Preserving and Annotating Dance Heritage Material through Deep Learning Tools: A Case Study on Rudolf Nureyev

Silvia Garzarella[1,†], Lorenzo Stacchio[2,*,†], Pasquale Cascarano[1], Allegra De Filippo[3], Elena Cervellati[1] and Gustavo Marfia[1]

[1]*Department of the Arts, University of Bologna, Italy*
[2]*Department of Political Sciences, Communication and International Relations, University of Macerata, Italy*
[3]*Department of Computer Science and Engineering University of Bologna, Italy*

Abstract

The cultural heritage of theatrical dance involves diverse sources requiring complex multi-modal approaches. Since manual analysis methods are labor-intensive and so limited to few data samples, we here discuss the use of the DanXe framework, which combines different AI paradigms for comprehensive dance material analysis and visualization. However, DanXe lacks models and datasets specific to dance domains. To address this, we propose a human-in-the-loop (HITL) extension to the DanXe to accelerate multi-modal data labeling through semi-automatic, high-quality data labeling. This approach aims to create detailed datasets providing humans with a set of user-friendly and effective tools for advancing multi-modal dance analysis and optimizing AI methodologies for dance heritage documentation. To this date, we designed a novel middleware that allows us to adapt data generated from visual Deep Learning (DL) models within DanXe to visual annotation tools, to empower domain experts with a user-friendly tool to preserve all the components included in the choreographic creation, enriching the process of metadata creation.

Keywords

Artificial Intelligence, Data Labeling, Deep Learning, Cultural Heritage, Dance

1. Introduction

The cultural heritage of theatrical dance consists of a multitude of sources, both tangible and intangible. These sources are diverse by nature and type, location, and preservation methods, creating a complex constellation that requires a diverse set of skills to be effectively enhanced [1]. Acknowledging this complexity is inherently tied to a comprehensive and integrated analysis of theory and practice, with significant implications in terms of accessibility [2]. Considering in particular choreography, while historiographical approaches are essential for working with written documentation, thorough analysis requires an understanding that often involves observing execution techniques [3] [4] [5].

International Workshop on Artificial Intelligence and Creativity (CREAI), co-located with ECAI 2024
*Corresponding author.
†These authors contributed equally.
silvia.garzarella3@unibo.it (S. Garzarella); lorenzo.stacchio@unimc.it (L. Stacchio); pasquale.cascarano2@unibo.it (P. Cascarano); allegra.defilippo@unibo.it (A. De Filippo); elena.cervellati@unibo.it (E. Cervellati); gustavo.marfia@unibo.it (G. Marfia)

Due to this challenge, we have attempted to envision a framework that allows for an integrated approach to theatrical dance's documentary assets, using the artistic and cultural legacy of dancer Rudolf Nureyev (1938-1993) as a case study [6]. The decision to focus on Nureyev stemmed from the distinctive nature of the documentary heritage associated with him. He was one of the first dancers to experience extensive and varied media coverage, given the period during which his career developed (the 1960s and 1980s). This widespread mediatization, during a transformative era for both dance and media, underscores the unique and multifaceted nature of his legacy, making him a pivotal example for the dance domain. However, it is worth noticing that in this case, as in others like it, the large amount of data available (e.g., dance videos, playbills, or biographical documents) and their international distribution often lead dance experts to apply multi-modal analysis to a limited number of samples. [7].

Such approaches exhibit three main limits: (a) even being an expert, the process of analyzing such data by hand is time-consuming; (b) the outcomes of such analyses would be hard to organize and visualize in an effective way (e.g., discover correlations); (c) it would prevent discovering semantical knowledge that could be only found by adopting a multi-modal analytical approach on a vast amount of data [8, 7, 9, 10]. To face all such challenges, Computational Dance (COMD) paradigms amount to a possible solution. However, COMD is underserved by comprehensive datasets, limiting the potential for in-depth research and development [7, 11, 12]. This lack is even greater when considering multi-modal dance datasets: the majority of datasets were collected for uni-modal analysis, in particular for the choreographic one [7, 11, 12]. Such datasets would be fundamental to optimizing AI methodologies capable of automatically extracting knowledge and labels from dance digital material [11].

In such a line, a recent work introduced a unified multi-modal analysis tool, DanXe, an Extended Artificial Intelligence framework that blends (i) AI algorithms for digitization and automated analysis of both tangible and intangible materials, with the goal of crafting a digital replica of dance cultural heritage, and (ii) XR solutions for immersive visualization of the derived insights. This framework introduces a novel space for the concurrent analysis of all elements that define the essence of dance. [12]. For the here considered use case, the AI analysis module of DanXe can be effectively used to extract knowledge for different kinds of dance heritage materials, since it employs different Deep Learning (DL)-based models to examine dance heritage materials, ranging from textual, audio, visual, and 3D data, providing a foundational framework for multi-modal dance analysis. However, such a framework does not resort to models specifically designed for the dance arena, for domains different from choreography, exhibiting again a lack of models and datasets.

