Rapid Online Quantification of Tip-Sample Interaction for High-Speed Dynamic-Mode Atomic Force Microscope Imaging

David Busch, Juan Ren, Qingze Zou, Baskar Ganapathysubramanian

Abstract—his paper presents a numerical-experimental integrated framework for rapid online estimation of the tip-sample interaction forces for high-speed AFM dynamic imaging. Quantifying the tip-sample interaction in AFM dynamic imaging is crucial towards achieving high-speed dynamic-mode AFM imaging with minimized tip-sample forces, particularly when imaging live biological samples in liquid media. Large tip-sample interactions can result in sample deformation or even destruction of the live biological samples. In this article, we propose an ultra-fast inversion strategy based on parallel algorithms implemented on Graphical Processing Units (GPUs). The tip-sample interaction estimation problem is posed as an inverse problem and is solved in near-real time using GPUs. We investigate several cost-functional formulations to ensure quality of reconstruction while maintaining a high rate of reconstruction. The computational scheme is verified with preliminary experimental results.

I. INTRODUCTION

In this article, a computational scheme is proposed for rapid online quantification of the tip-sample interaction force during dynamic-mode atomic force microscope (AFM) imaging. We note that dynamic-mode AFM imaging [1] has been established as the de facto mode to interrogate surface topography of soft samples [2], [3], particularly for live biological samples in their physiologically friendly liquid environment [4], [5]. Although by using dynamic-mode imaging, the detrimental lateral force on the sample has been largely reduced, the applied normal force can still be large and result in not only imaging distortion, but more seriously, sample deformation and damage that can completely modify the sample [2]. Large normal force, however, is needed in current dynamic-mode AFM imaging to ensure imaging quality (i.e., high signal to noise ratio), and/or to achieve high-speed imaging. With an aim of developing a novel high-speed dynamic-mode AFM imaging strategy with minimized tip-sample interaction forces, this article presents, as a first step, the development and experimental validation of a computational framework for rapid online quantification of tip-sample interaction force.

We note that high-speed dynamic-mode AFM imaging is needed in the frontiers of polymer [3], [6], [7] and biological [8], [9] science and engineering, where understanding nanoscale behavior and evolution is important. For instance, by using dynamic AFM imaging, time evolving phenomena like crystallization of polymers [6] and the dehydration process of collagen [10], [9] have been experimentally revealed for the first time. In such interrogations, however, large temporal errors exist as current dynamic AFM imaging—with tens of minutes of imaging time—is too slow to discern the temporal details that occur in the order of a few (tens of) seconds. In addition, high-speed dynamic-mode AFM imaging of soft samples without causing sample deformation (or sample destruction) is challenging. Although the speed of contact-mode AFM imaging can be substantially improved by using recently-developed control techniques [11], [12], additional hurdles exist in improving the speed of dynamic-mode AFM imaging due to the oscillatory nature of the tip-sample interaction and consequently, the much more complicated interaction dynamics involved [13], [14]. The challenge in tackling these hurdles lies in the need to maintain a small tip-sample interaction force. Such a need, however, has not been accounted for in recent efforts in achieving high-speed dynamic-model AFM imaging via either control approach [15] or hardware improvements [16]. Therefore, maintaining a small probe-sample interaction force is central to achieving high-speed dynamic-model AFM imaging of soft samples. Tip-sample interaction dynamics has attracted great interest recently. There has been some successes in post-processing (but not on-line estimation) the experimental results to extract the interaction. Examples include the linearized single mode model to reconstruct interaction forces from Chebyshev polynomials [17]. Recently, off-line inverse problems have been formulated to estimate tip-sample interaction forces using conjugate gradient optimization [18] with limited success.

This article details the development of an approach to achieve rapid quantification of tip-sample interaction force towards high-speed dynamic imaging with the long-term goal of minimizing the tip-sample interaction force. The main contributions of this article are the following: (a) formulating the
problem of estimating the tip-sample interaction as an inverse problem posed as an unconstrained optimization problem; (b) developing an ultra-fast predictive model for AFM dynamics based on parallel algorithms implemented on Graphical Processing Units (GPUs); (c) investigating several real-space and frequency-space cost-functionals for the solution of the inverse problem to ensure quadratic minimization and speed; and (d) experimentally linking and verifying the results.

