MAP ZDF Segmentation and Tracking using Active Stereo Vision: Hand Tracking Case Study

Andrew Dankers [a,1] Nick Barnes [a,1] Alex Zelinsky [b]

[a] *National ICT Australia, Locked Bag 8001, Canberra ACT Australia 2601*
[b] *CSIRO ICT Centre, Canberra ACT Australia 0200*

Abstract

A *maximum a posterior probability zero disparity filter* (MAP ZDF) ensures coordinated stereo fixation upon an arbitrarily moving, rotating, re-configuring hand, performing marker-less pixel-wise segmentation of the hand. Active stereo fixation permits real-time foveal hand tracking and segmentation over a large visual workspace, allowing investigation of unrestricted natural human gesturing. Hand segmentation is shown to be robust to lighting conditions, defocus, hand colour variation, foreground and background clutter including non-tracked hands, and partial or gross occlusions including those due to non-tracked hands. The system operates at approximately $27fps$ on a $3GHz$ single processor PC.

Key words: Active Stereo Vision, Zero Disparity Filter, Markov Random Field, Hand Segmentation and Tracking, Human-Computer Interaction
PACS: 01.30.-y

1 Introduction

Humans interact with each other efficiently using mutually understood words, gestures and actions. Intelligent artificial systems that gather information from

Email addresses: andrew.dankers@nicta.com.au (Andrew Dankers), nick.barnes@nicta.com.au (Nick Barnes), alex.zelinsky@csiro.au (Alex Zelinsky).

[1] National ICT Australia is funded by the Australian Department of Communications, Information Technology and the Arts and the Australian Research Council through *Backing Australia's ability* and the ICT Centre of Excellence Program.

the observation of a human subject can facilitate natural, intuitive and efficient *human-computer interactions* (HCI). They can also impact positively on the versatility and acceptance of the system amongst users. HCI systems can be used to understand gestured or spoken instructions and to direct attention intelligently. They may advance the development of intuitive interfaces and automated systems that interact with humans for task-oriented or assistive behaviors.

HCI systems have countless applications. In particular, systems that focus on non-invasive, marker-less hand gesture recognition form an important branch of visual HCI systems that are changing the way we communicate with computers. For example, stereo sensing and 3D model fitting have been combined to achieve visual gesture interfaces for virtual environment interactions [29]. Nevertheless, the pervasiveness of computer vision methods in the field has often been hindered by the lack of real-time, robust algorithms. Limitations in visual workspace size have also meant that observed human subjects must deliberately confine their hand motion, such that natural gesturing may be compromised.

We focus on robust, real-time localisation and segmentation of hands during natural gesturing to facilitate gesture recognition for use with HCI systems. Actual gesture discrimination is beyond the scope of this paper.

1.1 Existing Methods

When tracking objects such as hands under real-world conditions, three main problems are encountered: ambiguity, occlusion and motion discontinuity. Ambiguities arise due to distracting noise, mismatching of the tracked objects and the potential for multiple hands, or hand-like distractors, to overlap the tracked target. Occlusions are inevitable in realistic scenarios where the subject interacts with the environment. Certainly, in dynamic scenes, the line of site between the cameras and target is not always guaranteed. At usual frame rates ($30fps$), the motion of dexterous subjects such as a hand can seem erratic or discontinuous and motion models designed for tracking such subjects may be inadequate.

Existing methods for marker-less visual hand tracking can be categorised according to the measurements and models they incorporate [14]. Regardless of the approach, hand gesture recognition usually requires a final verification step to match a model to observations of the scene.

1.1.1 Cue-Based Methods

In terms of cue measurement methods, tracking usually relies on either intensity information such as edges [15,6,27,9], skin colour, and/or motion segmentation [42,21,17,26], or a combination of these with other cues [24,38,39] or depth information [19,42,30,3]. Fusion of cues at low levels of processing can be premature and may cause loss of information if image context is not taken into account. For example, motion information may occur only at the edges of a moving object, making the fused information sparse. Further, for non-spatial cue-based methods, occlusions from other body parts – such as the face or another hand – may become indistinguishable from the tracked hand.

Mean Shift and *Cam Shift* methods are enhanced manifestations of cue measurement techniques that rely on colour chrominance based tracking. For real-time performance, a single channel (chrominance) is usually considered in the color model. This heuristic is based on the assumption that skin has a uniform chrominance. Such trackers compute the probability that any given pixel value corresponds to the target color. Difficulty arises where the assumption of a single chrominance cannot be made. In particular, the algorithms may fail to track multi-hued objects or objects where chrominance alone cannot allow the object to be distinguished from the background, or other objects.

The *Mean Shift* algorithm is a non-parametric technique that ascends the gradient of a probability distribution to find the mode of the distribution [13,10]. Particle filtering based on color distributions and Mean Shift was pioneered by Isard and Blake [18] and extended by Nummiaro et al. [28]. *Cam Shift* was initially devised to perform efficient head and face tracking [8]. It is based on an adaptation of Mean Shift where the mode of the probability distribution is determined by iterating in the direction of maximum increase in probability density. The primary difference between the Cam Shift and the Mean Shift algorithms is that Cam Shift uses continuously adaptive probability distributions (recomputed each frame) while Mean Shift is based on static distributions. More recently, Shen has developed *Annealed Mean Shift* to counter the tendency for Mean Shift trackers to settle at local rather than global maxima [37].

Although very successful in tracking the vicinity of a known chrominance, *Shift* methods are not designed for direct target segmentation and background removal (for classification enhancement). In terms of output, they provide an estimation of a tracked target bounding box, in the form of an estimate of the 0th and 1st moments of the target probability distribution function. They are also not typically capable of dealing with instantaneous or unexpected changes in the target colour model (such as, for example, when a hand grasps another object such as a mug or pen). They do not incorporate spatial constraints when considering a target in a 3D scene, and are not inherently intended to deal with occlusions and other ambiguous tracking cases (for example, a tracked target passing in front of a visually similar distractor). In such circumstances, these trackers may shift between alternate subjects, focus on the center of gravity of the two subjects, or track the distracting object rather than the intended target. To alleviate this, motion models and classifiers can be incorporated, but they may rely upon weak and restrictive assumptions regarding target motion and appearance.

1.1.2 Spatiotemporal Methods

Spatial techniques use depth information and/or temporal dynamic models to overcome the occlusion problem [42,21]. The use of spatial (depth) information can introduce problems associated with multiple camera calibration, and depth data is notoriously sparse, computationally expensive, and can be inaccurate. Spatiotemporal continuity is not always a strong assumption in terms of hand tracking models. At frame rates, hand motion may appear discontinuous since the hands can move quickly and seemingly erratically, or undergo occlusion by other body parts. Methods such as Kalman filtered hand tracking [20] that are strongly reliant upon well-defined dynamics and temporal continuity may prove inadequate. Traditional segment-then-track (exhaustive search methods, e.g. dynamic template matching) approaches are subject to cumulative errors where inaccuracies in segmentation affect tracking quality, which in turn affect subsequent segmentations.

Hands are part of an articulated entity (the human body), so model-based methods incorporating domain knowledge can be used to resolve some of the ambiguities. Joint tracking of body parts can be performed with an exclusion principle on observations [33,25] to alleviate such problems. *A priori* knowledge such as 2D hand models may be used [26,17]. Alternatively, a 3D model of the hand and/or body may be used such that skeletal constraints can be exploited [42,30,9]. 2D projections of deformable 3D models can be matched to observed camera images [15,27]. Unfortunately, these methods can be computationally expensive, do not always resolve projection ambiguities, and performance depends heavily upon the accuracy of complex, subject dependent, articulated models and permitted motions.

