On-line Learning of Shape Information for Object Segmentation and Tracking

John Chiverton
jpchiverton@gmail.com
School of Information Technology,
Mae Fah Luang University, Thailand

Majid Mirmehdi
majid@cs.bris.ac.uk
Dept. Computer Science,
University of Bristol, UK

Xianghua Xie
x.xie@swansea.ac.uk
Dept. Computer Science,
Swansea University, Wales

Abstract

We present segmentation and tracking of deformable objects using non-linear on-line learning of high-level shape information in the form of a level set function. The emphasis is for successful tracking of objects that undergo smooth arbitrary deformations, but without the *a priori* learning of shape constraints. The high-level shape information is learnt on-line by defining a memory of object samples in a high-dimensional shape space. These shape samples are then used as weights via a locally defined shape space kernel function to define a template against which potential future shapes of the tracked object can be compared. Results for the successful tracking of a range of deformable motions are presented.

1 Introduction

Robust tracking of the contour of a moving object is made complex by numerous factors, such as the possibility of poor image contrast, arbitrary deformations e.g. for articulated objects, and 2D silhouette representations of 3D objects. To deal with such difficulties, object contour tracking is often assisted with the use of *a priori* learnt information, such as the shape of the object being tracked, which can reduce potential ambiguities and provide realistic object contour hypotheses. However, shape learning is inherently a pre-processing stage which is usually cantankerous, requiring supervised and often manual preparation.

An ideal medium for shape modelling is the active contour model which has been extensively investigated in conjunction with prior shape knowledge since at least [1, 2]. These spline based approaches are limited by topological constraints unlike level set based active contour approaches, e.g. [16]. PCA is often used in these techniques to compress and summarise the important components of a set of characteristic level sets [16, 23] or control points modelled using Active Shape Modelling [4].

These approaches typically assume the shape space (or some other representation) can be conveniently represented by a multivariate Gaussian density. This is often not a realistic assumption, particularly for shapes representative of 2D image based projections (i.e. silhouettes) of (articulated) 3D objects. In [7, 22], shape spaces were modelled non-linearly using Gaussian kernels, and e.g. in [8, 9], followed by extensive non-linear modelling of shape

BMVC 2009 doi:10.5244/C.23.21

Figure 1: Tracking results obtained for a *moving observer* where many frames possess similar shapes to those in other frames.

priors. The non-linear shape space can be modelled linearly when considering a dynamical model of the shape sequences as shown by [5]. Simpler models that do not impose sequential dependence may however allow for more diverse shape configurations, e.g. [22].

Many prior shape based tracking methods have been demonstrated to be quite robust, providing accurate outlines of the shape of the object being tracked rather than, e.g. a box around the object, see e.g. [5, 9, 21] and cf. e.g. [12, 14]. However, preparation of extensive prior shape knowledge is not always convenient and even cumbersome. Furthermore, many methods can encounter difficulties if the tracked object is protean and can not be easily approximated by the current reduced dimensional shape space - a realistic prospect for articulated objects and their 2D image projections. Two promising alternatives are available, Yilmaz et al. [25] proposed a tracking method that adapts to previously unseen shapes, but only utilised shape information when an occlusion was detected. Yezzi and Soatto [24] described a framework that used a moving average of the shape information without reference to shapes seen in much earlier frames that may otherwise have provided useful information for much later frames. The work described here also provides an extensive framework for tracking of shapes in video data but it places much more emphasis on past shape modelling defined and learnt on-line, simultaneous to the tracking process.

We are interested in segmentation and tracking using high-level shape information, particularly for objects that undergo arbitrary and smooth deformations (as illustrated in Fig. 1), but without the *a priori* learning of shape constraints. We introduce a shape based level set active contour framework that learns shape information on-line, simultaneously combining the newly learnt shape information into a probabilistic non-linear shape space via a localising kernel function. The approach described here requires, for the first frame, a broadly defined object outline and then object tracking is able to commence. Alternatively, the framework could be initialised via a bootstrap approach which detects foreground objects with independent motions from the dominant background, e.g. using the approach in [13].

The proposed method is described in detail in Section 2. The framework is assessed both qualitatively and quantitatively in Section 3. Conclusions then follow in Section 4.

2 Methodology

The model proposed here comprises three main parts: photometric image model, shape model, and implicit contour position re-estimation. Section 2.1 describes the main components of the overall probabilistic estimation process and includes a description of the pixel level, probabilistic, photometric model. This image model can be considered as performing *model based photometric competition* because the two image regions, foreground and background, are both represented by probabilistic models that are combined with the currently estimated photometric information to compete against each other. The model of the fore-

ground region shape is described in Section 2.2 where a probabilistic shape space is used to model the distribution of learnt tracked object shapes which are locally weighted via a kernel function in order to derive shape templates for future shape hypotheses during tracking. Section 2.3 then describes how, for every new frame, the position of the contour is estimated from the available photometric information using the optical flow constraint along the implicitly defined object boundary, weighted by confidence terms. This allows the active contour evolution to be initialised close to the true object boundary with the start of every new frame. Section 2.4 summarises the gradient descent level set based optimisation process.

