# Exploring the User Interaction with a Multimodal Web-Based Video Annotator

Rui Rodrigues[1,2(✉)], Rui Neves Madeira[1,2], and Nuno Correia[1]

[1] NOVA LINCS, NOVA School of Science and Technology, NOVA University Lisbon, Lisbon, Portugal
nmc@fct.unl.pt

[2] Sustain.RD, Setúbal School of Technology, Polytechnic Institute of Setúbal, Setúbal, Portugal
{rui.rodrigues,rui.madeira}@estsetubal.ips.pt

**Abstract.** People interact with their surroundings using several multimodal methods. Human-computer interaction is performed using these capabilities in order to provide, as much as possible, the most natural and productive experiences through speech, touch, vision, and gesture. The Web-based application used in this paper is a multi-platform video annotation tool that supports multimodal interaction. MotionNotes has the primary goal of fostering the creativity of both professional and amateur users. It is possible to interact with this tool using keyboard, touch, and voice, making it possible to add different types of annotations: voice, drawings, text, and marks. Furthermore, a feature of human poses identification in real-time was integrated into the annotation tool, enabling the identification of possible annotations. This paper presents and discusses results from a user study conducted to explore the user interaction with the tool, evaluating the prototype and its different interaction methods. User feedback shows that this approach to video annotation is stimulating and can enhance the user's creativity and productivity.

**Keywords:** Multimodal interfaces · Video annotations · Performing arts · User study · HCI

## 1 Introduction

Human poses and human motion are essential components in video footage, particularly in activities related to performing arts. The tool explored in this paper proposes applying multimodal annotation input/output with AI algorithms for pose estimation to improve Human-Computer Interaction [1–3]. MotionNotes development was part of the EU-funded project called CultureMoves [4], which follows a user-oriented approach and has the primary goal of developing software tools to "access and augment educational and cultural content" such as the one contained in Europeana [5]. MotionNotes enables its users to record, replay, and add new information to the video footage by working with multiple annotations and machine learning algorithms to increase productivity and creativity.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2022
Published by Springer Nature Switzerland AG 2022. All Rights Reserved
Z. Lv and H. Song (Eds.): INTETAIN 2021, LNICST 429, pp. 13–22, 2022.
https://doi.org/10.1007/978-3-030-99188-3_2

New AI techniques such as 2D human pose estimation are becoming very reliable in recent years [6], and they can be applied with promising results in Human-Computer Interaction and Multimodal Systems.

Therefore, three main research questions have been posed while developing and testing this Web-based tool as an input interface to add and manipulate multimodal time-based annotations over video:

1. Is it preferable to carry out annotation work during or after recording?
    a. Moreover, for each mode, what are the differences in annotation type usage?
2. Regarding the user devices used for annotation work, is there a preference between laptop and mobile devices?
    a. Additionally, for each device type, what are the differences in annotation type usage?
3. Could the human pose estimation feature be an asset to users when carrying out annotation work?

This work will contribute with a preliminary evaluation of the prototype and its interactions by answering these research questions, which brings insights regarding users' preferences. We collected this feedback through questionnaires and informal interviews. As a result, we concluded that our users accepted the general idea of replacing previous annotation methods with this web-based solution during our lab days. Moreover, we can state that people who work with video annotation are receptive to exploring different tools and interactions.

This paper is structured as follows. We start by analysing the related work, followed by the MotionNotes description. Afterwards, we present the testing environment and the results that were obtained. Finally, in the last section, we conclude with a summary, highlighting the tool's potential, and plan the future work.

## 2 Related Work

Video annotation is a valuable resource in different application areas, including analysing and studying human body motion. Furthermore, they are essential tools for encouraging collaborative teamwork by enabling information sharing [7]. These reasons motivated the development of several tools over the last years.