1.1.3 Zero Disparity Methods

Methods exist that do not require *a priori* models or target knowledge. Instead, the target is segmented using an un-calibrated semi-spatial response by detecting regions in images or cue maps that appear at the same pixel coordinates in the left and right stereo pairs. That is, regions that are at *zero disparity*[2] . To overcome pixel matching errors associated with gain differences between left and right views, these methods traditionally attempt to align vertical edges and/or feature points.

Rougeaux [34,35] investigated the use of *virtual horopters*[3] to test whether the tracked subject was moving away from or towards the cameras. One of the stereo pair images (e.g. the left image) was *virtually* shifted horizontally by a single pixel to the left (by purging the leftmost column of pixels) and then to the right (by adding an extra column at the left of the image), and the zero disparity response determined between each new image and the unaltered (right) image, for both cases. The virtual shift that yields largest zero disparity response area was deemed the correct tracking direction, and the cameras were then verged or diverged accordingly such that the horopter best aligned with the tracked subject.

Oshiro applied a similar edge extraction method to foveal log-polar cameras [31]. Yu used a wavelet representation to match broader image regions [43]. Rougeaux later revisited the approach, combining the edge-based ZDF with optical flow for broader segmentation [36]. Rae combined edge-based techniques with additional aligned point features such as corners, symmetry points and cue centroids [32].

Unfortunately, these methods do not cope well with bland subjects or backgrounds, and perform best when matching textured sites and features on textured backgrounds. The zero disparity class of segmentation forms the base upon which we develop our approach.

1.2 Overview

We aim to ensure coordinated active stereo fixation upon a hand target, and to facilitate its robust pixel-wise extraction. We propose a biologically inspired, conceptually simple method that segments and tracks the subject in parallel, eliminating problems associated with the separation of segmentation and tracking. The method inherently incorporates spatial considerations to disambiguate between, for example, multiple overlapping hands in the scene such that occlusions or distractions induced by non-tracked hands do not affect tracking of the selected hand. As we shall see, the method does not rely on imposing motion models on the commonly erratic trajectory of a hand, and

[2] A scene point is at *zero disparity* if it exists at the same image frame coordinates in the left and right images

[3] The *horopter* is the locus of zero disparity scene points that would project to identical left and right image coordinates if that scene point was occupied by a visual surface.

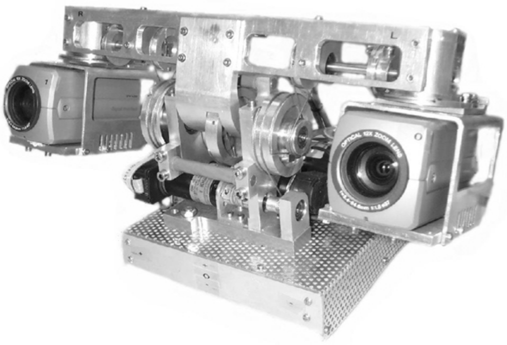

Fig. 1. CeDAR, active vision apparatus.

can cope with gross partial occlusions. In this regard, the three common problems of ambiguity, occlusion and motion discontinuity are addressed. Despite using stereo vision, the approach does not require stereo camera calibrations, intrinsic or extrinsic. The method utilises dynamic stereo foveal scene analysis, and we choose an active implementation that has the benefit of increasing the volume of the visual workspace.

We proceed by introducing the visual apparatus (Section 2). We motivate the active approach, in consideration of biological influences (Section 3). We present hand segmentation under the assumption that the hand is at the stereo fixation point. We formalise the approach as an energy minimisation problem and present an optimisation method used to solve the segmentation problem (Section 4), addressing issues of segmentation robustness. The hand tracking algorithm is then presented (Section 5). We present results (Section 6), and tracking performance and quality are evaluated (Section 7), including presentation of our results alongside those of other techniques for comparison. We finish with a brief discussion (Section 8) and conclusion (Section 9).

2 Platform

CeDAR (Fig. 1), the Cable-Drive Active-Vision Robot [40], is the experimental apparatus. It incorporates a common tilt axis and two independent pan axes separated by a baseline of $30cm$. All axes exhibit a range of motion of greater than 90^o, speed of greater than $600^o s^{-1}$ and angular resolution of 0.01^o. Syncronised images with a field of view of 45^o are obtained from each camera at $30fps$ at a resolution of 640x480 pixels. Images are down-sampled to 320x240 resolution before hand tracking processing occurs.

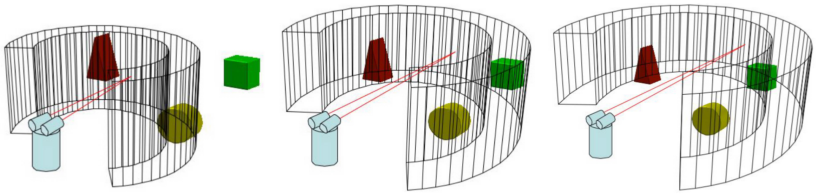

Fig. 2. Scanning the horopter over the scene: the locus of zero disparity points defines a plane known as the horopter. For a given camera geometry, searching for pixel matches between left and right stereo images over a small disparity range defines a volume about the horopter. By varying the geometry, this measurable volume can be scanned over the scene. In the first frame, only the circle lies within the searchable region. As the horopter is scanned outwards by varying the vergence point, the triangle, then the cube become detectable, and their spatial location becomes known.

3 Active Vision for Task-Oriented Behaviors

A vision system able to adjust its visual parameters to aid task-oriented behavior – an approach labeled *active* [2] or *animate* [5] vision – can be advantageous for scene analysis in realistic environments [4]. In terms of spatial (depth) analysis, rather than obtaining a depth map over a large disparity range (as per static depth mapping), active vision allows us to consider only points at or near zero disparity for a given camera geometry. Then, by actively varying the camera geometry, it is possible to place the horopter and/or fixation point over any of the locations of interest in a scene and thereby obtain relative local spatial information about those regions. Where a subject is moving, the horopter can be made to follow the subject. By scanning the horopter over the scene, we increase the volume of the scene that may be measured. Fig. 2 shows how the horopter can be scanned over the scene by varying the camera geometry for a stereo configuration. This approach is potentially more efficient than static spatial methods because a small (or zero) disparity search scanned over the scene is less computationally expensive than a large and un-scanable disparity search from a static configuration.

Foveal systems are able to align the region around the centre of the image (where more resources are allocated for processing) with a region of interest in the scene such that attention can be maintained upon a subject. Active systems increase the visual workspace while maintaining high subject resolution and maintaining a consistent level of computation. Indeed, much success has come from studying the benefits of active vision systems [35]. Alternatively, pseudo-active configurations are possible where either fixed cameras use horizontal pixel shifting of the entire images to simulate horopter reconfiguration, or where the fovea is permitted to shift within the image. Although feasible for the operations presented herein, relying on such *virtual* horopter shifting of the entire images reduces the useful width of the images by the number of pixels of shift and dramatically decreases the size of the visual workspace. Target contextual information is also reduced where a target moves away from the optical centers of the static cameras such that its surroundings cannot be seen. Valuable processing time could also be compromised in conducting whole image shifts or in re-configuring the fovea position. Both virtual horopter and virtual fovea approaches are simply methods to simulate true active stereo vision.

