Interactive Video Synthesis using Vibrational Modal Analysis

Navtegh Singh Gill, Stuti Wadhwa and Sehajpreet Kaur

Abstract—This report explores vibrational modal analysis for interactive video synthesis, building upon concepts introduced by [1] and [2]. We explore algorithms to learn vibration modes of deformable objects from videos and synthesize motion in response to unseen forces without any knowledge of scene geometry or material properties. This information is used to build image-space models of object dynamics around a rest state, letting us turn short video clips into physically-plausible, interactive animations. By adapting the methodology introduced in [1] to synthesize object dynamics from monocular video, we implemented and compared four approaches: the Lucas-Kanade optical flow method [3], the Complex Steerable Pyramid(CSP) approach used by [2], the RAFT deep learning-based optical flow technique [4], and a novel hybrid approach combining the Lucas-Kanade method with complex steerable pyramids. In addition, Frequency-domain analysis enables interactive selection of vibration modes for simulating plausible object responses. Our findings highlight the strengths and limitations of each method, with the hybrid approach offering a significant improvement in capturing realistic object dynamics while maintaining computational efficiency and balancing fine-grained motion tracking. This work provides insights into advanced motion analysis techniques and their applications in dynamic video synthesis.

Index Terms—Vibrational Modal Analysis, Interactive Video Synthesis, Optical Flow Methods and Phase-Based Motion Estimation

1 Introduction

Interactive video synthesis is an emerging field that seeks to bridge the gap between real-world object dynamics and computational simulations. By extracting vibrational characteristics from simple video footage, researchers can create physically plausible animations that respond realistically to user interactions. This capability holds promise for applications in fields such as virtual reality, structural analysis, material characterization, content creation, and augmented reality (AR). For instance, accurately predicting how bridges and other structures react to vibrations can inform design decisions that prevent catastrophic failure. In entertainment and AR, the ability to animate objects realistically based on their dynamic properties opens new possibilities for immersive experiences. Central to this advancement is the work of Abe Davis [1], who introduced the groundbreaking concept of image-space modal analysis. This method enables the extraction of vibrational modes directly from monocular video, effectively transforming static footage into dynamic, interactive simulations without requiring detailed knowledge of an object's geometry or material properties.

The present study builds on this foundation while incorporating insights from more recent advancements, such as *Visual Vibration Tomography* described in [2]. It highlighted the limitations of traditional optical flow methods for detecting small-scale vibrations and proposed a phase-shift-based approach utilizing complex steerable pyramids. While effective for capturing fine-grained motion, this method struggles under conditions of large-scale deformation or high external forces. Additionally, the advent of deep learning-based motion estimation techniques, such as RAFT (Recurrent All-Pairs Field Transforms), offers powerful tools for robust optical flow estimation but at the cost of high computational overhead.

In this work, we systematically investigate and compare four distinct approaches to modal analysis for interactive video synthesis:

1) **Lucas-Kanade Optical Flow:** A classic technique known for its computational efficiency but limited in capturing independent component motion due to its assumption of constant motion within a small, localized region.
2) **Complex Steerable Pyramid (CSP):** This phase-based approach, inspired by Wadhwa et al. [5], enables precise detection of localized motion through multi-scale and multi-orientation decompositions but also increases sensitivity to small-scale variations.
3) **RAFT Optical Flow:** A deep learning-based approach that, in our application, delivered results comparable to Lucas-Kanade but required greater computational resources.
4) **Hybrid Approach (Lucas-Kanade + CSP):** A novel combination of the Lucas-Kanade method and CSP, designed to integrate their respective strengths, enabling accurate and robust motion tracking.

Our primary objective is to overcome the limitations of current modal analysis techniques by systematically evaluating their strengths and weaknesses across a variety of scenarios. Through rigorous experimentation, we assess the performance of each method under different conditions, including subtle vibrations and high-force interactions.

This study makes two significant contributions: first, it provides a comprehensive comparative analysis of four distinct modal analysis methodologies for interactive video synthesis, delivering valuable insights into their applications and performance equipping researchers with the essential knowledge to select the most effective technique for their specific needs. Second, it introduces and validates an

innovative hybrid technique which effectively addresses the challenges posed by both subtle vibrations and large-scale deformations, setting a new benchmark for accuracy and reliability in the field.

