Integrating 3D Objects in Multimodal Video Annotation

Rui Rodrigues

NOVA LINCS, NOVA School of Science & Technology, NOVA University Lisbon, Portugal, and Sustain.RD, Setúbal School of Technology, Polytechnic Institute of Setúbal, Portugal, rui.rodrigues@estsetubal.ips.pt

Stephan Jurgens

Second author's affiliation, possibly the same institution, xxxx@gmail.com

Carla Fernandes

ICNOVA, FCSH, NOVA University of Lisboa, Portugal, carla.fernandes@fcsh.unl.pt

João Diogo

NOVA School of Science & Technology, NOVA University Lisbon, Portugal, jp.diogo@campus.fct.unl.pt

Nuno Correia

NOVA LINCS, NOVA School of Science & Technology, NOVA University Lisbon, Portugal, nmc@fct.unl.pt

This paper presents and discusses the introduction of 3D functionalities for an existing web-based multimodal video annotation tool. Over the past years, we have developed a multimodal web video annotation tool that now combines 3D models and 360º content with more traditional annotation types (e.g., text, drawings, images), offering users the possibility of adding extra information in their annotation work. We show how 3D models augment the annotation work and add advantages like viewing or exploring objects in detail and from different angles. The paper reports detailed feedback from a pilot study in form of a workshop with traditional dance experts to whom these new features were presented. We conclude with an outlook of future iterations of the video annotator based on the experts' feedback.

CCS CONCEPTS • Human-centered computing • Human-computer interaction (HCI) • HCI theory, concepts and models

Additional Keywords and Phrases: Video Annotation, Virtual 3D models, Performing Arts; HCI

ACM Reference Format:

First Author's Name, Initials, and Last Name, Second Author's Name, Initials, and Last Name, and Third Author's Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. In Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA, 10 pages. NOTE: This block will be automatically generated when manuscripts are processed after acceptance.

1 INTRODUCTION

This paper describes the application of 3D functionalities for an existing web-based multimodal video annotation tool, which has been developed in the context of a large-scale European research project. First, we present the previous versions of the annotation tool. Next, we focus on the current version, enhanced with the novel functionality of customizable 3D models ready to be used as a new and different annotation type. As an additional 3D functionality, we introduced the option to import and visualize 360º images and videos.

Although MotionNotes can be used in any context where the purpose is to annotate human motion in video, the current version has been developed with a focus on performing arts and their respective usability scenarios. The European project funding our work is dedicated to creating new tools and content for the analysis and documentation of tangible and intangible heritage, fostering innovative progress in applying 3D models to cultural and artistic contexts.

The use of multimedia elements is an excellent technique for increasing our communication capacity and improving the audience's interest. Analyzing video recordings positively supports the learning of multiple activities in different areas [7, 10, 15]. Hence, annotations are often used in various daily tasks since writing down our ideas is essential to retaining important information to help invoke memory in future needs.

Exploring virtual three-dimensional (3D) models enables the user to visualize and inspect a specific object; they can be observed from multiple angles, spun in different directions, and zoomed in/out. In the past, hardware and software limitations have restricted the usage of 3D models and the seamless integration with other media. Nonetheless, the rapid evolution of computer hardware, particularly graphics cards capable of rendering detailed 3D models, and the availability of high-speed internet connections have opened multiple possibilities in this area of research [1, 6].

Moreover, this paper introduces MotionNotes, a real-time multimodal web video annotation tool based on keyboard, touch, and voice inputs. Five different ways of adding annotations were previously developed: voice, draw, text, web URL, and mark annotations. Additionally, we conducted a pilot study in the form of a workshop with experts in traditional Portuguese dances to understand the real benefit of using 3D models in conjunction with the traditional video annotation modalities. As a result, the specialists were asked for their suggestions regarding the future development of our video annotator.

This paper presents the following main contributions:

- A web-based video annotation tool capable of managing 3D models as time-framed digital annotations.
- Workshop outcomes, where developers worked with traditional dance experts to assess the current software features and discuss further research directions.