ELAN [8] is one of the most well-known and used tools in manually annotating or transcribing non-verbal communication. The work of Goldman [9] explored video annotations with object tracking methods. However, this work does not support touch or pen-based annotations; the tracking feature could not perform in real-time. The Choreographer's Notebook [10] was designed specifically for Choreography workflow, allowing digital-ink and text annotations. The WML tool [11] is another Web-based tool specifically designed to annotate, archive, and search dance movements.

In contrast, a pen-based video annotation tool was developed by Cabral et al. [12] to track motion. Their solution used frame differences, and they later tried similar methods on video editing [13]. Silva et al. [14] presented a work that enables real-time object tracking using the same pen-based video annotations following the same path. After that, as part of the BlackBox project [15], a prototype was developed to experiment with annotations in a 3D environment using Microsoft Kinect. Commercial video annotation applications, such as Wipster [16], Camtasia [17], Frame.io [18], and Vimeo Pro Review [19], have simplified the process of annotating and sharing videos for users. However, none of them supports automatic human pose detection.

Human pose estimation is a valuable computer vision technique in several areas, such as gaming, virtual reality, and video surveillance. This technique seeks to detect the human body parts computationally from video frames, and the goal is to identify the head, elbows, shoulders, knees, hips and feet. To address this issue, a few approaches have been proposed over the years.

By the end of the 2000s, state of the art was based on algorithms using features selected by human specialists, like gradient histograms [20–22]. Later, deep learning techniques have motivated a great deal of attention over the AI community [23], and human pose estimation was no exception. Deep learning-based methods can extract more and better features from training data, being possible to find literature with superior results [24–28]. Our proposal will explore implementations based on this last technique.

## 3 MotionNotes

MotionNotes [29] is a web-based real-time multimodal video annotation tool based on keyboard, touch, and voice inputs. This tool can support professional and amateur users working on creative and exploratory processes. MotionNotes enables the capture of multimodal annotations while and after recording video. The annotation types available to be used can be text, ink strokes, audio, or user-configured marks.

### 3.1 MotionNotes Implementation Overview

The prototype was designed to run on any regular Web browser, exploring multiple input modes, such as keyboard and touch interaction. The interface is responsive in order to enable users with different screen sizes to enjoy adequate interaction. The MotionNotes user interface has a video display area in the centre of the screen where it is possible to add new annotations or update current ones. In order to improve user feedback, there is a graphical representation of all annotated moments right below the video area (Fig. 1). Moreover, we included a machine learning technique in MotionNotes to perform real-time human pose predictions. PoseNet [30], a pre-trained neural network, in conjunction with tensorFlow.js, is used to process the body part classification in the client's machine. Predicted points are drawn on an HTML canvas object located in the same position as the video but at a higher layer position. Finally, straight lines are calculated between the points, with the skeleton staying visible, giving the user another resource to identify possible annotations (Fig. 2).

