DIGITAL PERCEPTION LAB - RESEARCH OVERVIEW

David Suter

Institute for Vision Systems Engineering, Dept. Elect. & Comp. Syst. Eng.
Monash University, PO Box 35, Clayton 3800, Australia

ABSTRACT

This paper outlines recent work in the Digital Perception Lab. at Monash University.

1. INTRODUCTION

The Digital Perception lab was formed in 2001, It has a research staff comprising of about 10 researchers (students and postdocs) augmented with some affiliated former students and visitors (including students from Nagoya). The lab now "sits inside" the Monash Institute for Vision Systems Engineering, which was formed in 2004 as a faculty vehicle for collecting together computer vision researchers.
This paper provides an overview of the research activities, essentially over the last year.

2. VISUAL TRACKING

Visual tracking is the ability to associate a candidate region in one image, with the image of the target object in a previous image. As such, the varied tracking schemes usually have some form of the following components:

1. Background models (which help to identify candidate regions). This model has to be maintained-updated as well as initialised.
2. Foreground models (which also help to identify - but also match - candidate regions). These models have to be maintained-updated as well as initialised.
3. Similarity measure (between target and candidate)
4. Motion model or prediction/search strategy
5. More complex systems may have ways of maintaining multiple hypotheses about target candidate matches.

We have studied many of these components and made some contributions.

2.1. Foreground modelling and similarity measure

Many tracking methods use colour information (such as colour histograms) because such measures of the colour content are reasonably robust to illumination changes, viewpoint changes, orientation of the object etc. However, capturing the right combination of attributes (spatial distribution of the colour etc.) to be discriminative enough (reject false matches) whilst invariant and robust enough, is a difficult thing to perfect. Recently [1] we have proposed a foreground modelling technique that captures aspects of the joint colour-spatial distribution of the foreground object. We call this Spatial Mixture Of Gaussians (SMOG). A crucial component of our work has been controlling the computational cost of this richer model. Fusion of colour and edge cues has also been a feature of our work e.g., [2]

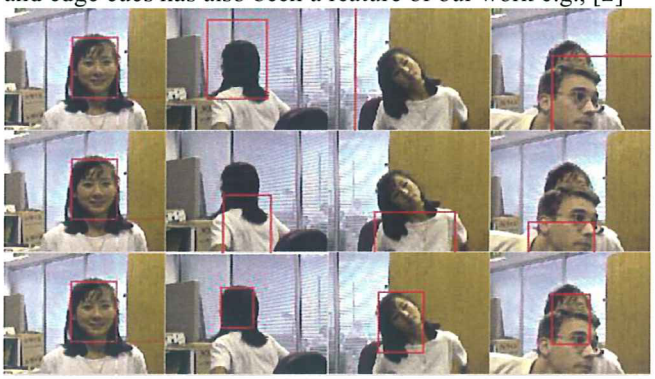

Figure 1 Fusion of cues: top colour only, middle edge only, bottom colour and edge.

Preceding work [3] investigated simple notions of foreground modeling based upon the idea of "sample consensus" (motivated by our considerable experience in robust statistical procedures). Basically, this method counts the number of pixels that "agree", based upon colour. Like our more recent work, some attempt is also made to incorporate spatial information in a non-brittle way - in this case by breaking the model into regions (roughly head, torso, legs) based upon a simple box model of humans. Updating the model requires some careful avoidance of contamination when objects (in this case people) get close to each other (we suspend updating in such circumstances).

2.2. Background modelling

Our sample consensus notions have also been used to devise a simple but effective background model (SACON) and then extended/modified [4]for foreground modelling [3]. This method essentially counts the agreement between pixels (in the case of background modelling with static camera - the history of a particular pixel position.

2.3. Other aspects

Of course, implementation of a real tracker requires some attention to details (such as colour space choice etc.). Our work [3] details some of these aspects.

