainty of the spatial states.

For the airport apron dataset, a total of 27 training segments and 14 test segments were prepared. For the person-ball verbs dataset, a total of 10 training segments and 5 test segments were prepared. In these segments, those pairs of tracks for which the spatial relationships change are first identified. These pairs are subjectively labelled, with the appropriate spatial relationship for each frame,

[8] This is because the purpose of the HMM is to learns a mapping from a pair of tracks to a corresponding sequence of spatial relationships.

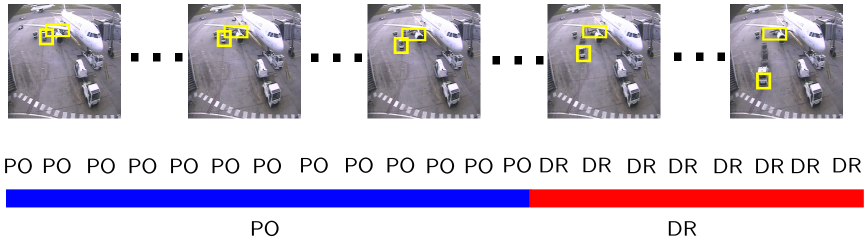
PO PO PO PO PO PO PO PO PO PO PO PO PO DR DR DR DR DR DR DR DR
PO
DR

Fig. 4. A segment with which the HMM based procedure for obtaining spatial relationships is trained and evaluated. Some images from a segment that has been manually annotated for training and evaluating the HMM are shown. An example of an annotation in the form of a sequence of spatial relationships is shown for a pair of tracks corresponding to a loader and trolley respectively.

in which both the tracks are observed. The segments are also provided with respective episodes[9]for this sequence of spatial relationships. One such segment is illustrated in figure 4.

The HMM is trained on the training segments for each random partition. The trained HMM is then applied on the test segments for the corresponding random partition. This gives rise to a sequence of spatial relationships between pairs of tracks on the test segments. A corresponding sequence of episodes are constructed from the inferred sequence of spatial relationships, for the sake of evaluation which is described below. Two such sequences for the verb *catch* and another event where the *trolley detaches from a loader* are illustrated in figure 6.

Qualitative Evaluation The performance of the HMM is evaluated qualitatively by examining the sequence of qualitative relationships obtained using the proposed approach and comparing it with the corresponding sequence obtained using the traditional point set intersection based computation. Figure 5 illustrates such a comparison, as explained in the corresponding caption. It can be seen that for this video sequence, the use of a HMM with the distance based semantics plays a significant role in eliminating noise arising from jitter of the bounding boxes. Other examples of correctly inferred spatial relationships for the airport and person-ball verbs datasets are shown in Figure 6.

Quantitative Evaluation The proposed approach can be quantitatively evaluated in two ways. The first involves evaluating the extent to which the HMM outputs a correct sequence of episodes. The accuracy is reported in terms of the mean and variance of the percentage of test segments for which the sequence

[9] Episodes are defined in [17] as sequence of spatial relationships such that within each episode the same spatial relation holds, but a different spatial relation holds immediately before and after the episode.

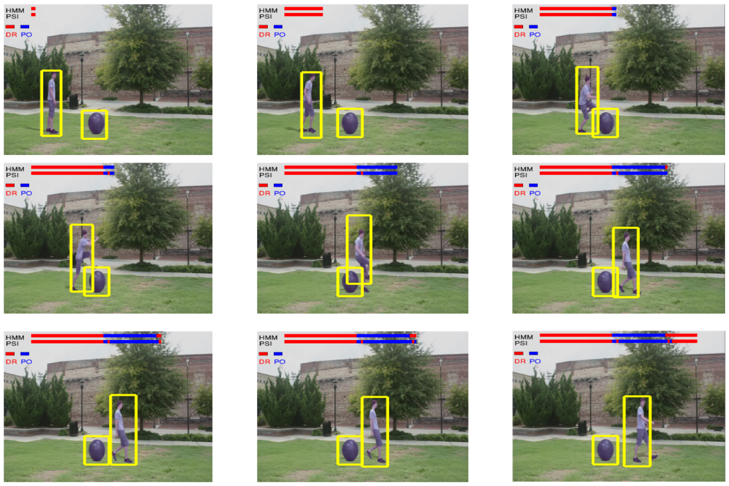

Fig. 5. Some images sampled from the video footage of a scene depicting a person jumping over a ball, with bounding boxes on them. At the top of each image are two bands showing the spatial relationships (PO in blue and DR in red) up until the time of the depicted frame. For this video sequence, it can be observed that the HMM based on the distance based semantics has eliminated noise arising from jitter of the bounding boxes. The noise is evident in the band corresponding to the traditional point set intersection (PSI) based approach. For example, in the middle frame in the bottom row, it can be seen that prior to this point, both the PSI and HMM have inferred a DRrelationship, but the PSI relationship *jitters* back to PO in this frame, whilst the HMM relationship is stable at DR.