2.1 Image model

Let $I^j_{\mathrm{x}} : \mathbb{R}^2 \times \mathbb{R}^+ \rightarrow \mathbb{R}^n$ be an n dimensional image intensity at pixel $\mathrm{x} \in \mathbb{R}^2$ in image frame $j \in \mathbb{R}^+$ where e.g. $n = 3$ for RGB colour images. Also, let a corresponding mask (for each image frame j) consist of foreground $\mathfrak{F}^j = \{\mathrm{x}|\mathfrak{f}^j_{\mathrm{x}}\}$ and background $\mathfrak{B}^j = \{\mathrm{x}|\mathfrak{b}^j_{\mathrm{x}}\}$ pixels only, with $\Omega = \mathfrak{F}^j \cup \mathfrak{B}^j$ the set of pixels in the image space and $\mathfrak{f}^j_{\mathrm{x}} \in \{0,1\}$, $\mathfrak{b}^j_{\mathrm{x}} = 1 - \mathfrak{f}^j_{\mathrm{x}}$ binary foreground and background labels, respectively. The foreground and background regions define a partition of the image space $q^j = \{\mathfrak{F}^j, \mathfrak{B}^j\}$. Furthermore, a high-level description of shape $g(\mathfrak{F}^j, \mathfrak{B}^j)$ is also considered here which is synonymous to the image space partition and is explained later in Section 2.2 where we consider the shape contribution to the model.

Considering all image frames j up to a current frame i, i.e. $\forall j,\ j = 0...i$ then we can use the Bayes theorem to calculate a conditional probability density $p(Q|I)$ for a set of image partitions Q given a set of image intensities I,

$$p\left(Q = \{q^j | \forall\ j,\ j = 0...i\} | I = \{I^j | \forall\ j,\ j = 0...i\}\right) = \frac{p(I,Q)}{p(I)}, \tag{1}$$

where $p(I)$ is the marginal data probability density which is not dependent on the image partition information and can therefore be ignored for the purposes of optimisation. $p(I,Q)$ is a joint probability density which is expanded with the assumption of a Markov first order dependence of the image data, thus $p(Q|I) \propto p(Q^{i-1}|q^i)p(q^i)p(I^0|q^0)\prod_{j=1}^{i} p(I^j|q^j, I^{j-1}, q^{j-1})$. $p(I^0|q^0)$ is the initial frame's data likelihood term, $p(q^i)$ is the prior probability density of the current frame partition, and $p(Q^{i-1}|q^i)$ is the probability density of all image partitions, except the current frame, i.e. $Q^{i-1} = \{q^j | j = 0...i-1\}$, given the current frame partition q^i. The partition estimates are treated as random variables and the past partition estimates are conditioned on the present indicating the potential for re-estimation, although this is not used here. Note that future estimates are not part of the formulation, which would otherwise produce a problematic formulation. The terms $p(Q^{i-1}|q^i)$ and $p(q^i)$ relate to shape and contour labelling smoothness respectively. The data likelihood term $p(I^j|q^j, I^{j-1}, q^{j-1})$ can then be divided into foreground and background terms (assuming conditional independent pixel intensities via the partition terms), hence (1) becomes

$$p(Q|I) \propto p(Q^{i-1}|q^i)p(q^i)p(I^0|q^0)\prod_{j=1}^{i}\prod_{\forall \mathrm{x} \in \Omega}\left[p_{\mathfrak{F}}(I^j_{\mathrm{x}}|q^j, \mathfrak{m}^j_{\mathfrak{F}})^{\mathfrak{f}^j_{\mathrm{x}}} p_{\mathfrak{B}}(I^j_{\mathrm{x}}|q^j, \mathfrak{m}^j_{\mathfrak{B}})^{\mathfrak{b}^j_{\mathrm{x}}}\right] \tag{2}$$

where $\mathfrak{m}^j_{\mathfrak{F}} = \{I^{j-1}_{\mathfrak{F}}, q^{j-1}\}$ and $\mathfrak{m}^j_{\mathfrak{B}} = \{I^{j-1}_{\mathfrak{B}}, q^{j-1}\}$ and the powers $\mathfrak{f}^j_{\mathrm{x}}$ and $\mathfrak{b}^j_{\mathrm{x}}$ act as mutually exclusive switches between the foreground and background. $p_{\mathfrak{F}}(.)$ and $p_{\mathfrak{B}}(.)$ in (2) thus correspond to two different PDFs for the foreground $I^j_{\mathfrak{F}} = \{I_{\mathrm{x}}|\mathrm{x} \in \mathfrak{F}^j\}$ and background $I^j_{\mathfrak{B}} =$

$\{I_{\mathrm{x}}|\mathrm{x}\in\mathfrak{B}^j\}$ image intensities respectively. The initial frame's data likelihood can also be similarly expanded, however, to save space the non-expanded form will be retained.