Tools like DanXe can be employed to digitalize dance heritage and at the same time accelerate the labeling of multi-modal dance data, which can be used to train multi-modal models, that can be employed to improve heritage preservation and analysis. Nevertheless, the integration of human experts is required to ensure the quality of generated data and provide novel and connected knowledge to those. For this reason, we here propose an extension of the DanXe framework to inject a human-in-the-loop (HITL) component that leverages the initial AI-inferred annotations as a foundation, enabling a semi-automatic approach to provide high-quality labels. This approach aims to facilitate the creation of richer and more accurate datasets to support the optimization of future heritage preservation models. We here contextualize such an approach for a multi-modal dance data annotation process, considering the specific case of choreography,

where there is a lack of datasets that capture fine-grained labels of specific dance moves, often focusing on the general style [13, 14, 11].

2. Materials and Methods

We here provide a detailed overview of the materials and methodologies employed in our study. We begin with the **Video Dataset** subsection, which describes the collection and characteristics of the video data used for analysis. Following this, the **AI Augmented Human Annotator** subsection outlines the HITL approach that leverages an AI Dance toolbox to enhance human annotation efficiency and accuracy. Finally, the **Visual Annotation Tool Integration Middleware** subsection discusses the middleware designed and implemented to seamlessly integrate the synthetic AI-generated annotations in a visual annotation tool, facilitating a cohesive and streamlined workflow for annotation domain experts. Each subsection aims to elucidate the integral components and techniques critical to our research process.

2.1. Video dataset

The dataset was created using materials from the case study, which were originally recorded on film, distributed in cinemas and on VHS, and later digitized. The original recording format often suffered from wear and tear (e.g., film damage, darkening). Additionally, the original intended use, designed for cinemas or home video viewing, included video direction elements such as close-ups, zooms, and fade-ins/outs. These elements often cover movements and are not ideal for a comprehensive recording of the performance. The process of selecting a video for building the initial basic dataset was therefore inevitably influenced by the need for well-lit footage, the highest possible definition, and minimal directorial interventions. To further reduce noise (e.g. background dancers, extras), it was decided to analyze a solo performance: Nureyev's adagio of Prince Siegfried in *Swan Lake* (Act I). In the analysis of this adagio, we've focused on the initial 20 seconds of choreography. Here, the dancer transitions from a static pose, embodying their character without movement, to a sequence of steps performed in place. These steps showcase a range of volumes and heights, adding depth and dimension to the performance.

2.2. AI Augmented Human Annotator

Considering our main use case, choreographic-related data, various labels, and information can be inferred, including music, dance styles, individual dance moves, and background descriptions. Some of this information could be inferred with a high degree of accuracy by modern DL approaches, like the ones introduced in the DanXe platform [12]. Despite this rich potential, it remains challenging for human experts to adjust, enhance, and integrate novel labels or information clearly and visually on top of this generated data. On this line, visual annotation tools (VATs) could be exploited [15]. In fact, the primary advantage of VATs is their ability to significantly reduce the manual effort required from users, even those who are non-experts. By incorporating various functionalities for manual, semi-automatic, and automatic annotations through advanced AI algorithms, VATs could accelerate high-quality data labeling [15], given also by the natural quantitative and qualitative approach introduced in such a process [16].

Given such consideration, we employed the DanXe visual annotation module as a black box capable of inferring different relevant data for the dance visual domain, such as textual data within pictures, human pose estimation, and semantic segmentation and defining a novel visual-annotation-based framework on top of it. This is visually represented in Figure 1.