3. MOTION SEGMENTATION AND SFM

Visual tracking has many applications. Some of these applications require the segmentation of tracks into those belonging to separate objects. Having done this, one can try to reconstruct the shape of the objects (structure from motion or SFM).

Our work [5] is the first reasonably general solution to the n-view case as it extends our previous 2-view approach [6]

4. FACE RECOGNITION

We have recently concentrated on face recognition from video (as opposed to still images). In such a context, as one gathers multiple images (from slightly different viewpoints and lighting conditions) one can improve the extracted model of the face. The aim is to decide when "enough images have been seen". Since we use a susbpsace-based model, we developed fast SVD updates [7], as well as a useful stopping criteria (the point at which we are comfortable making a judgement of the identity). In collaboration with Faggian and Papalinski (Monash Information Technology), we have used our incremental SVD technologies for speeding up Active Appearance Modelling of human faces [8]. This work has also been extended to incremental PCA [9, 10]

5. MULTI-CAMERA SYSTEM AND HUMAN MOTION CAPTURE

We built a multi-camera system (about 5 or 6 cameras currently) using firewire cameras and synchronization circuits. Since construction, we have used the system for a number of activities - including human motion capture. Color calibration of our multi-camera system was studied by Mr. K. Yamamoto on a visit from Nagoya [11-13].

6. HUMAN ACXTIVITY CLASSIFICATION

We have been working the classification of human actions (including gait and other simple movements/actions) based upon very simple extracted primitives. For example [14], we directly convert an associated sequence of human silhouettes derived from videos into two types of computationally efficient representations, i.e., average motion energy and mean motion shape, to characterize actions. Supervised pattern classification techniques using various distance measures are used for recognition.

7. URBAN SCANNING - CITY MODELLING

Figure 2 Large scale laser scanner - Riegl LMS420i - note the color Nikon camera on top of the laser to capture the color information augmenting the depth information from the laser.

The basic aim of this project is to produce 3D models automatically (realistically - *semi*-automatically) from the laser scan - see Figure 2 - (point cloud) data.

We first focused on some elementary side issues. The first being artefacts that occur because of the time lag of the scanning process and the camera operation. Moving objects create the following problems:

1. Where the object was when the laser scan was taken, there will be a point cloud representing the object. This will be texture mapped with what the camera sees when the object has moved to reveal the background (e.g., a human shape texture mapped with the grass from behind the person).
2. Where the object was when the camera was operated there will be a human "picture" texture mapped onto the background (e.g., a flat human lying on the road).

Thus we need to remove both types of artefact. The laser scan artefacts are relatively easy to remove (if not expensive/time consuming). Several scans can be taken and the final data is composed of the furthest point in each direction.

The image artefacts also use more than one "scan" (pass of the camera). The most obvious is to take 3 or more images and pick the median pixel values (assumes at least 50% of the images are un-occluded at each point. However, one can do better than that and we have modified Herley's method to this end [15]. In essence, the improvements we have made to Herley's method (requires only 2 images) involve: methods to handle overlapping occlusion objects (the occlusion boundary in the consensus image contains internal boundaries).

