# **PhD Proposal**

Stuart Eiffert

# **Project title**

Understanding and predicting animal behaviour from motion and pose using deep learning

# **Research supervisor**

Salah Sukkarieh

Table of Contents

[1 Proposed mode of research 2](#__RefHeading___Toc422_262613666)

[2 Synopsis 2](#__RefHeading___Toc424_262613666)

[3 Overall aims and objectives 3](#__RefHeading___Toc426_262613666)

[4 Background 4](#__RefHeading___Toc428_262613666)

[Object detection ......4](#__RefHeading___Toc964_262613666)

[Classification ......4](#__RefHeading___Toc966_262613666)

[Object Tracking ......5](#__RefHeading___Toc968_262613666)

[Feature mapping ......5](#__RefHeading___Toc1310_1639344015)

[3D model creation ......5](#__RefHeading___Toc1312_1639344015)

[Pose estimation ......6](#__RefHeading___Toc1314_1639344015)

[5 Expected research contribution 6](#__RefHeading___Toc430_262613666)

[6 Work plan 6](#__RefHeading___Toc434_262613666)

[7 Resources 7](#__RefHeading___Toc436_262613666)

[8 References 7](#__RefHeading___Toc438_262613666)

# Proposed mode of research

The research will take place within the Australian Centre for Field Robotics at the University of Sydney, as part of on ongoing project with Meat and Livestock Australia. This project is aiming to establish methods of objective measures of animal welfare and body state using computer vision techniques.

The primary modes of research will include quantitative data analysis of computer vision and sensing systems and field implementation of methods for comparison against existing state of the art techniques**.**

# Synopsis

For any robotic platform to be able to properly interact with either animals or humans it requires a basic understanding of behaviour and the ability to predict future behaviour of the animal, person, or group. This is an important challenge relevant to activities such animal herding, welfare monitoring, stress identification, estimation of body condition and other handling purposes.

This is an important challenge that is two fold in it’s aims, allowing the prediction of behaviours that require an intervention from the system, in activities such as herding and handling, and allowing the estimation of the current state of the animal with regards to it’s behaviour and physical condition, relevant to activities such as welfare monitoring, injury detection, stress identification and estimation of body composition and condition.

The goals of this research are to address the issue of tracking an animal’s motion and model through a series of observations and to investigate the use of data driven approaches, such as recurrent neural networks, in the estimation and prediction of animal state and behaviour from it’s current motion.

The task of tracking an animal to extract and link observational information is made more difficult by the fact that the animal may change its appearance between observations, due to lighting or environment conditions, or by deformation and rotation. As such, it becomes essential to develop a learnt model, or feature vector, of the animal or object instance to allow later recognition.

Current methods for the development of object feature models involve both offline-learning, in which it is necessary to train a classifier on prior labelled data, and online-learning, where features from the current object detection are added to the prior data, allowing the system to adapt to changing targets and surrounding background.

This is still an active field of research, with a major area of interest being the estimation and tracking of an object's pose, a necessary step for the development of a 3D model from an object. This issue is again made more difficult when taking into consideration non-rigid or semi-rigid objects, such as pedestrians or animals.

The ability to accurately estimate the current pose of a non-rigid object and extract a 3D model allows for a variety of further classification and behavioural analysis of the object, including: classification of objects of a given class within sub-categories which can be used for outlier detection in animals and pedestrians; analysis of object characteristics to select for certain traits such as height and volume; accurate estimation of object pose for handling purposes such as lifting.

Accurate pose estimation may also be relevant to the understanding and prediction of behaviour. Current methods of behaviour recognition use a large amount of data, not always attainable in practice. By limiting the input features to a pose estimation rather than raw image or depth cloud, it may be possible to achieve similar results with a much smaller dataset. This idea could allow the system to be easier applied to novel observations, such as extending its use to a variety of animals.

Building upon current methods of using temporal fusion of rgbd data to develop a reconstruction of non-rigid scenes and objects [1], I propose the use of an initial class-generic model, chosen based on classification of the object in rgb data, to allow faster reconstruction of the object instance model, and better estimation of object model in situations in which complete observations are not possible.

