Semantic Captioning for Bharatanatyam via *Mushti* Motion Classification

Abstract

This paper presents a semantic analysis aimed at enabling AI closed captioning for Bharatanatyam (BN), with a first computational component that detects the *Mushti* (closed-fist) gesture and maps its motion to two semiotic meanings: courage and steadiness. Steadiness is defined as a downward motion of a closed fist facing the self, while courage is defined as a rotation of the closed fist that starts facing downwards and turns upwards toward the self. The system is a browser-based demo that runs on webcam input using MediaPipe hand landmarks and exposes its thresholds as live, persistent controls, which keeps the analysis transparent and tunable. The dataset includes 60 short clips (30 for courage and 30 for steadiness), each 2 seconds long. We do not report metrics yet; instead we focus on the interpretive link between gesture and meaning, and show how the *Mushti* classifier can serve as one measurable anchor for future captioning systems that are sensitive to semantic nuance in classical dance.

1 Introduction

Bharatanatyam is a structured system of embodied meaning, where hand gestures, timing, and spatial orientation produce semantic content that a trained audience can read. The long-term goal of this project is AI closed captioning for BN, so that a model can recognize motion and meaning rather than only track pose. This draft focuses on a single semantic unit: *Mushti*, a closed fist gesture whose motion is commonly read as courage or steadiness in performance contexts. In our framing, steadiness is a downward motion of a closed fist facing the self, while courage is a rotation that starts facing downwards and turns upwards toward the self. These are not treated as lexical words in isolation, but as semiotic meanings that arise from embodied form in a conventional system.

The project's contribution is therefore twofold. First, it provides a clear semantic description of how a specific gesture is mapped to interpretable meaning within BN practice. Second, it implements that description in a transparent, controllable computational component that can be inspected and tuned during live use. This pairing is necessary for captioning: without a formalized mapping between gesture and meaning, a captioning system has no stable target; without a runnable system, the mapping remains descriptive rather than operational.

The emphasis on a single gesture is deliberate. BN meaning is layered, and a captioning system must respect that layering rather than collapse it into generic labels. Starting with *Mushti* allows the project to define a precise motion-to-meaning link and to test whether a classifier can detect it consistently. This provides a foundation for scaling to additional gestures while keeping the semantic analysis grounded in the performance tradition.

2 Background and Motivation

BN hand gestures are organized as *hasta vinayogas*, the established uses and meanings of hand positions (Tandon, 2020). In this system, a handshape or mudra is a conventionalized sign that contributes meaning when placed in a specific orientation, location, and movement. The communicative unit is not the handshape alone, but the handshape-in-motion: a mudra gains semantic force through its alignment with body position, gaze, and rhythm. As a result, the same hand configuration can carry different meanings when its orientation or trajectory changes. This makes BN a natural setting for semantic analysis and for computational modeling that is sensitive to the structure of gesture (**?**).

We adopt a linguistically grounded view of gesture composition. Mudras can be treated as the phonological elements of dance: the smallest contrastive units of form (Tecelão, n.d.). When a mu-

dra is combined with position/location and motion, the system creates a word- or phrase-like unit, analogous to morphosyntax in language. When multiple gestures are sequenced, typically alongside facial expression and affect, the performance supports sentence- and discourse-level interpretations. This is not a claim that dance is reducible to language, but a claim that dance has a systematic compositional structure that can be described with the same analytic tools used in semantics and pragmatics (Patel-Grosz et al., 2018).

This compositional view aligns with standard descriptions of Bharatanatyam abhinaya, where meaning is conveyed through coordinated hand gestures, eye movement, facial expression, and rhythm (**?**). Hand gestures are a primary channel, but their interpretation depends on how they are embedded in a phrase of movement and expression. For captioning, this implies that a system must be able to anchor meaning at the level of individual gestures while remaining sensitive to motion and orientation cues that shift interpretation. The goal is to define these cues in a way that can be annotated and detected, rather than relying on vague descriptors.