A smooth labelling and a smooth boundary (defined as synonymous to each other here) separating the foreground and background regions are desirable properties of an image space partition for object tracking and segmentation applications. These properties can be achieved by minimising the length of the boundary of the partition q^i. Therefore $p(q^i)\triangleq p(\mathfrak{L})\propto\exp(-\lambda_\kappa\mathfrak{L})$ where λ_κ is an exponential rate parameter and $\mathfrak{L}$ is the length of the contour defining the partition (cf. [6]). Substituting this term into (2) gives

$$p(\mathcal{Q}|I)\propto p(\mathcal{Q}^{i-1}|q^i)p(\mathfrak{L})p(I^0|q^0)\prod_{j=1}^{i}\prod_{\forall\mathrm{x}\in\Omega}\overbrace{\left[p_{\mathfrak{F}}(I^j_{\mathrm{x}}|q^j,\mathfrak{m}^j_{\mathfrak{F}})^{\mathfrak{f}^j_{\mathrm{x}}}p_{\mathfrak{B}}(I^j_{\mathrm{x}}|q^j,\mathfrak{m}^j_{\mathfrak{B}})^{\mathfrak{b}^j_{\mathrm{x}}}\right]}^{\textit{model based photometric competition}}. \tag{3}$$

The foreground and background terms in this model work in competition with each other as indicated. Foreground/background competition is the basis of many active contour techniques, e.g. see [19, 27]. However, the photometric components used here are first order Markovian, i.e. they remember photometric information from the preceding frame via $\mathfrak{m}^j_{\mathfrak{F}}$ and $\mathfrak{m}^j_{\mathfrak{B}}$. We now define the form of the shape model represented in (3) using $p(\mathcal{Q}^{i-1}|q^i)$.

2.2 Shape model

The primary focus of the work here is that each image partition q^j is equivalent to a term describing the shape of the object i.e. $\mathsf{S}^j=g(\mathfrak{F}^j,\mathfrak{B}^j)$. We use the distribution $p(\mathcal{Q}^{i-1}|q^i)$ in (3) to model this shape information. We therefore now describe an approach to on-line estimation of a template shape that can be used for comparison with the currently evolving shape. Shapes from all previous frames similar to the object shape in the preceding frame are used to derive the template via a shape space kernel function. This template shape has useful properties in that it represents the important components of the shape we are tracking and may contain elements of the tracked shape from potential future frames and past frames.

We consider the level set $\phi^j\triangleq\mathsf{S}^j$ as the primary representation of shape information in our model. ϕ^j enables important geometric information to be conveniently incorporated into the modelling process and it can be considered synonymous to the image partition q^j where pixel level labelling information is fully encapsulated by the level set representation. This can be seen from the properties of the level set which include: $\phi^j_{\mathrm{x}_c}=0$ on the coterminous foreground and background regions for contour points x_c and $\phi^j_{\mathrm{x}}=\pm\min|\mathrm{x}-\mathrm{x}_c|\forall\mathrm{x}_c|\phi^j_{\mathrm{x}_c}=0$, i.e. the contour point with minimum Euclidean distance, see e.g. [18]. Also (here in this work) $\phi^j_{\mathrm{x}}\leq 0$ for $\mathrm{x}\in\mathfrak{F}^j$ and $\phi^j_{\mathrm{x}}>0$ for $\mathrm{x}\in\mathfrak{B}^j$.

The shape ϕ^i of the tracked object at the current frame can be controlled via comparisons with a set of shapes Φ^{i-1} from a dynamically built space of good shape hypotheses from previous frames. The comparison of ϕ^i with Φ^{i-1} should be invariant to translation $\mathbf{t}^i_{\mathrm{s}}$, scale s^i_{s} and rotation $\mathbf{R}^i_{\mathrm{s}}$ to enable meaningful comparison resulting in a normalised shape space Ω_{s} representation. Thus, Φ^{i-1} and ϕ^i have equivalent shape space forms given by $\Theta^{i-1}_{\mathrm{x}}=\Phi^{i-1}_{\mathbf{a}^{i-1}_{\mathrm{s}}(\mathrm{x})}$ and $\theta^i_{\mathrm{x}}=\phi^i_{\mathbf{a}^i_{\mathrm{s}}(\mathrm{x})}\ \forall\ \mathrm{x}$ where

$$\mathbf{a}^i_{\mathrm{s}}(\mathrm{x})=s^i_{\mathrm{s}}\mathbf{R}^i_{\mathrm{s}}\mathrm{x}+\mathbf{t}^i_{\mathrm{s}} \tag{4}$$

is the similarity transformation from image space $\mathrm{x}\in\Omega$ to shape space $\mathbf{a}^i_{\mathrm{s}}(\mathrm{x})\in\Omega_{\mathrm{s}}$ for the object shape in frame i. Shape comparisons also have to be in the current image space requiring the inverse transformation of (4), i.e. $\mathbf{a}^i_{\hat{\mathrm{s}}}(\mathrm{x})=s^i_{\hat{\mathrm{s}}}\mathbf{R}^i_{\hat{\mathrm{s}}}\mathrm{x}+\mathbf{t}^i_{\hat{\mathrm{s}}}$ where $\mathbf{a}^i_{\hat{\mathrm{s}}}(\mathrm{x})\in\Omega$.