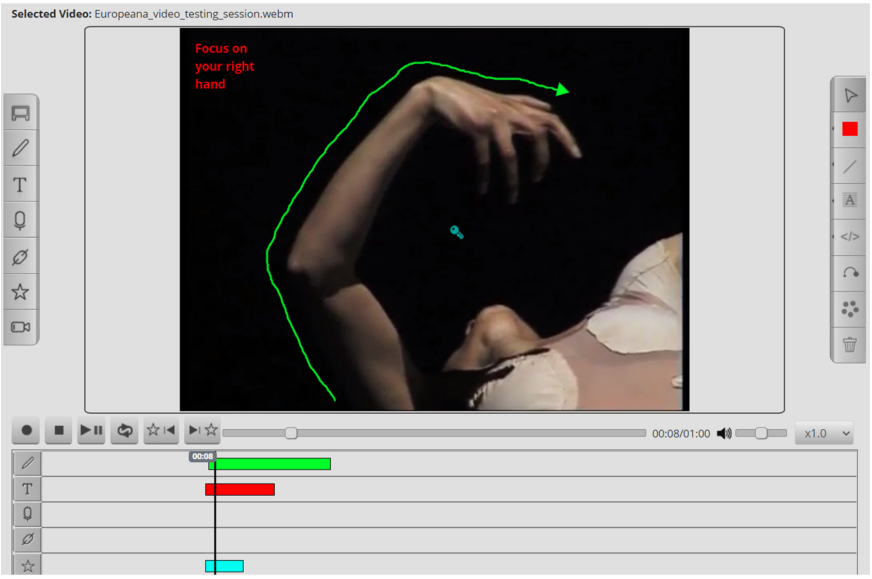


**Fig. 1.** Tool GUI and annotation types (Green: drawing; Red: text; Blue: mark) (Color figure online).

### 3.2 MotionNotes Interaction Example

After an initial interaction and reflecting on testing scenarios, a procedure was created for MotionNotes. In order to better understand the interaction, we follow Mary while she annotates her video with the MotionNotes.

First, Mary opens MotionNotes as she would like to load a video that she recorded in a recent dance competition. Mary goes to the File menu and clicks on the import video option; MotionNotes immediately opens a new window. Mary browses and selects her video. Once done, she needs to click on the play button, so the video instantly starts playing. She notices that the right-hand movement could be better and feels the need to highlight this. Mary pauses the video and then thinks about which annotation types are appropriate to express her thoughts. She decides to add one text annotation; to do that, she selects the text annotation type in the left menu and clicks above the video on the preferred location. MotionNotes immediately creates a new textbox and give focus to it. Next, Mary starts typing, "Right-hand position improvements were needed here", and clicks enter. MotionNotes saves several details about the annotation, for instance, the text, font, colour, position over the video, and the exact timestamp. To give more details about the objective, Mary decides that a draw annotation could help. She clicks on the draw annotation option, and MotionNotes activates the draw functionality; Then, she creates a line across the location where the arm and hand should be. Finally, Mary believes that this section is crucial for the performance and adds a mark annotation. For that, she selects the mark annotation, and MotionNotes opens a popup with the predefined images. She selects a key icon and concludes the procedure by clicking above the video on the desired location where the icon should be (Fig. 1).


**Fig. 2.** MotionNotes with pose estimation. (Green: pose predictions; Red: manual annotation) (Color figure online).

## 4 User Study

The study was composed of three phases, each one to understand a different question. The study's first phase focused on collecting data regarding the annotation experience both while and after recording. The second phase was focused on understanding the user behaviour regarding different devices. Finally, the last one assessed if human pose estimation could add value to video annotation software.

### 4.1 Design and Participants

We performed a user study with 27 participants. The user test started with the users watching a 15-min tutorial. The next step was for participants to interact with the software, which was achieved by asking them to complete a set of proposed tasks.

Regarding the participants, the most representative age interval was between 25 and 34 years old. The gender representation was nearly even, with 51% female and 49% male. Regarding the testing group education levels, 37% had a master's degree, 29.6% had a bachelor's degree, 18.5% had studied until high school, and the remaining 14.8% held a PhD degree. Most of the participants reported they frequently annotate their work in some way (77.8%). The preferred method used to annotate is a regular paper notebook (63.6%), with the laptop devices being the second most popular way (36.4%), and mobile phones appearing right after (27.3%).

The questionnaire included 23 questions using the five-point Likert Scale. We used paired-samples t-tests and one-way ANOVA to analyse the feedback.

### 4.2 Results and Discussion

The three tables present in this section summarise the results, including descriptive statistics, t-test, and ANOVA regarding each one of the research questions posed in the introduction.

Table 1 focuses on results related to RQ1. Participants classified the tool regarding the annotation experience during and after a recording. The t-test returned a significant difference, showing preferences for using the tool in playback mode. Moreover, the ANOVA test has shown a substantial difference concerning this after recording mode, indicating that text was the most popular annotation type, followed by drawing.