This project treats BN as a system where form carries semantic content that can be described and annotated. The goal is not to reinterpret gesture in the abstract, but to specify concrete motion patterns that can be operationalized for captioning. *Mushti* is an appropriate starting point because it is visually distinctive and has a stable interpretive link to two specific semiotic meanings. That pairing makes it possible to define motion criteria, capture short clips, and test whether a classifier can track the same distinction that a human viewer would articulate in words. The broader motivation is to build a vocabulary of such mappings that can support future captioning beyond a single gesture.

3 System Overview

The system is a lightweight, browser-based app that uses MediaPipe hand landmarks to detect *Mushti* and classify vertical wrist motion. The interface pairs a live demo with an explanation panel, and it provides controls for thresholds and timing, with settings stored in local storage so calibration survives refreshes. Requirements are centralized in JSON files, including motion thresholds in `movements.json` and finger curl parameters in `mushti-requirements.json`. The controls make the model behavior visible and adjustable during live use, which aligns with the explainability focus of ACL 2026 and keeps the semantic claims tied to observable signals rather than black-box behavior (**?**). This positioning also reflects the broader field of dance informatics, where computational systems are developed to analyze and preserve dance structure (Joshi and Jadhav, 2019).

4 *Mushti* Detection and Semantic Mapping

Mushti is detected by comparing ratios of wrist-to-tip and wrist-to-MCP distances for each finger. A finger is treated as curled when its ratio falls below a threshold, and the gesture is marked as *Mushti* when a required count of fingers, including the thumb, meet their thresholds. This operational definition follows canonical descriptions of *Mushti* as a closed fist with all fingers bent into the palm (Tandon, 2020). Once *Mushti* is active, the system considers both vertical motion and rotation. Steadiness is defined as a downward motion of a closed fist facing the self, while courage is defined as a rotation of the closed fist that starts facing downwards and turns upwards toward the self. This distinction matters for captioning because it links a measurable motion pattern to a stable interpretive contrast, and it aligns the gesture with the traditional association of *Mushti* with *Veera* (heroism) and a steady state of mind (Rama, 2021).

The semantic mapping is intentionally explicit. Rather than treating courage and steadiness as class labels detached from meaning, we treat them as semiotic meanings grounded in embodied form. The mapping remains configurable so that the interpretation can be verified, discussed, and revised if additional BN sources motivate a different alignment. The system logs each detected label with a confidence score derived from motion magnitude, making it possible to compare predicted meanings with human judgments over short clips.

The mapping also clarifies the data requirements. A clip must contain a full motion arc for courage and a clear downward trajectory for steadiness, which guides how examples are selected and annotated. This is important for captioning because subtle differences in orientation or direction can change meaning. By encoding the distinction explicitly, the classifier can be evaluated against criteria that are interpretable to dancers and annotators, rather than only to model developers.

5 Data

The dataset used in this draft consists of 60 short clips: 30 two-second videos labeled as courage and 30 two-second videos labeled as steadiness. These clips serve as a minimal dataset for examining whether the *Mushti*-based mapping can support a semantic distinction that is meaningful in performance practice. At this stage we do not report evaluation metrics, since the dataset is intended to validate the semantic framing rather than optimize classification accuracy. Future work will add more gestures and longer sequences, which is necessary for a full captioning system.

6 Discussion

The main contribution is a clear link between a specific BN gesture and two semiotic meanings that are relevant for closed captioning. The *Mushti* classifier is therefore not an end in itself; it is one piece of a semantic vocabulary that can eventually cover a larger portion of the dance lexicon. The system also demonstrates how a transparent, adjustable pipeline can support semantic analysis by letting researchers tune thresholds in real time and observe how meaning assignments shift under controlled changes.

