Diogo Nuno Crespo Ribeiro Cabral

Mestre em Engenharia Informática

Video Interaction using Pen-Based Technology

Dissertação para obtenção do Grau de Doutor em Informática

Orientador : Professor Doutor Nuno Manuel Robalo Correia,
Professor Catedrático,
Universidade Nova de Lisboa

Júri:

Presidente: Professor Doutor Luís Manuel Marques da Costa Caires

Arguentes: Doutor Pablo Cesar
Professor Doutor Joaquim Armando Pires Jorge

Vogais: Professor Doutor Nuno Manuel Robalo Correia
Professora Doutora Teresa Isabel Lopes Romão
Professora Doutora Maria Teresa Caeiro Chambel

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

February, 2014

Video Interaction using Pen-Based Technology

Acknowledgements

I first wish to thank to my supervisor, Prof. Dr. Nuno Correia, for his advice and immeasurable support. A special thank you goes to the Creation-Tool development team João Gaspar Valente, João Silva and Urândia Aragão, for their important contributions. I would also like to thank to the rest of the TKB project team and collaborators, particularity to Dr. Carla Fernandes, Stephan Jürgens, Carlos Oliveira, Filipe Cruz, David Santos Rui Horta, the staff from "O Espaço do Tempo" and the dancers from "Forum Dança PEPCC 2010/2012", for their help, support and comments.

I would like to thank my present and former colleagues at the Interactive Multimedia Group (IMG) at CITI/FCT/UNL, Rui Nóbrega, Rui Jesus, Sofia Reis, André Sabino, Filipa Peleja, Bruno Cardoso, Rossana Santos, Pedro Centieiro and Rui Madeira, for their help, suggestions and support. I also would like to thank to Prof. Dr. Carmen Morgado for her presence in the IMG group coffee breaks with awesome cakes and cookies.

I would like to thank to Prof. Dr. Sharon Strover, Prof. Dr. Luis Francisco-Revilla and Prof. Anne S. Lewis, form University of Texas at Austin, for their comments and suggestions.

I also would like to thank Ricardo Mano, Nuno N. Correia and Pedro Maurício Costa for sending me important material for this research work as well as all the participants in the usability tests.

I am especially grateful to my parents and sisters for their help and support.

A special thank you to my lovely spouse for her support and patience.

This work was partially funded by the UTAustin-Portugal, Digital Media, Program (Ph.D. grant: SFRH/BD/42662/2007 - FCT/MCTES); by the HP Technology for Teaching Grant Initiative 2006; by the project "TKB - A Transmedia Knowledge Base for contemporary dance" (PTDC/EAT/AVP/098220/2008 funded by FCT/MCTES); and by CITI/DI/FCT/UNL (PEst-OE/EEI/UI0527/2011).

Abstract

Video can be considered one of the most complete and complex media and its manipulating is still a difficult and tedious task. This research applies pen-based technology to video manipulation, with the goal to improve this interaction. Even though the human familiarity with pen-based devices, how they can be used on video interaction, in order to improve it, making it more natural and at the same time fostering the user's creativity is an open question.

Two types of interaction with video were considered in this work: video annotation and video editing. Each interaction type allows the study of one of the interaction modes of using pen-based technology: indirectly, through digital ink, or directly, trough pen gestures or pressure. This research contributes with two approaches for pen-based video interaction: pen-based video annotations and video as ink.

The first uses pen-based annotations combined with motion tracking algorithms, in order to augment video content with sketches or handwritten notes. It aims to study how pen-based technology can be used to annotate a moving objects and how to maintain the association between a pen-based annotations and the annotated moving object

The second concept replaces digital ink by video content, studding how pen gestures and pressure can be used on video editing and what kind of changes are needed in the interface, in order to provide a more familiar and creative interaction in this usage context.