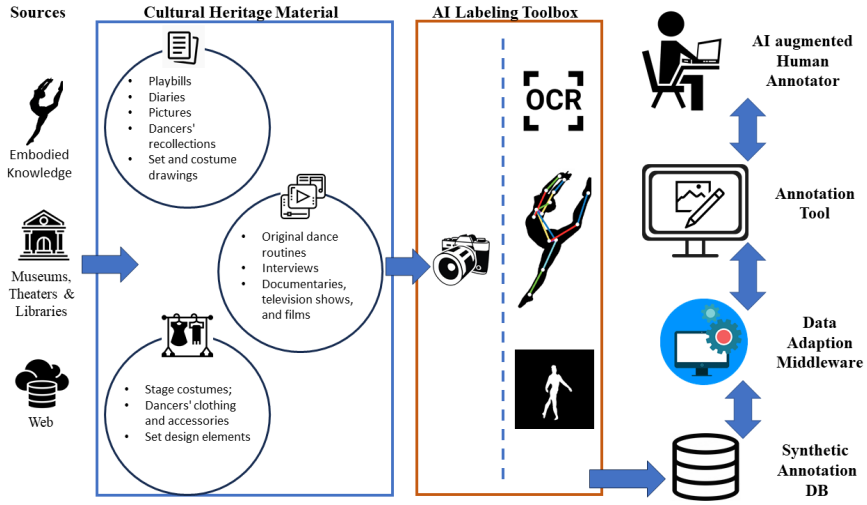

Figure 1: HITL Augmented Semi-automatic Annotation Architecture.

This annotation layer assumes that all synthesized AI label data are stored in a local database after their inference. A data adaptation middleware ingests and transforms the various data formats inferred by different AI models, ensuring compatibility with the visual annotation tool at hand. This setup enables human annotators to use the tool to correct and add new labels on top of the existing information. Subsequently, the updated annotations are re-adapted and stored in the database, following the inverse chain of processes. This iterative approach facilitates the efficient enhancement of dance video annotations, leveraging both AI and human expertise. To implement such a framework, a fundamental step amounts to defining a smart middleware able to bridge different file formats and data structures coming from AI models and make them interpretable from different visual annotation tools. For this reason, in the following, we will describe the general architecture of the middleware we defined to ingest and adapt annotations coming from different AI tools to visual annotation tools.

2.3. Visual Annotation Tool Integration Middleware

In response to the growing complexity of data formats and interpretation regarding different tasks (e.g. Human Pose Estimation), a middleware solution has been developed to foster interoperability between diverse AI models and various visual annotation tools. This middleware serves as a bridge, facilitating seamless communication and data exchange between different components of the annotation pipeline. Its architecture is reported in Figure 2. Implementing standardized interfaces and protocols, enables the integration of multiple deep learning models, each specializing in different aspects of visual analysis, such as pose estimation or object detection. Simultaneously, the middleware performs a conversion of the ingested data respecting

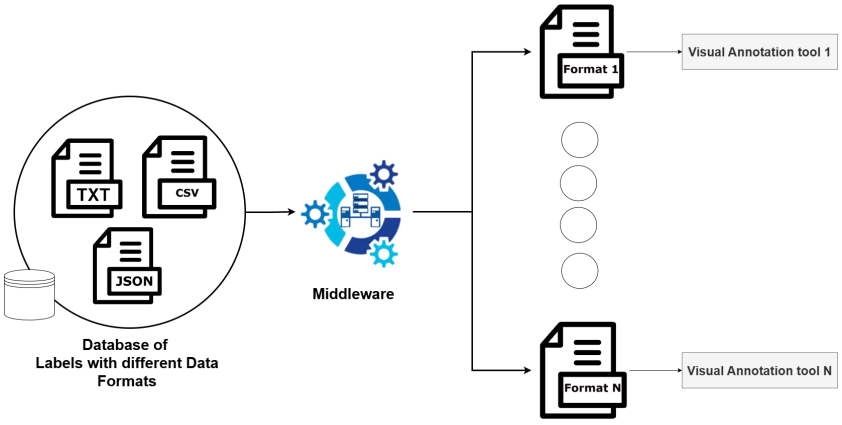

Figure 2: Middleware Architecture.

a range of different visual annotation tools interfaces, providing a unified platform for annotators to interact with and refine the output of these models. Through this interoperability, the annotation process is augmented, offering annotators the flexibility to leverage the strengths of different models while at the same time having user-friendly interfaces. Moreover, by automating certain aspects of annotation and providing semi-automatic functionalities, the middleware accelerates the annotation workflow, significantly reducing the time and effort required to generate high-quality annotated datasets.

3. Results

We concretely applied our introduced methodology to accelerate single dance moves annotation from a multi-modal perspective (i.e., linking human pose estimation and single dance moves). To the best of our knowledge, this is the first attempt to do so through a custom-defined middleware and semi-automatic approach.

Figure 3: Visual Keypoints inferred by AlphaPose.

In particular, we took as a use case choreographical human pose estimation by using the AlphaPose models [1] that were included in the DanXe pipeline. AlphaPose allows to extract and

[1]https://github.com/MVIG-SJTU/AlphaPose

track multi-person poses, codified into 17 body key points when used in the model trained on the COCO dataset [17]. In our case, we applied it for a variation from *Swan Lake* performed by Rudolf Nureyev in 1967. The human pose estimation extracted from AlphaPose is stored in a JSON file which contains one record per each frame where a person was detected. Some visual representations of the inferred key points are reported in Figure 3 while an example of the resulting JSON file is provided in Listing 1.

