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
University of Oxford
Parks Road, Oxford, UK

Abstract

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, 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. We demonstrate the ideas using a level-set based tracking system that provides shape and pose, however, the framework could be embedded in any segmentation-based tracking system. Our experiments demonstrate the efficacy of the method on a variety of objects using both real data and artificial data.

1 Introduction

In recent years, there has been substantial research in segmentation-based tracking [1, 4, 15]. These methods extract an active contour at each frame (often using level sets [6]) and use it to update the shape of a tracked object. This process results in the efficient tracking of previously unseen objects. However, a challenge for these systems is occlusion, because, unless the shape is constrained in some way, the resulting contour will have an incorrect shape. Our aim in this paper, then, is to show how to learn the set of legal shapes of a potentially deformable object incrementally, online, and then how to use this learned model to detect occlusion and recover the original shape of the object at each frame. We focus on level set-based segmentation but the concepts could be applied to any other types of segmentation.

A typical solution to recover the complete shape in the presence of occlusion is to put constraints on the minimization of the level set energy function. Such methods roughly fall into two categories: the first category comprises methods which try to learn the space of legal shapes by learning either a space of embedding functions (*e.g.* [8, 12]) or a space of

BMVC 2011 http://dx.doi.org/10.5244/C.25.11

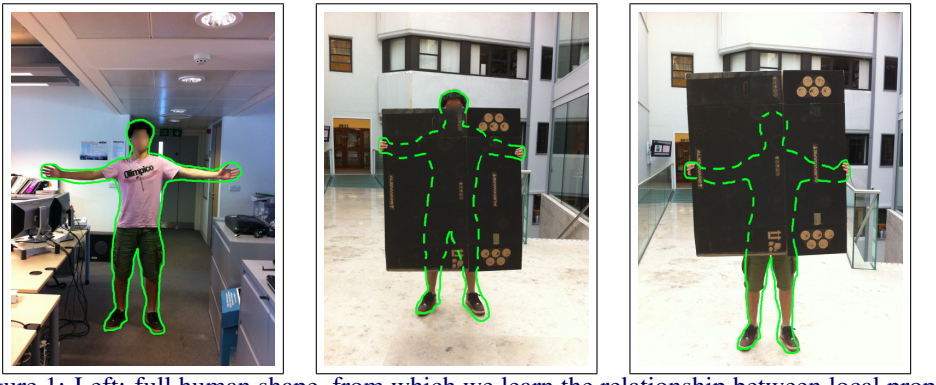

Figure 1: Left: full human shape, from which we learn the relationship between local properties and global ones. Middle & Right: when occlusion happens, we can reconstruct the global shape from observed local properties based on the learnt relationship. (This an illustrative example of our idea, for real examples, please refer to Figure 5 and 6)

contours (i.e. the zero level-sets of the embedding functions), *e.g.* [7]. [12] applies principle component analysis (PCA) to the set of embedding functions to determine a low dimensional space of embedding functions while [7] represents the explicit contours with elliptic Fourier descriptors and uses Gaussian Process Latent Variable Models (GP-LVM) [8] to achieve the dimensionality reduction. [8] combines the two approaches, by using GP-LVM to learn spaces of embedding functions compressed with the discrete cosine transform (essentially a 2D Fourier transform). Their method thus allows for holes to be correctly modeled and still provides a tractable way to learn and use nonlinear shape spaces. In the above-mentioned methods, the level set energy function is minimized w.r.t. the position in the learned lower dimensional space. While these methods are indeed robust to occlusions (since the evolution of the contour is limited to the space of possible shapes), none of them explicitly consider occlusion modelling or recovery. Furthermore, all of them are designed for offline training. When new contours are added, the model must be re-trained.