This paper is structured as follows: we start by analyzing the related work, followed by contextualizing MotionNotes and our previous work. After that, we introduce the tool's new features and technologies used in their development. Subsequently, we present the feedback obtained in a workshop with experts in traditional Portuguese dances. Lastly, in the conclusion section of the paper, we highlight the tool's potential for a more holistic perspective of the annotated work and provide an outlook regarding future work.

2 RELATED WORK

Several international projects have been investing in the development of innovative tools for digital cultural heritage documentation, storage, and public accessibility, as well as in creating and compiling novel content to be added to large European archives and databases [5]. Video annotation is a valuable resource in many different application areas.

Furthermore, user-friendly video annotators are essential tools to assist and analyze video documentation and knowledge transfer in general. This has motivated the development of several annotation tools over the last few years.

ELAN [16] is one of the most well-known and used tools in manually annotating video content, and it is commonly used for the transcription and analysis of human interaction and communication, especially regarding gesture studies in conversation. Cabral et al. [2, 3] presented Creation-Tool, a concept of pen-based video annotations using frame differences in order to track motion in video features, and where the object tracking method was fast enough for real-time annotations. The Choreographer's Notebook [14] was designed specifically for choreographers and dancers, allowing only two types of annotations: digital-ink and text annotations. Piecemaker [4], by the Motion Bank project, is also a video annotation tool for dance analysis. As part of the BlackBox project [11], a prototype was developed to experiment with annotations in a virtual reality environment using the Microsoft Kinect. The web-based Movement Library [9] is another tool specifically designed to archive and search dance movements in the framework of the WholoDance project. However, none of these previous tools allows real-time recordings and multimodal annotations simultaneously while supporting 3D models to enhance user perception, allow the exploration of particular details in a 3D environment and analyze the object's characteristics.

Before the current version of MotionNotes, we have developed other prototypes from scratch, being amongst the pioneer developers of real-time video annotator prototypes at an international level, especially regarding cultural heritage and performing arts-education contexts. Our tool was designed to integrate different types of technologies, multimedia data, multimodal interaction, AI, and 3D modeling. It combines several of the functionalities present in other tools, adds other annotation types and modes, and is available in a single web-based environment [2].

MotionNotes [13] can assist both professional and amateur users working in any creative and exploratory setting where the analysis and improvement of human motion performance is the focus. Since the first version (originally developed as a stand-alone application), the available annotation types are text notes, ink strokes, short audio instructions, user-configured marks, and URL hyperlinking. In the newest version, 3D models have been integrated, and the details will be described below.

3 PREVIOUS LAB DAYS

Prior to our most recent experiment, and in order to evaluate the tool's potential and understand the procedures used in the video annotation, a testing session was performed in 2020, using a previous version of MotionNotes. We organized a user study where most participants reported that they frequently annotate their work. Briefly, throughout their interaction, users had close contact with the tool using all the annotations available [12].

From our results, we concluded that participants preferred to annotate the video recording session in a post-production phase and not in real-time, i.e., while the scene is being recorded. Additionally, they preferred to work in a traditional hardware with a larger screen instead of using the currently popular mobile devices. Regarding the annotation types, the most discussed one was the customizable marks where icon images are used to provide semantic meaning to a specific video frame. Participants showed great interest, given its novelty. Consequently, different ideas were brought forth, and 3D became the subject of debate as a possible new annotation type. This has led us to a new discussion topic, namely on how relevant it would be to upload 3D models and add them to a scene as if they were annotations *per se*.

4 ANNOTATION TOOL

The design guidelines were preserved from the previous versions since they have been very well accepted. The menus are on the top, and the multiple-input modalities are on the left, with their respective properties being displayed on the right.