Introspection of human vision provides motivation for coordinated foveal fixation. Humans find it difficult to fixate on *unoccupied space*. Empty space contains little information; we are more concerned with interactions with objects or surfaces and direct our gaze accordingly. The human visual system exhibits its highest resolution at the fovea where higher-level cognition such as object recognition has been shown to operate [41]. The extent of the fovea covers a retinal area of approximately the size of a fist at arms length [41], conceptually in line with task-oriented interactions with the real world.

We limit foveal processing resources to the region of the images immediately surrounding the image centres. The region beyond the fovea is considered only for an estimate of where the foveas are to fixate next (for tracking purposes). For the resolution of our cameras, the fovea corresponds to a region of about 60x60 pixels and an approximate area of $0.5m^2$ at $2m$ distance. Actively moving this region over the scene facilitates a large visual workspace.

For humans, the boundaries of an object upon which we have fixated emerge effortlessly because the object is centred and appears with similar retinal coverage in our left and right eyes, whereas the rest of the scene usually does not. For synthetic vision, the approach is the same. The object upon which fixation has occurred will appear with identical pixel coordinates in the left and right images, that is, it will have zero disparity. For a pair of cameras with suitably similar intrinsic parameters, this condition does not require epipolar or barrel distortion rectification of the images. Camera calibration, intrinsic or extrinsic, is not required.

4 Hand Segmentation

4.1 MAP ZDF Formulation

We begin by assuming short baseline stereo fixation upon the hand. A *zero disparity filter* (ZDF) is formulated to identify the projection of the hand as it maps to identical image frame pixel coordinates in the left and right foveas. Fig. 7 shows example ZDF output. Simply comparing the intensities

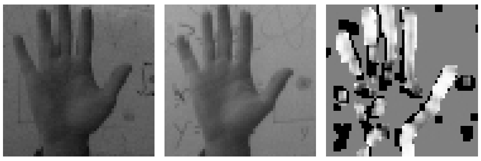

Fig. 3. NCC of 3x3 pixel regions at same coordinates in left and right images. Correlation results with higher values shown more white.

of pixels in the left and right images at the same coordinates is not adequate due to inconsistencies in (for example) saturation, contrast and intensity gains between the two cameras, as well as focus differences and noise.

A human can easily distinguish the boundaries of the object upon which fixation has occurred even if one eye looks through a tinted lens. Accordingly, the regime should be robust enough to cope with these types of inconsistencies. One approach is to *normalised cross-correlate* (NCC) small templates in one image with pixels in the same template locations in the other image. The NCC function is shown in Eq.1:

$$NCC(I_1, I_2) = \frac{\sum_{(u,v)\in W} I_1(u,v) \cdot I_2(x+u, y+v)}{\sqrt{\sum_{(u,v)\in W} I_1^2(u,v) \cdot \sum_{(u,v)\in W} I_2^2(x+u, y+v)}}, \qquad (1)$$

where I_1, I_2 are the compared left and right image templates of size W and u, v are coordinates within the template. Fig. 3 shows the output of this approach. Bland areas in the images have been suppressed (set to 0.5) using *difference of Gaussians*[4] (DOG) pre-processing. The 2D DOG kernel is constructed using symmetric separable 1D convolutions. The 1D DOG function is shown in Eq.2:

$$DOG(I) = G_1(I) - G_2(I), \qquad (2)$$

where $G_1(), G_2()$ are Gaussians with different standard deviations σ_1, σ_2 according to:

$$G(x) = \frac{e^{-x^2}}{2\sigma^2}, \qquad (3)$$

DOG pre-processing is used to suppress untextured regions that always return a high NCC response whether they are at zero disparity or not. As Fig. 3 shows, the output is sparse and noisy. The palm is positioned at zero disparity but is not categorised as such.

To improve results, image context needs to be taken into account. Contextual information can assist by assigning similar labels to visually similar neighbourhoods. Most importantly, contextual refinement allows slight relaxation of the zero disparity assumption such that non-planar surfaces or surfaces that are

[4] The *difference of Gaussians* function approximates the *Laplacian of Gaussians* function. Convolving a 2D DOG kernel with an image suppresses bland regions.

not perpendicular to the camera optical axes – but appear visually similar to the dominantly zero disparity region – can be segmented as the same object.

For these reasons, we adopt a Markov Random Field [16] (MRF) approach. The MRF formulation defines that the value of a random variable at the set of sites (pixel locations) S depends on the random variable configuration field f (labels at all sites) only through its neighbours $N \in S$. For a ZDF, the set of possible labels at any pixel in the configuration field is binary, that is, sites can take either the label *zero disparity* ($f(S) = l_z$) or *non-zero disparity* ($f(S) = l_{nz}$). For an observation O (in this case an image pair), Bayes' law states that the *a posterior* probability $P(f \mid O)$ of field configuration f is proportional to the product of the likelihood $P(O \mid f)$ of that field configuration given the observation and the prior probability $P(f)$ of realisation of that configuration:

$$P(f \mid O) \propto P(O \mid f) \cdot P(f). \tag{4}$$

The problem is thus posed as a *maximum a posterior probability* (MAP) optimisation where we want to find the configuration field $f(l_z, l_{nz})$ that maximises the a posterior probability $P(f \mid O)$. In the following two sections, we adapt the approach of [7] to construct the terms in Eq. 4 suitable for ZDF hand tracking.

4.1.1 Prior term $P(f)$

The prior encodes the properties of the MAP configuration we seek. It is intuitive that the borders of zero disparity regions coincide with edges (or intensity transitions) in the image. The Hammersly-Clifford theorem, a key result of MRF theory, is used to represent this property:

$$P(f) \propto e^{-\sum_C V_C(f)}. \tag{5}$$

Clique potential V_C describes the prior probability of a particular realisation of the elements of the clique C. For our neighbourhood system, MRF theory defines cliques as pairs of horizontally or vertically adjacent pixels. Eq. 5 reduces to:

$$P(f) \propto e^{-\sum_p \sum_{q \in N_p} V_{p,q}(f_p, f_q)}. \tag{6}$$

In accordance with [7], we assign clique potentials using the *Generalised Potts Model* where clique potentials resemble a well with depth u:

$$V_{p,q}(f_p, f_q) = u_{p,q} \cdot (1 - \delta(f_p - f_q)), \tag{7}$$

where δ is the unit impulse function. Clique potentials are isotropic ($V_{p,q} = V_{q,p}$), so $P(f)$ reduces to:

$$P(f) \propto e^{-\sum_{\{p,q\} \in \varepsilon_N} \begin{cases} 2u & \forall f_p \neq f_q, \\ 0 & otherwise. \end{cases}} \tag{8}$$

V_C can be interpreted as a cost of discontinuity between neighbouring pixels p, q. In practice, we assign the clique potentials according to how continuous the image is over the clique using the Gaussian function:

$$V_c = \frac{e^{-(\Delta I_C)^2}}{2\sigma^2}, \tag{9}$$

where ΔI_C is the change in intensity across the clique, and σ is selected such that 3σ approximates the minimum intensity variation that is considered smooth.

Note that at this stage we have looked at one image independently of the other. Stereo properties have not been considered in constructing the prior term.

4.1.2 Likelihood term $P(O \mid f)$

This term describes how likely it is that an observation O matches a hypothesized configuration f and involves incorporating stereo information for assessing how well the observed images fit the configuration field. It can be equivalently represented as:

$$P(O \mid f) = P(I_A \mid f, I_B), \tag{10}$$

where I_A is the primary image and I_B the secondary (chosen arbitrarily) and f is the hypothesized configuration field. In terms of image sites S (pixels), Eq. 10 becomes:

$$P(O \mid f) \propto \prod_S g(i_A, i_B, l_S), \tag{11}$$

where $g()$ is some symmetric function [7] that describes how well label l_S fits the image evidence $i_A \in I_A$ and $i_B \in I_B$ corresponding to site S. It could for instance be a Gaussian function of the difference in observed left and right image intensities at S; we evaluate this instance – Eq. 15 – and propose alternatives later.