The linguistic framing is central to the project, not decorative. BN gesture is compositional: a mudra functions as a contrastive unit, but meaning is realized through its integration with motion and placement. This view allows a captioning system to aim at the level of meaning rather than raw kinematics. It also clarifies what counts as evidence for a caption. A classifier that detects a handshape without motion is incomplete in this framework, because the semiotic meaning depends on trajectory and orientation. The explicit operationalization of steadiness and courage reflects this principle.

This framing keeps a balance between computational clarity and linguistic interpretation, which reflects the project's dual goal: to remain precise in computational terms while still respecting the interpretive structure of performance. It provides a template for expanding the system to additional gestures, where the same method can be used to define motion criteria, label clips, and validate mappings against expert interpretation. Future extensions will need to address multi-gesture sequences where meaning is distributed across several hand configurations, and the present work can serve as a grounded starting point for that broader compositional modeling.

Limitations

This draft covers only one gesture and two semantic labels, and it uses a small dataset of short clips. The system has not yet been evaluated on longer sequences or on multiple performers, and it cannot capture the full range of meaning conveyed by timing or narrative context, nor does it model facial expression. These constraints are expected at this stage, but they limit claims about general captioning performance.

These limits have direct implications for the study. First, the current mapping should be read as a proof of concept for semantic anchoring rather than a comprehensive model of BN meaning. Second, because the data are short and tightly scoped, the system cannot yet address discourse-level interpretation, where meaning is distributed across sequences of gestures and expressions. Third, performance variability is not yet modeled, so the classifier may not generalize across different lineages, tempos, or stylistic choices without additional calibration. Finally, the present focus on a single gesture means that the broader significance of captioning for BN remains aspirational; the project demonstrates feasibility but does not yet deliver coverage of a performance's full semantic content. These limitations define the next steps for scaling the semantic inventory and evaluating how compositional meanings can be detected over longer spans.

Ethical Considerations

The system operates on local webcam input and does not transmit video off-device. Users should be informed when the camera is active and can stop the stream at any time. Future dataset expansion should include explicit consent from performers and clear policies for annotation use.

AI Assistance Disclosure

If generative tools are used for writing or coding, their use will be disclosed in the Responsible NLP checklist and acknowledgements, following ARR and ACL policy. The authors remain responsible for correctness and for proper citation.

References

Aishika Chakraborty, Syed Wasif Moin, Arpita Dey, and Ankita Bose. n.d. Dance (Bharatanatyam): The art of non verbal communication. *International Journal of English Learning & Teaching Skills*.

Manish Joshi and Sangeeta Jadhav. 2019. An extensive review of computational dance automation techniques and applications. arXiv preprint.

Tanwi Mallick, Patha Pratim Das, and Arun Kumar Majumdar. 2020. Bharatanatyam dance transcription using multimedia ontology and machine learning. arXiv preprint arXiv:2004.11994.

Pritty Patel-Grosz, Patrick Georg Grosz, Tejaswinee Kelkar, and Alexander Refsum Jensenius. 2018. Coreference and disjoint reference in the semantics of narrative dance. In Uli Sauerland and Stephanie Solt (eds.), *Proceedings of Sinn und Bedeutung 22*, vol. 2, 199–216. Berlin: ZAS.

Siri Rama. 2021. Multivector model analysis of hastas or hand gestures as bio semiotic conveyors of rasa. *Indica*.

Svetlana Ryzhakova. 2019. Hasta, mudra, viniyoga. Hand gestures in Indian culture: Problems of origin and sense making. *Scientific Articles*.

K. Sahayaraani. n.d. Semiotics of Indian classical dance: A study of Bharatanatyam. *International Journal on Science and Technology*.

Garima Tandon. 2020. The asamyuta hastas of Abhianyadarpana and Natyashastra: (In context of text and performing tradition). *Sangeet Galaxy* 9(2): 20–33.

André-Luiz Tecelão. n.d. The structure of hand gestures in Indian dancing according to Bharata's *Nāṭya-Śāstra*. Scribd.