MotionNotes uses a canvas layer overlaid on top of the displayed video frame to enable users to select and customize the position of any annotation (see Figure 1). To give a concrete example, users can sketch on top of the video and MotionNotes saves the ink strokes in the particular region with the specific timestamp. Editing and customizing all annotation types is also possible. The new 3D features were carefully integrated into this new MotionNotes version.

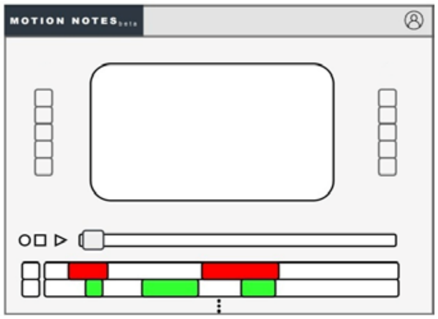

Figure 1- MotionNotes user interface mockup.

We developed support for the 3D model's management, e.g., uploading new models, selecting between available models, or deleting them. Moreover, it is possible to associate different 3D models to our videos in a specific timestamp. The objective was to create a new annotation type to provide users with even more possibilities during the annotation work. One example is the possibility of adding 3D annotations of elements contained in the video, while offering the possibility to interact with the objects and gain access to details that were impossible to identify by simply watching the raw video.

Therefore, we developed multiple new features for this new version:

- A 3D model Importer interface. Users should be able to upload their 3D models to MotionNotes servers and have them available while creating new annotations.
- A 3D model Viewer. Each user will have his collection of 3D models available in the respective account. Therefore, it is essential to be able to look at the model, rotate it, and zoom it in/out (see Figure 2).
- A 3D annotation mechanism. After selecting which 3D model to use, users can annotate any part of the video frame. MotionNotes will then save the timestamp and respective position in which the 3D model was added (see Figure 3).

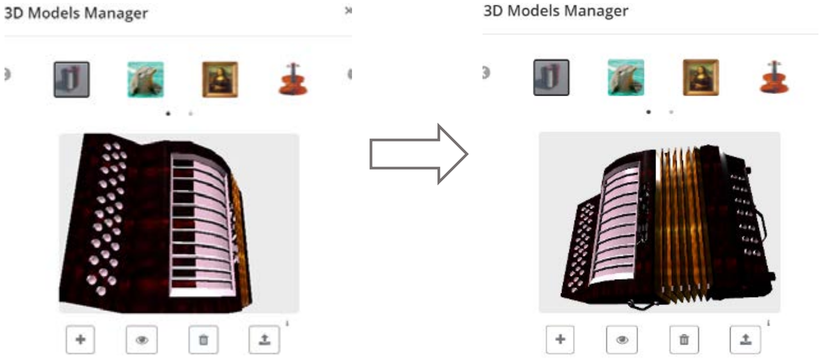

Figure 2- MotionNotes 3D Model Manager with user rotating the accordion.

4.1 3D Annotations Implementation Details

MotionNotes is a single-page application (SPA) developed using a client-server architecture to enable users to share their videos and annotations. The web browser works as the application client due to wide availability and compatibility with almost every user device with internet access.

Regarding the 3D functionalities in the browser, WebGL [8] is the technology behind it. This Javascript API, available under the canvas HTML element, supports interactive 3D graphics rendering without the need to install additional browser plugins. In order to improve developers' productivity, a solution able to abstract WebGL low-level details was considered. As a result, this approach simultaneously prevents long periods of development and ensures coding quality with shorter testing periods. We do so by using Three.js, an open-source library capable of handling GPU-accelerated 3D animations using the Javascript language. Consequently, creating complex 3D computer animations that display in the browser becomes substantially more attainable.

Three.js offers excellent compatibility with GLTF [17], a flexible format for efficient transmission/loading of 3D content. Assets may be provided in JSON (.gltf) or binary (.glb) format. This library also gives us support to work with lights, cameras, and controls.