II. ONLINE ESTIMATION OF TIP-SAMPLE INTERACTION FORCE IN DYNAMIC AFM IMAGING

The destruction of biological or other soft samples is invariably caused by excessive tip-sample interaction forces. As alluded to in the introduction, several attempts to overcome these excessive forces have been made, none really solving the problem. The true problem, lies in the complex nature of these interaction forces. If the interaction forces could be calculated, corrective action could be taken to avoid excessive force. This section addresses the issue of calculating tip-sample interaction forces in real time. Given that the force calculation is an ill-posed problem, a method for solving ill-posed inverse problems is needed. This requires (a) an efficient, accurate and fast forward model that models the AFM dynamics as well as (b) appropriately defined cost-functionals that aid rapid inversion.

A. Nonlinear Dynamic Model of the Tip-Sample Interaction Dynamics in Dynamic Imaging

In AFM dynamic imaging (e.g., tapping mode imaging), a micro-machined cantilever beam with a nanoscale-probe on one free end (the other end is fixed) is driven under a piezo-actuator to oscillate and tap on the sample surface. The oscillation amplitude and phase are then measured and the tapping amplitude is maintained around the setpoint value through feedback control.

![AFM tip sample interaction model](image)

We start with modeling the cantilever as a clamped-free vibration beam based on the Euler-Bernoulli Beam theory [19], [20], [21]:

\[ EIu''' + \mu u'' = F'(t) - \mu \ddot{Y}(t), \]

where \( u \) is the cantilever deflection as a function of both time and the location on the cantilever beam relative to its base, \( x \); \( F' \) is the tip sample interaction force; \( Y \) is the displacement of the cantilever base; \( E, I, \mu \) are Young’s modulus, moment of inertia and mass per unit length of the cantilever, respectively, and \( \dot{u} \) and \( u' \) denote the derivative of time \( t \) and position \( x \), respectively. Note that in the above model, we assume, as widely adopted in previous works [20], [22], that the cross-section of the beam is uniform and rectangular, the length of the cantilever is much larger than the width, and the width is much larger than the thickness, the vibration amplitude is much smaller, and the cantilever beam is fully elastic.

We utilize the separation of variables approach [21] to convert this partial differential equation into a set of ordinary differential equations. The spatio-temporal variation in deflection is represented as:

\[ u(x, t) = \sum_{i=1}^{N} \eta_i(t) \Phi_i(x) \]

The (orthogonal set of) modal shape functions, \( \Phi_i(x) \), are computed by solving a homogeneous eigen-value problem [21]. The unknowns are now the temporal variables (or coefficients), \( \eta_i(t) \). Define an inner product as,

\[ \langle f, g \rangle = \int_L fg dx. \]

where \( L \) is the length of the cantilever. Solving the PDE system Eqn. 1 is equivalent to solving the following set of ODEs.

\[ \ddot{\eta}_i + \omega_i^2 \eta_i = F_i(t) + Y_i(t), \]

where \( F_i(t) = \langle F'(x, t), \Phi_i(x) \rangle \), \( Y_i(t) = \langle Y(t), \Phi_i(x) \rangle \). We use the fact that the modal functions satisfy \( \mu \frac{d^4\Phi(x)}{dx^4} = \omega_i^4 \Phi_i(x) \) as well as orthogonality \( \langle \Phi_i, \Phi_j \rangle = \delta_{ij}/\mu \).

B. Formulating the inverse problem as an unconstrained optimization problem

An efficient computational implementation of the forward problem is a necessary prerequisite for fast inversion of the tip-sample interaction. We pose the inverse problem as follows: Given the experimentally measured deflections at a finite number of positions along the cantilever \( U(x_j, t) \), \( j = 1, \ldots, k \), and the base deflection, \( Y(t) \), estimate the tip-sample interaction force, \( F'(t) \) such that the deflections, \( u(x, t) \), computed by solving Eqn. 1 match the experimental results. The inverse problem can be reformulated into an unconstrained optimization problem [23], [24] by defining an appropriate cost-functional, \( \mathcal{J} \), that quantifies the mismatch between the experimentally estimated deflections, \( U \) and the numerically computed deflections, \( u \):

\[ \text{Find } F'(t) \in L_2[0, t_{max}] \ni \mathcal{J}(F') \leq \mathcal{J}(F), \forall F \in L_2[0, t_{max}] \]