Keywords: Pen-based Video Interaction, Pen-based Video Annotation, Pen-based Interfaces, Video Interfaces

Resumo

O vídeo pode ser considerado um dos media mais completos e complexos e sua manipulação ainda é uma tarefa difícil e aborrecida. Esta pesquisa aplica tecnologia baseada em caneta para a manipulação de vídeo, tendo como objetivo melhorar esta interação. Apesar da familiaridade humana com dispositivos baseados em caneta, como estes podem ser utilizados na interação de vídeo de modo melhorá-la, tornando-a mais natural e ao mesmo tempo estimulando a criatividade do utilizador é uma questão em aberto.

Dois tipos de interação com vídeo foram considerados neste trabalho: anotação de vídeo e edição de vídeo. Cada um dos tipos de interação permite o estudo de um dos modos de interação da tecnologia baseada em caneta: indiretamente, por meio de tinta digital, ou diretamente, usando gestos ou pressão. Esta pesquisa contribui com duas abordagens para interação com vídeo baseada em caneta: anotações de vídeo e caneta e vídeo como tinta.

O primeiro usa anotações feitas com caneta e combinadas com algoritmos de deteção de movimento, a fim de aumentar o conteúdo de vídeo com desenhos ou anotações manuscritas. Procura estudar como a tecnologia da caneta pode ser usada para anotar objetos em movimento e como manter a associação entre as anotações feitas com caneta e o objeto dinâmico anotado.

O segundo conceito substitui a tinta digital por conteúdo de vídeo, estudando assim como os gestos e pressão feitos com a tecnologia da caneta podem ser usados em edição de vídeo e que tipos de mudanças são necessárias na interface, de modo a proporcionar uma interação mais familiar e criativa neste contexto.

Palavras-chave: Interação com vídeo baseada em caneta, Anotações de vídeo manuscritas, Interfaces baseadas em caneta, Interfaces de vídeo

Contents

List of Figures

List of Tables

Acronyms and Notation

CD	Compact Disc
CRT	Cathode Ray Tube
CSI	Creative Support Index
DVD	Digital Videodisk
GUI	Graphical User Interface
HD	High-Definition
IP	Internet Protocol
OSC	Open Sound Control
PC	Personal Computer
PDA	Personal Digital Assistant
PNG	Portable Network Graphics
UDP	User Datagram Protocol
URL	Uniform Resource Locator
VJ	Video-Jockey
Wi-Fi	Wireless Local Area Network
XML	eXtensible Markup Language

$\bar{x}$	Mean
σ	Standard Deviation
$\tilde{x}$	Median
$Var(X)$	Variance

1

Introduction

The motivation, problem statement, solution overview and the main contributions of this work are presented in this chapter.

1.1 Motivation

Video can be considered one of the most complete and complex media, as Yeo and Yeung [YY97] discuss in their research about video database management systems and as Goldman [Gol07] mentions in his research about video annotation, visualization and interaction. The use of digital platforms not only changed video editing and visualization processes but also proven to be necessary to improve video content browsing and searching [MD89; YY97; Cha04; Dan11; Cha12]. Nonetheless, manipulating digital video is still a difficult task [GBCDFGUW00; CLMBSDYC02; RB03; Cha04; DE06; Cha12]. Recent developments in digital video technology, like video sharing platforms (e.g., YouTube[1], Vimeo[2]) or the integration of video cameras in mobile devices[3], have increased the production, distribution and access of video content, making video manipulation not only an important issue for professionals, but also for non-expert users. Therefore, it is crucial to find more natural ways to interact with video content.

Video manipulation and interaction, by a single user (i.e., excluding group sharing tasks), can be divided in three main tasks, besides the regular playback: video browsing or navigation; video data analysis, which includes different subtasks, like video browsing or video annotation; and video editing.