2 Related Work

Many physically-based animation techniques leverage modal analysis to reduce the degrees of freedom in deformable body simulations [6] [7]. Traditional methods rely on detailed geometric information or material properties, and rely on finite element model (FEM) methods to derive a modal basis for simulation. But, we wanted to implement a method that enables the generation of dynamic simulations without requiring this prior knowledge of an object, thereby enhancing flexibility and applicability across a broader range of scenarios.

We built upon the foundational work of Abe Davis et al. [1] that introduced the concept of image-space modal analysis in their groundbreaking study *Interactive Dynamic Video*. This method demonstrated how vibrational modes of objects could be extracted directly from monocular video footage, allowing for physically plausible simulations of dynamic object behavior. The Lucas-Kanade method [3] is a widely recognized standard for generating optical flow vectors which are employed for modal analysis. Recently, deep learning-based optical flow techniques, such as RAFT [4], have emerged, employing novel architectures that leverage all-pairs correlations and recurrent networks to generate dense motion fields. These methods create smooth, consistent motion fields but fail to capture complex deformations.

Berthy Feng et al. [2] expanded the field by proposing *Visual Vibration Tomography*, an advanced methodology for inferring interior material properties through vibrational analysis. Their work focused on using phase-based motion estimation through CSP, a technique designed to accurately detect and quantify small-scale vibrations, providing non-invasive estimation of material properties. This phase-based approach effectively detects fine-grained vibrational patterns, outperforming traditional optical flow techniques in scenarios requiring subtle motion detection, but is susceptible to inaccuracies when subjected to large forces.

This study contributes to the field by systematically evaluating the strengths and weaknesses of existing techniques and introducing a hybrid approach that combines Lucas-Kanade and CSP. By addressing their individual limitations and integrating insights from foundational work, our method provides an enhanced and practical solution for achieving realistic dynamic simulations.

3 Proposed Method

This research introduces a novel approach for synthesizing interactive video animations through vibrational modal analysis, enabling realistic motion generation in response to unseen forces. Below, we provide an in-depth explanation of each method, culminating in our hybrid approach, which offers a balanced solution to the limitations of the individual techniques.

3.1 Feature Detection

We use the feature points extracted from Good Features [8] to track the feature points through the video. Before tracking the detected feature points, we can manually add more feature points as well as fixed points to the reference frame. The fixed points are the ones we do not want to track, like the background or other immovable parts. Fig. 1 shows how the feature points should be arranged - the blue points are the detected and/or added feature points, while the red points are the manually selected fixed points. To ensure spatial coherence, Delaunay triangulation [9] is applied to connect feature points into a network of triangles. This setup preserves relationships between points, enabling accurate motion tracking and dynamic simulations. At simulation time, each fixed point moves the average amount of its neighboring detected feature points, and the points without neighboring feature points would stay fixed.

3.2 Mode Extraction

Modes are simply periodic motions occurring at particular frequencies, so we would expect them to appear as peaks in the power spectrum of motion amplitude. As high frequency modes generally contribute less to an object's deformation, they can often be discarded to obtain a lower dimensional basis for faster simulation. We use this reduced modal basis to simulate objects in video, but assume no knowledge of scene geometry and cannot therefore use traditional techniques to compute vibration modes. These techniques work by first deriving orthogonal vibration modes from known geometry using FEM approaches. Instead, we observe non-orthogonal projections of an object's vibration modes directly in video. For this we explore numerous mode extraction algorithms as well as propose a new approach that is more robust to independent component motion and large-force scenarios.