# Overall aims and objectives

* Comparison of machine learning techniques for object detection and classification in both rgb and rgbd data.
* Comparison of sensor technologies for object tracking uses in real world environments.
* Implementation of object tracking techniques for individual object instance tracking during multiple object observations.
* Implementation of machine learning techniques for initial detection and classification of objects of interest in rgb data.
* Implementation of object feature model development using current online-learning techniques.
* Development of a canonical feature model of an observed object instance, based on initial class-generic model and updated instance specific features from observation.
* Application of the canonical model development to 3D model development using structure from motion techniques.
* Development of a complete computer vision system to allow real-time understanding of the behaviour of non-rigid and semi-rigid (not really a term, but a person or a horse only has so many DOF, allowing for easier estimation of model changes through frames, so it’s not the same idea as for instance a non-rigid towel) objects through frames.
* Development and implementation of techniques for the classification of the developed object model into sub-classes.
* Temporal analysis of the observed object model pose estimation for behaviour prediction in pedestrians and animals.

# Background

Analysing a visual scene and recognising all of the constituent objects is a task in computer vision that has been on going for decades and covers a broad area of research.

However, the process of identifying any object in a camera frame, whether it be a 2D image, point cloud or any other sensor image, follows the same general principle of : image segmentation, feature extraction, feature vector creation, and classification. The real world is of course not as simple, with

Alongside this is the process of tracking an individual object instance across a series of camera frames, which carries its own challenges, such as occlusion and background clutter.

## Object detection

Recognition of objects in computer vision can be broken down depending on the type of object we are trying to detect, for instance, whether we are trying to find a particular known object, or whether we are looking for a general category/class of object.

All versions of detection make use of some sort of feature descriptor, whether they be a ‘bag of words’ approach at interest point locations, wherein an object is described by a set of image features and their relative positions, such as in Haar [17] , SIFT, SURF, ORB [8] etc., or by a more generalised feature vector, such as HOG [3], LBP (local binary patterns), contour or gradient based descriptors.

The extracted feature descriptor is then used to perform a level of discrimination on the observation, usually through the means of a classifier trained on prior labelled data.

## Classification

There exist many supervised learning techniques to classify visual datasets, with the choice of classifier usually dependent on the dimensionality of the feature vector being used.

Examples of classifier used in object detection include support vector machines (SVM), k-nearest neighbour and decision tree learning. All of these techniques have been successfully used in computer vision classification applications to a high degree of accuracy. The HOG descriptor used by Dalal and Triggs in their seminal 2005 work [3] has been used with linear SVM to create very accurate object detectors, used extensively in human detection applications even now.

The traditional recognition approach, of developing better ‘hand-designed’ feature descriptors is being taken over by the development of better classifiers, which learn features directly.

The majority of research in image classification is currently focused on the use of neural networks, an idea that was developed decades ago [2] but is only recently finding useful applications in computer vision [18]. These networks are trained directly on labelled data, rather than extracted feature vectors, and so develop their own feature descriptor. This approach is applicable to both rgb and rgbd data directly [6].

## Object Tracking

A number of issues are faced in the tracking of any object in a series of images, including clutter, occlusion, and appearance variations [4]

Current methods incorporate online learning to provide an appearance model which is able to adapt to the target object and its surrounding background during tracking. This approach can suffer from the ‘template-update problem’ [5] which summaries the idea that if a model, or template, of an object is not updated it will not be able to remain an accurate representation in all environments, but that by updating the model with all observations, errors can accumulate, leading to a drift away from the objects correct representation.

Some recent works, such as the Tracking-Learning-Detection approach [7], have aimed to minimise this issue by updating the model using observations based on the object's trajectory alongside actual detection of the object, however the issue still remains to a significant extent.

## Feature mapping

Important to both tracking and detection is the idea of feature mapping, in which image features, corners, edges, SIFT, SURF, ORB [8] etc., are matched between a known representation of an object, whether that be a template or a previous image frame, and the current image frame. This process is followed by geometric verification, using techniques such as RANSAC, to find the 6D transformation, or homography of the object between frames.

This same method is used in SLAM approaches and 3D model creation.

## 3D model creation

Current methods of dynamic object model extraction tend to focus upon rgb-d data [1] [9]. These methods have developed from early works developed using the KinectFusion methods [10] and allow the accurate recreation of 3D models of arbitrary objects in near real-time. These methods are limited in their application due to the nature of the IR structures patterns used to determine depth in ambient IR lighting, such as any time ever outdoors. New stereoscopic cameras, such as the ZED camera, are aiming to overcome these limitations but still have a significant amount of development required before they are comparable to the accuracy or IR rgbd cameras in a structured lighting environment.