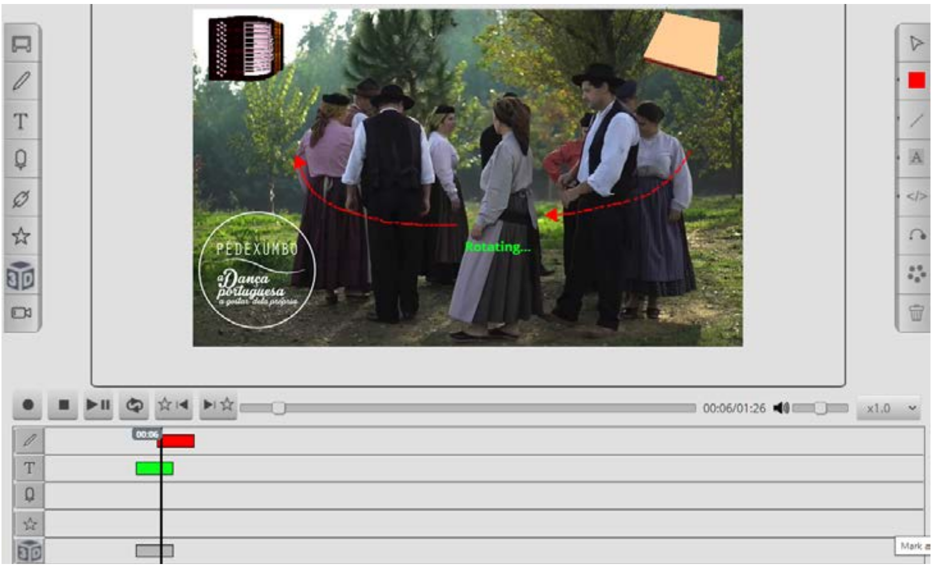

Figure 3- MotionNotes containing three annotation types: Top - 3D Models; Center - Drawing and text annotations.

5 WORKSHOP FEEDBACK

We organized a workshop session with Portuguese traditional dance experts, who are also partners in the EU project funding this research (see Figure 4). The workshop's goal was to obtain feedback from the experts regarding the novel 3D functionalities of our new video annotation software in the context of traditional dance forms and cultural heritage in general. Additionally, we elicited suggestions for further design iterations that could benefit the use case scenarios in the context of their work as researchers, educators, and practitioners in the field.

The participants work as founders and directors of a cultural association dedicated to Portuguese traditional dance forms. On the one hand, they are researching and preserving dances that have fallen into oblivion and are no longer danced in their original communities. On the other hand, they work on documenting dances that are still regularly re-enacted in their respective communities. Notably, this cultural association also supports the artistic creation of new dance formats by exploring the concept of social dances and related themes.

Figure 4 - Workshop session with specialists in Portuguese traditional dance

5.1 3D object annotations

5.1.1 Musical instruments as 3D annotations

The first type of 3D annotation presented to the workshop participants was the addition of smaller-scale 3D models of musical instruments on top of a 2D video layer. In our example, the appearance of an accordion, a triangle, and an *adufe* (a unique Portuguese type of tambourine) marked those moments in the video of a traditional dance called *Rancho*, in which these instruments started to play. Our goal here was to enhance the perception of the musical accompaniment of the dance, as well as to let the users explore a particular instrument in a 3D environment to understand its characteristics. In the case of the *adufe,* this intention was particularly justified since that instrument is fundamental in the context of Portuguese traditional music.

Using musical instruments as 3D annotations made a lot of sense for the experts, who pointed out that the relation between dance and music in some traditional dances can be rather complex, for instance when one of the dancers assumes the role of the *mandador*, usually a senior leader who organizes the dance in real-time. The *mandador* uses verbal instructions, clapping hands, and musical instruments to indicate a new choreography of the dancers for the following dance sequence. From the specialists' perspective, a 3D annotation of these types of dances could be very useful to understand and teach such dances, which are partly improvised and rely on an acquired movement vocabulary shared by the community.