**Table 1.** Descriptive statistics, t-test, and ANOVA for the different annotation type experience.

| Question | A[1] | B | C | D | E | M | SD | t-test / ANOVA |
|---|---|---|---|---|---|---|---|---|
| 1. Consider the annotation experience | | | | | | | | |
| 1.1. Annotating while recording | 2 | 3 | 10 | 6 | 6 | 2.59 | 1.16 | t=-3.56; p<0.05 |
| 1.2. Annotating in playback mode | 12 | 4 | 5 | 2 | 4 | 3.67 | 1.47 | |
| 2. Consider the recording phase | | | | | | | | |
| 2.1. Classify the sketch usage | 8 | 4 | 4 | 3 | 8 | 3.04 | 1.62 | |
| 2.2. Classify the text usage | 6 | 4 | 4 | 5 | 8 | 2.81 | 1.54 | f=0.32; p>0.05 |
| 2.3. Classify the audio usage | 5 | 6 | 5 | 3 | 8 | 2.89 | 1.49 | |
| 2.4. Classify the marks usage | 3 | 5 | 7 | 3 | 9 | 2.63 | 1.39 | |
| 3. Consider the after recording | | | | | | | | |
| 3.1. Classify the sketch usage | 9 | 7 | 4 | 3 | 4 | 3.52 | 1.42 | |
| 3.2. Classify the text usage | 14 | 7 | 3 | 2 | 1 | 4.15 | 1.11 | f=2.89; p<0.05 |
| 3.3. Classify the audio usage | 5 | 8 | 4 | 7 | 3 | 3.19 | 1.31 | |
| 3.4. Classify the marks usage | 6 | 9 | 3 | 3 | 6 | 3.22 | 1.47 | |

[1] A: strongly useful, B: useful, C: ok, D: not useful, E: strongly not useful

Regarding RQ2, classifying the experience when using different devices, a mobile touch-based device (less than 576 px wide) was compared to a regular laptop. Again, the t-test returned a significant difference, showing preferences for using the tool in a regular laptop with larger resolutions. The ANOVA test did not show significant differences regarding the usage of different annotation types. However, it is possible to verify that text annotation is slightly more prevalent when using laptops, while mark annotation leads in mobile. Table 2 summarises the results.

Regarding RQ3, which addresses classifying the overall experience with the human pose estimation feature active, the feedback was positive, as shown in Table 3. However, when users were asked if they could consider using this feature during annotation work, the results were just ok. Users' comments about this feature were collected, which let us understand they were expecting more options to work with it, such as recording only the pose or reproducing the movements in isolation (e.g., without video and sound) and adding annotations in this mode.

**Table 2.** Descriptive statistics, t-test, and ANOVA for the device interaction experience.

| Question | A[1] | B | C | D | E | M | SD | t-test / ANOVA |
|---|---|---|---|---|---|---|---|---|
| 1. Consider the tool interaction | | | | | | | | |
| 1.1. Classify when using a laptop | 11 | 11 | 3 | 2 | 1 | 4.11 | 0.99 | t=6.93; p<0.05 |
| 1.2. Classify when using a mobile device | 1 | 3 | 16 | 3 | 4 | 2.78 | 0.96 | |
| 2. User experience using a laptop | | | | | | | | |
| 2.1. Classify the sketch usage | 9 | 5 | 7 | 3 | 3 | 3.52 | 1.34 | |
| 2.2. Classify the text usage | 13 | 10 | 3 | 0 | 1 | 4.26 | 0.93 | f=2.06; p>0.05 |
| 2.3. Classify the audio usage | 14 | 6 | 4 | 2 | 1 | 4.11 | 1.13 | |
| 2.4. Classify the marks usage | 12 | 5 | 8 | 1 | 1 | 3.96 | 1.10 | |
| 3. User experience using mobile | | | | | | | | |
| 3.1. Classify the sketch usage | 3 | 3 | 15 | 2 | 4 | 2.96 | 1.10 | |
| 3.2. Classify the text usage | 1 | 2 | 15 | 7 | 2 | 2.74 | 0.84 | f=1.77; p>0.05 |
| 3.3. Classify the audio usage | 2 | 5 | 16 | 2 | 2 | 3.11 | 0.92 | |
| 3.4. Classify the marks usage | 4 | 5 | 15 | 2 | 1 | 3.33 | 0.94 | |