To bias the likelihood term towards hand-like objects, we include a hand cue term H_S, Eq. 12. This term is not required for the system to operate, it merely provides a greater propensity for the MAP ZDF detector to track hand-like scene objects (rather than any arbitrary object), as required by the task. In our tuned implementation, the hand cue term enumerates (assigns a probability to site S in each image) how hand-like a pixel site is in terms of its colour and texture. However, formulation of the hand cue term is beyond the scope of this paper, and to show the generality of this body of work, we have set this term to zero throughout this paper, including results section. The reader may formulate this term to best suit their tracking application, or modulate this term dynamically to intelligently select the tracked/attended object.

$$P(O \mid f) \propto \prod_S g(i_A, i_B, l_S, H_S) \tag{12}$$

4.1.3 Energy minimisation

We have assembled the terms in Eq. 4 necessary to define the MAP optimisation problem:

$$P(f \mid O) \propto e^{-\sum_p \sum_{q \in N_p} V_{p,q}(f_{p,f_q})} \cdot \prod_S g(i_A, i_B, l_S). \tag{13}$$

Maximising $P(f \mid O)$ is equivalent to minimising the energy function:

$$E = \sum_p \sum_{q \in N_p} V_{p,q}(f_{p,f_q}) - \sum_S ln(g(i_A, i_B, l_S)). \tag{14}$$

4.1.4 Optimisation

A variety of methods can be used to optimise the above energy function including, amongst others, *simulated annealing* and *graph cuts*. For active vision, high-speed performance is a priority. At present, a graph cut technique is the preferred optimisation technique, and is validated for this class of optimisation as per [23]. We adopt the method used in [22] for MAP stereo disparity optimisation (we omit their use of α*–expansion* as we consider a purely binary field). In this formulation, the problem is that of finding the *minimum cut* on a *weighted graph*:

A weighted graph G comprising of vertices V and edges E is constructed with two distinct terminals l_{zd}, l_{nzd} (the source and sink). A cut $C = V^s, V^t$ is defined as a partition of the vertices into two sets $s \in V^s$ and $t \in V^t$. Edges t, s are added such that the cost of any cut is equal to the energy of the corresponding configuration. The cost of a cut $|C|$ equals the sum of the weights of the edges between a vertex in V^s and a vertex in V^t.

The goal is to find the cut with the smallest cost, or equivalently, compute the *maximum flow* between terminals according to the Ford Fulkerson algorithm

[12]. The minimum cut yields the configuration that minimises the energy function. Details of the method can be found in [22]. It has been shown to perform (as worst) in low order polynomial time, but in practice performs in near linear time for graphs with many short paths between the source and sink, such as this [23].

4.1.5 Robustness

We now look at the situations where the MAP ZDF formulation performs poorly, and provide methods to combat these weaknesses. Fig. 7a shows ZDF output for typical input images where the likelihood term has been defined using intensity comparison. Output was obtained at approximately $27fps$ for the 60x60 pixel fovea on a standard $3GHz$ single processor PC. For this case, $g()$ in Eq. 11 has been defined as:

$$g(i_A, i_B, f) = \begin{cases} \frac{e^{-(\Delta I_C)^2}}{2\sigma^2} & \forall f = l_z \\ 1 - \frac{e^{-(\Delta I_C)^2}}{2\sigma^2} & \forall f = l_{nz} \end{cases} \tag{15}$$

The variation in intensity at corresponding pixel locations in the left and right images is significant enough that the ZDF has not labeled all pixels on the hand as being at zero disparity. To combat such variations, NCC is instead used (Fig. 7b). Whilst the ZDF output improved slightly, processing time per frame was significantly increased ($\sim 12fps$). As well as being slow, this approach requires much parameter tuning. Bland regions return a high correlation whether they are at zero disparity or not, and so the correlations that return the highest results cannot be trusted. A threshold must be chosen above which correlations are disregarded, which also has the consequence of disregarding the strongest correct correlations. Additionally, a histogram of correlation output results is not symmetric (Fig. 5, left). There is difficulty in converting such output to a probability distribution about a 0.5 mean, or converting it to an energy function penalty.

To combat the thresholding problem with the NCC approach, the images can be pre-processed with a DOG kernel. The output using this technique (Fig. 7c) is good, but is much slower than all previous methods ($\sim 8fps$) and requires yet more tuning at the DOG stage. It is still susceptible to the problem of non-symmetric output.

We prefer a comparator whose output histogram resembles a symmetric distribution, so that these problems could be alleviated. For this reason we chose a simple *neighbourhood descriptor transform* (NDT) that preserves the relative intensity relations between neighbouring pixels (in a fashion similar to but less rigidly than that of the *Rank* transform), and is unaffected by brightness

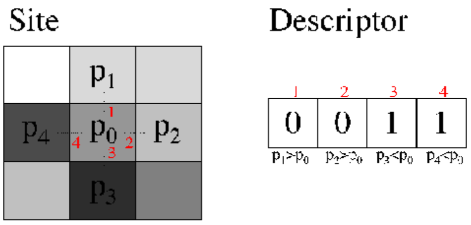

Fig. 4. NDT descriptor construction, four comparisons.

or contrast variations between image pairs. Fig. 4 depicts the definition of the NDT transform.

In this approach, we assign a boolean descriptor string to each site and then compare the descriptors. The descriptor is assembled by comparing pixel intensity relations in the 3x3 neighbourhood around each site (Fig. 4). In its simplest form, for example, we first compare the central pixel at a site in the primary image to one of its four-connected neighbours, assigning a *1* to the descriptor string if the pixel intensity at the centre is greater than that of its northern neighbour and a *0* otherwise. This is done for its southern, eastern and western neighbours also. This is repeated at the same pixel site in the secondary image. The order of construction of all descriptors is necessarily the same. A more complicated descriptor would be constructed using more than merely four relations[5]. Comparison of the descriptors for a particular site is trivial, the result being equal to the sum of entries in the primary image site descriptor that match the descriptor entries at the same positions in the string for the secondary image site descriptor, divided by the length of the descriptor string.

Fig. 5 shows histograms of the output of individual neighborhood comparisons using the NCC DOG approach (left) and NDT approach (right) over a series of sequential image pairs. The histogram of NDT results is a symmetric distribution about a mean of 0.5, and hence is easily converted to a penalty for the energy function.

Fig. 7d shows NDT output for typical images. Assignment and comparison of descriptors is faster than NCC DOG, ($\sim 27fps$) yet requires no parameter tuning. In Fig. 7e, the left camera gain was maximised, and the right camera contrast was maximised. In Fig. 7f, the left camera was defocussed and saturated. The segmentation retained it's good performance under these artificial extremes.

[5] Experiment has shown that a four neighbour comparator gives results that compare favorably (in terms of trade-offs between performance and processing time) to more complicated descriptors.