5.1.2 3D object annotations as historical and cultural references

Moreover, the experts suggested that 3D object annotations could be used to refer to the cultural and historical background of the dances. One aspect in this context is the use of 3D object annotations as historical references, which could help to understand better how the dance was performed: what kind of costumes were worn, which objects were used in the performance, and on what occasions the dance was carried out. A different yet related aspect could be the origin of a danced movement: during harvest time, for example, the picking of olives is a gesture that has been incorporated as an arm movement in the traditional dances. Another example is the *Baile para fazer o chão das casas* (Dance to make the floor of houses), a dance performed by neighbors and friends to even out the clay floor during the construction of a new house. Or the *Fandango*, a dance only performed by men (most often in taverns), who took on male and female roles. The workshop participants further suggested that animated 3D object annotation could be very helpful, particularly for quick tap dances, where it is often challenging to learn the correct combinations of steps.

5.1.3 Geographical mapping through the use of 3D object annotations

Last but not least, our workshop participants mentioned that some traditional dances (especially those at risk of disappearance) could nowadays only be performed in a specific region of the country, or differ in the way they are taught and performed in different regions. Hence, they suggested using 3D culture-bound objects that would help locate those dances in specific geographical regions of the respective countries.

5.2 360º images and video

Following our discussion of 3D object annotation, we presented an additional 3D functionality to our workshop participants, namely the possibility to import 360º images and videos (in the equirectangular format), which can be visualized and navigated in the tool (see Figure 5) and combined with the other annotation modalities. Both experts saw great potential in using this 3D functionality for creative processes and visualization of technical aspects of dance performances.

A frequent situation that choreographers are facing today is the limited rehearsal space and time available during a creative process. Being able to acquaint themselves with the dimensions and characteristics of a studio space or theatre stage (through 360º imagery) could help them prepare for a rehearsal period beforehand. The same holds true for theatre technicians, who need to visualize a studio space or stage, ideally combined with technical riders and floor plans. Our workshop participants felt that 360º imagery could be beneficial for technicians, if combined with measuring tools and floor plans (as available in some commercial virtual tour software packages today). However, the specialists also suggested that further development of these 3D functionalities could enhance the creative process. Being able to simulate or even rehearse parts of a dance could be very exciting. As a good example, they mentioned the *dança de mastro* (Dance of the pole), in which a huge pole decorated with several ribbons is maneuvered skillfully by the dancers in such a way that the ribbons are woven into a colorful fabric. Studying those movement patterns beforehand, in a 3D space, would save the dancers hours of rehearsal time when they get to the real studios.

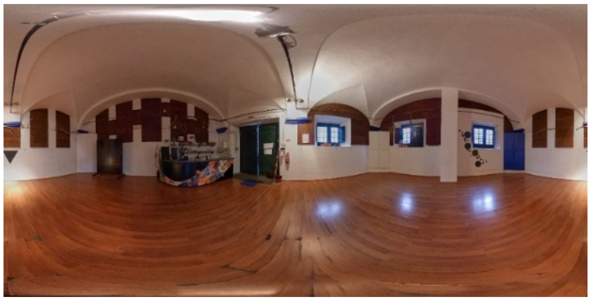

Figure 5 - 360º image of cultural association's studio (in equirectangular format).

6 FUTURE RESEARCH AND DEVELOPMENTS

The feedback we obtained during the workshop with our invited experts will certainly inform the design and development of the 3D functionalities described above. In particular, we plan to include 1) the use of animated 3D objects; 2) the combination of different scaled 3D objects (for example, a large-scale model of a historic building in combination with a mid-scale model of a stage and several small-scale models, such as the musical instruments); and 3) the merging of 360º video with 2D videos, from which performers have been isolated through subtraction so that a dance performance can be visualized in different environments.

7 CONCLUSION

In this paper, we presented the introduction of novel 3D functionalities for an existing multimodal web-based annotation tool. This tool enables users to work on multimedia content by combining various types of annotations such as text, ink strokes, audio, personalized imagery (marks), and most recently, 3D models and 360º content. We briefly described the technologies that support these features and the new possibilities that this novel type of 3D annotation opens up, particularly because the users of our annotation tool will be able to directly access substantial libraries of 3D content developed by our European project partners.

