Regressing Local to Global Shape Properties for Online Segmentation and Tracking

Carl Yuheng Ren
carl@robots.ox.ac.uk
Victor Adrian Prisacariu
victor@robots.ox.ac.uk
Ian Reid
ian@robots.ox.ac.uk

Department of Engineering Science,
Oxford University

We propose a regression based learning framework that learns a set of shapes online, which can then be used to recover occluded object shapes. We represent shapes using their 2D discrete cosine transforms (DCT), and the key insight we propose is to regress low frequency harmonics, which represent the global properties of the shape, from high frequency harmonics, that encode the details of the object's shape. We learn the regression model using Locally Weighted Projection Regression (LWPR) which expedites online, incremental learning. After sufficient observation of a set of unoccluded shapes, the learned model can detect occlusion and recover the full shapes from the occluded ones.

Our shape regression method is linked to the pixel-wise posteriors (PWP) level set-based tracker of [1]. The PWP tracker obtains the target pose (a 6 DoF 2D affinity or 4 DoF 2D similarity transform) and figure/ground segmentation at each frame. We use the pose to align the shapes and then add them to the learning framework. After a burn-in period, the framework is able to recover occluded shapes at real time. We demonstrate the ideas using PWP tracker, however, the framework could be embedded in any segmentation-based tracking system.

We use the DCT to represent a silhouette mask image (i.e. a binary image of the figure/ground segmentation, with 1 for foreground and -1 for background), so that the shape representation becomes a set of DCT coefficients. The transform yields a natural hierarchical representation of a shape in which the top-left, low frequency coefficients in the DCT capture the overall shape, while the high frequency coefficients (further away from top-left) capture the details of the shape.

We use Locally Weighted Projection Regression (LWPR) [3] as our regression model. LWPR is based on the hypothesis that high dimensional data are characterized by locally low dimensional distribution. A learned LWPR has K local models, each comprising a Receptive Field (RF) characterized by a field center $\mathbf{c}_k$ and a positive semi-definite distance metric $\mathbf{D}_k$ that determines the size and shape of the neighborhood contributing to the local model; and a locally weighted partial least square (LWPLS) regression model characterized by a set of projections $\mathbf{u}_k$ and respective their weights β_k. Given a set of high frequency DCT coefficients as input $\mathbf{x}^{hf}$, the RF weight, also known as the activation, of the k^{th} local model is computed as:

$$w_k = exp(-\frac{1}{2}(\mathbf{x}^{hf}-\mathbf{c}_k)^T\mathbf{D}_k(\mathbf{x}^{hf}-\mathbf{c}_k)) \quad (1)$$

Given an input high frequency DCT coefficient vector $\mathbf{x}^{hf}$, every linear model calculates a prediction of low frequency DCT coefficient vector $\hat{\mathbf{x}}_k^{lf}(\mathbf{x}^{hf})$ (as is described in Table 3 [3]). The final output (i.e. a set of low frequency DCT coefficients) is given by the weighted mean of all K local outputs:

$$\hat{\mathbf{x}}^{lf} = \frac{\sum_{k=1}^{K} w_k \hat{\mathbf{x}}_k^{lf}}{\sum_{k=1}^{K} w_k} \quad (2)$$

The whole LWPR-DCT learning algorithm is outlined in Table 1. $w_{gen} \leq 1$ is a threshold that determines when to create a new RF: the closer w_{gen} is set to 1, the more overlap local models will have. $\mathbf{D}_{def}$ is the initial distance metric in Equation 1, which controls the shape of the RF and is adapted during learning. The details of updating distance metric and local models are lengthy so the reader is referred to [3]. The learning algorithm also has a simple mechanism to determine when to add a new projection to current local model, by recursively keeping track of the mean-square error (MSE), as a function of the number of projections in a local model. In the 'burn-in' period of our method (when we assume the shapes adopted by a object are clear and unoccluded and proper aligned by the PWP tracker [1]), we transform the observed shape into high frequency and low frequency DCT coefficients $(\mathbf{x}^{hf}, \mathbf{x}^{lf})$, and train LWPR on this sequence of observations.

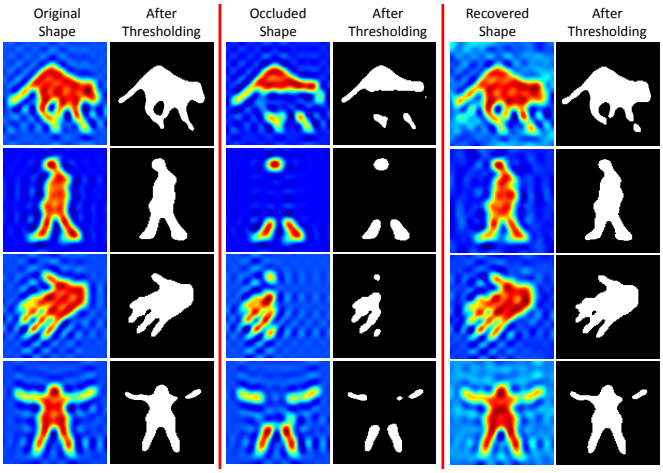

Figure 1: Examples of recovered shapes from artificial occlusion.

- Initialize the LWPR with no receptive field.
- For each training shape Φ
  - compute its compute it's $1 \sim N$ DCT coefficients as low frequency harmonics $\mathbf{x}^{lf}$ and $N+1 \sim M$ DCT coefficients as high frequency harmonics $\mathbf{x}^{lf}$.
  - For the k^{th} out of K existing receptive fields:
    - Calculate the activation using Equation 1.
    - Update $\mathbf{u}_k$ and β_k of the k^{th} LWPLS according to Table 3 in [3].
    - Update the distance metric $\mathbf{D}_k$ according to Table 4 in [3].
    - Check the decreasing rate of MSE at each projection to see if the number of projections needs to be increased.
  - If no RF was activated by more than w_{gen}:
    - Create a new RF with initial number of projections $R = 2$, RF center with $\mathbf{c}_{K+1} = \mathbf{x}^{hf}$ and $\mathbf{D}_{K+1} = \mathbf{D}_{def}$, $K \leftarrow K+1$.

Table 1: Pseudo code for the learning part of the LWPR-DCT algorithm.

When a shape is observed, we first compute the activation using Equation 1. We assume that we are observing a previously unseen shape if none of the existing RFs is activated by more than w_{gen} and proceed no further. Activation of any RF in the current LWPR model indicates that the high frequency details of the current shape have been observed before. The system then makes a prediction of the low frequency components for the shape and calculates the difference between the observation and prediction. If the mean square error (MSE) between the observation and the prediction is larger than twice the MSE in the training data (empirically defined threshold), we consider the shape as being known but occluded and update it according to our prediction.

In the paper we showcase the efficacy of our shape regression method to occlusion detection and recovery with a set of comprehensive experiments on both real and artificial occlusion videos (Figure 1 shows examples of recovering occlusion). We also compare our method to the shape prior method of [2] to show our method's efficiency.

[1] Charles Bibby and Ian Reid. Robust real-time visual tracking using pixel-wise posteriors. In *ECCV 2008*, pages 831–844, 2008.

[2] Victor Prisacariu and Ian Reid. Shared shape spaces. In *ICCV 2011*, 2011.

[3] Sethu Vijayakumar, Aaron D'Souza, and Stefan Schaal. Incremental online learning in high dimensions. *NECO*, 17:2602–2634, 2005.