As the shape information is learnt on-line, without supervision, the resulting estimated shapes will not be perfect representations and can be considered to be inherently noisy. Thus, we define a probabilistic shape space with probability distribution $p_m(\theta^{i-1}|\Theta^{i-1})$ that represents the distribution of the learnt noisy shapes over the normalised shape space $\Theta^{i-1} = \{\theta^j | 0 \leq j \leq i-1\}$ consisting of shapes up to the current frame. We can then define a locally weighted shape space expectation $\bar{\Theta}^{i-1}$ to provide a best estimate over the shape distribution $p_m(\theta^{i-1}|\Theta^{i-1})$ (which acts as a prior) and a local weighting distribution $p_w(\theta|\theta^{i-1})$ (acting as a local shape space kernel). This best estimate can then be used to compare the currently evolving shape rather than a global mean or one based on assumptions on the linearity of the shape space or even one based on local integrity. The local weighting is given here by a Gaussian distribution $p_w(\theta|\theta^{i-1})$ centred on the previous frame's object shape θ^{i-1}, so that

$$\bar{\Theta}^{i-1} = \int_{\Omega} \theta \; p_m(\theta^{i-1}|\Theta^{i-1}) \; p_w(\theta|\theta^{i-1}) \; \mathrm{d}\theta. \tag{5}$$

The normalised shapes θ^j for $j = 0...i-1$ are considered to be distributed according to $p_m(\theta^{i-1}|\Theta^{i-1})$, so that the expectation is approximated here via

$$\bar{\Theta}^{i-1} = \frac{\sum_{j=0}^{i-1} \left[\theta^j \mathbf{W}^{i-1,j}\right]}{\sum_{j=0}^{i-1} \mathbf{W}^{i-1,j}} \tag{6}$$

where previously identified object shapes are θ^j for frame j and the local weighting for θ^{i-1} and θ^j frame shapes is $\mathbf{W}^{i-1,j} = \exp(\frac{1}{|\Omega_{\mathfrak{s}}|}\sum_{\mathrm{x}\in\Omega_{\mathfrak{s}}}(\theta^{i-1}-\theta^j)^2)$. Each weight is an element in a weight matrix that encapsulates the similarity of shapes at different frames. This can be seen by the strong diagonals for the example weight matrices illustrated in the results in Section 3. A simple comparison between ϕ^i and Φ^{i-1} can then be the sum of squared differences in the current image space:

$$\mathscr{C}_{\mathfrak{s}}(\phi^i, \Phi^{i-1}) = \sum_{\mathrm{x}\in\Omega} \left(\phi^i_{\mathrm{x}} - \bar{\Phi}^{i-1}_{\mathbf{a}^i_{\mathfrak{s}}(\mathrm{x})}\right)^2, \tag{7}$$

where $\bar{\Phi}^{i-1}_{\mathbf{a}^i_{\mathfrak{s}}(\mathrm{x})} = \bar{\Theta}^{i-1}_{\mathrm{x}}$, and $\mathbf{a}^i_{\mathfrak{s}}(\mathrm{x})$, defined earlier, transforms the shape space estimate $\bar{\Theta}^{i-1}$ to a current image space estimate $\bar{\Phi}^{i-1}_{\mathbf{a}^i_{\mathfrak{s}}(\mathrm{x})}$. A sum of squared differences calculation implicitly assumes a Gaussian distribution. Therefore, taking the exponential of (7) results in a Gaussian distribution and considering the partition representation then $p(q^i|Q^{i-1})$ can be used to symbolise the distributional form of (7), hence $p(q^i|Q^{i-1}) \propto \exp\left(-\mathscr{C}_{\mathfrak{s}}(\phi^i, \Phi^{i-1})\right)$. Furthermore, as a Gaussian distribution is symmetric about the mean, then we can define $p(q^i|Q^{i-1}) \triangleq p(Q^{i-1}|q^i)$ which is of the form found in (3). Also $p(q^i|Q^{i-1}) = p_{\mathfrak{s}}(q^i|Q^{i-1})$ is in a more intuitive form for inference where $\mathfrak{s}$ is indicative of shape so that (3) can be written as

$$p(Q|I) \propto p(\mathfrak{L}) \prod_{\forall \mathrm{x}\in\Omega} \left[p_{\mathfrak{s}}(q^i_{\mathrm{x}}|Q^{i-1}_{\mathrm{x}}) p(I^0_{\mathrm{x}}|q^0) \prod_{j=1}^{i} \left[p_{\mathfrak{F}}(I^j_{\mathrm{x}}|q^j, \mathfrak{m}^j_{\mathfrak{F}})^{\mathfrak{f}^j_{\mathrm{x}}} p_{\mathfrak{B}}(I^j_{\mathrm{x}}|q^j, \mathfrak{m}^j_{\mathfrak{B}})^{\mathfrak{b}^j_{\mathrm{x}}} \right] \right]. \tag{8}$$

This probabilistic model now incorporates memory based photometric competition terms, a spatial smoothness term and a shape based term. The current image frame partition q^i is estimated here using a gradient descent level set based optimisation process, described later in Section 2.4. Section 2.3 next describes the approach used here to initialise the position of the contour for each new image frame to assist in the optimisation of (8).