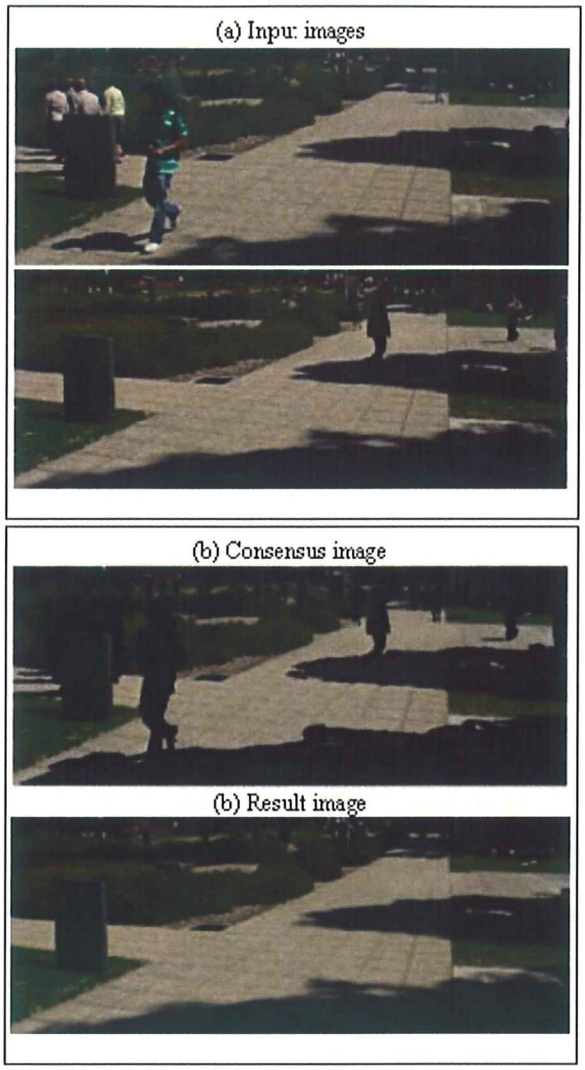

Figure 3 Top pair - two input images. Bottom pair - consensus image and result.

Ultimately we will be using forms of (robust) fitting to extract man-made structure[1]. However. the amount of data and especially the amount of clutter (non-building points), makes such processing consuming and unreliable. Thus we have been investigating pre-processing techniques to classify the data in building data and non-building. We are investigating methods based upon the 2D image data (alone) [16] - see Figure 3 and 3D image data (alone) [17] - see Figure 6 and Figure 5. We intend to also look at using both sources of information together. The 2D image classification uses a multi-scale random field and Gaussian mixture modelling (based on simple measures of local linearity) augmented with a fractal pre-classification stage. The result performs as well as previous schemes using much richer feature sets. The 3D classification uses adaptive (for the sampling density) measures of linearity, and planarity classified using a conditional random field and Gaussian mixture modelling.

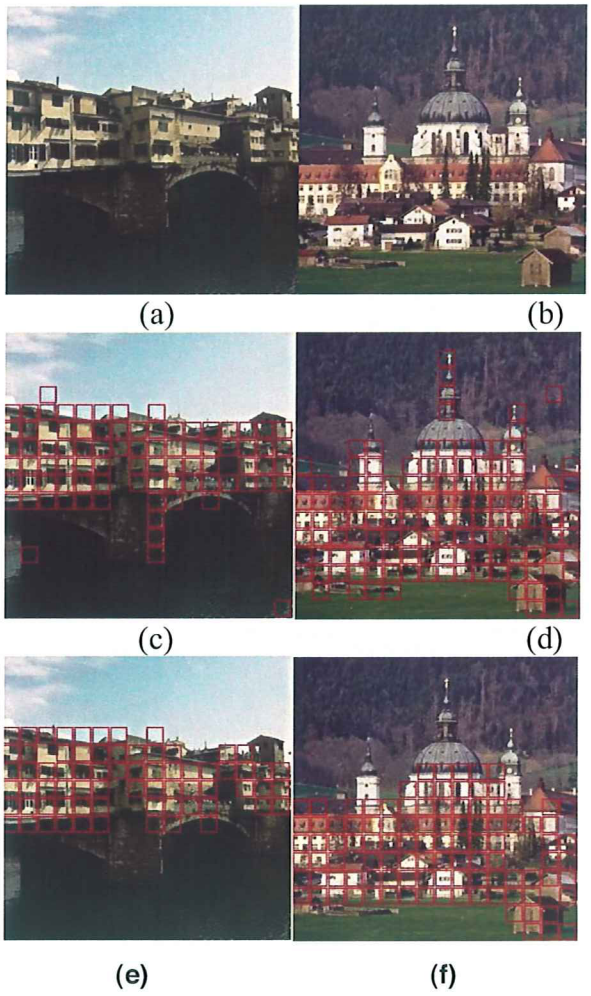

Figure 4. MSRF segmentation results. (a) (b) original images, (c)(d) GMM results, (e)(f) GMM+MSRF

[1] Mr. J. Nagao, of Nagoya worked on modelling tree trunks by fitting circles during his visit.