Listing 1: Human pose estimation JSON data generated by AlphaPose on a single image.

```
{   "image_id": "0.jpg",
    "category_id": 1,
    "keypoints": [
        311.9952087402344, 307.96734619140625, // nose
        314.58056640625, 305.3819580078125, // right eye
        311.9952087402344,304.08929443359375, // left eye
        322.336669921875, 305.3819580078125, // right ear
        309.40985107421875, 305.3819580078125, // left ear
        330.0927734375, 322.1868591308594, // right shoulder
        306.8244934082031,322.1868591308594, // left  shoulder
        333.9708251953125,340.2843933105469, // right elbow
        299.0683898925781 338.9917297363281, // left elbow
        327.5074157714844,357.0892639160156, // right wrist
        286.1415710449219,350.6258544921875, // left wrist
        322.336669921875, 357.0892639160156, // right hip
        310.7025451660156,355.7966003417969, // left hip
        323.6293640136719,386.82098388671875, // right knee
        314.58056640625, 386.82098388671875, // left knee
        323.6293640136719, 415.260009765625, // right ankle
        318.4586181640625,411.3819580078125 // left ankle
    ], "score": 3.0010504722595215,
    ... }
```

In this example, there was only one person identified (category ID 1), indicating the human ID within the considered frame. The key points array contains precise x and y coordinates along with confidence scores for various body joints, exemplified by the first key point positioned at (311.995, 307.967). We do not include the confidence score provided for each key point inferred for description simplicity. As mentioned, those are 17 key points, corresponding to the nose, eyes, ears, shoulders, elbows, wrists, hips, knees, and ankles. Each key point serves as a precise indicator of a specific body part's location within the image frame. The overall confidence in the pose estimation is quantified by a score of 3.001. Also, information related to the bounding box enclosing the detected human figure is inferred but was not reported for simplicity.

Starting from this representation, we then considered adapting it for our target visual annotation tool Vidat [2], which could be exploited. Vidat is a high-quality video annotation tool for computer vision and machine learning applications that is simple and efficient to use for

[2]https://github.com/anucvml/vidat

a non-expert and supports multiple annotation types including temporal segments, object bounding boxes, semantic and instance regions, and human pose (skeleton). Moreover, it is completely data-driven: all the data can be stored and loaded by encoding them in a predefined key-value structure (i.e., a JSON file). Our goal was to load the annotated data from AlphaPose in a format readable by Vidat. However, the Vidat skeleton structure description does not take into account the elbow data. This means that we first filtered out the data per each detection and then re-adapt the remaining information to match the reading structure of the Vidat tool. The resulting JSON is reported in Listing 2.

Listing 2: JSON representation of video annotations and configurations.

```
{   ...,
    "objectAnnotationListMap": {},
    "regionAnnotationListMap": {},
    "actionAnnotationList": [],
    "skeletonAnnotationListMap": {
        "0": [{  ...,
                "pointList": [
                    {   "id": 0, "name": "nose",
                         "x": 312.0, "y": 308.0},
                    {   "id": 1, "name": "left eye",
                        "x": 312.0, "y": 304.0},
                    {   "id": 2, "name": "right eye",
                        "x": 315.0, "y": 305.0},
                    ...
                ], "centerX": 315.67, "centerY": 346.13}
        ]
    },
    "config": {
        "objectLabelData": [...],
        "actionLabelData": [...],
        "skeletonTypeData": [...]
    }}}
```

The provided JSON encapsulates metadata crucial for video annotation and analysis. Within its structure, key parameters such as video dimensions, frame rate, and duration are outlined, essential for Vidat temporal analysis and processing (not reported in the example for simplicity). The inclusion of keyframe listings offers strategic markers for video segmentation and analysis, facilitating efficient data handling. Furthermore, the presence of object and region annotation maps anticipates future expansion into object detection and spatial characterization. The delineation of action annotation lists underscores the intention to annotate dynamic data. Particularly noteworthy is the skeleton annotation list, which furnishes detailed skeletal representations. The configuration segment provides an extensive catalog of object and action label data, coupled with skeleton-type specifications, forming the cornerstone for semantic understanding and classification in video content.