2.3 Implicit contour position re-estimation

We track the moving object in the video data by estimating the optical flow $\mathbf{v}_{\mathrm{x}}^{i}$ (using work from [17]) along the implicitly defined boundary in our active contour framework $\mathrm{x}_c|\phi(\mathrm{x}_c) = 0$. We assume that there will be errors associated with the estimated position of the object contour and in the optical flow estimates. We therefore treat the optical flow based tracking process in a probabilistic manner, taking into account two potential sources of error: (a) a prior distribution to model the random error associated with the estimated contour location, modelled here as dependent on the image gradient magnitude, i.e. $p_v(\phi_{\mathrm{x}}^{i}|\,|\nabla I_{\mathrm{x}}^{i}|)$; and (b) a systematic error distribution associated with the estimated contour location $p_c(\mathrm{x}|\phi_{\mathrm{x}}^{i})$. The term $p_v(\phi_{\mathrm{x}}^{i}|\,|\nabla I_{\mathrm{x}}^{i}|)$ also implicitly models errors associated with the optical flow estimates. Thus, the pixel mean movement $\bar{\mathbf{v}}^{i}$ of the object can be estimated taking into account these two distributions (via the Law of the Unconscious Statistician)

$$\bar{\mathbf{v}}^{i} = \int_{\Omega} \mathbf{v}_{\mathrm{x}}^{i} p_c(\mathrm{x}|\phi_{\mathrm{x}}^{i}) p_v(\phi_{\mathrm{x}}^{i}|\,|\nabla I_{\mathrm{x}}^{i}|)\mathrm{dx} \tag{9}$$

where $\mathbf{v}_{\mathrm{x}}^{i}$ is an optical flow estimate at point x. This estimated mean is then used to update the position of the signed distance function for each new image frame $\hat{\phi}_{\mathrm{x}}^{i} = \phi_{\mathrm{x}+\bar{\mathbf{v}}^{i-1}}^{i-1}$. In practise $p_c(\mathrm{x}|\phi^{i})$ is implemented as a binary distribution, i.e. a binary mask in which the mean optical flow estimate is determined on or surrounding the tracked object's estimated boundary weighted by the (spatially normalised) image gradient magnitude using $p_v(\phi_{\mathrm{x}}^{i}|\,|\nabla I_{\mathrm{x}}^{i}|)$.

2.4 Optimisation process

A gradient descent level set based approach is used (see e.g. [3, 20]) to optimise (8) and (4). The optimisation of (8) is made possible via its variational derivative

$$\frac{\partial \phi_{\mathrm{x}}^{i}}{\partial t} = -2\lambda_{\mathfrak{s}}\left(\phi_{\mathrm{x}}^{i} - \bar{\Phi}_{\mathbf{a}_{\mathfrak{s}}^{i}(\mathrm{x})}^{i-1}\right) + \delta_0(\phi_{\mathrm{x}}^{i})\left(\lambda_{\kappa}\mathscr{K}_{\mathrm{x}} - \ln\left(\frac{p_{\mathfrak{F}}(I_{\mathrm{x}}^{j}|q^{j},\mathrm{m}_{\mathfrak{F}}^{j})}{p_{\mathfrak{B}}(I_{\mathrm{x}}^{j}|q^{j},\mathrm{m}_{\mathfrak{B}}^{j})}\right)\right) \tag{10}$$

where $\lambda_{\mathfrak{s}}$ is a shape term parameter which corresponds to the inverse variance of $p_{\mathfrak{s}}(q^{i}|Q^{i-1})$ and $\mathscr{K}_{\mathrm{x}} = -\nabla\cdot(\nabla\phi_{\mathrm{x}}/|\nabla\phi_{\mathrm{x}}|)$ is the curvature of the level set at x. The curvature result follows $p(\mathfrak{L}) \propto \exp(-\lambda_{\kappa}\mathfrak{L})$ (in (8)) and using the definition of length defined in [3], i.e. $\mathfrak{L} \triangleq \int_{\Omega}|\nabla H(\phi_{\mathrm{x}}^{i})|\mathrm{dx}$ where $H(\cdot)$ is the Heaviside function. In common with many active contour techniques, manual parameter adjustment is required to control the relative contribution of the individual components as well as the similarity alignment weight parameters, although these latter weights can be kept constant once suitable values have been determined.

The expectation-maximisation algorithm [10] in combination with a Finite Gaussian Mixture model is used to learn the photometric properties of the foreground and background regions. Empirical tests have shown that the background region for this learning process is usually best defined as the set of pixels within a limited distance of the object contour.