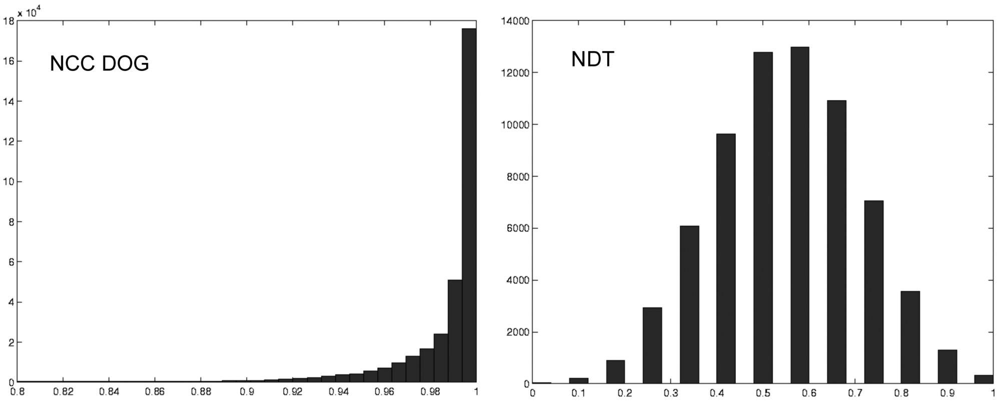

Fig. 5. Histograms of individual NCC DOG (left) and NDT (right) neighborhood comparisons for a series of observations.

5 Tracking and Segmentation

Hand tracking is implemented using a combination of virtual and physical retinal shifts. Fig. 6 describes the four steps of the tracking algorithm. Initialisation of the system is simple. The operator merely passes their hand through the area surrounding the arbitrary initial stereo fixation point. At a fixation point $2m$ from the cameras, the initial search window defines a receptive volume of about $0.5m^3$. Once tracking begins, segmentation of the zero disparity region induced by the hand is followed by continual NCC alignment of the horopter such that the zero disparity segmentation area is maximised. The NCC search window is sufficient to cope with the upper limits of typical hand motions between successive frames. The MAP ZDF process reduces the segmented area to that associated with a 2D projection of the object on the horopter, such that occlusions or secondary hands do not distract track unless they are essentially touching the tracked hand (see section 7.2.1). If track is lost, it will resume on the zero disparity region induced by the subject closest to the fixation point. In this manner, if track is lost, the subject need only return their hand to the volume surrounding the current fixation point (where track was lost).

The method of virtual verification followed by physical motion copes with rapid movement of the hand, providing an awareness of whether the hand has moved towards or away from the cameras so that the physical horopter can be shifted to the location that maximises the zero disparity area associated with the hand. It is emphasised that template matching is not used to track the hand, it is only used to estimate the pixel shift required to align the virtual horopter over the hand. Tracking is performed by extracting the zero disparity region at the virtual horopter, and physically moving the cameras to point at the centre of gravity of the segmented zero disparity region, if it is significantly non-zero. The virtual horopter alignment is successful if any part of the hand

MAP ZDF Tracking Algorithm:

(1) Determine virtual shift required to approximately align virtual horopter over subject: the pixel distance d between a small template (approximately 30x30 pixels) at the centre of the left image and its location of best match in the right image is determined using NCC. We conduct the search in a window a few pixels above and below the template location in the left image and up to 10 pixels to the left and right in the right image. In this manner, the NCC will only return a high correlation result if the subject in the template is located near the 3D scene fixation point.

(2) Perform a virtual shift of the left fovea by $d/2$ and the right fovea by $-d/2$ to approximately align the location of best correlation in the virtual centre of the left and right foveas. If the NCC result was not sufficiently high, no physical shift will be conducted and the process returns to the first step.

(3) MAP ZDF segmentation extracts the zero disparity pixels associated with a 2D projection of the hand from the virtually aligned foveas. If there is indeed a hand at the virtual fixation point, the area of the segmented region will be significantly beyond zero.

(4) If the area is greater than a minimum threshold, the virtual shift has aligned the centre of the images over the hand. In this case, a physical movement of the cameras is executed that reduces the virtual shift to zero pixels, and aligns the centres of the cameras with the centre of gravity of the segmented area. If the area is below the threshold, there is little likelihood that a hand or object is at the virtual fixation point, and no physical shifting is justified.

The process then cycles, continuing from step (1).

Fig. 6. Hand tracking algorithm using *maximum a posterior probability zero disparity filter* (MAP ZDF) segmentation.

is selected as the template, and does not depend on the centre of the hand being aligned in the template.

6 Results

Hand tracking and segmentation for the purpose of real-time HCI gesture recognition and classification must exhibit robustness to arbitrary lighting variations over time and between the cameras, poorly focussed cameras, hand orientation, hand velocity, varying backgrounds, foreground and background

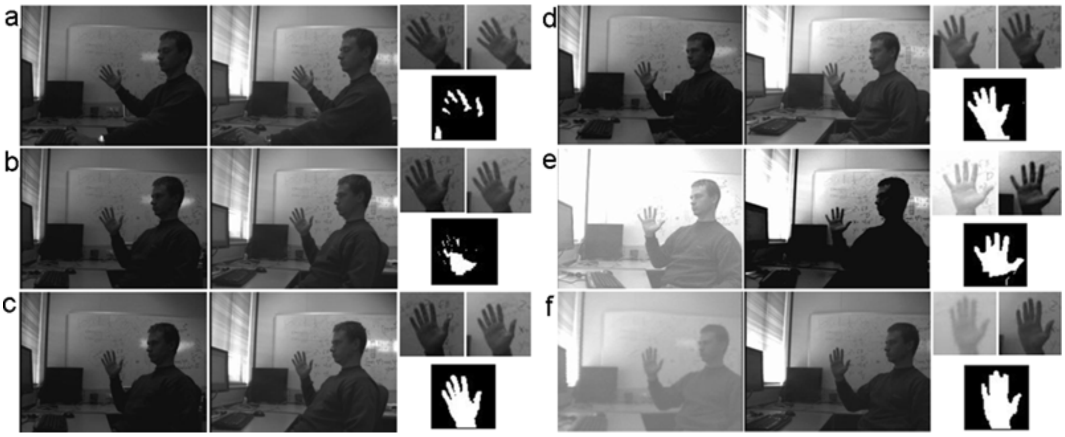

Fig. 7. MAP ZDF hand segmentation. The left and right images and their respective foveas are shown with ZDF output (bottom right) for each case *a-f*. Result *a* involves intensity comparison, *b* involves NCC, and *c* DOG NCC for typical image pairs. Result *d-f* show superior NDT output for typical images *d*, and extreme adverse conditions *e,f*.

distractors including non-tracked hands and skin regions, and hand appearance such as skin or hand covering colour. System performance must also be adequate to allow natural hand motion in HCI observations. The quality of the segmentation must be sufficient that it does not depart from the hand over time. Ideally the method should find the hand in its entirety in every frame, and segment adequately for gesture recognition. For recognition, segmentation need not necessarily be perfect for every frame because if track is maintained, real-time classification is still possible based on classification results that are validated over several frames. Frames that are segmented with some error still usually provide useful segmentation information to the classifier.

Fig. 7 shows snapshots from online MAP ZDF hand segmentation sequences. Segmentations on the right (d-f) show robust performance of the NDT comparator under extreme lighting, contrast and focus conditions. Fig. 8 shows the robust performance of the system in difficult situations including foreground and background distractors. As desired, segmentation of the tracked hand continues. Fig. 9 shows a variety of hand segmentations under typical circumstances including reconfiguring, rotating and moving hands as they perform a sequence of conceptually symbolic gestures in real time.