[1] `http://www.youtube.com/`, with 4 billion hours of video [You].
[2] `http://vimeo.com/`
[3] 3 hours of video are uploaded per minute to YouTube from mobile devices [You]

- **Video Browsing or Navigation:** Yeo and Yeung [YY97] defined video browsing as "a technique users employ to view information rapidly in order to decide whether the content is relevant to their needs". Video browsing or navigation is implicitly linked to video visualization and the playback of video segments. Video browsing can be used on video data analysis, on video editing, or by someone who simply wants to re-watch his favorite film scene. Video browsing has been a fairly studied research topic, with plenty of published literature presenting a full range of solutions [SHMBJ10] depending on the video content [ED94; TMDSK98; DRBNBS08; KWLB08], interaction input (including pen-based interactions) or device [RB03; RB05; HG08; WMYL12].

- **Video Data Analysis:** Ratcliff [Rat03] defines video data analysis as the process of making the video information meaningful. Video data analysis is useful in different areas, like social and design studies, education, performative arts, medicine, media production or sports. Video allows one to capture and re-watch a version of an event as it happens and extract useful information from it, which depends on the subject of study and its context [HHL10]. Different subtasks, such as video browsing; video annotation; information extraction combined with quantitative measurements (e.g, measuring the time of a recorded process or task, counting the number of repeated scenes) and slow motion or frame-by-frame playback can be involved in video data analysis. All these subtasks can be performed by humans or computers, depending on the objectives and context of analysis. Video annotation is a useful feature of video analysis, saving time [Mac89; MD89; WP94] and fostering a process that combines watching and critical thinking, named active watching [CC99].

  - **Video Annotation:** Bottoni et al [BCLOPT04] defined digital annotations of a multimedia document as "the production of additional information related to the document or some parts of it". Therefore, it is possible to infer that video annotations can be defined as *additional information related to the video or some parts of it*. Digital annotations have the advantage that can be archived, shared, searched, filtered and easily manipulated when compared with paper annotations [BM03; SW04]. Bulterman [Bul04] defines two categories of annotations: hierarchical and peer-level. Hierarchical annotations include "document markup (including metadata) that provides an abstract classification of media content for a given use and ontology" whereas peer-level annotations include "document markup that provides companion information and which results in augmented media content". Annotation supports different cognitive functions such as remembering, thinking and clarifying [BCLOPT04]. Marshall [Mar97], on her study about annotations made on textbooks, went further by defining annotations as procedural signals; placemarkings and aids to memory; as in situ locations for problem-working; a record of interpretive

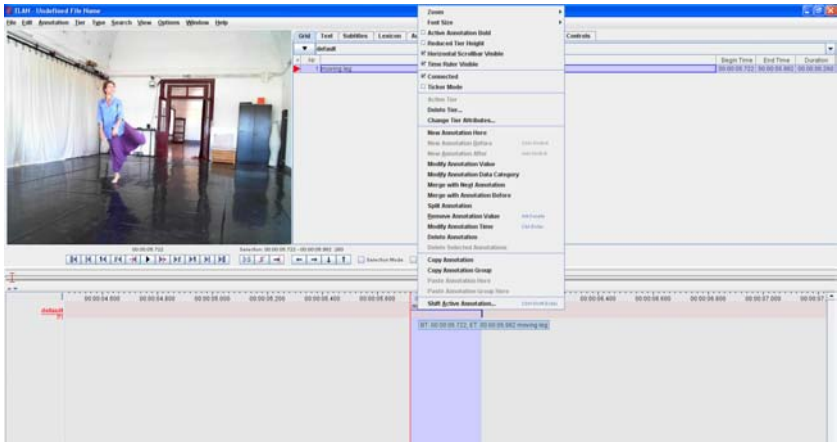
activity; a visible trace of the reader's attention and as incidental reflections of the material circumstances. Later, Marshall [Mar10] defined three elements which compose an annotation: a body, an anchor, and a marker. The annotation body is any content that the user wants to add to it; the anchor delimits the scope of an annotation, i.e., the content of the original document associated with the annotation, and the marker defines how the anchor should be rendered when it is displayed. These issues combined with video dynamic properties, generate an additional level of complexity, when trying to link other media to dynamic visual content [Bux07; Gol07].