3 Experimental Results

We present results including the tracking of arbitrary motions, rigid and deforming, whilst also showing how critical the shape modelling is when it is switched off during tracking. We will also show results for cyclic and non-cyclic human motion. Comparative results

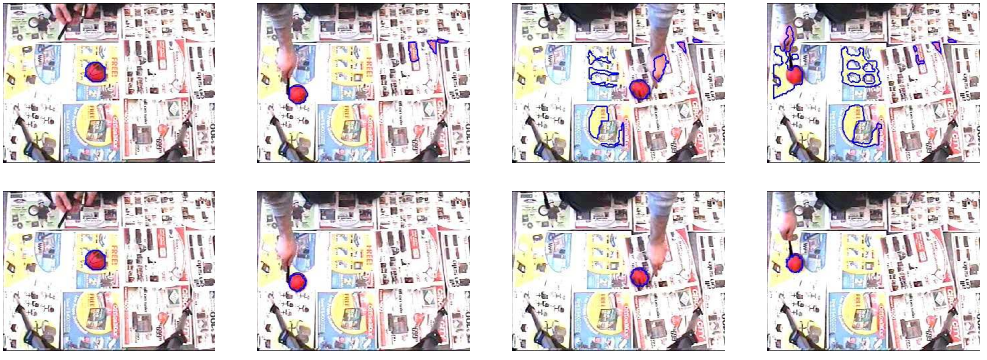


Figure 2: Tracking a ball without (top row) and with shape information (bottom row). It is successful in the latter despite the busy and often disturbed background where the on-line shape information eliminates confusion with the background. Data from [26].

are provided against a simple foreground/background competition active contour approach, similar to parts of [19, 27], except modified to include our implicit contour position re-estimation process (Section 2.3) for object tracking.

Tracking arbitrary motions and deformations - Fig. 1 presents a simple well-contrasting scene of an arbitrarily moving animal, undergoing a variety of motions and deformations for a moving observer. A relatively close fit is achieved despite the complex shapes.

Rigid shape tracking - In Fig. 2 we present the tracking of a small red ball in a very cluttered scene. The top row shows exemplar frames using region competition, the spatial smoothness constraint and the tracking component, but without high level shape modelling or the photometric memory. This experiment therefore concentrates on the photometric information only based on foreground/background competition. The tracking fails quite quickly and extraneous regions of initially relatively similar photometric properties gradually reduce the integrity of the foreground model and add to the computational burden. In the bottom row, the full proposed model is employed.

Cyclic human motions - Results for various human motions can be seen in Fig. 3 i.e. running, jumping, and walking. Each sequence demonstrates cyclic motions, thus providing significant opportunity to re-use existing information in the shape space from previous frames. Weight matrices are also shown for each tracking result which clearly illustrate the cyclic commonality in the shape information across different frames. The forward and backward diagonal structures in the weighting matrices are dependent on the scale of the matrix and the cyclic changes in shape through which the object undergoes. The weight matrices in the 2nd and 3rd rows contain strong backward diagonals in comparison to the 1st row weight matrix. This indicates greater similarity of the object shape within the same cycle, where shape information is further reused. The 1st row result demonstrates less similarity for within cycle motions which can be understood from the coarser sampling in relation to the relatively faster changes in shape.

We performed quantitative performance characterisation to compare our results with those of the groundtruth provided by [11] obtained via background subtraction. The results are shown in Fig. 4(a) using the Dice coefficient: $D = \frac{2|\mathfrak{F}_{\mathrm{tf}} \cap \mathfrak{F}_{\mathrm{gt}}|}{|\mathfrak{F}_{\mathrm{tf}}|+|\mathfrak{F}_{\mathrm{gt}}|}$. This quantifies the

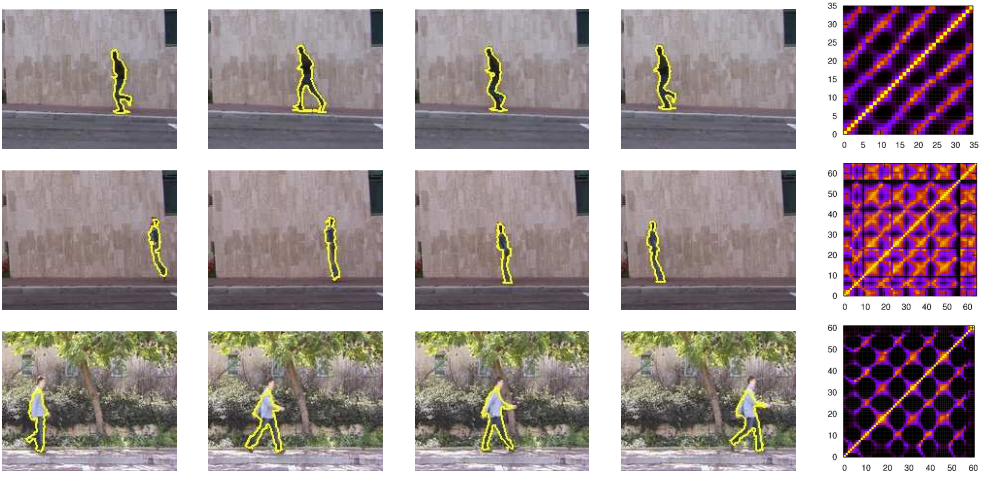
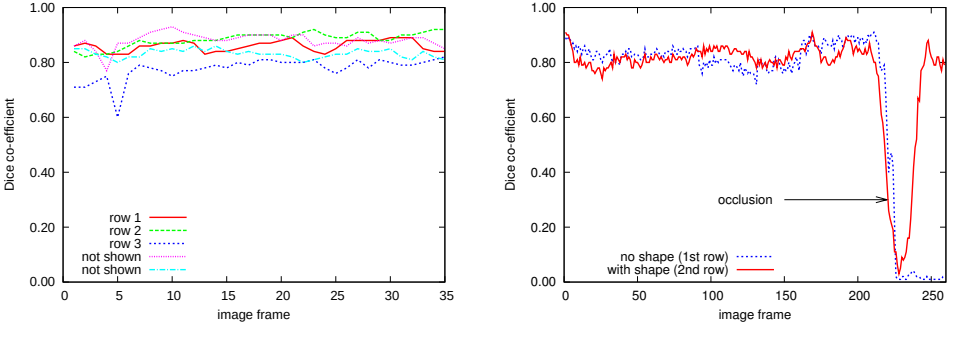


