Regression-based Hand Pose Estimation from Multiple Cameras

Teófilo E. de Campos and David W. Murray

Department of Engineering Science, University of Oxford

www.robots.ox.ac.uk/ActiveVision

Introduction

- **Objective:** to estimate hand pose in 3D for real-time applications.
- **Current Approaches:**
 - *Model-based generative methods:* require (re-)initialisation and tracking can easily get lost with sudden movements;
 - *Classification-based methods:* require a large training set to be loaded into memory for classification;
 - *Regression-based methods:* provide a continuous map from image to 3D pose. They are fast and can work with global image descriptors.
- **Problem:** hands present a high degree of self-occlusion.
- **Our approach:** extend a regression-based method for multiple views.

Rotation Invariant Shape Contexts

- **Pre-processing:** skin is detected using a histogram-based classifier in the CbCr colour space (from YCbCr).
- **Shape Contexts**

input image silhouette contour a shape context

Methods for Rotation Invariance

Classification Results

Multiple View Image Descriptors

Image Descriptor Manifolds

Evaluating the Descriptor with Real Images

Regressor Learning

- Given a set of pairs of image measurements x_i and 3D poses y_i , the system is described as:

$$y_i = \sum_{k=1}^p a_k \phi_k(x_i) + \epsilon \equiv A f(x_i) + \epsilon \quad (1)$$

where ϕ is a basis function and a is a weights vector which determines the relevance of each basis function.

- Using a matrix representation, the problem is described as:

$$A := \arg \min_A \{ \|A F - Y\|^2 + R(A) \} \quad (2)$$

- To solve this for multi-dimensional target spaces y , we use Agarwal and Triggs' method, which first initialises A with *ridge regression* and iteratively adjusts vectors of A using a quadratic residual function $R(A)$ that penalises columns of A with large norm.

Weights Matrices

In contrast to Tipping's original RVM, Agarwal & Triggs' method use columnwise priors, leading to a sparse set that allows feature selection or sample selection, depending on the kernels used.

Experiments and Results

Number of Relevance Vectors

Feature Selection

Quantitative Results with Synthetic Images

Open-close data set

Complex data set

Summary of the Quantitative Results

# Views	Data Set	Kernel	Ftrs.	Spls.	Avg. Error	STD	Worst Result	Which dof
1	Open-Close	Linear	3	273	8.6	6.9	23.2	θ_z
		Gaussian Both	90	10	5.6	4.3	14.5	T IP
	Complex	Linear	31	839	3.0	2.7	11.4	M DIP
		Gaussian Both	90	42	2.9	2.5	9.9	M DIP
3	Open-Close	Linear	31	839	2.5	2.1	8.9	M DIP
		Gaussian both	90	10	3.6	2.7	14.9	T IP
	Complex	Linear	31	839	2.5	2.1	8.9	M DIP
		Gaussian Both	90	41	2.4	2.0	8.3	M DIP

Qualitative Results with Real Images

Using all Features and Samples with Gaussian Kernels

Using 32 Features and 38 Samples

Time Measurements

Training Phase

Kernel	Input Dim.	Selected RVs	Iters.	Total Time (s)
Linear	90	32	4	6.7
Gaussian	90	38	4	327.7
Both	32	38	5	305.4

Application

Kernel	Features	Training Samples	Time (μ s)
Linear	32	1679	7.2
Gaussian	90	38	35.7
Both	32	38	25.4

Conclusion

The proposed method to combine multiple views for regression improves the results for estimation of hand pose in 3D. This method also increases the sparsity of the regressor.

Acknowledgements: This work was supported by the Brazilian government CAPES foundation and by EPSRC (UK).

Contact: [teo,dwm]@robots.ox.ac.uk