**Table 3.** Descriptive statistics, t-test for the pose estimation experience.

| Question | A[1] | B | C | D | E | M | SD | t-test |
|---|---|---|---|---|---|---|---|---|
| 1. Classify the overall experience | 2 | 8 | 11 | 3 | 2 | 3.19 | 0.98 | |
| 2. Consider the annotation work | | | | | | | | |
| 2.1. Classify it using pose estimation | 4 | 5 | 10 | 4 | 4 | 3.0 | 1.23 | t=-0.9; p<0.05 |
| 2.2. Classify it without pose estimation | 3 | 6 | 13 | 2 | 3 | 3.1 | 1.07 | |

The user statements during the test were mostly positive. One user (U3) said: “Easy to learn; the multiple annotation types complement each other very well.” Another user (U15) stated: “The marks was a good idea, very fast to apply, even in small screen devices”. Regarding human pose estimation, another user (U24) stated the following: “There were scenarios where having the pose helped in the creation of new annotations”.

## 5 Future Work

The feedback obtained while testing MotionNotes was positive. However, we discussed and collected a couple of new ideas for additional developments.

Participants mentioned a few scenarios in which they considered human pose estimation and MotionNotes could benefit in a future version. The first scenario discussed by some participants was the background subtraction. That means, for instance, reproducing the body parts motion on a skeleton format in the same timeframe as the source video, but with a clean background. Additionally, the skeleton colour, background colour

and audio should allow personalisation. The second scenario was based on having a particular type of annotation associated with the pose. This type of annotation should be optional and could be activated or deactivated depending on motion tracking status. This scenario brings several advantages like users could concentrate only on pose in a specific annotation iteration, leaving other elements for future work; another advantage is the movement correction, where users could edit several frames by drawing the correct pose.

Regarding the annotation types, the most discussed were the marks, where participants showed great interest given its novelty. Again, we stimulated the participants to give suggestions and ideas, and 3D was the subject of debate for this type of annotations. Right now, these marks are predefined 2D icons or images uploaded by users, and participants commented about how interesting it could be to upload 3D models and add them to a scene as annotations. We think most of these ideas could foster the users' creativity, which is one of our main goals, and we are already designing a new MotionNotes version containing some of these features.

## 6 Conclusion

The multimodal Web video annotation tool MotionNotes described in this paper enables users to add different annotation types and identify human poses in real-time. The tool was tested in order to address three main research questions concerning user preferences and interaction.

From our results, we can conclude that annotation software users prefer to work after the recording session and not during it. Additionally, they preferred to work in a traditional environment with a larger screen over the more modern and popular mobile devices. Finally, we observed a significant curiosity about both automatic human pose recognition and marks annotation type. Future work should focus on these features providing additional research.

**Acknowledgements.** This work is funded by Fundação para a Ciência e Tecnologia through a Ph.D. Studentship grant (2020.09417.BD). It is Supported by the project CultureMoves, Grant Agreement Number: INEA/CEF/ICT/A2017/1568369. It is also supported by NOVA LINCS RC, partially funded by project UID/CEC/04516/2020 granted by FCT.