The second school of methods attempts to control the shape in the current frame by comparing it with a number of recently observed shapes. [4] incrementally builds a dynamic space of good shape hypotheses from frames up to the current one. The shape of the current frame is thus constrained by minimizing its distance from a locally Gaussian weighted shape expectation of the learned space. By continually updating a weight matrix, this method can incrementally update the space of good shapes without re-training. However, in practice, both the size of the weight matrix and the time it takes to update it grows as n^2 (where n is the number of observed good shapes), and, in order to keep track of this matrix, all previously observed shapes need to be stored. Alternatively, by using a fixed size weight matrix, the method suffers from rapid forgetting. The authors also note that this method is very slow, making it unsuitable for real-time operation. Another method, [15], embeds a dense level set in the shape, with the background area set to zero. A variance for each grid point on the level set is modeled with a single Gaussian, which is updated only where no occlusion is present. Once occlusion is detected (using area and distance heuristics), the method uses the Gaussian model on each grid point to cast an expansion force on the level set, to recover the missing parts. However, the correct updating rate is difficult to tune when the shape of a deformable object is learned: updating too fast will result in recovering the current shape simply based on the previous shape, while updating too slowly will suffer from large uncertainty.

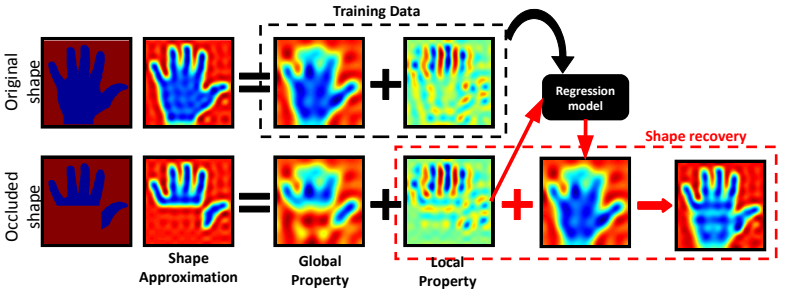

Figure 2: Overview of our shape recovery algorithm.

In this work, we consider the problem of occlusion detection and shape recovery using a different approach, by modeling the relationship between the local and global properties of shape. The motivation behind our idea is illustrated in Figure 1, where we show an occluded human (with only legs visible). Even though the bulk of the person is occluded, a human observer can reconstruct the shape (i.e. the global property) from the relationship between the hands, arms, legs etc. (i.e. the local properties). In this paper, we describe a method to formalize this insight by learning the relationship between the local and global properties. Specially, we show how Locally Weighted Projection Regression (LWPR) can be used to learn a regression from the high frequency harmonics to the low frequency ones of a shape, and how this regression can be used to detect and recover occlusions on previously seen shapes. We link our shape regression to the pixel-wise posteriors (PWP) level set-based tracker of [10]. 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, as they are received. After a burn-in period, the framework is able to recover occluded shapes at real time.

The remainder of the paper is structured as follows: we begin in Section 2 by discussing the discrete cosine transform shape representation and its advantages. Section 3 gives details of the LWPR algorithm and describes how we detect occlusion, discriminate between occlusion and a new shape, and recover occluded shapes. We show qualitative and quantitative evaluations of our method in Section 4, and conclude in Section 5.

2 Shape representation via DCT

The 2D discrete cosine transform (DCT) [14] is a special case of the discrete Fourier transform, which represents an image using a series of orthogonal cosine basis functions (harmonics), each with its own frequency and amplitude. A common use for the DCT is image compression, it being the basis for the JPEG format. Similarly, [8] used it to compress level set embedding functions. Our work is based on a different property of the DCT, namely the fact that the low frequency harmonics contain the coarse bulk properties of the information in the signal, while high frequency ones contain the "details". When applied to shapes, this means that, often, when an object is occluded, parts of its main body may be missing, but many high frequency details remain. Our experiments suggest that occlusions introduce relatively minor changes to the high frequency DCT coefficients. Based on this observation, we train a regression model from higher frequency harmonics to lower harmonics using previously observed complete shapes. Thus, when an occluded shape is observed, we compute its high-frequency harmonics and use the regressor to determine the expected low frequency