Starting from a guessed value of the tip-sample interaction, an optimization framework is used to minimize the cost-functional, \( \mathcal{J} \) to find out the optimal tip-sample interaction. The choice of the cost functional is very important: (a) to ensure that a minima exists, (b) an appropriate initial guess can be estimated to ensure convergence. In particular, quadratic cost-functionals with an initial guess that lies in the basin of attraction of the cost-functional are helpful to guarantee convergence. We investigated the following
choices of the cost-functional:
\[ J_{L2}(F) = \sqrt{\int_0^{t_{max}} (u(x, t) - U(x, t))^2 dt} \] (6)
\[ J_{PSD}(F) = \sqrt{\int_0^{f_{Nyquist}} \left( \tilde{u}(f) - \tilde{U}(f) \right)^2 df} \] (7)
\[ J_{C-PSD}(F) = \sqrt{\int_0^{f_{Nyquist}} \left( \int_0^{f_{Nyquist}} (\tilde{u}(f) - \tilde{U}(f))^2 df \right)^{0.5} df'} \] (8)
\[ J_{KL}(F) = \left[ \int_0^{f_{Nyquist}} \left( \log \left( \frac{\tilde{u}(f)}{\tilde{U}(f)} \right) \right)^2 df \right]^{0.5} \] (9)

where \( u(x, t) \) and \( U(x, t) \) are the calculated and experimental cantilever deflection respectively and \( \tilde{u} \), \( \tilde{U} \) represent the power spectral density of the deflections. As a first step to inversion, we looked at the feasibility of using each of these cost-functionals in the context of two issues (a) occurrence of a convex quadratic form with one single minima, (b) speed of calculating on the GPU based framework. Results are detailed in Section III.

C. Computational Framework for the direct problem

Since both the modal functions, \( \Phi_i(x) \) and the modal coefficients, \( \eta_i(t) \) are needed for fast calculation of deflection, we develop accelerated strategies for concurrently solving for \( \Phi_i(x) \), and \( \eta_i(t) \). Fig. 2 shows a schematic of the modal function calculation framework.

![Modal Function Calculation Flowchart](image)

An optimized Newton root solver and Simpson integration modules used. One difficulty in implementation is the fact that the hyperbolic trigonometric functions (in the Modal function evaluation) grow very quickly and cause over-flow errors. This limits the number of modes to around the first ten modes, if the hyperbolic trig form is used. To resolve this issue, we recast the evaluation of the modal functions, \( \Phi_i(x) \) [21] into the following equivalent form:

\[ 2 \sin(\beta_i x) - 2G \cos(\beta_i x) + (G - 1) \exp^{\beta_i x} + \exp^{-\beta_i x} \]

where,

\[ G = \frac{2 \exp^{-\beta_i L} \sin(\beta_i L) - \exp^{-2 \beta_i L} + 1}{2 \exp^{-\beta_i L} \cos(\beta_i L) - \exp^{-2 \beta_i L} + 1}. \] (11)

The solution of the ODE Eqn. 4 requires a user-defined \( F'(x, t) \) (which will become the unknown function in the inverse problem), and pre-computed \( \omega_i^2 \) and \( \Phi_i(x) \) from the modal function program. The value of \( \omega_i^2 \) grows very quickly with increasing \( i \), making Eqn. 4 a stiff ODE. We utilize an explicit fourth order Runge-Kutta method. Because the solver is explicit, it is very efficient. However explicit schemes require small time steps to ensure stability. This is however, not an issue since to mimic the experimental sampling rate, very small time steps are required.

Results of the computational framework have been checked against analytical solutions for simple forced vibration and constant forces. This also allowed us to perform extensive runtime analysis to pick an optimal time-step, spatial discretization as well as the number of modes to ensure maximum accuracy while guaranteeing computational speed. Fig. 3, and 4 show the run-times with increasing spatial discretization, and number of modes. Extensive numerical analysis revealed that the optimal spatial discretization was \( n_x = 101 \), while \( m = 8 \) modes prove more than necessary to resolve most dynamics.