3.2.1 *Lucas-Kanade Optical Flow*

The Lucas-Kanade (L-K) method is a foundational approach to motion estimation, leveraging the **brightness constancy assumption**, which states that the intensity of the same point remains constant across consecutive frames of a video. This assumption is mathematically expressed as:

$$I(x, y, t) = I(x + dx, y + dy, t + dt)$$

where $I(x, y, t)$ is the pixel intensity at location (x, y) and time t, and (dx, dy) represents the pixel displacement over time interval dt. Expanding the right-hand side using a Taylor series, neglecting higher-order terms, and rearranging yields the **optical flow equation**:

$$f_x u + f_y v + f_t = 0$$

Here:

- $f_x = \frac{\partial f}{\partial x}$ and $f_y = \frac{\partial f}{\partial y}$ are the spatial image gradients,
- $f_t = \frac{\partial f}{\partial t}$ is the temporal gradient,
- (u, v) are the optical flow components (motion in the x- and y-directions).

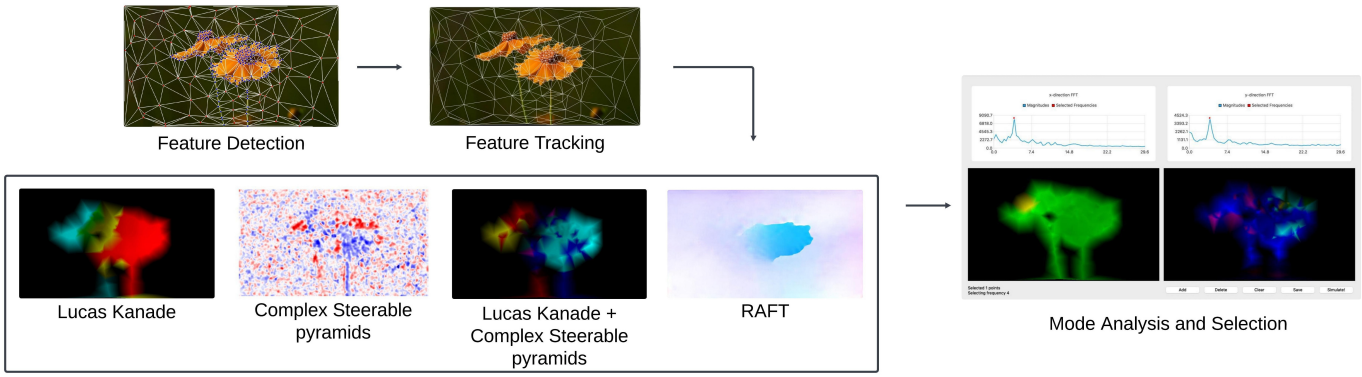

Fig. 1: Flow diagram of the methodology.

This equation is under-determined with one equation and two unknowns (u, v). The Lucas-Kanade method addresses this by assuming that motion is locally smooth and can be approximated as constant within a small neighborhood (e.g., a 3×3 pixel patch). By applying the optical flow equation to all pixels within this patch, we obtain an over-determined system of equations. Using the least squares method, the optical flow components are computed as:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_i f_{xi}^2 & \sum_i f_{xi} f_{yi} \\ \sum_i f_{xi} f_{yi} & \sum_i f_{yi}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i f_{xi} f_{ti} \\ -\sum_i f_{yi} f_{ti} \end{bmatrix}$$

This system highlights that corners or textured regions with significant gradients provide more reliable optical flow estimates, a property also emphasized in the Harris corner detector.

Methodology in the project:

- The Lucas-Kanade algorithm is applied to estimate the motion of feature points between consecutive video frames.
- Optical flow vectors (u, v) are derived for all feature points, representing their displacements in the x- and y-directions.
- The algorithm assumes that motion is locally smooth and can be approximated by a linear model. For each pixel, the optical flow is calculated by solving a system of linear equations derived from image gradients.
- By separately analyzing x- and y-direction vectors, we obtain two modal shapes for each vibration mode.

Strengths:

- High computational efficiency due to its simplicity and low computational overhead.
- Suitable for scenarios with small, localized and uniform motion across an object.

Limitations:

- Struggles with large displacements or non-rigid deformations.
- Ineffective for independently moving components of an object.

3.2.2 *Complex Steerable Pyramids*

The extraction of vibrational modes using complex steerable pyramids is a robust approach for analyzing small, often imperceptible motions in videos. This method relies on local phase shifts, offering subpixel motion sensitivity that surpasses traditional techniques like optical flow. Introduced by [5] and extended by [2], this approach facilitates accurate frequency-based analysis of object dynamics.