Figure 3: Tracking results for a variety of different video sequences taken from [▭] with corresponding weight matrices. The application of the data in [▭] was for activity recognition and the results obtained here are of sufficient quality to be useful for such an application.

amount of overlap between the tracking framework's definition of the foreground, $\mathfrak{F}_{\text{tf}}$, and the background subtraction defined foreground $\mathfrak{F}_{\text{gt}}$ of [▭]. The groundtruth data are not



(a) (b)

Figure 4: Dice coefficient performance characterisation of (a) results in rows 1 to 3 of Fig. 3 and two further results not pictorially illustrated here; and (b) results in Fig. 5.

perfect, but they are expected to sufficiently represent the desired characteristics of shape. Quantitative comparison with the groundtruth data are therefore not expected to indicate perfect segmentations. The Dice coefficient values illustrated in Fig. 4(a) consistently suggest good agreement with the background subtraction templates.

Multi-phase human motion Inter-frame shape affinity appears to be useful for the on-line learning of the shape information to augment the tracking process, although high integrity shape information may not always be available for image sequences consisting of more ambiguous photometric properties, such as highly textured backgrounds. Nevertheless, the affinity of the tracked object across frames, and not necessarily sequentially is useful even

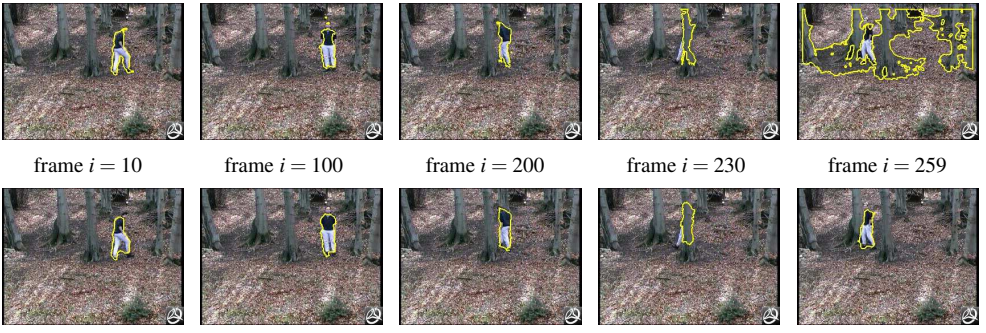


Figure 5: Comparative tracking result for person walking with a close-to-complete occlusion: (top row) results using region competition, the spatial smoothness constraint and the tracking component, but without high level shape modelling or the photometric memory; (bottom row) results of the proposed method. Data from [15].

in more complicated tracking scenarios. This is demonstrated by the comparative tracking result in Fig. 5 where the person walks behind a tree, resulting in a close-to-complete occlusion. The top row illustrates a result obtained that does not use shape information which is not able to track past the tree unlike the full on-line shape tracking approach proposed here (bottom row). The sequence contains similarity in the shape information between many frames and tracking of the person is successful despite the occlusion and the highly textured background. The shape model is able to retain sufficient information regarding the tracked object despite the photometric properties of the tracked object completely changing for the respective frames where the occlusion occurs. This is illustrated further by the performance characterisation in Fig. 4(b) where the Dice coefficient has been used again except here it is used to quantify the overlap with manually defined groundtruth. The tracking process is particularly aided here by the use of the optical flow estimates along the implicitly defined boundary, see (9), where sufficient information is drawn from the non-occluded object region to enable the shape model to be updated to the relevant position for every frame.

4 Conclusions

A new active contour based tracking framework has been presented. This utilises high-level shape information that is learnt on-line, adapting to new shape configurations whilst constraining the evolution of the active contour. Tracking is performed via optical flow estimation along the implicitly defined boundary of the active contour rather than relying on the extraction of salient points in the image and associating those points with the object being tracked. Results have shown that the combined framework is able to track objects undergoing complex deformations of shape with small changes in inter-frame object position. Tracking under close-to-complete occlusion with complex background photometric information has also been demonstrated. The main shortcoming of the method is that it is too slow for real-time purposes and, similar to many active contour tracking frameworks, successful tracking is dependent on an empirical selection of parameter values that control the relative contribution of the different model components.