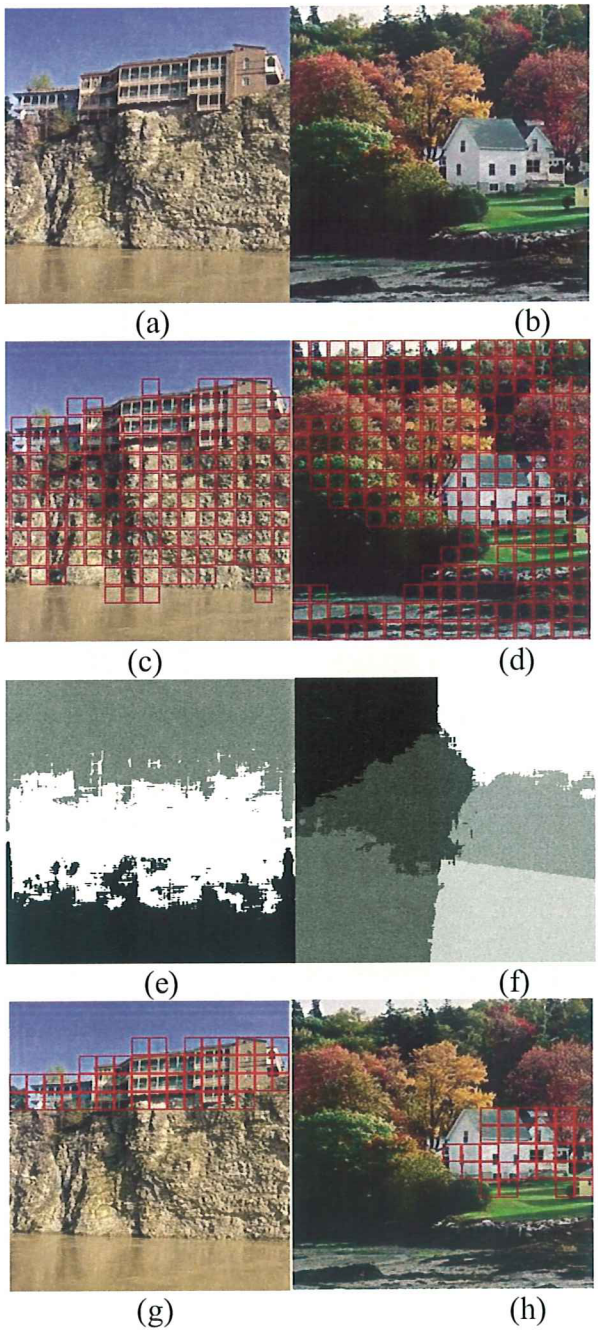

Figure 5. Fractal+MSRF segmentation results. (a)(b) original images, (c)(d) GMM+MSRF results, (e)(f) fractal based clustering (g)(h) Fractal+MSRF segmentation.

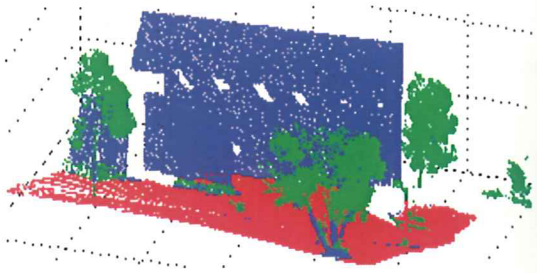

Figure 6 Classification of point cloud data in vegetation (green), terrain (red) and building (blue).