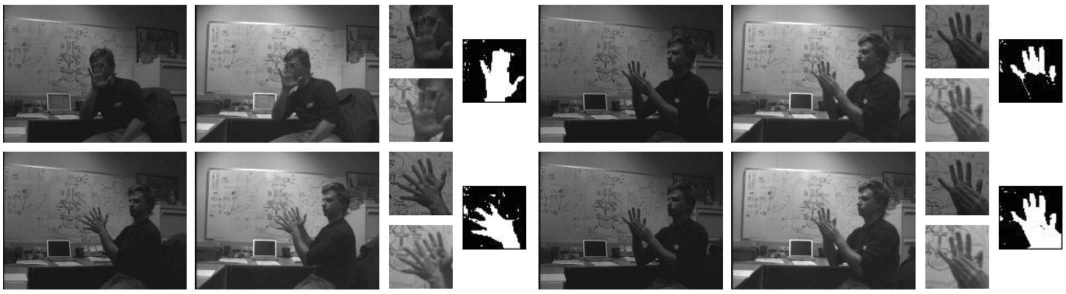

Fig. 8. Robust performance in difficult situations: Segmentation of the tracked hand from a face in the near background (top left); from a second distracting hand in the background (bottom left); and from a distracting occluding hand in the immediate foreground, a distance of $3cm$ from the tracked hand at a distance of $2m$ from the cameras (top right). Once the hands are closer together than $3cm$, they are segmented as the same object (bottom right).

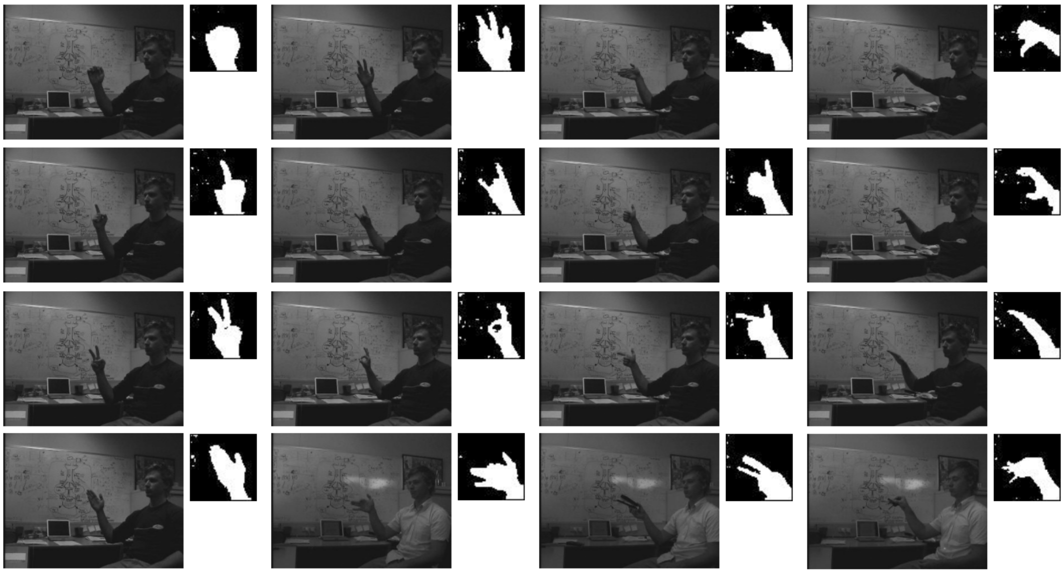

Fig. 9. Segmentation of conceptually symbolic gestures.

7 Performance

7.1 Speed

On average, the system is able to fixate and track hands at $27fps$, including display. Acquiring the initial segmentation takes a little longer ($23 - 25fps$ for the first few frames) after which successive MAP ZDF optimisation results do not vary significantly so using the previous segmentation as an initialisation for the current frame accelerates MRF labeling. Similarly, the change in segmentation area between consecutive frames at $30fps$ is typically small, allowing sustained high frame rates after initial segmentation. The frame rate remains above $20fps$ and is normally up to the full $30fps$ camera frame rate.

7.2 Quality

In typical tracking of a reconfiguring, moving hand over 100 consecutive frames, inaccurate segmentation of the hand typically occurs in around 15 frames. We describe a frame as *inaccurate* if the segmentation result has incorrectly labeled more than 10% of the pixels associated with the hand segmentation (either miss-labeling pixels on the hand as not being on the hand, or vice versa). These figures have been determined by recording segmentation output for typical gesturing sequences and having a human arbitrator review and estimate the percentage-wise inaccuracies in each frame.

Segmentation success also depends on the complexity of hand posture. For example, if the hand is posed in a highly non-planar fashion or a pose whose dominant plane is severely non-perpendicular to the camera optical axes, non-successful segmentation can degrade to up to around 50 frames in 100. In these situations, the zero disparity assumption is violated over some parts of the hand. The induced relaxation of the zero disparity assumption due to MRF contextual refinement is not always sufficient to segment the hand. Under such circumstances, methods reliant on prior knowledge could conceivably assist segmentation. For example, if the colour or appearance of the hand was known prior to segmentation and incorporated using the H_S term from Eq. 12. Nevertheless, despite some inaccurately segmented frames, track is rarely lost for natural motions and gestures.

The approach compares favorably to other ZDF approaches that have not incorporated MRF contextual refinement, allowing relaxation and refinement of the zero disparity assumption such that surfaces that are not perpendicular to the camera axis can be segmented.

7.2.1 Foreground and Background Robustness

Fig. 8 shows examples of segmentations where subject-like distractors such as skin areas, nearby objects, or other hands are present. For the case where the tracked hand passes closely in front of a face (that has the same skin colour and texture as the tracked hand) the system successfully distinguishes the tracked hand from the nearby face distractor (Fig. 8, top left). Similarly, when the tracked hand passes in front of a nearby hand, segmentation is not affected (Fig. 8, bottom left). Cue- or model-based methods are likely to have difficulty distinguishing between the tracked hand and background hand.

The right side images in Fig. 8 show the case where a tracked hand is occluded by an incoming distractor hand. The hands are located approximately $2m$ from the cameras in this example. Reliable segmentation of the tracked hand (behind) from the occluding distractor hand (in front) remains until the distractor hand is a distance of approximately $3cm$ from the tracked hand. Closer than this the hands are segmented as a connected object, which is conceptually valid.

7.3 Tracking Constraints

The hand can be tracked as long as it does not move entirely out of the fovea between consecutive frames. This is because no predictive tracking is incorporated (such as a Kalman filter). In practice, we find that the hand must be moved unnaturally quickly to escape track, such that it leaves the fovea completely between consecutive frames. Tracking a target as it moves in the depth direction (towards or away from the cameras) is sufficiently rapid that loss of track does not occur. In interacting with the system, we find that track was not lost for natural hand motions (see demonstration footage, Section 9).

The visual workspace for the system remains within a conic whose arc angle is around 100^o. Performance remains good to a workspace depth (along the camera axis) of $5m$, for the resolution, baseline, and zoom settings of our stereo apparatus. Higher resolution or more camera zoom would increase disparity sensitivity, permitting zero disparity filtering at larger scene depths.

7.3.1 Segmentation for Gesture Recognition

This work can give a basis segmentation to facilitate gesture validation. Fig. 9 shows various segmentations for conceivable symbolic gestures. Segmentation quality is such that the hand is extracted from its surroundings which has significant benefits in classification processes because the operation is not tainted by background features. In order that greater restriction on the segmentation of hand-like regions be ensured, an intuitive step would be to combine MAP ZDF segmentation with other cues to ensure “hand-ness” of the subject. Appearance classification or model verification could also be used. The framework, however, provides the means to incorporate probabilistic hand-ness of the segmentation. By inserting knowledge of the hand into the prior term in the ZDF formulation (for example a skin colour cue or shape/size cue), a measure of hand-ness could be incorporated into the segmentation process itself. In this instance, reliance on a final verification step is reduced or eliminated.

