Available online at www.sciencedirect.com
ScienceDirect
Image and Vision Computing 26 (2008) 67–81
image AND vision COMPUTING
www.elsevier.com/locate/imavis

# Enhanced tracking and recognition of moving objects by reasoning about spatio-temporal continuity

Brandon Bennett, Derek R. Magee, Anthony G. Cohn *, David C. Hogg

*School of Computing, University of Leeds, Leeds LS2 9JT, UK*

Received 16 July 2004; received in revised form 29 July 2005; accepted 15 August 2005

**Abstract**

A framework for the logical and statistical analysis and annotation of dynamic scenes containing occlusion and other uncertainties is presented. This framework consists of three elements; an object tracker module, an object recognition/classification module and a logical consistency, ambiguity and error reasoning engine. The principle behind the object tracker and object recognition modules is to reduce error by increasing ambiguity (by merging objects in close proximity and presenting multiple hypotheses). The reasoning engine deals with error, ambiguity and occlusion in a unified framework to produce a hypothesis that satisfies fundamental constraints on the spatio-temporal continuity of objects. Our algorithm finds a globally consistent model of an extended video sequence that is maximally supported by a voting function based on the output of a statistical classifier. The system results in an annotation that is significantly more accurate than what would be obtained by by-frame evaluation of the classifier output. The framework has been implemented and applied successfully to the analysis of team sports with a single camera.


*Keywords:* Visual surveillance; Spatial reasoning; Temporal reasoning; Resolving ambiguity; Continuity

## 1. Introduction

No computer vision algorithm for tracking or object recognition is perfect under real-world operating conditions. Object trackers have difficulty with complex occlusions (e.g. in crowded pedestrian scenes or on the sports field) and object recognition algorithms rarely give 100% accuracy, even on well posed data sets, let alone under unconstrained circumstances. This lack of reliability is one of the reasons for the slow commercial uptake of visual surveillance systems based on object tracking. In this paper, we propose a framework for enhancing the imperfect output of an object tracker by enforcing principles of logical consistency and spatio-temporal continuity of physical objects. This results in a scene annotation that is far more accurate than the raw output from the tracker.

Our tracker processes a video recording of a dynamic situation, taken with a single fixed camera (i.e. a sequence of 2D images). Using statistical techniques, the tracker detects and classifies moving objects in the scene. The tracking and recognition systems explicitly model the possibility of ambiguity and error by assigning probabilities for the presence of objects within bounding boxes in each video frame. However, the tracker output is unreliable in that: (a) the object detected with the highest probability may not actually be present in the box; (b) there may be multiple overlapping or occluding objects within any box, and the tracker output does not tell us how many objects are present.

This imperfect tracker output is passed to a reasoning engine which constructs a ranked set of possible labellings for the whole video sequence, that are consistent with the requirements of object continuity. The final output is then a globally consistent spatio-temporal description of the scene which is maximally supported by probabilistic information given by the classifiers.

A number of researchers have attempted to deal with object occlusion (and the resultant tracking problems) by attempting to track through occlusion. This can involve reasoning about object ordering along the camera optical axis,

* Corresponding author.
*E-mail address:* agc@comp.leeds.ac.uk (A.G. Cohn).

0262-8856/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2005.08.012

either using ground-plane information [1–3] or simply reasoning about relative spatial ordering [4]. 3D [3] or 2D (planar) [1] object models may be used with a known camera-to-ground-plane transformation to identify occlusion for a given set of object configurations. This can then be used to exclude subsets of image information from the model fitting process. This works well for 'hypothesise-and-test' type tracking (e.g. [2]). However, complete occlusion leads to zero information from the image, and weak constraints on object position/configuration. [5] takes a conservative approach of not tracking in uncertain situations, such as object occlusion. The ends of broken tracks are then joined based on spatio-temporal similarity measure.

Another approach has been to use multiple cameras in an attempt to circumvent the occlusion problem. In [6], multiple football players are tracked using eight cameras. Each view is tracked separately and overlapping/occluding objects are associated with the same blob (as in our system). Each blob from each camera is projected to the ground-plane (using a camera calibration), and associated with one, or more, players using a 'closed-world' assumption, and a stochastic constraint optimisation procedure.