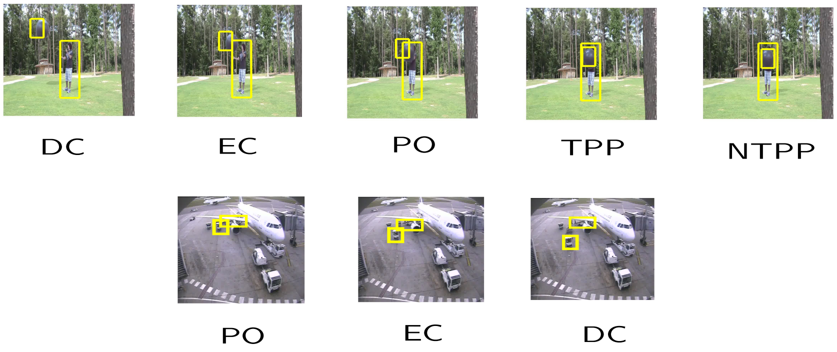

Fig. 6. Examples of correctly inferred spatial relationships for the airport and person-ball verbs datasets. Each image represents a sample from the sequence of images, corresponding to the interval, during which the spatial relationships given below hold. These spatial relationships have been inferred by the proposed approach. Note that the relation in the bottom row is PO according to PSI semantics, but is inferred as EC by the HMM.

of episodes exactly correspond to those segments in the ground truth, across the 10 random train-test partitions. These results for the proposed HMM based approach and the traditional point set intersection based approach for the two datasets are reported in Table 1.

	Aircraft Apron	Person-Ball Verbs
HMM	82.3%, 7.1%	90.6%, 6.2%
Point Set Intersection	30.8%, 13.2%	67.9%, 10.4%

Table 1. Results evaluating the extent to which the inferred sequence of episodes exactly correspond to those segments in the ground truth. Each entry consists of the mean and variance respectively.

	Aircraft Apron	Person-Ball Verbs
HMM	66.1%, 2.2%	81.1%, 2.0%
Point Set Intersection	27.8%, 11.2%	57.8%, 8.3%

Table 2. Results evaluating the extent to which the outputted episodes temporally align with those of the ground truth. Each entry consists of the mean and variance respectively.

The second evaluation involves evaluating the extent to which the outputted episodes temporally align with those of the ground truth. This evaluation is restricted only to those those segments whose sequence of episodes obtained from the HMM matches the ground truth. This is because the purpose is to understand the extent of deviation in temporal alignment, despite the fact that the episodes have been matched correctly. A good alignment ensures a reduced chance of structural difference in temporal relationships (amongst the episodes) between the ground truth and the output of the HMM. Accuracy is measured in terms of the mean and variance of the percentage of temporal overlap, between the outcome of the HMM and the ground truth, across the 10 random partitions. These results for the proposed HMM based approach and the traditional point set intersection based approach for the two datasets are reported in Table 2.

It can be concluded that the HMM significantly outperforms the traditional point intersection based technique. In particular, the potential advantage of using the HMM based approach using the RBDs on RCC8, for inducing stable sequence of qualitative spatial relationships from video data, has been demonstrated.

5 Summary

This paper explores this idea of a distance based interpretation for mereotopology. By introducing a distance measure between two regions x and y, and for various Boolean combinations of x and y, we show that all the RCC8 relations can be distinguished. We then introduce a distance measure which combines these individual measures which we show reflect different paths through the RCC8 conceptual neighbourhood (i.e. the measure decreases/increases monotonically given certain monotonic transitions, such as one region expanding). In contrast to traditional definitions of mereotopologies, in terms of point set intersections, our region based distance measures the degree to which the different mereotopological relations hold. We have demonstrated how a Hidden Markov Model can be used to exploit this distance based semantics to yield improved interpretation of video data at a qualitative level.

6 Limitations and Future Work

There are a number of avenues of further work which might be fruitfully explored. For simplicity in this work we limited the regions considered to bounding boxes aligned to orthogonal axes. If this assumption is relaxed then the RBD measure can fail to work as expected. Figure 7 illustrates three such regions, where c_5 remains constant as the orange shape translates from left to right. One possible solution that we are currently investigating is to formulate a semantics that is based on a separate analysis of projections of a region along the horizontal and vertical coordinate axes respectively.

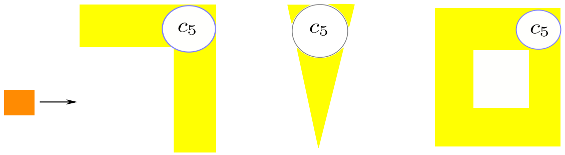

Fig. 7. Examples of regions where the proposed distance measure between regions can fail, as the smaller orange shape translates left to right. All though c_2 and c_4 will change as expected, c_5 will remain constant!