## References

1. Turk, M.: Multimodal interaction: a review. Pattern Recogn. Lett. **36**, 189–195 (2014)
2. Dumas, B., Lalanne, D., Oviatt, S.: Multimodal interfaces: a survey of principles, models and frameworks. In: Lalanne, D., Kohlas, J. (eds.) Human Machine Interaction. LNCS, vol. 5440, pp. 3–26. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00437-7_1
3. Abuczki, Á., Esfandiari Baiat, G.: An overview of multimodal corpora, annotation tools and schemes. Argumentu **9**, 86–98 (2013)
4. CultureMoves: Culture Moves. https://culturemoves.eu/. Accessed 17 Jun 2021
5. Europeana: Europeana. www.europeana.eu. Accessed 16 May 2021
6. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 (2017)

7. Cabral, D., Valente, J., Silva, J., Aragão, U., Fernandes, C., Correia, N.: A creation-tool for contemporary dance using multimodal video annotation. In: Proceedings of the 2011 ACM Multimedia Conference and Workshops, MM 2011 (2011)
8. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: ELAN: a professional framework for multimodality research. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006 (2006)
9. Goldman, D.B., Gonterman, C., Curless, B., Salesin, D., Seitz, S.M.: Video object annotation, navigation, and composition. In: Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology, UIST 2008 (2008)
10. Singh, V., Latulipe, C., Carroll, E., Lottridge, D.: The choreographer's notebook-a video annotation system for dancers and choreographers. In: Proceedings of the 8th ACM Conference on Creativity and Cognition, C and C 2011 (2011)
11. El Raheb, K., Kasomoulis, A., Katifori, A., Rezkalla, M., Ioannidis, Y.: A web-based system for annotation of dance multimodal recordings by dance practitioners and experts. In: ACM International Conference Proceeding Series (2018)
12. Cabral, D., Valente, J.G., Aragão, U., Fernandes, C., Correia, N.: Evaluation of a multimodal video annotator for contemporary dance. In: Proceedings of the Workshop on Advanced Visual Interfaces AVI (2012)
13. Cabral, D., Correia, N.: Video editing with pen-based technology. Multimedia Tools Appl. **76**(5), 6889–6914 (2016)
14. Silva, J., Fernandes, C., Cabral, D., Correia, N.: Real-time annotation of video objects on tablet computers. In: Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia, MUM 2012 (2012)
15. Ribeiro, C., Kuffner, R., Fernandes, C., Pereira, J.: 3D annotation in contemporary dance: Enhancing the creation-tool video annotator. In: ACM International Conference Proceeding Series (2016)
16. Wipster | Review Software. https://wipster.io/. Accessed 15 Jun 2021
17. Camtasia. https://www.techsmith.com/video-editor.html. Accessed 2 Jun 2021
18. Frame.io. https://www.frame.io/. Accessed 25 May 2021
19. Vimeo. https://vimeo.com/features/video-collaboration. Accessed 5 Jun 2021
20. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2008). https://doi.org/10.1109/CVPR.2008.4587597
21. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: people detection and articulated pose estimation, pp. 1014–1021 (2010). https://doi.org/10.1109/CVPR.2009.5206754
22. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1385–1392 (2011). https://doi.org/10.1109/CVPR.2011.5995741
23. Markoff, J.: Scientists See Promise in Deep-Learning Program. Nyt. (2012)
24. Toshev, A., Szegedy, C.: DeepPose: human pose estimation via deep neural networks (2014)
25. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: DeeperCut: a deeper, stronger, and faster multi-person pose estimation model. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 34–50. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_3
26. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 5686–5696 (2019)
27. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 4724–4732 (2016)

28. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
29. Rodrigues, R., Madeira, R.N., Correia, N., Fernandes, C., Ribeiro, S.: Multimodal web based video annotator with real-time human pose estimation. In: Yin, H., Camacho, D., Tino, P., Tallón-Ballesteros, A.J., Menezes, R., Allmendinger, R. (eds.) IDEAL 2019. LNCS, vol. 11872, pp. 23–30. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33617-2_3
30. PoseNet. https://learn.ml5js.org/#/reference/posenet?id=posenet. Accessed 15 Nov 2021