Dynamic models such as the Kalman filter are often used to model the position of occluded objects [7,8], under the assumption of known dynamics (e.g. linear motion), when no visual information is available. In [9], the authors use both motion models and 2D dynamic appearance models to track through occlusion. This work is interesting as it is a composite of a moving region ('blob') segmentation/extraction algorithm (that has no explicit notion of occlusion), and a model based tracker, that can track through partial and (short-term) complete occlusion using motion and appearance models.

The success (or otherwise) of tracking through occlusion with motion and appearance models depends on a number of factors: the degree to which the models accurately model object dynamics and appearance, the time over which the occlusion occurs, the complexity of object behaviour during occlusion, the degree to which objects are occluded (partially or completely), and the similarity (or otherwise) of the occluding objects. Our proposed method (see later) is independent of all these factors (except object similarity). Multiple cameras have also been used to bypass the occlusion problem [10,11]. However, this is not always possible or practicable, and it is not necessarily a complete solution.

Our approach to occlusion handling differs from this body of work and has more similarity with the methods of McKenna et al. [12] and Sherrah and Gong [13]. These works do not attempt to disambiguate occluding objects, but instead reason about the occlusion taking place. McKenna et al. track 'blobs' that may be groups or individuals. In their work, it is initially assumed all objects are separate (an assumption we do not make); and when blobs merge, the resultant blob is recorded as a 'group' made up of the contributing individuals. A dynamically updated model of global object colour is used to disambiguate objects at the point at which blobs split. This model is also used to reason about object occlusion within a group that makes up a single blob. This is useful when a split group consists of more than two individuals; however, it relies on an assumption that no object is completely occluded during the split. Sherrah and Gong [13] present work in a highly constrained scenario where the head and hands of a single individual are tracked as blobs. The hands may occlude each other or the face (to form a single blob). A hand-built Bayesian network is used to perform frame-by-frame occlusion reasoning, based on available data (blob positions, velocities, number of blobs, etc.). Perhaps the closest work to ours was presented recently by Yang et al. [14]. This system uses multiple cameras to provide a top view of the 'visual hull' of a crowd scene. Constraints on the number of pedestrians represented by each observed blob are determined according to the size of the blob's bounding box. These constraints are propagated from frame to frame to give an upper and lower limit on the number of objects present. All observed moving objects are assumed to be pedestrians, and no attempt is made to localise or identify individual pedestrians. Lipton et al. [15] present a system that uses simple object classification (pedestrian vs. car) to aid object tracking. Simple temporal consistency rules are used to prune transient objects resulting from noise. None of these systems performs more than frame-by-frame reasoning or allows for the possibility of error in the underlying low-level tracking and recognition algorithms. Our system performs long-term reasoning about object-blob associations over extended sequences of frames. By maintaining spatio-temporal consistency over sequences, many local imperfections and ambiguities in the low-level data are eliminated.

Our system is based on 'blob tracking' and object classification methods. Much prior work has been presented in these areas. 'Blob tracking' may be described as tracking with a weak object model. This has the advantage that trackers based on this approach (e.g. [9,16–18]) are able to track a wide range of objects. Such trackers are often based on background modelling methods (e.g. [9,16,19]), which are used to extract foreground regions that are associated over time. The disadvantage is a lower grade of information is obtained; typically only position, scale and 2D shape are extracted. In contrast, model-based trackers may extract pose [3], posture [20] or identity information [21], for a particular (known) class of objects only. The acquisition of good models for model based tracking is also a non-trivial problem.

The extraction of a wider range of parameters from 'blob tracker' output may be seen as a separate post-processing operation. In [22], neural networks are used to classify vehicle type (small, medium, or large) based on edge information in the region around an object detection. In [23], a $k$-nearest neighbours classifier based on vector quantisation (similar to that used in our work) is used on various simple object features (height/width ratio, area, etc.) to classify tracked objects as pedestrian or bicycle. In [24], the output of a blob tracker is classified in an unsupervised manner into various categories (representing colour, and texture, etc.). From this temporal protocols are learnt. In [12], histograms of foreground pixel colour are used to parameterise appearance and recognise individuals at a later time, using histogram intersection. We also use the colour histogram approach. However, any of the classification approaches described (or others) could be used equally well