Methodology:

- **Phase Computation:** A complex steerable pyramid representation [10] [11] is constructed for each video frame. This multiscale, multi-orientation transform captures local phase and amplitude information at various spatial scales, where local phase shifts indicate horizontal and vertical displacements at pixel-level.
- **Noise Reduction:**
  1) *Outlier Removal:* Pixels with the top 1% of displacement magnitudes are treated as outliers and removed.
  2) *Amplitude-Weighted Gaussian Blur:* This smoothing technique reduces noise while preserving motion details.
- **Phase-to-Displacement Conversion:** Local phase shifts are converted into pixel displacements using [11] [12], which ensures sensitivity to motions as small as 0.001 pixels.
- **Motion Field Construction:** The processed displacements are assembled into a motion field for each frame, quantifying horizontal (Δx) and vertical (Δy) displacements at every pixel relative to the reference frame.

Strengths:

- Excellent sensitivity to fine-grained, small-scale motions.
- Particularly effective for scenarios involving subtle vibrations.

Limitations:

- Prone to excessive deformation under high-force scenarios.

- Lacks robustness in tracking large displacements or complex object interactions.

3.2.3 *RAFT Optical Flow*

Recurrent All-Pairs Field Transforms (RAFT) [4] is a state-of-the-art deep-learning based approach to optical flow estimation. It computes dense per-pixel motion by maintaining a single high-resolution flow field updated iteratively. Traditional methods, such as Lucas-Kanade, rely on assumptions like local smoothness and gradient-based optimizations, often struggling with large displacements or textureless regions. RAFT overcomes these limitations by using a recurrent neural network and multi-scale correlation volumes to model pixel-wise motion effectively. This allows RAFT to efficiently capture both small and large displacements with unprecedented accuracy. By incorporating RAFT, our project explores the performance of deep learning-based optical flow methods for vibrational modal analysis, providing insights beyond traditional techniques like Lucas-Kanade.

Methodology:

- **Feature Extraction and Encoding:**
  1) RAFT uses a convolutional encoder to extract dense feature maps from the input video frames at $1/8$ of the original resolution, capturing critical spatial details.
  2) A context encoder further enriches the features of the first frame with spatial information, providing motion priors that guide flow updates.
- **Correlation Volumes:**
  1) A dense 4D correlation volume is constructed for all pixel pairs across frames by computing the dot product of feature vectors, generating a $W \times H \times W \times H$ correlation volume.
  2) To balance computational efficiency and detail, RAFT pools this volume into a multiscale pyramid, where each level captures information about both small and large displacements.
- **Iterative Refinement:**
  1) Optical flow, initially zero, is updated iteratively using a lightweight gated recurrent unit (GRU)-based update operator.
  2) Each iteration refines the flow field by retrieving relevant information from the correlation volumes and combining it with context and flow features.
- **Upsampling:** The predicted flow field is upsampled to full resolution using a convex combination of coarse flow predictions, ensuring high accuracy near motion boundaries.

In this study, RAFT is used to compute optical flow fields between consecutive video frames. These fields represent the per-pixel displacements (u, v) and serve as the basis for extracting vibrational modes.

Strengths:

- Can handle large displacements effectively.

Limitations:

- Computationally expensive and resource-intensive.
- Requires pre-trained models, limiting adaptability to novel scenarios.
- Since we are dealing with low frequency motion, this approach fails to capture small scale motion effectively as compared to our proposed approach.

3.2.4 *Hybrid Approach: Lucas-Kanade + Complex Steerable Pyramid*

To overcome the limitations of the individual methods, we propose a hybrid approach that combines the computational efficiency of Lucas-Kanade with the sensitivity of the complex steerable pyramid. This approach aims to capture both fine-grained details and large-scale deformations, ensuring robustness and accuracy.