8. MEDICAL IMAGING - KNEE CARTILAGE VOLUME ESTIMATION

In collaboration with Cicuttini (Monash Dept. of Epidemiology), we have been working on automated (or semi-automated) methods of measurement of knee cartilage volume (useful in tracking diseases such as Osteoarthritis (OA). We have a semi-automated method that produces good results (saving much time and effort on the part of the operator).

Early work focussed on a mode of operation where a semni-skilled operator would quickly lay down a few points in the region of the cartilage. The system would then find the cartilage boundaries with reasonable accuracy and repeatability - saving much operator time and fatigue. We have recently looked at ways to automatically find the cartilage region - thereby obviating the need for an operator. Recently, some very encouraging results have been obtained (see Figure 7).

GOF	Sensitivity		
0.66	0.81	Avg	Lynch's Method
0.69	0.88	Median	
0.13	0.16	St Deviation	
0.79	0.88	Avg	Manual - operator 1
0.08	0.07	Median	
0.81	0.90	St Deviation	
0.84	0.91	Avg	Manual operator 2
0.07	0.06	Median	
0.85	0.92	St Deviation	
0.42	0.61	Avg	Lynch's Method Initialised with Felzenswalb's method
0.45	0.70	Median	
0.25	15.83	St Deviation	

Table 1 Measurement of knee cartilage volume.

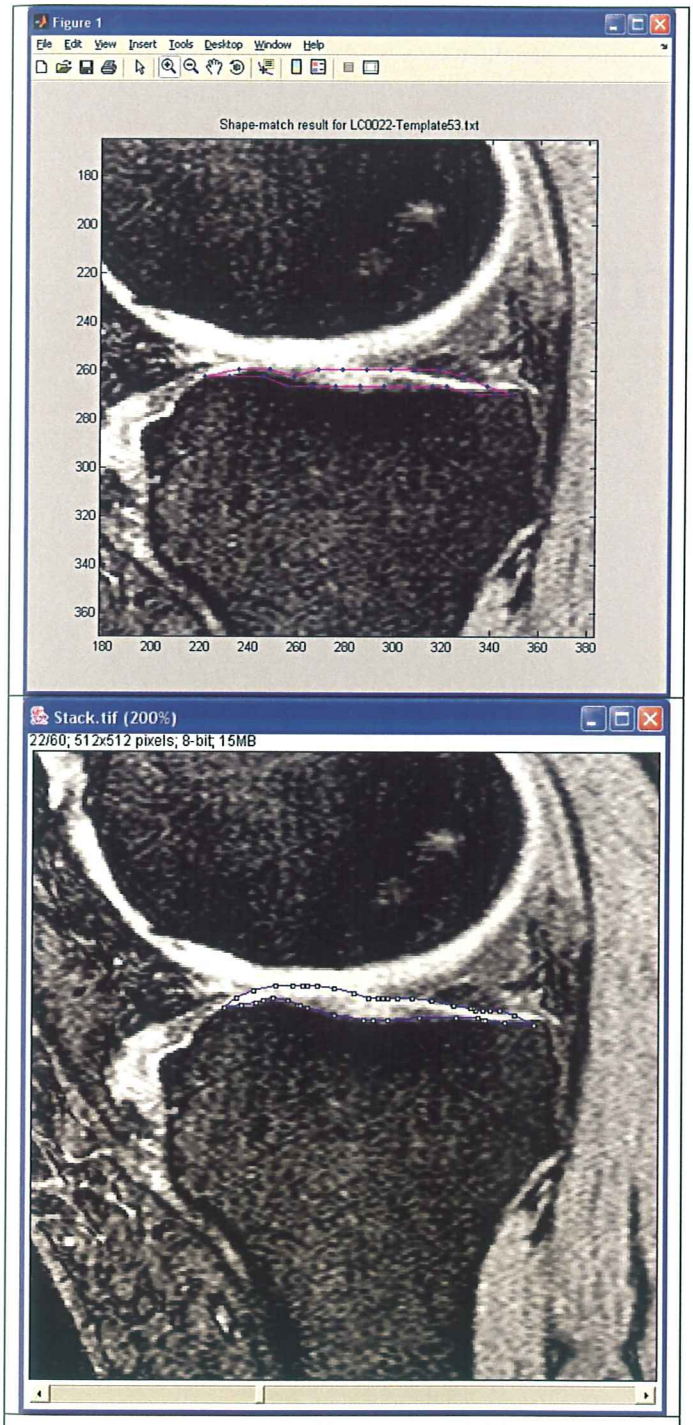

Figure 7 Top - initial automatic placement (using Felzenswalb's method) Bottom auto-refined using Lynch's method.