- **Video Editing:** In the *Dictionnaire mondial du Cinèma* [VBGLS+11] is defined that "film editing is the phase of the manufacture of a film in which are assembled and arranged images and sounds of the film". Dancyger [Dan11] mentions that "what can be said about the craft and art of film editing can also be said about video editing". This idea is reinforced also in the *Dictionnaire mondial du Cinèma* [VBGLS+11], by claiming that digital technology made the editing process equal on film and video. Therefore, it is possible to use the same definition of film editing for video: *editing is the phase of the manufacture of a video in which are assembled and arranged images and sounds of the video*. Editing not only allows to shoot different scenes in a non-linear order, which can be properly selected and arranged on a post-production process, but can also be used for narrative purposes. Dancyger [Dan11] refers as goals for editing: narrative clarity (e.g., narratives with multiple characters); dramatic emphasis (e.g., the combination of close-ups in a scene); subtext (e.g., imposing a particular pace, in order to transmit an idea to the audience) and aesthetics (e.g., providing an intentional aesthetic). This idea is also reinforced by Thompon and Bowen [TB09], in their definition of editing: "Editing for motion pictures is the process of organizing, reviewing, selecting, and assembling the picture and sound "footage" captured during production. The result of these editing efforts should be a coherent and meaningful story or visual presentation that comes as close as possible to achieving the goals behind the original intent of the work - to entertain, to inform, to inspire, etc."

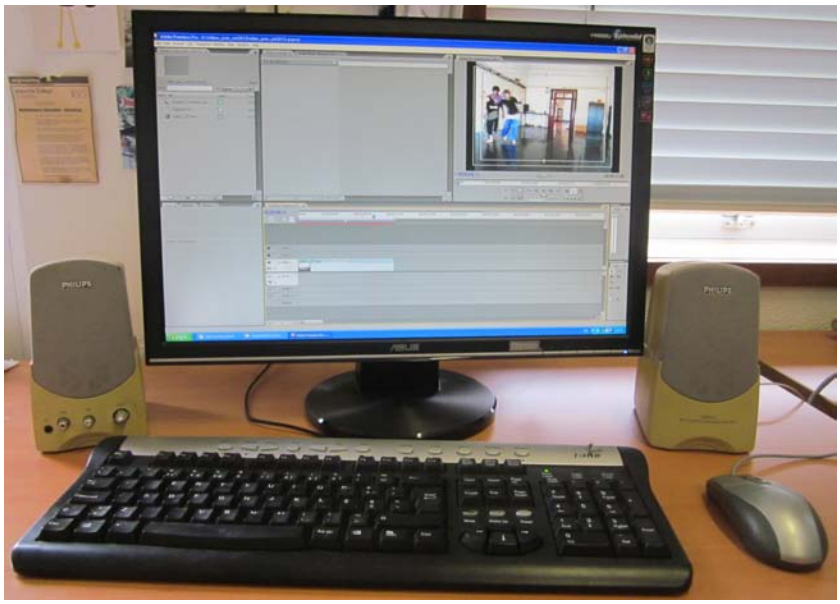
  Dancyger [Dan11] also considers three aspects of the editing process: the *technique*, the physical joining of two disparate pieces of film; the *craft*, the joining of two pieces of film together to yield a meaning that is not apparent from one or the other shot and the *art*, when the combination of two or more shots takes meaning to the next level-excitement, insight, shock, or the epiphany of discovery. As previously mentioned, recent technology developments, like video sharing platforms and mobile devices with integrated cameras, made the editing process important not only for film industry professionals but also for regular consumers.