Finally, since this JSON is aligned with the original video frames, it can be loaded into the Vidat visual annotation tool. Our dance domain expert used the inferred human key-point labels to add new dance move labels, supported by the already-generated dance poses. Each label corresponds to a name (e.g., arabesque) and a time interval, representing the duration of the step execution. After completing the label descriptions, the next step would normally amount to the labeling of human movement frame by frame, but those were already labeled automatically generated, so the domain expert only corrected minor interpolations or mismatches between the skeleton and the video image. Finally, the dance domain experts annotated dance moves linked with one or more inferred poses. The resulting JSON can be stored at any moment, and will now include both skeleton and dance move label data. The outputs of such a process are visually reported in Figure 4.

Figure 4: Action labeled with VIDAT

4. Discussion and Conclusion

The introduction of the DanXe framework represents a significant leap forward in digitizing and analyzing dance heritage materials, offering promising capabilities for the automatic annotation of archive videos. Supported by human oversight and augmented by XR technologies, the proposed multi-modal, semi-automatic annotation framework signifies a substantial advancement in cultural heritage conservation, especially in cases involving intangible heritage alongside tangible assets. Given the unique nature of the analyzed case study (that of an archival collection related to a dancer's legacy), the annotations cannot be limited to just recognizing steps but must also allow for tracking props, stage settings, and performers involved. This would enable the preservation of all scenic components that contributed to a choreographic creation, ensuring better preservation, facilitating restaging processes, and enriching the process of metadata creation, which is typically limited to principal performers or even just the choreographer. Providing a tool that can support the work of scholars and archivists, without replacing their expertise but leveraging it to validate semi-automatic acquisitions, not only represents a valuable contribution in expediting their work but also enriches the metadata associated with archival sources, thus enabling user research. This approach promises to generate richer, more accurate datasets, ultimately fostering a deeper understanding and appreciation of the art form.

Acknowledgments

This work was partly funded by: (i) the PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 8 "Pervasive AI", funded by the European Commission under the NextGeneration EU program.

References

[1] J. Adshead-Lansdale, J. Layson, Dance history: An introduction, Routledge, 2006.
[2] M. De Marinis, Il corpo dello spettatore. performance studies e nuova teatrologia, Sezione di Lettere (2014) 188–201.
[3] E. Giannasca, Dance in the ontological perspective of a document theory of art, Danza e ricerca. laboratorio di studi, scritture, visioni 10 (2018) 325–346.
[4] E. Randi, Primi appunti per un progetto di edizione critica coreica, SigMa-Rivista di Letterature comparate, Teatro e Arti dello spettacolo 4 (2020) 755–771.
[5] S. Franco, Corpo-archivio: mappatura di una nozione tra incorporazione e pratica coreografica (2019).
[6] J. Kavanagh, Rudolf Nureyev: the life, Penguin UK, 2013.
[7] K. El Raheb, Y. Ioannidis, Dance in the world of data and objects, in: International Conference on Information Technologies for Performing Arts, Media Access, and Entertainment, Springer, 2013, pp. 192–204.
[8] L. A. Naveda, M. Leman, Representation of samba dance gestures, using a multi-modal analysis approach, in: Enactive08, Edizione ETS, 2008, pp. 68–74.
[9] N. Li, Q. Shen, R. Song, Y. Chi, H. Xu, Medukg: a deep-learning-based approach for multi-modal educational knowledge graph construction, Information 13 (2022) 91.
[10] L. Church, N. Rothwell, M. Downie, S. DeLahunta, A. F. Blackwell, Sketching by programming in the choreographic language agent., in: PPIG, 2012, p. 16.
[11] R. Li, J. Zhao, Y. Zhang, M. Su, Z. Ren, H. Zhang, Y. Tang, X. Li, Finedance: A fine-grained choreography dataset for 3d full body dance generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10234–10243.
[12] L. Stacchio, S. Garzarella, P. Cascarano, A. De Filippo, E. Cervellati, G. Marfia, Danxe: an extended artificial intelligence framework to analyze and promote dance heritage, Digital Applications in Archaeology and Cultural Heritage (2024) e00343.
[13] O. Alemi, J. Françoise, P. Pasquier, Groovenet: Real-time music-driven dance movement generation using artificial neural networks, networks 8 (2017) 26.
[14] T. Tang, J. Jia, H. Mao, Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis, in: Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 1598–1606.
[15] S. Bianco, G. Ciocca, P. Napoletano, R. Schettini, An interactive tool for manual, semi-automatic and automatic video annotation, Computer Vision and Image Understanding 131 (2015) 88–99.
[16] L. Stacchio, A. Angeli, G. Lisanti, G. Marfia, Applying deep learning approaches to

mixed quantitative-qualitative analyses, in: Proceedings of the 2022 ACM Conference on Information Technology for Social Good, 2022, pp. 161–166.
[17] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, C. Lu, Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).