Quantitative results can be found in Table 1 where:
TP is total number of pixels/voxels, FN is false negative, FP is false positive and
Goodness of Fit (GOF) is TP/(TP+FP+FN)
Sensitivity is TP/(number of pixels in manual segmentation).

9. REFERENCES

[1] H. Wang and D. Suter, "Effective Appearance Model and Similarity Measure for Particle Filtering and Visual Tracking," in *ECCV2006*, Graz, Austria, 2006, pp. 606-618

[2] H. Wang and D. Suter, "Efficient Visual Tracking by Probabilistic Fusion of Multiple Cues," in *ICPR2006*, 2006, pp. 892-895.

[3] H. Wang and D. Suter, "A Consensus Based Method for Tracking: Modelling Background Scenario and Foreground Appearance," *Pattern Recognition,* p. to appear, 2006 2006.

[4] H. Wang and D. Suter, "Background subtraction based on a robust consensus method," in *ICPR2006*, 2006, pp. 223-226.

[5] H. Wang, J. U, and K. Schindler, "Perspective n-view Multibody Structure-and-Motion through Model Selection.," in *ECCV2006*, Graz, Austria, 2006, pp. 606-619

[6] K. Schindler and D. Suter, "Two-view multibody structure-and-motion with outliers through model selection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, pp. 983-995, 2006.

[7] T.-J. Chin, K. Schindler, and D. Suter, "Incremental Kernel SVD for Face Recognition with Image Sets," in *FGR '06: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pp. 461-466.

[8] N. Faggian, A. Paplinski, and T.-J. Chin, "Face Recognition from Video using Active Appearance Model Segmentation," in *ICPR2006*, 2006, pp. 287-290.

[9] T.-J. Chin and D. Suter, "Incremental kernel pca for efficient non-linear feature extraction. ," in *BMVC2006*, Edinburgh, UK, 2006, pp. 939-948.

[10] T.-J. Chin and D. Suter, "Improving the Speed of Kernel PCA on Large Scale Datasets " in *IEEE Int. Conf. on Advanced Video and Signal-based Surveillance*, Sydney, Aust., 2006, p. (to appear).

[11] K. Yamamoto, T. Yendo, T. Fujii, M. Tanimoto, and D. Suter, "Color Correction for Multi-Camera System by Using Correspondences," in *SIGRAPH 2006 Research Posters (no. 73)*, 2006.

[12] K. Yamamoto and J. U, "Color Calibration for Multi-Camera System without Color Pattern Board," Dept. Elect. & Comp. Syst. Eng., Monash University, Clayton, 3800, Australia MECSE-4-2006, 2006.

[13] K. Yamamoto and J. U, "Color Calibration for Multi-Camera System by using Color Pattern Board," Dept. Elect. & Comp. Syst. Eng., Monash University, Clayton, 3800, Australia MECSE-3-2006, 2006.

[14] L. Wang and D. Suter, "Informative shape representations for human action recognition," in *ICPR2006*, 2006, pp. 1266-1269.

[15] E.-H. Lim and D. Suter, Occlusion Removal for 3D Urban Modelling, (in review), 2006

[16] H. Zhou and D. Suter, A Hybrid Approach to Man-Made Structure Extraction from Natural Scenes (in review), 2006

[17] E.-H. Lim and D. Suter, Classification of 3D LIDAR Point Clouds for Urban Modelling, (in review), 2006