![Number of space steps vs. run-time](image)

![Number of modes vs. run-time](image)

D. Parallel implementation on GPUs: Complexity analysis

A complexity analysis of the various stages of the direct problem is performed. This allowed us to develop and deploy an optimized algorithm implemented on a Graphical Processing Unit (GPU). The calculation of modal functions, \( \Phi_i(x) \) involves:

1. Newton-Raphson Root finding, (2) Simpson Integration, (3) Equation Evaluation. These calculation have a complexity of \( O([k + n_x] m + n_x) \), where \( k \) is the number of digits of accuracy required for the root finding method, \( m \) is the number of modes, and \( n_x \) is the spatial discretization. This
is a one-time calculation and in practice adds little to the overall complexity.

The calculations involved in calculating the modal coefficients, \( \eta_i(t) \) are:
1) Simpson integration for force, \( F_i(t) \); 2) Marching forward in time using a fourth order Runge-Kutta ODE Solver. The runtime complexity is \( O((n_x n_t + n_t)m) \), where \( n_t \) is the number of time steps. The cost-functional calculation (an integral over time) has a runtime complexity of \( O(n_t) \). Thus, a single forward problem with associated cost-functional calculation has a runtime complexity of \( O((n_x n_t + n_t)m + n_t) \).

In practice, for a typical run of \( n_x = 101, n_t = 10001 \), and \( m = 8 \), this results in an average runtime of 1.29 seconds. To get close to real-time inversion (which takes multiple hundreds of forward problem evaluations), this runtime must be reduced by at least two orders of magnitude. To that end, the use of Graphical Processing Units is considered.

The CUDA framework is used to implement a parallel algorithm [25] on the GPU computing platform. The only limitation of using GPUs is memory availability [25]. This necessitates the development of algorithms with a focus on memory management and efficient running of streaming multiprocessors. The major memory needs of the forward problem and cost-functional calculations are: (a) storing the experimental and calculated deflection data \( 4n_x k \) Bytes (approximately 12 KB of memory). Here \( k \) is the number of deflection points measured experimentally (usually 3); (b) storing modal functions, \( \Phi_i(x) \) requires \( 4n_x m \) Bytes (approximately 3.2 KB of memory). Ideally, all the data for calculation on a GPU should be stored in the shared memory in order to maintain fast memory accesses. However, in this case the memory needs are such that the 16 KB shared memory is not enough.

Using all the shared memory is a possibility, but that would leave no room for other data requiring subsequent data paging. An alternative to shared memory that can also provide fast memory access is the constant memory. We use this memory to store the modal function and the experimental deflection data since these are one time transfers. The CPU would have to calculate the modal functions before the GPU routine is called. Thus, room is left in the shared memory for the calculated deflection data, prefix sum for integration, and other smaller data needs. This will lead to only one interaction with the slower global memory to store the metric value. For the GPU based algorithm, the modal functions are calculated the same way as the serial algorithm and stored in constant memory at one time computational cost as explained above. In order to calculate the modal coefficients, a force integral, \( F_i(t) \) (Eqn. 4) needs to be solved. This is a good candidate for parallel prefix. The next step is the time-marching of the ODE. Given that ODE solvers have a strong dependence on the previous time step, it is difficult to parallelize solving the ODE’s. Since multiple modes, \( m \) are computed, several force integrals and ODE need to be solved in the same time step; thus parallelizing across modes is another possibility.

We considered both approaches in developing our parallel framework. When only parallel prefix is used, a thread is assigned to every space point of the modal functions in order to put the most threads to work. This results in a runtime complexity of \( O((n_x \log(n_x) + n_t)m + n_t) \) compared to the serial runtime complexity of \( O((n_x n_t + n_t)m + n_t) \). However, this method requires a large number of threads, which on the GPU is limited to 512. In contrast, when only parallelism across modes is used, a runtime complexity of \( O(n_x n_t + 2n_t) \) is achieved. If only one thread was used per mode, each thread would have the same work as the serial algorithm. In practice, the number of modes is small. Having a small number of threads running on a GPU is a problem because streaming multiprocessors require a large number of threads to work efficiently. To get the most out of the GPU, the best solution should be to use a combination of the two methods. Choosing to operate as many threads as possible, there are two ways the 512 threads could be divided, either maximizing the number of threads towards parallel prefix or maximizing the number of threads to the modes. The runtime analysis of the combined method, with maximizing the number of threads towards parallel prefix is \( O((n_x \log(n_x) + n_t)\frac{m}{\text{threads}} + \log(n_t)) \). The end result of these combined methods is that each CUDA block solves one cost-functional and the threads are focused on the forward problem. With this implementation each forward problem is solved in 0.002 seconds. This is two orders of magnitude reduction in time compared with the serial algorithm (1.29 seconds). However, there is significant scope for further improvement as the current implementation only achieves 13% occupancy. We anticipate an occupancy of 41% with further optimization cutting the compute time by another order of magnitude. Fig. 5 shows the GPU runtime as a function of the number of cost-functionals calculated.