## Pose estimation

Pose estimation of objects is essential for purposes such as object handling and intention prediction, for instance in pedestrians alongside a road. For well textured rigid objects with prior models this becomes a simple process of feature mapping, however the issue is confounded when using a non-rigid object or when the object’s model is not known prior to observation.

Techniques for use on ‘semi-rigid- objects’, such as pedestrians include the use of limb joint detection and association, using trained neural classifiers [11] or stereo perception [12]

# Expected research contribution

A better understanding of the relationship between body motion and behaviour will help enable robotic systems to interact with humans and animals in future. By being able to predict the intent of the animal with regards to it’s motion this information can then be used in its decision making process for controlling the robotic system itself in order to achieve a required state or position of the animal.

This ability to predict intent and behaviour will also enable robotic systems to operate in such a way that undesirable states, such as high stress levels can be minimised in the animals it interacts with. This information would also be able to be used to extract objective measurements of an animals welfare to be used in classification systems aiming to identify such things as injury, lameness or abnormal behaviour in animals.

# **Work plan**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Task** | **Semester** | | | | | |
| **1** | **2** | **3** | **4** | **5** | **6** |
| Review of current methodologies regarding object detection, tracking and pose estimation |  |  |  |  |  |  |
| Development of skills regarding quantitative data analysis and handling |  |  |  |  |  |  |
| Implementation of current detection and tracking techniques |  |  |  |  |  |  |
| Development of online-learning methods for object pose estimation |  |  |  |  |  |  |
| Implementation of developed techniques in field tests |  |  |  |  |  |  |

# Resources

* Video data of non-rigid objects
* Labelled image datasets for objects to be detected
* GPU computing for classifier training
* Embedded robotic platform for field testing

# References

1. Newcombe et al, 2015 *DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time 2015*
2. Y. LeCun, L. Bottou, Y. Bengio and P. Haffner: Gradient-Based Learning Applied to Document Recognition, Proceedings of the IEEE, 86(11):2278-2324, November 1998,
3. Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’2005), pp. 886–893, San Diego, CA..
4. E. Maggio and A. Cavallaro. Video Tracking: Theory and Practice. Wiley, 2011
5. L. Matthews, T. Ishikawa, and S. Baker. The template update problem. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(6):810–815, 2004
6. C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. CoRR, abs/1612.00593, 2016
7. Z. Kalal, K. Mikolajczyk, and J. Matas, “Tracking-learningdetection,” TPAMI, 2012.
8. E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: an efficient alternative to SIFT or SURF,” in IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, November 2011, pp. 2564– 2571.
9. M. Innmann, M. Zollhofer, M. Nießner, C. Theobalt, and ¨ M. Stamminger. VolumeDeform: Real-time volumetric non-rigid reconstruction. arXiv preprint arXiv:1603.08161, 2016. 2
10. R. Newcombe, A. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, D. Molyneaux, S. Hodges, D. Kim, and A. Fitzgibbon, “KinectFusion: Real-time dense surface mapping and tracking,” in Proc. IEEE Int. Symp. Mixed Augmented Reality, 2011, pp. 127–136.
11. Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. arXiv:1611.08050v1, 2016
12. J. Almeida, V. Santos, Pedestrian pose estimation using stereo perception, in: Second Iberian Robotics Conference, Vol. 417, Robot 2015, Springer, 2016, pp. 491–502
13. Y. Xiang, A. Alahi, and S. Savarese. Learning to track: Online multi-object tracking by decision making. In ICCV, 2015. 7
14. Yang, H., Shao, L., Zheng, F., Wang, L., Song, Z.: Recent advances and trends in visual tracking: A review. Neurocomputing 74(18), 3823–3831 (2011)
15. N. Dalal, and B. Triggs, Histograms of oriented gradients for pedestrian detection. In CVPR, 2005.
16. R. Szeliski, Computer Vision: Algorithms and Applications, Springer, 2011
17. P. Viola and M. J. Jones, Robust real-time face detection, International Journal of Computer Vision, 57 (2004), pp. 137–154.
18. A. Krizhevsky, I. Sutskever, and G. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012