An Efficient Implementation of Reid's Multiple Hypothesis Tracking Algorithm and Its Evaluation for the Purpose of Visual Tracking

Ingemar J. Cox and Sunita L. Hingorani

Abstract—An efficient implementation of Reid's multiple hypothesis tracking (MHT) algorithm is presented in which the k-best hypotheses are determined in polynomial time using an algorithm due to Murty [24]. The MHT algorithm is then applied to several motion sequences. The MHT capabilities of track initiation, termination, and continuation are demonstrated together with the latter's capability to provide low level support of temporary occlusion of tracks. Between 50 and 150 corner features are simultaneously tracked in the image plane over a sequence of up to 51 frames. Each corner is tracked using a simple linear Kalman filter and any data association uncertainty is resolved by the MHT. Kalman filter parameter estimation is discussed, and experimental results show that the algorithm is robust to errors in the motion model. An investigation of the performance of the algorithm as a function of look-ahead (tree depth) indicates that high accuracy can be obtained for tree depths as shallow as three. Experimental results suggest that a real-time MHT solution to the motion correspondence problem is possible for certain classes of scenes.

Index Terms—Multiple hypothesis tracking, motion correspondence, data association, tracking, visual tracking, ranked bipartite graph matching.

1 Introduction[1]

THE analysis of image sequences for purposes of estimating camera motion and/or 3-D scene geometry often requires the tracking of geometric features over long image sequences. Typically, predictions are first made as to the expected locations of the current set of features of interest. These predictions are then matched to actual measurements. At this stage, ambiguities may arise. Predictions may not be supported by measurements—have these objects ceased to exist or were they simply occluded? There may be unexpected measurements—do these measurements originate from newly visible objects or are they spurious readings from noisy sensors? More than one measurement may match a predicted feature—which measurement is the correct one and what is the origin of the other measurements? Or a single measurement may match to more than one feature—which feature should the measurement be assigned to? These ambiguities must be resolved in order to solve the motion correspondence problem.

Visual tracking has been extensively studied in recent years. However, almost all such work has assumed that the motion correspondence problem has been solved or is trivial so that a nearest neighbor strategy is effective. In some cases, a nearest neighbor strategy is indeed adequate. For example, Tomasi and Kanade [30]. track corner features over very many frames using such an approach. A nearest neighbor strategy usually relies on the frame-to-frame image motion being extremely small. Much more data must then be processed than if a sparser sampling were used. However, if significant frame to frame motions are present, then ambiguities can quickly arise. Zheng and Chellappa [34] minimize these ambiguities by using a weighted correlation window to detected tracked features in the next frame. While correlation techniques can significantly reduce the motion correspondence ambiguity, our experiments suggest that partial occlusion and significant changes in the background can be problematic for such methods. Moreover, such techniques are only appropriate to the detection of measurements from *existing* tracking, not for the detection of *new* tracks. Many researchers have used the Kalman filter to track geometric features such as lines [1], [17] and corners [5], [4] in a scene, under the assumption that motion correspondence is straightforward. The motivation and significance of this work was in designing stable and reliable algorithms to infer the 3-D structure and motion from 2-D image plane measurements. Shapiro et al. [28] describe tracking corners in the image plane. Their system has several similarities to the one described herein, specifically, the use of Kalman filtering and a cross correlation measure to compare corners. However, the motion correspondence problem is not rigorously addressed: correspondences are determined between two consecutive frames based on a similarity measure between corners. Correspondences are determined without looking at subsequent frames and there is no mechanism for dealing with ambiguous motion correspondences.

[1] Portions of Sections 1 and 2 are taken from [8] and are reprinted courtesy of Kluwer Academic Publishers.

- *I.J. Cox is with NEC Research Institute, 4 Independence Way, Princeton, NJ 08540. E-mail: ingemar@research.nj.nec.com.*
- *S.L. Hingorani is with AT&T Bell Laboratories, 184 Liberty Corner Road, Warren, NJ 07059. E-mail: sunita@cartoon.lc.att.com.*

Manuscript received Aug. 25, 1994; revised Aug. 18, 1995.
Recommended for acceptance by: A. Singh.
For information on obtaining reprints of this article, please send e-mail to: transactions@computer.org, and reference IEEECS Log Number P95149.