![Fig. 5. Number of cost-functional evaluations vs. run-time in sections on GPU implementation](image-url)

### III. Experimental Validation and Preliminary Results

In this section, we demonstrate some preliminary results. We investigate the applicability of the cost-functional defined in the previous sections and show-case the near-real time inversion capabilities of the framework in estimating the base deflection given tip deflections for simple sinusoidal and chirp input signals. We start by describing the AFM system employed and the experimental procedure.
A. Experimental AFM System

The highlight of the experiments is the operation speed which is contributed by the hardware set-up. For the purpose of obtaining highly sensitive AFM deflection output, the frequency of the driving input signal was around the resonant frequency of the AFM cantilever, which was 25.53KHz. The usual DA Q cards are not able to convert such high speed signal. In order to handle the high speed signal, an Altera Stratix FPGA development kit was used during the experiments. The control signals, both sinusoidal signal and chirp signal were sent to the AFM piezo drive channel through FPGA card and its daughter card. The sampling frequency of the FPGA used in the experiments was 25MHz.

![Graph of 25KHz sinusoidal control input signal](image1)

![Graph of Deflection of different points on cantilever](image2)

Fig. 6. Top: sinusoidal input signal; bottom: deflection of sinusoidal drive

B. Procedure

The AFM probe used for testing is a Veeco MPP-31200-10 tapping mode probe, which is made of 0.01 - 0.025 Ω cm Antimony (n) doped Si. The tip specifications are as follows:

- Geometry: Anisotropic
- Tip Height: 15 - 20μm
- Tip Radius: 8nm
- Spring Constant: 0.9N/m
- Length: 450μm
- Width: 35μm
- Thickness: 3.0μm
- Resonant Frequency: 25.53KHz

During the experiment, a sinusoidal input signal frequency at 25KHz and a voltage amplitude of 0.4V was used. A chirp input drive signal with a frequency range from 10KHz to 50KHz, and voltage amplitude 0.35V was also used. Duration of both kinds of input drive signal is 2s, and data capture sampling frequency is 25MHz.

C. Results and Discussion

Figure 6 plots the sinusoidal control input and the cantilever deflection without tip sample interaction, respectively.

![Graph of chirp control input and deflection](image3)

Fig. 7. Top: chirp input signal; bottom: deflection of chirp drive

The computational framework estimated the frequency and amplitude of the base deflections within 0.2 seconds by rapidly constructing the phase space and recognizing the minima. Only two of the cost-functionals satisfied the two criterion for rapid inversion (existence of only one minima and most initial guess should lie in the minimal basin). The $J_{L2}(F)$ and $J_{PSD}(F)$ cost-functional both demonstrated minima as shown in Fig. 8 and Fig. 9, while the other two cost-functionals ($J_{C−PSD}(F)$, and $J_{KL}(F)$) did not show a global minima as seen in Fig. 10. The $J_{L2}(F)$ showed a minima at an amplitude of 1.3 nm and a frequency of 25.590 kHz. $J_{PSD}(F)$ demonstrated a minima at an amplitude of 1.6 nm and a frequency of 25.610 kHz. Both these values are
in close agreement to the actual values (of frequency). Similarly, the computational framework estimated the frequency and amplitude characteristics of the base deflections within 0.22 seconds for the chirp signal by rapidly constructing the phase space and recognizing the minima. For this case also, the $J_{L2}(F)$ and $J_{PSD}(F)$ cost-functionals performed well.

IV. CONCLUSIONS

We have developed an ultra-fast inversion strategy based on parallel algorithms implemented on Graphical Processing Units (GPUs). The estimation problem is posed as an inverse problem and is solved in near-real time using GPUs. We investigated several cost-functional formulations to ensure quality of reconstruction while maintaining a high rate of reconstruction. The computational scheme is verified with experimental results. We have show that reconstruction of simple, preliminary inputs is possible within fractions of seconds (0.2, 0.22 seconds). We are currently extending this framework to estimate more realistic tip-sample interaction forces.

REFERENCES