The likelihood term described in the MAP ZDF formulation does not incorporate the hand-ness term so that we are able to accurately segment the hand

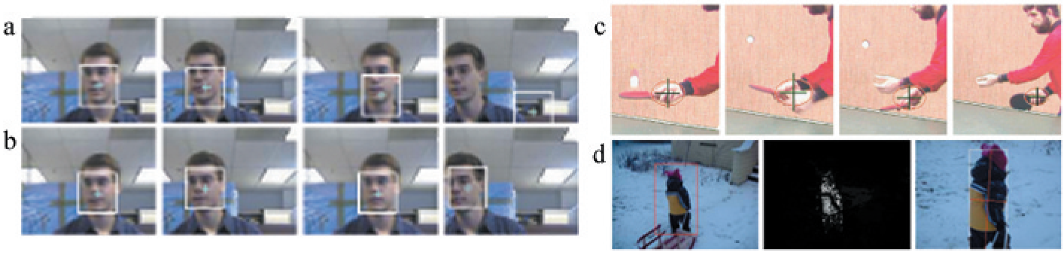

Fig. 10. Comparision to other methods: example output. Images reproduced from: a) Shen (Mean Shift) [37], b) Shen (Annealed Mean Shift)[37], c) Comaniciu (Cam Shift) [11], d) Allen (Cam Shift) [1].

and any hand-held object. The last two examples in Fig. 9 show the segmentation of a hand holding a set of keys, and a hand holding a stapler. The term has also been excluded for performance comparision with other ZDF tracking filters that do not incorporate biasing for task-dependent tracking of specific features such as hands (Section 7.4.2).

7.4 Comparison to state-of-art

Our method is based on active vision hardware, and as such, it is difficult to find a performance metric such that numerical comparison between method such as ours and methods that do not use active vision mechanisms can be conducted. Additionally, implementation details for other ZDF methods are difficult to obtain, and are usually hardware and calibration dependent such that reproduction is not viable. Methods that do not use contextual refinement for direct segmentation cannot be party to a segmentation performance comparison. Having said that, we provide samples of output from other implementations to allow the reader to assess performance visually.

7.4.1 Comparision with colour-based methods

We provide tracking output from recent methods for empirical evaluation (Fig.10). These methods provide bounding box output only, and as such do not deal with segmenting – for example – two overlapping hands (Fig.10c).

7.4.2 Comparision with other ZDF-based methods

Fig.11 shows sample ZDF output from existing methods for comparison. These methods provide probability distribution and bounding box outputs. The underlying probability maps may be suitable for MRF refinement such as ours, but they do not inherently provide segmentation.

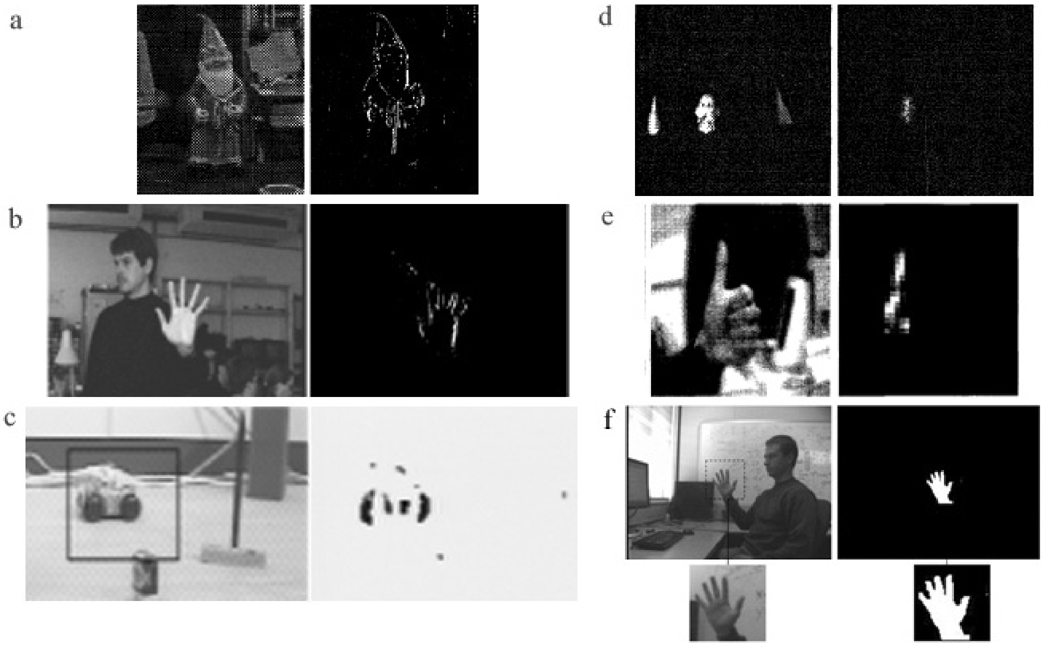

Fig. 11. ZDF performance comparision. Images reproduced from: a) Oshiro [31], b) Rae [32], c) Rougeaux [34], d) Yu [43], e) Rougeaux [36], f) This paper.

7.4.3 Comparision to non-MRF methods

Fig. 3 shows sample ZDF output from our system without the incorporation of MRF contextual refinement. Fig. 7c shows output using the same algorithm as in Fig. 3, but incorporates MAP MRF contextual refinement from the original images. Any attempt to use the output in Fig. 3 alone for segmentation (via any, perhaps complex, method of thresholding), or for tracking, would not yield results comparable to those achievable by using the output in Fig. 7c. The underlying non-MRF processes may or may not produce ZDF probability maps comparable to those produced by others (Section 7.4.2). However, the tracking quality achievable by incorporating MRF contextual image information refinement is better than is possible by the underlying ZDF process.

8 Discussion

It is critical that the MAP ZDF refinement operates at or near frame rate. This is because we consider only the 60x60 pixel fovea when extracting the zero disparity region. At slower frame rates, a subject could more easily escape the fovea, resulting in loss of track. Increasing the fovea size could help prevent this occurring, but would have the consequence of increasing processing time per frame.

Our method uses all image information, it does not match only edges, features or blobs extracted from single or multiple cues. The strongest labeling evidence does indeed come from textured and feature rich regions of the image, but the Markov assumption propagates strongly labeled pixels through pixel neighbourhoods that are visually similar until edges or transitions in the images are reached. The framework deals with the trade-off between edge strengths and neighbourhood similarity in the MAP formulation.

In contrast to many motion based methods, where motion models are used to estimate target location based on previous trajectories and motion models (eg, Kalman filtering), the implementation does not rely upon complex spatiotemporal models to track objects. It merely conducts a continual search for the maximal area of ZDF output, in the vicinity of the previous successful segmentation. The segmentations can subsequently be used for spatial localisation of the tracked object, but spatiotemporal dynamics do not form part of the tracking mechanism.

In this paper, we have operated on intensity images only. We have already begun experiments where colour channel data is used to enhance segmentation. For this implementation, the NDT comparison operation is conducted on the intensity channel as well as RGB channels, and all comparison results are incorporated into the Likelihood term in eq 12.

The focus of this paper has been on hand tracking, but this is just one example of the general usefulness of robust zero disparity filtering.