Subsequently, this paper presented a pilot study with experts in the context of traditional Portuguese dances, which has shown that using 3D annotations is definitely useful, as they enable a richer and more holistic perspective of the work at hand. As a result, we concluded that further work towards the use of animated 3D objects, combinations of differently scaled 3D models, and the usage of 360º content could generate the next impactful iteration in our tool's development cycle. Providing a more refined and additional level of annotation detail will bring multimodal video annotation closer to real-world experiences.

REFERENCES

[1] Baji, T. 2018. Evolution of the GPU Device widely used in AI and Massive Parallel Processing. *2018 IEEE Electron Devices Technology and Manufacturing Conference, EDTM 2018 - Proceedings*. (Jul. 2018), 7–9. DOI:https://doi.org/10.1109/EDTM.2018.8421507.

[2] Cabral, D. et al. 2011. A creation-tool for contemporary dance using multimodal video annotation. *MM'11 - Proceedings of the 2011 ACM Multimedia Conference and Workshops* (2011).

[3] Cabral, D. et al. 2012. Evaluation of a multimodal video annotator for contemporary dance. *Proc. of the Workshop on Advanced Visual Interfaces AVI* (2012).

[4] DeLahunta, S. and Jenett, F. 2015. Making digital choreographic objects interrelate : a focus on coding practices. 6, (2015).

[5] Europeana: 2020. *www.europeana.eu*. Accessed: 2021-05-16.

[6] Evans, A. et al. 2014. 3D graphics on the web: A survey. *Computers & Graphics*. 41, 1 (Jun. 2014), 43–61. DOI:https://doi.org/10.1016/J.CAG.2014.02.002.

[7] Goldman, R. et al. 2014. *Video Research in the Learning Sciences*. Routledge.

[8] Parisi, T. 2012. WebGL : up and running. (2012), 213.

[9] El Raheb, K. et al. 2018. A web-based system for annotation of dance multimodal recordings by dance practitioners and experts. *ACM International Conference Proceeding Series* (2018).

[10] El Raheb, K. et al. 2019. Dance interactive learning systems: A study on interaction workflow and teaching approaches. *ACM Computing Surveys*. 52, 3 (Jun. 2019). DOI:https://doi.org/10.1145/3323335.

[11] Ribeiro, C. et al. 2016. 3D Annotation in Contemporary Dance. *Proceedings of the 3rd International Symposium on Movement and Computing* (New York, NY, USA, Jul. 2016), 1–4.

[12] Rodrigues, R. et al. 2021. Exploring the User Interaction with a Multimodal Web- Based Video Annotator. *13th EAI International Conference on Intelligent Technologies for Interactive En-tertainment (Intetain 2021)* (Qingdao, People's Republic of China, 2021), 1–10.

[13] Rodrigues, R. et al. 2019. Multimodal Web Based Video Annotator with Real-Time Human Pose Estimation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 23–30.

[14] Singh, V. et al. 2011. The choreographer's notebook-a video annotation system for dancers and choreographers. *C and C 2011 - Proceedings of the 8th ACM Conference on Creativity and Cognition* (2011).

[15] Towey, D. et al. 2017. Students as partners in a multimedia note-taking app development: Best practices. *Proceedings - 2017 IEEE/ACM 39th International Conference on Software Engineering Companion, ICSE-C 2017*. (Jun. 2017), 334–335. DOI:https://doi.org/10.1109/ICSE-C.2017.58.

[16] Wittenburg, P. et al. 2006. ELAN: A professional framework for multimodality research. *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006* (2006).

[17] 2014. glTF: Designing an Open-Standard Runtime Asset Format. *GPU Pro 5*. (May 2014), 393–410. DOI:https://doi.org/10.1201/B16721-30.