Methodology:

- **Initial Motion Estimation:** The Lucas-Kanade method is first applied to estimate broad, coarse-grained displacements. Its computational efficiency and suitability for large-scale motions provide a robust foundation for motion tracking.
- **Fine-Grained Analysis:** The CSP method is then utilized to refine these estimates, focusing on localized, small-scale vibrations. Its high sensitivity to subpixel motions ensures the accurate detection of fine-grained details.
- **Fusion and Normalization:**
  1) The displacement data from Lucas-Kanade and CSP are combined in the frequency domain.
  2) During normalization, we carefully analyzed the relative contributions of the two methods. Experimental evaluations revealed that a weightage ratio of **4:6** (Lucas-Kanade:CSP) is the most effective. This ratio optimally leverages Lucas-Kanade for its efficiency in large displacements and CSP for its precision in capturing subtle motions.

Strengths:

- **Efficiency and Precision:** Balances the computational efficiency of Lucas-Kanade with the ability of CSP to capture independent component motion.
- **Robustness:** Handles both subtle vibrations and large-scale deformations effectively.
- **Realistic Simulations:** Produces accurate and physically plausible results, as demonstrated by experimental outcomes.

Limitations:

- The hybrid approach is marginally more computationally intensive than using either Lucas-Kanade or CSP individually.

3.3 Identifying Image-Space Modes

To extract these modes, we perform a **Discrete Fourier Transform (DFT)** on the temporal motion fields. The workflow is as follows:

- **Frequency Space Representation:** Let $\Delta x_t(x, y)$ and $\Delta y_t(x, y)$ denote horizontal and vertical displacements at pixel (x, y) in frame t. The 1D FFT of these fields across time produces complex-valued representations, $\Delta c_x^\ell(x, y)$ and $\Delta c_y^\ell(x, y)$, corresponding to frequencies $f_\ell = \frac{\text{FPS}\cdot\ell}{T}$.
- **Power Spectrum Analysis:** The motion power at frequency f_ℓ is computed as:

$$\sqrt{|\Delta c_x^\ell|^2 + |\Delta c_y^\ell|^2}.$$

  Peaks in the power spectrum (ℓ^*) identify the dominant natural frequencies (f_{ℓ^*}) and their corresponding image-space modes, given by:

$$\left[\text{Re}(\Delta c_x^{\ell^*}), \text{Re}(\Delta c_y^{\ell^*})\right].$$

- **Mode Representation:** The modes are represented as image-space motion fields corresponding to the real parts of $\Delta c_x^{\ell^*}$ and $\Delta c_y^{\ell^*}$ at the identified peak frequency f_{ℓ^*}.

Analysis Pipeline
For all methods, the motion estimation process is followed by displacement analysis in the frequency domain using FFT. Vibrational modes were extracted, and users could interactively select single or composite desired modes to simulate plausible object dynamics. These modes were then integrated into the simulation environment, enabling dynamic, interactive visualizations.

4 Experimental Results

The performance of the four motion estimation methods—Lucas-Kanade, CSP, RAFT, and the hybrid approach (LK + CSP)—was evaluated using qualitative visual results on different videos out of which results on two videos are shown in the Fig. 2 and user preferences shown in Table 1. Forty eight participants assessed the realism, accuracy, and overall effectiveness of the dynamic simulations generated by each method. The analysis focused on participants' rankings of the methods based on their qualitative impressions.

4.1 Qualitative Results

Simulations generated using the hybrid approach (LK + CSP) were consistently rated as the most visually realistic and accurate across a variety of scenarios. This method captured both subtle vibrational details and large-scale deformations, producing natural and plausible motion. In contrast, simulations using CSP often appeared overly exaggerated in high-force scenarios, while those using Lucas-Kanade alone lacked the detail necessary for finer motion nuances. RAFT produced visually acceptable results but was criticized for occasional artifacts in small-scale vibration scenarios.

4.2 User Preference Analysis

Participants were asked to rank the methods from 1 (most preferred) to 4 (least preferred) based on their qualitative assessment of the simulated results. The preference distribution is shown in Table 1.