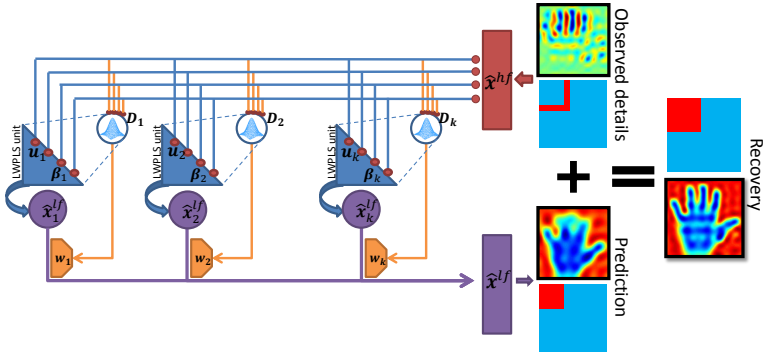

Figure 3: Structure and work flow of LWPR, inspired by Figure 3 in [13].

harmonics, and hence recover the whole shape by adding the low frequency harmonics to the high frequency ones. Figure 2 gives an overview of our framework.

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. Taking the inverse transform of only the first N harmonics and thresholding at zero yields an approximation for the silhouette.

3 Incremental online learning

In this work we aim to recover the missing part of a shape directly using an online trained regression (as opposed to learning a shape space), from the high frequency DCT coefficients to the low frequency ones. We therefore need to learn an incremental approximation of a highly nonlinear and high dimensional function. Gaussian Process Regression [9] or Support Vector Machine Regression [11] are both well established methods that fit non-linear functions globally, but they are not the most suitable solutions for online learning in high dimensional spaces. First, they require a priori determination of the right basis or kernel functions. Second, both methods are developed primarily for offline batch training, rather than for incremental learning, making the addition of a new point computationally expensive.

Instead, we use Locally Weighted Projection Regression (LWPR) [13] as our regression model. LWPR is a nonlinear function approximator that learns rapidly from incrementally acquired data, without needing to store the training data. The computational complexity grows linearly with the number of inputs. LWPR can also deal with a large number of possibly redundant inputs, which is often the case when tracking rigid objects.

Figure 3 shows the workflow of LWPR. 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

- 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 [13].
    - Update the distance metric $\mathbf{D}_k$ according to Table 4 in [13].
    - 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 algorithm.

as the activation, of the k^{th} local model is computed as:

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

Given an input vector $\mathbf{x}^{hf}$, every linear model calculates a prediction $\hat{\mathbf{x}}_k^{lf}(\mathbf{x}^{hf})$ (as is described in Table 3 [13]). The final output (i.e. a set of low frequency DCT coefficients) is given by the weighted mean of all K local outputs:

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

The LWPR 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 [13]. 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 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.

The occlusion detection and shape recovery mechanism of LWPR operates as follows: 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. For an occluded, known (i.e. previously learnt) shape, we expect agreement between prediction and measurement. So 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.

