Continuous Ultrasound Speckle Tracking with Gaussian Mixtures

Colas Schretter$^{1,2}$, Jianyong Sun$^3$, Shaun Bundervoet$^{1,2}$, Ann Dooms$^{1,2}$, Peter Schelkens$^{1,2}$
Catarina de Brito Carvalho$^{2,4}$, Pieter Slagmolen$^{2,4}$ and Jan D’hooge$^{2,5}$

Abstract—Speckle tracking echocardiography (STE) is now widely used for measuring strain, deformations, and motion in cardiology. STE involves three successive steps: acquisition of individual frames, speckle detection, and image registration using speckles as landmarks. This work proposes to avoid explicit detection and registration by representing dynamic ultrasound images as sparse collections of moving Gaussian elements in the continuous joint space-time volume. Individual speckles or local clusters of speckles are approximated by a single multivariate Gaussian kernel with associated linear trajectory over a short time span. A hierarchical tree-structured model is fitted to sampled input data such that predicted image estimates can be retrieved by regression after reconstruction, allowing a (bias-variance) trade-off between model complexity and image resolution. The inverse image reconstruction problem is solved with an online Bayesian statistical estimation algorithm. Experiments on clinical data could estimate subtle sub-pixel accurate motion that is difficult to capture with frame-to-frame elastic image registration techniques.

Index Terms—Speckle tracking echocardiography (STE), Gaussian mixture model (GMM), local motion estimation

I. INTRODUCTION

Speckle tracking echocardiography (STE) [1] is a motion capture method that emerged from cardiology and is now in the midst of being translated to new medical applications such as measuring strain and stress of small structural tendons and ligaments in orthopedics [2]. Image-guided surgical applications are also foreseen by correlating dynamic image observations with advanced model-based biomechanics simulations of deformation [3].

In STE, motion estimation can be obtained with elastic image registration methods that use strong regularization priors to ensure spatial smoothness and local invertibility of the deformation vector field [4]. As a consequence, these techniques are not yet capable of capturing local motion occurring in small scale soft tissues, despite their robustness [5]. This work proposes to bridge this gap by avoiding computing explicit registration. Instead, the dynamic ultrasound image and motion of speckles is represented in a single unified spatio-temporal model. Estimated local linear motion trajectories are associated to each moving image element and can be retrieved from the model retrospectively.

We represent dynamic ultrasound images as a multi-resolution mixture of weighted multivariate Gaussians in 2D+T dimensions, as illustrated in Figure 1. The image and motion is therefore stored explicitly at varying levels of details in the scale-space [6]. Starting from uniform initialization, maximum likelihood values of parameters are estimated by an online conditional expectation-maximization method [7]. After estimation, the subtle motion of speckles is captured in the sparse model, using few mixture components. The rationale of our approach is that solving together dynamic image formation and motion estimation as a single problem may be better conditioned than performing successively frame-by-frame speckle identifications and tracking from the ultrasound image sequence.

The remainder of this paper is structured as follows. In section II, we define the input image data and a continuous statistical model to represent joint image and motion. Section III describes a Bayesian method for estimating maximum-likelihood values of model’s parameters. In section IV, we discuss results from experiments on an acquisition of a clamped and stretched tendon. Finally, we conclude and pinpoint next research steps in section V.

II. DATA AND IMAGE MODELS

We collect temporally apodized raw B-mode images in a discrete spatio-temporal amplitude field $A(x, y, t)$, with $x \in \{1, \ldots, N_x \}$, $y \in \{1, \ldots, N_y \}$ and $t \in \{1, \ldots, N_t \}$. This 2D+T matrix of real positive amplitude values is first normalized for every time frame $t$ such that

$$\sum_{x=1}^{N_x} \sum_{y=1}^{N_y} A(x, y, t) = 1, \forall t \in \{1, \ldots, N_t \}.$$ 

We approximate $A(x, y, t)$ with a continuous and smooth Gaussian mixture model (GMM) $G(x, y, t)$ having the form

$$A(x, y, t) \approx G(x, y, t) = \sum_{k=1}^{K} w_k \frac{g(x, y, t; \mu_k, \Sigma_k)}{g(t; \mu, \Sigma)},$$

with the set of component’s weights, means, and symmetric variances-covariances matrices

$$\Theta = \left\{ w_k, \mu_k = \begin{bmatrix} \mu_x \\ \mu_y \\ \mu_t \end{bmatrix}, \Sigma_k = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} & \sigma_{xt} \\ \sigma_{xy} & \sigma_y^2 & \sigma_{yt} \\ \sigma_{xt} & \sigma_{yt} & \sigma_t^2 \end{bmatrix} \right\}_{k=1}^{K}$$

where $g(x, y, t; \mu_k, \Sigma_k)$ is the product of a 1D marginal (temporal) and the 2D conditional (spatial) distributions:

$$g(x, y, t; \mu_k, \Sigma_k) = g(t; \mu_t, \sigma_t^2) \times g(x, y; \mu, \Sigma).$$
which updates parameters for increasing the data likelihood. We use an online expectation-maximization (EM) method for non-uniform sampling [8].

such that samples are drawn from the apodized 2D+T linear. After apodization, we sample non-uniformly points of consecutive frames where the motion is assumed to be stationary speckle.