The work featured in this research, applies pen computing to video manipulation, aiming to improve user media interaction. The lack of familiar interactions for video

annotation and for video editing as well as the tasks involved in both topics (Figure 1.1), make them natural candidates for the usage of pen-based technology, which can be considered a familiar computer input interface for humans [Mey95]. In the context of video annotation, it is possible to find a wide set of solutions but still remains the need of a fluid video annotation method, which incorporates the dynamic dimension of the content [RB03; Bur06; Bux07; SLCL11]. Whereas, video editing misses solutions that make the process more efficient [GBCDFGUW00; CLMBSDYC02; RB03; DE06], easier to learn [Cha04; Cha12], but at the same time foster the user's creativity [Dan11].



(a)



(b)

Figure 1.1: Video interaction. (a) Annotating a video. (b) Editing a video.

Pen-based technology can be used indirectly, through digital ink or directly, using pen gestures or pressure. The familiarity of pen-based technology for the note taking task was reported in Vannevar Bush's 1945 Memex [Bus45]: "He can add marginal notes and comments, taking advantage of one possible type of dry photography, and it could

even be arranged so that he can do this by a stylus scheme, such as is now employed in the telautograph seen in railroad waiting rooms, just as though he had the physical page before him.". More than 50 years later, a field study about reading and writing on a Tablet computer [MPGS99] observed this interaction familiarity with pen-based technology on note taking tasks. Moggridge, in his 2006 book *Designing Interaction* [Mog06] discusses the familiarity of pen-based technology with Bert Keely, former responsible of tablet developments of Silicon Graphics and Microsoft, "you can sketch and enter text with one hand with a fluency that you have been developing since childhood" (Figure 1.2). In addition, pen-based technology has the advantages that can give a more accurate positioning and avoid dirty screens, compared to touch [DFAB04].

Figure 1.2: Child using pen and paper.

In order to cover the two modes of pen-based interaction, this work uses digital ink for video annotations and pen gestures and pressure as input commands for video editing. Although, in this research these topics were studied separately they can be combined: annotations, made by professionals, can be used during the video editing process [MP94; Cha04; Bux07]; video segments can be automatically generated and combined based on annotations [MD89; CLMBSDYC02; Dav03; VDJ03; VD04; GCSS06] and pen-based commands can be useful for annotation tasks [PGS98; SGP98; BM03; HZSBCST07; LGHH08] as well as for video browsing [RB03; RB05; HG08].

The first attempts of using pen computing occurred in the 1950s and 1960s [Mey95; Bux12]. The most recent and widely used development of pen-based technology is the Tablet, a computer equipped with stylus and a touch screen. The size of the displays, almost the same as a regular sheet of paper, and its computational power, comparable to a regular laptop, make the Tablets an interesting approach to electronic paper and a precursor to interactions that will be possible in the future [DFAB04]. A similar idea was expressed by Bert Keely [Mog06], "... pen-based computer will grow up [...], trying to simulate pen and paper, and then evolving unique characteristics that have special

advantages in the digital context." (Figure 1.3).

Figure 1.3: Pen: paper and computer.

Although there is not a clear definition of electronic paper [HDYK11], it can be considered as a combination of paper and digital technology. The affordances of each are described by Sellen and Harper in the book *The Myth of the Paperless Office* [SH03] and shown on Table 1.1.

Affordances of Paper
Quick, flexible navigation through and around documents
Reading across more than one document at once
Marking up a document while reading
Interweaving reading and writing

Affordances of Digital Technology
Storing and accessing large amounts of information
Displaying multimedia documents
Fast full-text searching
Quick links to related materials
Dynamically modifying or updating content

Table 1.1: Affordances of paper and of digital technology for reading [SH03].