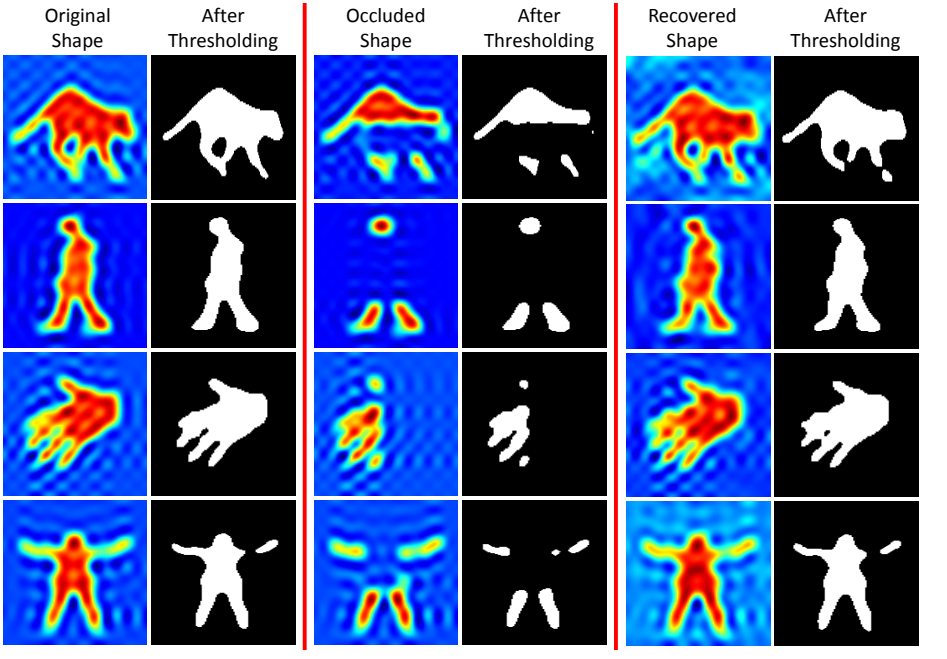

Figure 4: Examples of recovered shapes from artificially occluded images. The left column of each pair shows false color images(blue=-1, red=1) of the inverse "truncated DCT" (ie. the approximation of the silhouette via the first 10 to 15 harmonics), while the right column shows the silhouette obtained from thresholding the approximation at zero. From top row to bottom row: *Cat running*, *Man walking* , *Hand*, *Woman jumping*

4 Experiments and performance analysis

We tested our method both qualitatively and quantitatively, on several video sequences and data sets. We used an Intel Core i7-870 (2.93GHz) machine to run all our experiments. We denote our method with *LWPR-DCT*. We begin with the qualitative analysis.

Examples of successful shape recovery using artificially generated occlusions are shown in Figure 4. The results show that the regression model is capable of recovering the shape in presence of artificially introduced occlusion. We begin with the inverse truncated DCT representation of the silhouette, then show the recovered inverse truncated DCT images and the thresholded forms to compare with original silhouettes. Note that the output silhouettes match the original ones, demonstrating that the regression is recovering the low frequency harmonics well. In Figures 5 and 6, we compare our algorithm to the standard pixel-wise posteriors tracker of [1] on real video sequences and show that we are able to successfully recover the correct contour, in spite of heavy occlusions. In the first 2 frames of Figure 5 there are no occlusions, so both our method and the standard PWP tracker yield similar results. When the hand is occluded, in the other 4 frames, the PWP segmentation is corrupted, while ours is still correct. Similarly for Figure 6.

We show two failure cases of our method in Figure 7. LWPR-DCT can fail in two ways: (i) when too many noisy high frequency harmonics are introduced by the partial occlusion, as is shown in the upper row of Figure 7; (ii) when too much detail is occluded, as is shown in the lower row of Figure 7.

We designed two sets of experiments to evaluate the performance of our LWPR-DCT

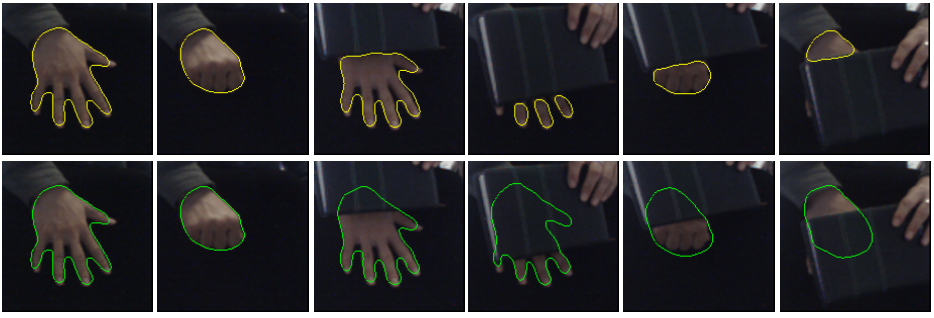

Figure 5: Example frames from a video tracking a hand, comparing our method to the PWP tracker of [11]. When no occlusions are present, both method produce similar results. However, as soon as the hand is occluded, the PWP tracker produces an incorrect segmentation, while our method still generates correct contours.