The univariate Gaussian kernel $g(x; \mu, \sigma^2) \equiv g_1$ is

$$g_1 = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right],$$

and the bivariate Gaussian kernel $g(x, y; \mu, \Sigma) \equiv g_2$ is

$$g_2 = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp \left[ -\frac{1}{2} \left( \begin{array}{c} x \\ y \end{array} \right) - \mu \right] \Sigma^{-1} \left( \begin{array}{c} x \\ y \end{array} \right) - \mu \right].$$

III. RECONSTRUCTION

For estimating the mixture parameters $\Theta$ for the given reference time $t$, we first multiply 2D frames from the input sequence $A$ with $g(t; t, \sigma^2)$ with $\sigma$ set to half the number of consecutive frames where the motion is assumed to be linear. After apodization, we sample non-uniformly points in the domain of $A(x, y, t)$ by generating a source sequence $D = \{(x_m, y_m, t_m)\}_{m=1}^M$ such that samples are drawn from the apodized 2D+T discrete probability density function $A$. We use a single inversion method for non-uniform sampling [8].

The set $\Theta$ of all weights and parameters of the mixture $G(x, y, t)$ is now estimated by maximizing the log-likelihood

$$P(D; \Theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right] \frac{1}{2\pi\sqrt{|\Sigma|}} \exp \left[ -\frac{1}{2} \left( \begin{array}{c} x \\ y \end{array} \right) - \mu \right] \Sigma^{-1} \left( \begin{array}{c} x \\ y \end{array} \right) - \mu \right].$$

A. Online expectation-maximization

In the E-step, we evaluate the model from the information carried by the $m$-th sample $(x_m, y_m, t_m)$ by computing the membership probability $p_k(x_m, y_m|t_m)$, expressing the probability that the sample is drawn from the conditional distribution of the component $k$ at given time $t_m$ using

$$p_k(x_m, y_m|t_m) = \frac{w_k g(x_m, y_m, t_m; \mu_k, \Sigma) \Sigma^{-1} \left( \begin{array}{c} x_k \\ y_k \end{array} \right) \Sigma^{-1} \left( \begin{array}{c} x_k \\ y_k \end{array} \right) - \mu_k \right]}{\sum_{k=1}^K w_k g(x_m, y_m, t_m; \mu_k, \Sigma) \Sigma^{-1} \left( \begin{array}{c} x_k \\ y_k \end{array} \right) \Sigma^{-1} \left( \begin{array}{c} x_k \\ y_k \end{array} \right) - \mu_k \right]}.$$

with the parameters $\mu_k$ and $\Sigma_k$ of the conditional distribution.

In the second M-step, the weight $w_k$ is incremented for each $k \in \{1, \ldots, K\}$ with $w_k \leftarrow w_k + w_k \alpha$ and kernel parameters $\mu_k$ and $\Sigma_k$ are updated such that the likelihood of observing the new sample is improved as follows:

$$\mu_k \leftarrow \mu_k + \delta_k \alpha \quad \text{and} \quad \Sigma_k \leftarrow [\Sigma_k + \delta_k \delta_k^T] (1 - \alpha).$$

with $\alpha = p_k(x_m, y_m|t_m)$ and $\delta_k = x_m - y_m - t_m - \mu_k$. In order to adapt smoothly to new data, a mechanism should be used to adjust parameters to the most recent data samples. In order to guarantee that every data sample in a finite sliding window over the recent stream of data samples has equal importance, we stored the history of all incremental weight updates in a sliding window of given size $T = 256$, according to Neal and Hinton [7]. The value of $T$ should be large enough to guarantee a stable statistical estimates of component’s mean and covariance matrix. Therefore $T$ is controlling the bias-variance tradeoff. The circular history buffer is segmented in 64 memory pages that accumulate batches of successive updates. No significant convergence speed improvement was observed when using more pages.

B. Tree-structured initialization

We use a multi-resolution binary tree structure as shown in Figure 2 to represent the mixture model [9] instead of choosing a priori the number of mixture component $K$ and initializing parameters with ad hoc values that are not specifically dependent on the input data. For creating the tree hierarchy, the model is initialized with $K = 1$, and maximum-likelihood parameters are computed for this single component using the first data samples from $D$. 

Fig. 2. Example of image formation from few data samples marked by grey discs. Four successive levels of a binary hierarchy are shown in frames and the multi-resolution tree structure is drawn on the right. The iso-contours of multivariate Gaussians show that each component approximates a local subset of the input data.
Once a mixture component $k$ is supported by more that $T = 256$ data samples, the conditional Gaussian kernel is split in two and siblings are displaced in opposite direction along its normalized dominant eigenvector $c_k = [e_x, e_y]'$ corresponding to the principal eigenvalue $\lambda_k$ w.r.t. $\Sigma_k$. Figure 3 shows an example of this binary splitting procedure.