TABLE 1: Preference Distribution of Methods (in Percentages)

Preference Rank	LK (%)	CSP (%)	RAFT (%)	LK+CSP (%)
1st preference	8.33	0.00	4.17	87.50
2nd preference	79.17	4.17	4.17	8.33
3rd preference	12.50	37.50	50.00	4.17
4th preference	0.00	54.17	41.67	0.00

The hybrid method (LK + CSP) received an overwhelming 87.50% as the first preference, highlighting its ability to produce the most realistic and robust simulations. Participants noted its effectiveness in balancing subtle independent component motion with broader displacement accuracy, making it superior in both small and large-scale deformation scenarios.

Lucas-Kanade was the second-most preferred method, with 79.17% ranking it second. While it lacked fine-grained detail, its stability and ability to capture large scale motion contributed to positive feedback. CSP, despite its ability to capture fine-grained motion, was less favored due to its poor performance in high-force scenarios, receiving a majority of third and fourth-rank preferences. RAFT, while more preferred than CSP, ranked lower than expected due to its computational expense and occasional artifacts in the background, particularly in scenarios involving subtle vibrations.

4.3 Discussion

The experimental results strongly favor the hybrid approach (L-K + CSP) as the optimal method for interactive video synthesis. Its dominance in first-preference rankings underscores its ability to integrate the strengths of Lucas-Kanade and CSP while mitigating their individual weaknesses. The qualitative feedback highlights that users value methods that strike a balance between detail, robustness, and realism, attributes that the hybrid approach delivers effectively.

These findings reinforce the hybrid approach's potential to set a new benchmark for vibrational modal analysis and interactive simulations. The qualitative assessment provides valuable insights into user perceptions of visual fidelity, emphasizing the importance of combining complementary techniques in motion estimation.

5 Conclusion

This report explored advanced techniques for vibrational modal analysis and interactive video synthesis, building on foundational work by [1] and [2]. Our findings demonstrate that the hybrid approach outperforms individual methods. Qualitative assessments, supported by a user preference study, revealed that the hybrid method was overwhelmingly favored, with 87.50% ranking it as their top choice for producing visually realistic and interactive simulations.