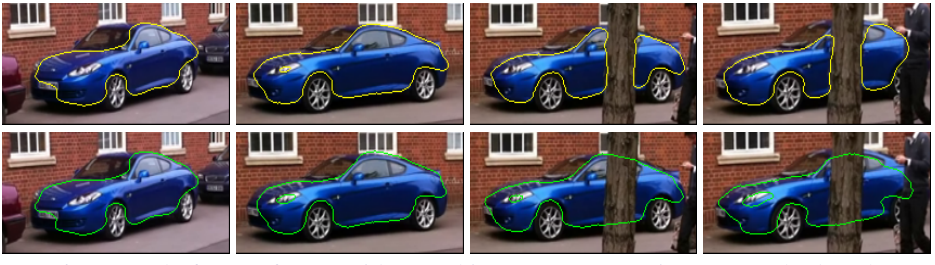

Figure 6: Example frames from a video tracking a car, comparing our method to the PWP tracker of [11]. When the car is not occluded both methods produce similar results. When the tree is in front of the car the segmentation produced by the PWP tracker is corrupted, while the one produced by our tracker is not.

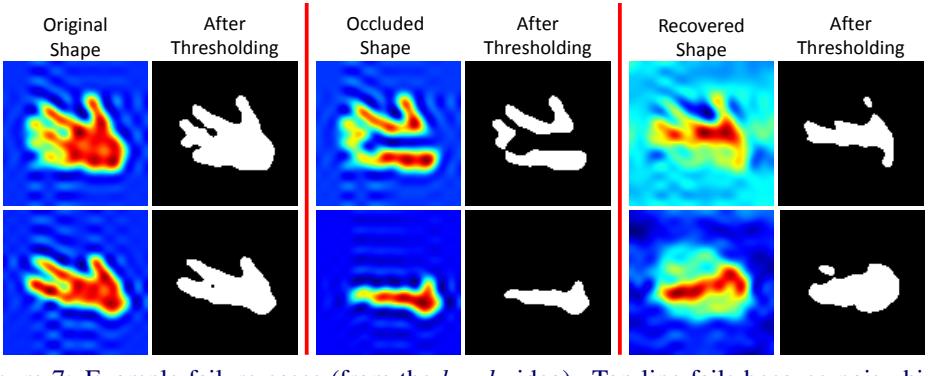

Figure 7: Example failure cases (from the *hand* video). Top line fails because noisy high frequency harmonics are introduced, while bottom line fails because details are missing.