Acknowledgments

This research was funded by the UK Leverhulme Trust and was carried out at the Department of Computer Science, University of Bristol.

References

[1] A. Baumberg and D. Hogg. An efficient method for contour tracking using active shape models. In IEEE Workshop on motion of non-rigid and articulated objects, pages 194–199, 1994.

[2] A. Blake, M. Isard, and D. Reynard. Learning to track curves in motion. In 33rd IEEE Conf. Decision and Control, volume 4, pages 4:3788–3793, 1994.

[3] T.F. Chan and L.A. Vese. Active contours without edges. IEEE Trans. Image Processing, 10(2):266–277, 2001.

[4] T. Cootes, D. Cooper, C. Taylor, and J. Graham. A trainable method of parametric shape description. Image Vision Comput., 10:289–294, 1992.

[5] D. Cremers. Dynamical statistical shape priors for level set based tracking. IEEE Trans. Pattern Analysis and Machine Intelligence, 28(8):1262–1273, 2006.

[6] D. Cremers. Nonlinear dynamical shape priors for level set segmentation. In Proc. IEEE CS Conf. Computer Vision and Pattern Recognition (CVPR'07), 2007.

[7] D. Cremers, T. Kohlberger, and C. Schnörr. Nonlinear shape statistics via kernel spaces. In Pattern Recognition, 2001.

[8] D. Cremers, S. Osher, and S. Soatto. Kernel density estimation and intrinsic alignment for knowledge-driven segmentation: teaching level sets to walk. In Proc. 26th DAGM Patt. Recog. Symp., pages 36–44. Springer, 2004.

[9] S. Dambreville, Y. Rathi, and A. Tannenbaum. A framework for image segmentation using shape models and kernel space shape priors. IEEE Trans. Pattern Analysis and Machine Intelligence, 30(8):1385–1399, 2008.

[10] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc., B39(1):1–38, 1977.

[11] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and Ronen Basri. Actions as space-time shapes. IEEE Trans. Pattern Analysis and Machine Intelligence, 29(12):2247–2253, 2007.

[12] B. Han, Y. Zhu, D. Comaniciu, and L. Davis. Kernel-based Bayesian filtering for object tracking. In CVPR, volume 1, pages 227–234. IEEE, 2005.

[13] M. Irani, B. Rousso, and S. Peleg. Detecting and tracking multiple moving objects using temporal integration. In ECCV, pages 282–287, 1992.

[14] A.D. Jepson, D.J. Fleet, and T.F. El-Maraghi. Robust online appearance models for visual tracking. IEEE Trans. Pattern Analysis and Machine Intelligence, 25(10):1296–1311, Oct. 2003.

[15] F. Korč and V. Hlaváč. Detection and tracking of humans in single view sequences using 2D articulated model. In Human Motion. Springer, 2007.

[16] M. Leventon, W. Grimson, and O. Faugeras. Statistical shape influence in geodesic active contours. In Proc. IEEE CS Conf. Computer Vision and Pattern Recognition (CVPR'00), pages 316–323, 2000.

[17] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In Proc. 7th IJCAI, pages 674–679, 1981.

[18] S. Osher and J.A. Sethian. Fronts propagating with curvature-dependent speed. J. Comp. Phys., 79:12–49, 1988.

[19] N. Paragios and R. Deriche. Geodesic active regions: a new framework to deal with frame partition problems in computer vision. J. Vis. Comm. Image Rep., 13(1-2):249–268, 2002.

[20] N. Paragios, M. Taron, X. Huang, M. Rousson, and D. Metaxas. On the representation of shapes using implicit functions. In Statistics and Analysis of Shapes. Springer, 2006.

[21] Y. Rathi, N. Vaswani, and A. Tannenbaum. A generic framework for tracking using particle filter with dynamic shape prior. IEEE Trans. Image Processing, 16(5):1370–1382, May 2007.

[22] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In Proc. Int'l Conf. Computer Vision, volume 2, pages 50–57, 2001.

[23] A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, W.E. Grimson, and A. Willsky. A shape-based approach to the segmentation of medical imagery using level sets. IEEE Trans. Medical Imaging, 22(2):137–154, Feb. 2003.

[24] A.J. Yezzi and S. Soatto. Deformotion: Deforming motion, shape average and the joint registration and approximation of structures in images. Int. J. Comp. Vis., 53(2): 153–167, 2003.

[25] A. Yilmaz, X. Li, and M. Shah. Contour-based object tracking with occlusion handling in video acquired using mobile cameras. IEEE Trans. Pattern Analysis and Machine Intelligence, 26(11):1531–1536, Nov. 2004.

[26] T. Zhang and D. Freedman. Improving performance of distribution tracking through background mismatch. IEEE Trans. Pattern Analysis and Machine Intelligence, 27(2): 282–287, 2005.

[27] S. Zhu and A. Yuille. Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. IEEE Trans. Pattern Analysis and Machine Intelligence, 18(9):884–900, 1996.