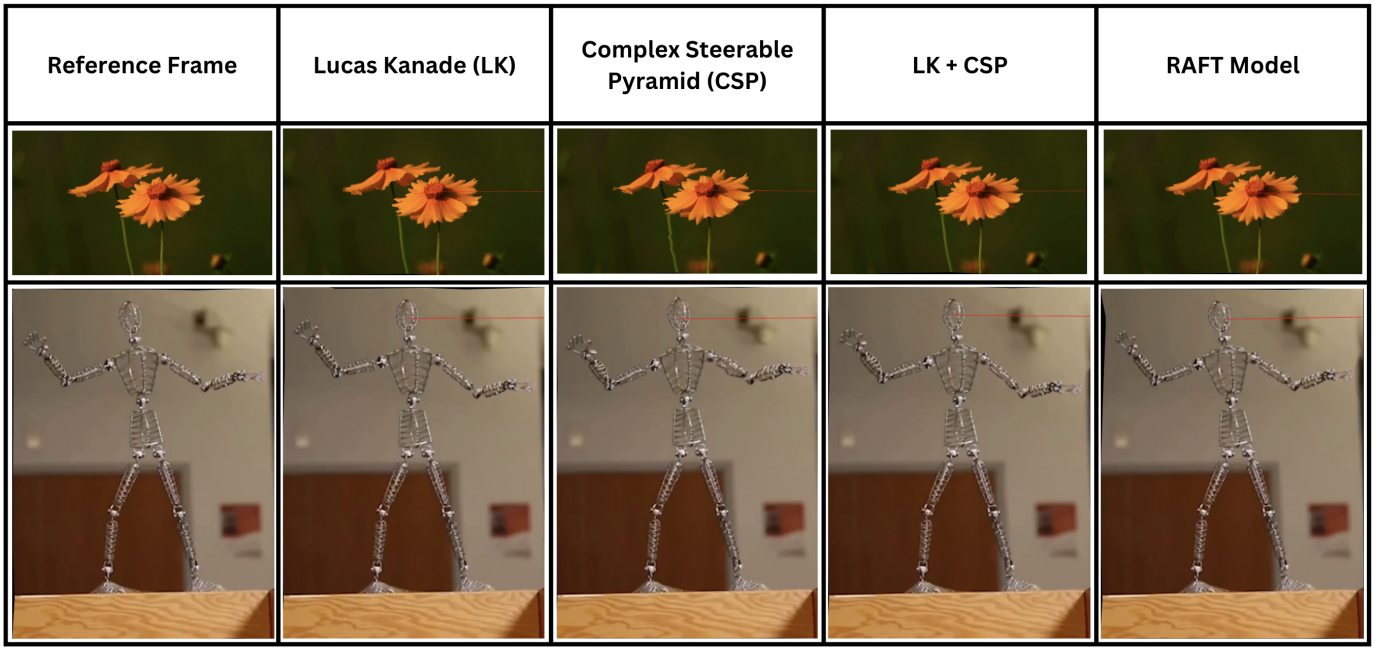

Fig. 2: Deformation of objects on applying force for all four methods

While Lucas-Kanade excelled in capturing large scale motions and CSP was highly sensitive to small-scale motions, both methods were limited in their standalone applications. RAFT provided results that produced artifacts in the background, making it less appealing visually.

By integrating the strengths of both Lucas-Kanade and CSP, the hybrid method provides a reliable and versatile solution for vibrational modal analysis. These results highlight the effectiveness of integrating multiple motion analysis techniques, with the hybrid method outperforming others in user rankings, making it a robust and generalizable solution for interactive video synthesis.

Future work could explore extending the hybrid approach to handle more complex scenes, incorporating machine learning to adaptively optimize performance, and integrating additional features to enhance real-time interactivity. This study serves as a stepping stone for advancing motion estimation techniques and unlocking new possibilities in video-based dynamic simulation.

Acknowledgments

We would like to express our appreciation to Shayan Shekarforoush for offering valuable feedback on our project. Our thanks also go to Dr. David Lindell, Dr. Aviad Levis and all the teaching assistants for their contributions to curating this exceptional course.

References

[1] A. Davis, J. G. Chen, and F. Durand, "Image-space modal bases for plausible manipulation of objects in video," *ACM Trans. Graph.*, vol. 34, no. 6, Nov. 2015. [Online]. Available: https://doi.org/10.1145/2816795.2818095

[2] B. T. Feng, A. C. Ogren, C. Daraio, and K. L. Bouman, "Visual vibration tomography: Estimating interior material properties from monocular video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 16 231–16 240.

[3] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI'81: 7th international joint conference on Artificial intelligence*, vol. 2, 1981, pp. 674–679.

[4] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 402–419.

[5] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman, "Phase-based video motion processing," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 4, pp. 1–10, 2013.

[6] S. Li, J. Huang, F. de Goes, X. Jin, H. Bao, and M. Desbrun, "Space-time editing of elastic motion through material optimization and reduction," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, pp. 1–10, 2014.

[7] D. L. James and D. K. Pai, "Multiresolution green's function methods for interactive simulation of large-scale elastostatic objects," *ACM Transactions on Graphics (TOG)*, vol. 22, no. 1, pp. 47–82, 2003.

[8] J. Shi and Tomasi, "Good features to track," in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 593–600.

[9] C. L. Lawson, "Transforming triangulations," *Discrete Mathematics*, vol. 3, no. 4, pp. 365–372, 1972. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0012365X72900933

[10] E. Simoncelli and W. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," vol. 3, 11 1995, pp. 444 – 447 vol.3.

[11] D. Fleet and A. Jepson, "Computation of component image velocity from local phase information," *International Journal of Computer Vision*, vol. 5, pp. 77–104, 08 1990.

[12] N. Wadhwa, J. G. Chen, J. B. Sellon, D. Wei, M. Rubinstein, R. Ghaffari, D. M. Freeman, O. Büyüköztürk, P. Wang, S. Sun, S. H. Kang, K. Bertoldi, F. Durand, and W. T. Freeman, "Motion microscopy for visualizing and quantifying small motions," *Proc. Natl. Acad. Sci. USA*, vol. 114, no. 44, pp. 11 639–11 644, 2017. [Online]. Available: https://doi.org/10.1073/pnas.1703715114