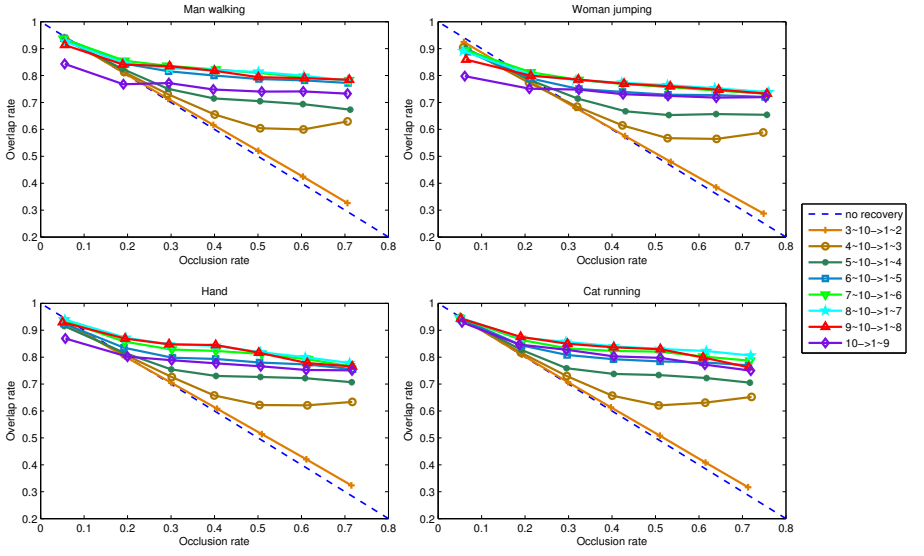

Figure 8: Shape recovery performance evaluation of LWPR-DCT on four datasets using different number of harmonics as input and output.

framework quantitatively: first we measure on the effectiveness of the shape recovery using LWPR-DCT and show how many high frequency harmonics should be used as input for occlusion recovery. Then we compare our algorithm with a state-of-the-art shape prior based method of [8] (denoted by GPLVM-DCT) on the performance of occlusion recovery and average processing time.

We used 4 datasets to evaluate the effectiveness of LWPR-DCT: *Cat running* (artificial video with few distinct poses, 398 frames), *Woman jumping* (real video with an average number of distinct poses, 410 frames), *Man walking* (real video with many distinct poses, 411 frames, the subject 2 walk of the HumanEva I dataset [10])) and *Hand* (real video with many distinct poses, 408 frames). For each video, all frames are segmented and aligned using the PWP tracker, then added to LWPR-DCT as training data. Then we add different sizes of artificial occlusions (where each occlusion is rectangular and in a random location) to each frame. We chose to to generate occlusions artificially both in order to control the percentage of occlusion and to know ground ground truth. For each frame, we generate 7 levels of occlusion, ranging from 0.1 (10%) to 0.8 (80%). We use the overlap rate $R = \frac{S_{gt} \bigcap S_{rcv}}{S_{gt} \bigcup S_{rcv}}$ as our performance criteria, where S_{gt} is the ground truth shape and S_{rcv} is the recovered shape. We use the first 10 harmonics to approximate the segmented shapes and run tests on all possible combinations of the numbers of input and output harmonics (harmonic 10 generating 1 to 9, 9 and 10 generating 1 to 8, etc.). Figure 8 shows the results. Our method gives reasonable results just by using the 10^{th} harmonic to regress all $1\sim9$ harmonics. Using the harmonics $8\sim10$ to regress harmonics $1\sim7$ gives the best performance in all cases. Performance decreases as we increase the number of known harmonics, since small occlusions introduce extra details, most of which are captured in the highest frequency harmonics.

In the second quantitative experiment, we compare our algorithm to the shape prior method of [8], which generates embedding functions from a 2 dimensional GPLVM latent space. Here, segmentation (i.e. the recovering of the unoccluded shape), is an iterative nonlinear minimization in the learned latent space. In our experiment, for each occluded shape,

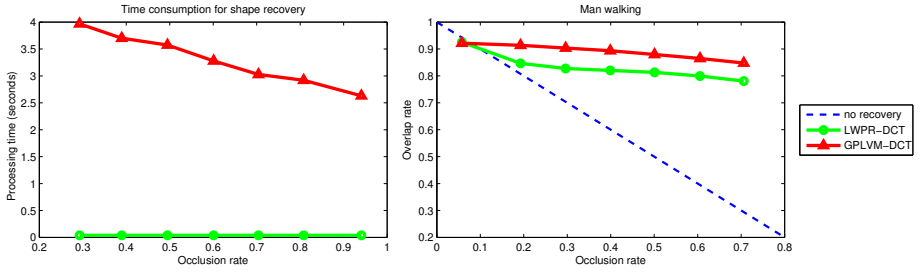

Figure 9: Comparing LWPR-DCT to GPLVM-DCT on processing time (left) and occlusion recovery performance (right).