0162-8828/96$05.00 ©1996 IEEE

The target tracking and surveillance community has extensively studied the motion correspondence problem [2] and a number of statistical data association techniques have been developed. These algorithms are now receiving wider attention, especially within the computer vision community [8]. For example, Chang and Aggarwal [6] have applied the joint probabilistic data association (JPDA) filter [18] to the problem of 3-D structure reconstruction from an ego motion sequence. However, the JPDA is only appropriate if the number of tracks is known a priori and remains fixed throughout the motion sequence. Zhang and Faugeras [32] have used the track splitting filter of Smith and Buechler [29] for dynamic motion analysis. The track splitting filter is similar to multiple hypothesis tracking in its use of track trees to delay correspondence decisions until more evidence is available. However, the track splitting filter allows measurements to be shared between tracks. This is physically unrealistic. More reasonable, is that a measurement originates from only a single source feature, e.g., a single measurement might originate from either a wall or corner feature but not from both. The motion correspondence now becomes one of partitioning measurements into *disjoint* tracks (or sets). Disjointness is also a common constraint in human vision where in stereo correspondence it is called uniqueness [22] and in motion correspondence it is called the element integrity principle [16]. It may also be reasonable to assume that a geometric feature gives rise to only a single measurement vector within a time frame. The track splitting algorithm cannot cope with these constraints and it is necessary to use an MHT approach. Moreover, one is unable to develop and efficient implementation, as discussed in Section 2.3, without the disjointness constraint.

This paper describes an efficient implementation of the multiple hypothesis tracking (MHT) algorithm originally proposed by Reid [27] and evaluates its usefulness in the context of visual tracking and motion correspondence. Our interest in the MHT is motivated by the fact that the MHT is the only statistical data association algorithm that integrates all the capabilities of

1) *Track Initiation*. The automatic creation of new tracks as new geometric features enter the field of view.
2) *Track Termination*. The automatic termination of a track when the geometric feature is no longer visible for an extended period of time
3) *Track Continuation*. The continuation of a track over several frames in the absence of measurements. Thus, the algorithm is capable of providing a level of support for temporary occlusion.
4) *Explicit Modeling of Spurious Measurements*.
5) *Explicit Modeling of Uniqueness Constraints*. A measurement may only be assigned to a single track and a track may only be the source of a single measurement per frame.

The multiple hypothesis tracking (MHT) algorithm is outlined in Section 2. Unfortunately, the MHT algorithm is computationally exponential both in time and memory. An approximation to the algorithm must therefore be implemented. Section 2.3 describes an efficient approximation to the MHT algorithm, the key contribution being the use of an algorithm due to Murty [24] to generate directly the k-best hypotheses in polynomial time [12] without explicitly enumerating all possible hypotheses. This is a significant contribution to the practical application of the MHT methodology which has recently been shown to be approximately three orders of magnitude faster than previous hypothesis generation strategies [13].

Section 3 then describes experimental results on three motion sequences. In each motion sequence, corner features are automatically detected using a variant of the Lucas and Kanade corner detector [21]. The MHT then tracks these corners over the sequence of frames. Each corner is tracked in the image plane using a simple linear Kalman filter. Section 3.4 demonstrates that the algorithm is robust to errors in the motion model. The most significant experimental problem encountered was that of track initiation during the first two or three frames of the sequence. Section 3.2 describes the approach used to reduce this problem. Section 3.4.1 investigates how the performance of the MHT varies as a function of the depth of the hypothesis tree. Finally, Section 4 summarizes the experimental results and suggests several promising lines of future work.

2 Multiple Hypothesis Algorithm

The multiple hypothesis tracking algorithm was originally developed by Reid [27] in the context of multi-target tracking. Recently, Cox and Leonard [9], [10][2] demonstrated its utility in the context of building and maintaining a map of a mobile robot's environment using acoustic sensors. Fig. 1 outlines the basic operation of the MHT algorithm. An iteration begins with the set of current hypotheses from iteration $(k - 1)$. Each hypothesis represents a different set of assignments of measurements to features, i.e., it is a collection of *disjoint tracks*. A track is defined to be a sequence of measurements that are assumed to originate from the same geometric feature. A dummy track in each global hypothesis denotes spurious measurements.

Different sets of assignments expect to see different sets of measurements. Thus, each hypothesis predicts the location (in the image plane) of a set of expected geometric features (specifically corners) and these are compared with actual measurements detected in the next camera frame on the basis of their Mahalanobis distance.[3] These comparisons are represented in the form of an *ambiguity matrix*,[4] defined

[2] The interested reader is also directed to Cox et al. [14] who applied the MHT to the problem of contour grouping and segmentation.

[3] For normally distributed measurements, the Mahalanobis distance is chi-squared distributed with number of degrees of freedom equal to the dimension n_z of the measurement vector. The probability that the distance is less than the parameter γ can, therefore, be obtained from χ^2 distribution tables. For example, if the measurement vector is two dimensional, $n_z = 2$, and a validation or search volume is to be established in which there is a 95% probability of finding the measurement, i.e., $P(\mathbf{z}(k+1) \in \tilde{V}(\gamma)) = 0.95$, then γ is set to $\gamma = 5.99$. Conversely, if a measurement fails the inequality test then there is a 5% or less chance that it is associated with the geometric feature.

[4] The ambiguity matrix is more often referred to as a hypothesis matrix. However, we feel that this is somewhat confusing since many hypotheses can be generated from a single (ambiguity) matrix.