The eigenvector $c_k$ is aligned with the longest axis of the first standard deviation contour ellipse and $\lambda_k$ is the squared radius. Using the orthogonal squashing operator

$$Q = \begin{bmatrix} 1 - \frac{3}{4}e_x^2 & \frac{3}{4}e_x e_y \\ \frac{3}{4}e_x e_y & 1 - \frac{3}{4}e_y^2 \end{bmatrix}$$

and $\Sigma_{xy} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$, we modify the variances-covariances matrices of children:

$$\Sigma_{2k} = \Sigma_{2k+1} = \begin{bmatrix} S_{1,1} & S_{1,2} & s_{xt} \\ S_{2,1} & S_{2,2} & s_{yt} \\ s_{xt} & s_{yt} & \sigma_t^2 \end{bmatrix},$$

with $S = Q\Sigma_{xy}$ and $[s_{xt} \ s_{yt}]^\top = Q \begin{bmatrix} \sigma_{xt} \\ \sigma_{yt} \end{bmatrix}$.

Then, we displace centers of the two new components with

$$\delta = \begin{bmatrix} e_x & e_y \\ \sigma_{xt} & \sigma_{yt} \end{bmatrix} \Sigma_{xy}^{-1} \begin{bmatrix} e_x & e_y \end{bmatrix}^\top \sqrt{\frac{1}{2}\lambda_k},$$

such that $\mu_{2k} = \mu_k + \delta$ and $\mu_{2k+1} = \mu_k - \delta$. Finally, we share the weight in two: $w_{2k} = w_{2k+1} = w_k/2$.

As we can see, the indices correspond to the conventional storage of a binary tree in a vector, with breadth-first traversal. After convergence, the leaf components provides a fully-resolved image model. However we also keep the intermediate nodes in a binary tree. This structure allows us to recover model estimates at different bias-variance trade-offs. Since the conditional image regression is made of Gaussian components, the structure is equivalent to a scale-space representation [6]. There is an advantage of using the tree structure for accelerating substantially computations by pruning branches that do not overlap significantly with the data sample or the query location during regression.

**IV. Experiments**

We conducted experiments on an 2D+T acquisition of an isolated digital flexor tendon of a sheep using a Vero 2100 ultrasound machine (FUJIFILM VisualSonics, Inc, Toronto, CA) with a 256 elements linear array transducer. The central frequency was 20MHz and images were acquired at 50 frames per second. The tendon was mechanically stretched on a loading platform in a continuous strain between 2% to 6%. The apparent motion of speckles is relatively smooth in the sequence of 128 frames. Since speckles are intrinsically three-dimensional shapes and only a planar cut over time is captured, the out of plane motion translates into smooth fading in and out of 2D cross sections through speckles. From a sequence of 2D frames over time, it was possible to infer 3D motion in the vicinity of the imaged plane.

The input sequence of B-mode images contains 128 frames of $496 \times 256$ pixels and the motion of speckles is visible from subtle image intensity changes. In order to meet the local linear motion hypothesis, the input dataset is smoothly apodized over time using a Gaussian kernel at the central reference frame 64 with standard deviation $\sigma = 32$. This parameter is trading-off the temporal resolution of the motion model and the robustness (or stability) of the statistical motion estimate. The effect of apodization weighting can
be seen in the left optical flow views in Figure 4. We reconstructed the whole dataset with $K = 7936$ components that correspond to a density of 256 components for the close-up $64 \times 64$ pixels window used for optical flows in Figure 5.

Figure 6 shows the effect of the model complexity on the smoothness of the image approximation. Selecting a small number of components regularizes the estimation problem in a natural way. We observed that about 256 mixture components per regions of $64 \times 64$ pixels was very satisfactory for capturing motion of speckles. In contrast to pixelized images, linear trajectories of speckle image elements are explicitly stored in the covariance parameters of variances-covariances matrices that are estimated individually for each component.

Figure 7 illustrates the precision of predicted frames using linear displacements of speckles that are identified in the reference frame. In Figure 8, we compared line profiles in the ground truth frames with the prediction of past and future frames using the sparse Gaussian mixture. Sub-pixel accurate relative displacements are measured from the horizontal shift of reference landmarks in the center of speckle bumps. For an offset of 16 frames, the smooth drift of speckles is still successfully predicted. With large offsets, the linear trajectories are crossing and spurious clusters appear. Prediction accuracy drops when the hypothesis of local linear trajectories does not hold anymore. These results on clinical data confirm our first study on 1D profiles from a numerical phantom [10].

V. CONCLUSION

We represent dynamic ultrasound images as a mixture of moving smooth ellipsoids that are approximated by multivariate Gaussian kernels. Bayesian parametric estimation is carried out a fast online expectation-maximization method and locally linear approximations of speckle’s motion are retrieved with conditional regression from the model. The potential of this new image representation for motion estimation in ultrasound speckle tracking has been experimentally verified on a clinical acquisition of an elongated tendon mounted in a motorized clamper. Very sparse mixture models could reconstruct fair approximations of the smooth motion vector field. Subtle linear motion can be represented exactly in the model. Future work will accelerate reconstructions by using the tree structure for approximating computations by updating only the few most significant mixture components.

REFERENCES