we run three separate minimizations, we compute the recovery rate for each resulting shape, and we take an average of those values. We run multiple minimizations (rather than a single one) because each one can converge to a different shape, so to accurately measure the performance of [8] on our test data we need to consider all these results. As starting points for the minimization, we use the three points that generate the shapes most similar to the ground truth from the previous frame. We run both methods on the training and testing data from the *man walking* sequence from last experiment. Figure 9 shows the time consumption and recovery rate of both methods. As a well trained, shape prior based method, GPLVM-DCT outperforms our method by an average of 10%. But, as is shown in the timings chart, the time consumption for LWPR-DCT stays constant at around 35ms per shape, while the processing time required by GPLVM-DCT increases with the occlusion rate and it is much larger than LWPR-DCT (up to 114 times higher). This happens because, when using LWPR-DCT, each shape recovery is a single (closed form) regression, while, in the GPLVM-DCT case, segmentation is an iterative process with the number of iterations being proportional to the percentage of occlusion in the image. Note that the GPLVM-DCT timings shown in Figure 9 are for a single mode search. Since we use three such searches, the actual processing time per frame it three times as large. In this experiment we used the harmonics 8∼10 to regress the other 1∼7 harmonics.

5 Conclusions

In this paper, we have presented a novel regression based framework for online shape learning and recovery. Shapes are represented by discrete cosine transform harmonics and the set of object shapes is modeled by a regression from the high frequency harmonics to the low frequency harmonics. Our method incrementally learns a shape model for an observed object and detects/recovers occlusions at real time. We have integrated our method with a level-set based tracker, but it could be potentially linked to other types of segmentation and tracking.

Our method currently has two limitations. First, the DCT representation of shape is rotation and scaling sensitive, i.e small rotation of a shape will make the high frequency coefficient change greatly, resulting in very different prediction results. Currently we are relying on the PWP tracker (which obtains camera pose and segmentation at each frame) to align the shapes. Secondly, some special types of occlusion are very difficult for LWPR-DCT to handle: 1) when noisy high frequency components are introduced by small occlusion and 2) when the details of the shape are occluded. In these two cases, LWPR-DCT might give incorrect predictions, while shape prior based methods would be more applicable.

While we have demonstrated the value of LWPR for shape recovery under occlusion, we believe that this general idea has wider application. For example, we could consider regressing local appearance to global positions, which would have similarity to [2] and [3], or more ambitiously regress local appearance to global appearance.

References

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

[2] Matthew B. Blaschko and Christoph H. Lampert. Learning to localize objects with structured output regression. In *ECCV 2008*, pages 2–15, 2008.

[3] Mario Fritz, Bastian Leibe, Barbara Caputo, and Bernt Schiele. Integrating representative and discriminant models for object category detection. In *ICCV 2008*, pages 1363–1370, 2005.

[4] Majid Mirmehdi John Chiverton and Xianghua Xie. On-line learning of shape information for object segmentation and tracking. In *BMVC 2009*, 2009.

[5] Neil Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *JMLR*, 6:1783–1816, 2005.

[6] Stanley Osher and James A Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations. *JCP*, 79(1):12–49, 1988.

[7] Victor Prisacariu and Ian Reid. Nonlinear shape manifolds as shape priors in level set segmentation and tracking. In *CVPR 2010*, 2010.

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

[9] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[10] Leonid Sigal, Alexandru Balan, and Michael Black. HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. *IJCV*, 87:4–27, 2010.

[11] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression, 2004.

[12] Andy Tsai, Anthony Yezzi, William Wells, Clare Tempany, Dewey Tucker, Ayres Fan, Eric Grimson, and Alan Willsky. A shape-based approach to the segmentation of medical imagery using level sets. *T-MI*, 22(2):137–154, 2003.

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

[14] Andrew B. Watson. Image compression using the discrete cosine transform. *Mathematica Journal*, 4:81–88, 1994.

[15] Alper Yilmaz, Xin Li, and Mubarak Shah. Contour-based object tracking with occlusion handling in video acquired using mobile cameras. *T-PAMI*, 26(11):1531–1536, 2004.