9 Conclusion

A MAP ZDF has been formulated and used to segment and track an arbitrarily moving, rotating and re-configuring hand, performing accurate marker-less pixel-wise segmentation of the hand. A large visual workspace is achieved by the use of active vision. Hand extraction is robust to lighting changes, defocus, hand colour, foreground and background clutter including non-tracked hands, and partial or gross occlusions including those by non-tracked hands. Good system performance is achieved in the context of HCI systems. It operates at approximately $27fps$ on a $3GHz$ single processor PC.

Demonstration Footage

Real-time sequences of the system in operation are available at:

`http://rsise.anu.edu.au/~andrew/cviu05`

References

[1] J G Allen, R Y D Xu, and J S Jin. Object tracking using camshift algorithm and multiple quantized feature spaces. In *Conf. in Research and Practice in Inf. Tech.*, 2003.

[2] J Aloimonos, I Weiss, and A Bandyopadhyay. Active vision. In *IEEE International Journal on Computer Vision*, pages 333–356, 1988.

[3] Y. Azoz, L. Devi, and R. Sharma. Tracking hand dynamics in unconstrained environments. In *IEEE International Conference on Face and Gesture Recognition Nara, Japan*, 1998.

[4] R Bajczy. Active perception. In *IEEE International Journal on Computer Vision*, pages 8:996–1005, 1988.

[5] D Ballard. Animate vision. In *Artificial Intelligence*, pages 48(1):57–86, 1991.

[6] A. Blake and M. Isard. Active contours. In *Springer-Verlag*, 1998.

[7] Y Boykov, O Veksler, and R Zabih. Markov random fields with efficient approximations. Technical Report TR97-1658, Computer Science Department, Cornell University Ithaca, NY 14853, 3 1997.

[8] G R Bradski. Computer vision face tracking for use in a perceptual user interface. In *Intel. Tech Journ.*, 1998.

[9] T. Cham and J. Rehg. Dynamic feature ordering for efficient registration. In *IEEE International Conference on Computer Vision, volume 2, Corfu, Greece*, 1999.

[10] Y Cheng. Mean shift, mode seeking, and clustering. In *IEEE Trans. Pattern and Machine Intelligence*, pages 17:790–799, 1995.

[11] D Comaniciu, V Ramesh, and P Meer. Kernel-based object tracking. In *IEEE Trans. Patt Anal. and Mach. Int.*, pages 25:5:564–575, 2003.

[12] L Ford and D Fulkerson. *Flows in Networks.* Princeton University Press, 1962.

[13] K Fukunaga. Introduction to statistical pattern recognition, 2nd edition. In *Academic Press*, 1990.

[14] D. M. Gavrila. The visual analysis of human movement - a survey. In *Computer Vision and Image Understanding*, 1999.

[15] D.M. Gavrila and L.S. Davis. 3-d model-based tracking of human motion in action. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1996.

[16] S Geman and D Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 721–741, 1984.

[17] K. Imagawa, S. Lu, , and S. Igi. Color-based hands tracking system for sign language recognition. In *IEEE International Conference on Face and Gesture Recognition, Nara, Japan*, 1998.

[18] M Isard and A Blake. Condensation: conditional density propatation for visual tracking. In *Int. Journal of Comp. Vis.*, pages 29:1:5–28, 1998.

[19] C. Jennings. Robust finger tracking with multiple cameras. In *Proceedings of the International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems, Corfu, Greece*, 1999.

[20] J.Joseph and Jr LaViola. A comparision of unscented and extended kalman filtering for estimation quaternion motion. In *Proceedings of American Control Conference*, 2003.

[21] N. Jojic, M. Turk, and T. Huang. Tracking self-occluding articulated objects in dense disparity maps. In *IEEE International Conference on Computer Vision, volume 1, Corfu, Greece*, 1999.

[22] V Kolmogorov and R Zabih. Multi-camera scene reconstruction via graph cuts. In *Europuan Conference on Comupter Vision*, pages 82–96, 2002.

[23] V Kolmogorov and R Zabih. What energy functions can be minimized via graph cuts? In *Europuan Conf. on Comupter Vision*, pages 65–81, 2002.

[24] Gareth Loy, Luke Fletcher, Nicholas Apostoloff, and Alexander Zelinsky. An adaptive fusion architecture for target tracking. In *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, page 261, 2002.

[25] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *IEEE International Conference on Computer Vision, volume 1, Corfu, Greece*, 1999.

[26] J. Martin, V. Devin, and J. Crowley. Active hand tracking. In *IEEE International Conference on Face and Gesture Recognition, Nara, Japan*, 1998.

[27] D. Metaxas. Deformable model and hmm-based tracking, analysis and recognition of gestures and faces. In *Proceedings of the International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems, pages 136-140, Corfu, Greece*, 1999.

[28] K Nummiaro, E Koller-Meier, and L V Gool. A colour-based particle filter. In *Proc. Int. Workshop on Generative Model Based Vis. in conj. ECCV*, pages 53–60, 2002.

[29] R O'Hagan, S. Rougeaux, and A. Zelinsky. Visual gesture interfaces for virtual environments. In *Interacting with Computers 14, (2002) 231-250*, 2002.

[30] E. Ong and S. Gong. A dynamic human model using hybrid 2d-3d representations in hierarchical pca space. In *British Machine Vision Conference, volume 1, pages 33-42, Nottingham, UK, BMVA*, 1999.

[31] N Oshiro, N Maru, A Nishikawa, and F Miyazaki. Binocular tracking using log polar mapping. In *IROS*, pages 791–798, 1996.

[32] R Rae and H Ritter. 3d real-time tracking of points of interest based on zero-disparity filtering. In *Workshop Dynamische Perzeption, Proceedings in Artificial Intelligence*, pages 105–111, 1998.

[33] C. Rasmussen and G. Hager. Joint probabilistic techniques for tracking multi-part objects. In *IEEE Conference on Computer Vision and Pattern Recognition, pages 16-21, Santa Barbara, CA*, 1998.

[34] S Rougeaux, N Kita, Y Kuniyoshi, S Sakano, and F Chavand. Binocular tracking based on virtual horopters. In *IROS*, 1994.

[35] S Rougeaux and Y Kuniyoshi. Robust real-time tracking on an active vision head. In *IEEE International Conference on Intelligent Robots and Systems*, pages 873–879, 1997.

[36] S Rougeaux and Y Kuniyoshi. Velocity and disparity cues for robust real-time binocular tracking. In *IEEE International CVPR*, 1997.

[37] C Shen, M J Brooks, and A Hengel. Fast global kernel density mode seeking with application to localisation and tracking. In *IEEE Int. Conf. on Comp. Vis*, 2005.

[38] K. Toyama and E. Horvitz. Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *Asian Conference on Computer Vision, Tapei, Taiwan*, 2000.

[39] J. Triesch and C. von der Malsburg. Self-organized integration of adaptive visual cues for face tracking. In *IEEE International Conference on Face and Gesture Recognition Grenoble, France*, 2000.

[40] H Truong, S Abdallah, S Rougeaux, and A Zelinsky. A novel mechanism for stereo active vision. In *Australian Conf. on Robotics and Automation*, 2000.

[41] Brian A Wandell. *Foundations of vision.* Sunderland, 1995.

[42] C. Wren, B. Clarkson, and A. Pentland. Understanding purposeful human motion. In *IEEE International Conference on Face and Gesture Recognition, pages 378-383, Grenoble, France*, 2000.

[43] H Yu and Y Baozong. Zero disparity filter based on wavelet representation in the active vision system. In *Proc. Int. Conf. Signal Proc.*, pages 279–282, 1996.