Again for simplicity, in this paper we only considered computing the most probably RCC8 relation for each pair regions considered in isolation. In general, there will be multiple objects and it is necessary to ensure that their spatial relationships are globally consistent. For example, consider the relations between three regions x, y, z and the three relations $R_1(x, y), R_2(y, z), R_3(x, z)$. The proposed approach does not ensure *path consistency* (i.e. the most probable interpretation of $R_1(x, y)$, $R_2(x, y)$ and $R_3(x, y)$ might not be mutually consistent e.g. $R_1 = R_2 = \mathsf{TPP}$, whilst $R_3 = \mathsf{TPPi}$). Thus an interesting possibility for future work is to couple HMMs between each pair of objects and introduce constraints between them. It would be worthwhile to evaluate evaluate whether such a coupling improves performance in abstracting stable qualitative spatial relations between multiple objects. Am alternative way of ensuring a globally consistent set of spatial relations is to check for this property on the most probable set of spatial relations, and if inconsistency is detected, then to choose the most probable set with this property.

Another direction for future work is to carefully analyse all the different processes over the conceptual neighbourhood graph to be found in [4] in terms of the RBD measure defined here, and to build a system which can recognise processes reliably from video data. Another avenue of research is to explore different formulations of the RBD from the six individual circle metrics: should they be weighted (are some not relevant/useful) – it might appear that c_2 is less useful since c_5 and c_6 capture much of what c_2 does, except in very particular cases of transition.

It will also be interesting to investigate if the HMM can be used to learn topological relationships[7] and to what extent these learnt relationships are qualitatively interesting or how they compare to qualitatively interesting topologies such as RCC8, when applied to a video event analysis tasks [19]. Mereotopologies such as RCC8 are not the only kind of qualitative spatial calculi, and another avenue of research would be to use HMMs to learn when to transition between the relations of other qualitative spatial calculi other than mereotopologies. Finally we note that if a probabilistic approach to QSR is desired, whereby for example, RCC8 relations have probabilities attached to them (for use in stochastic logic programming), then the HMM could provide such probabilities.

References

1. Cohn, A.G., Gotts, N.M.: Representing spatial vagueness: a mereological approach. In: Proceedings of the 5th Conference on Principles of Knowledge Representation and Reasoning (KR) (1996)
2. Cohn, A.G., Varzi, A.: Mereotopological connection. Journal of Philosophical Logic 32, 357–390 (2003)
3. Egenhofer, M.J.: Reasoning about binary topological relations. In: Proceedings of the Second International Symposium on Advances in Spatial Databases (SSD). Springer-Verlag (1991)
4. Egenhofer, M.J., Al-Taha, K.K.: Reasoning about gradual changes of topological relationships. In: Theory and Methods of Spatio-Temporal Reasoning in Geographic Space. Lecture Notes in Computer Science, vol. 639, pp. 196–219. Springer-Verlag
5. Egenhofer, M.J., Dube, M.P.: Topological relations from metric refinements. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (2009)
6. Egenhofer, M.J.: A formal definition of binary topological relationships. In: 3rd International Conference, Foundations of Data Organization and Algorithms (FODO). Springer-Verlag (1989)

7. Galata, A., Cohn, A.G., Magee, D.R., Hogg, D.C.: Modeling interaction using learnt qualitative spatio-temporal relations and variable length Markov models. In: Proceedings of the European Conference on Artifical Intelligence (ECAI) (2002)
8. Galton, A.: Towards a qualitative theory of movement. In: COSIT. Lecture Notes in Computer Science, vol. 988. Springer-Verlag (1995)
9. Galton, A.: Modes of overlap. Journal of Visual Languages and Computing 9, 61–79 (1998)
10. de Laguna, T.: Point, line and surface as sets of solids. The Journal of Philosophy 19, 449–461 (1922)
11. Li, S., Cohn, A.G.: Reasoning with topological and directional spatial information. Computational Intelligence p. to appear
12. Ott, P., Everingham, M.: Implicit color segmentation features for pedestrian and object detection. In: Proceedings of the International Conference on Computer Vision (ICCV) (2009)
13. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings IEEE pp. 257–286 (1989)
14. Randell, D., Cui, Z., Cohn, A.: A spatial logic based on regions and connection. In: Proc. 3rd Int. Conf. on Knowledge Representation and Reasoning (1992)
15. Renz, J.: A canonical model of the region connection calculus. In: Journal of Applied Non-Classical Logics (JANCL). vol. 12, pp. 469–494 (2002)
16. Schockaert, S., Cock, M.D., Kerre, E.E.: Spatial reasoning in a fuzzy region connection calculus. Artif. Intell. 173(2), 258–298 (2009)
17. Sridhar, M., Cohn, A.G., Hogg, D.C.: Learning functional object-categories from a relational spatio-temporal representation. In: Proceedings of the European Conference on Artifical Intelligence (ECAI) (2008)
18. Sridhar, M., Cohn, A.G., Hogg, D.C.: Discovering an event taxonomy from video using qualitative spatio-temporal graphs. In: Proceedings of the European Conference on Artifical Intelligence (ECAI). IOS Press (2010)
19. Sridhar, M., Cohn, A.G., Hogg, D.C.: Unsupervised learning of event classes from video. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). pp. 1631–1638. AAAI (2010)
20. Welch, G., Bishop, G.: An introduction to the Kalman filter. Tech. rep., University of North Carolina at Chapel Hill (1995)
21. Yu, Q., Medioni, G.: Integrated detection and tracking for multiple moving objects using data-driven MCMC data association. In: IEEE Workshop on Motion and Video Computing (2008)