From the combination of paper and digital technology affordances, results that the display, navigation, annotation and modification of multimedia documents, particularly those involving large amounts of information like video, are important aspects for future electronic paper. Therefore, the affordances for reading, described by Sellen and Harper, can be easily transposed for “watching”. This fact increases the relevance of applying the same human behaviors when interacting with paper, like using pen or touch, to multimedia manipulation. Although the focus of this thesis is on pen-based interaction, both touch and pen interfaces should be combined for a more familiar human-computer interaction, emulating paper interactions [SH03; BFWHS08; HYPCRWBB10]. This bimanual interaction, with pen and touch, is superior in terms of speed, accuracy and user preference, to the other bimanual combinations, i.e., using two hands or two

pens [BFWHS08] as well as to the different mode switching techniques of pen-based user interfaces [LHGL05]. Nonetheless, Vogel and Balakrishnan [VB10], in their study about the difficulties of direct pen interaction with a conventional graphical user interfaces (GUI), alerted for the need of improvement of hardware, base interaction and widgets behavior. This need of user interface adaptations for pen-based interactions was also expressed by Marshal [Mar10] in her book about reading and writing on electronic books. Considering these issues, it is possible to conclude that pen-based interactions should be applied to video manipulation, in order to achieve more familiar video interactions.

1.2 Research Questions

Considering the complexity of video data and the difficulty of its manipulation, this research has the aim to improve it by using pen-based technology as an input interaction device. Even though the human familiarity with pen-based devices makes this technology a potential candidate for video interaction improvement, how it can be used remains an open question. Therefore, the main research question of this work is:

How pen-based technology can be applied to video interaction, in order to improve it, making it more natural and at the same time fostering the user's creativity?

As mentioned in the previous section, pen-based technology can be used indirectly, through digital ink. The use of pen as an input technology (physical or digital) for annotations is usually associated with activities that involve static content, like writing or sketching, linked to static media, e.g., text and images. The usage of pen-based technology for note taking actions associated with video content presents different challenges: Regarding the video annotation using pen-based technology two main research question can be made:

How to annotate a moving object using pen-based technology?

How to maintain the association between a pen-based annotation and the annotated moving object?

Although the recognition of handwritten annotations is a relevant research topic and important for the pen usefulness, it is independent from the type of media, static or dynamic, associated with the annotations. Therefore, the recognition of handwritten annotations is not considered as a focus of this thesis.

On the other hand, pen-based technology can also be used directly, through pen gestures and pressure, working as an interface device for command input. By applying pen-based technology as an input interface to the manipulation and editing of a time-based visual media like video, additional questions can be considered. Regarding video manipulation and editing two main research questions can also be made:

How to improve video manipulation and editing using pen gestures or pressure?

What kind of changes are needed in a video editor interface, in order to achieve this improvement?

1.3 Research Overview

This research presents two approaches for pen-based video interaction: pen-based video annotations and video as ink. The first one is focused on adding pen-based annotations to video content, which can also be used for video navigation or for the generation of a new video stream. The second approach replaces the digital ink by video frames or segments, which are used for video editing.

In the context of pen-based video annotations two applications were developed. The first one was a proof-of-concept prototype and the second a more complete video annotator. The later was implemented and applied to contemporary dance as a Creation Tool. This application, developed for Tablet PCs, includes annotations in the form of ink, text, audio, hyperlinks and marks. The tool was conceived to assist the creative processes of choreographers, working as a digital notebook.

The development followed an iterative design process involving two choreographers, and a usability study was carried out, involving international dance performers. The development and evaluation of the video annotator was made as part of the "TKB: A Transmedia Knowledge Base for contemporary dance" project [FJ09], in collaboration with the School of Social Sciences and Humanities (FCSH), Universidade Nova de Lisboa (UNL). In the context of the TKB project, two master students collaborated in this research: João Gaspar Valente and João Silva. João Gaspar Valente focused his thesis on the development of the multimodal video annotator [Val11], whereas João Silva focused his work on motion tracking algorithms applied to video annotation [Sil12]. In addition, the designer Urândia Aragão contributed in the graphical proposal and helped in the design interaction of the tool.

In the context of the second approach, video as ink, a pen-based proof-of-concept prototype was also developed. The prototype also uses a Tablet PC as a platform for video editing. The editing features include operation features such as add, move and erase video content. Transitions effects, like fade, and two selection modes, one using pen gestures and the other using a lasso tool, were also developed. In addition, zoom features based on pen pressure were implemented.

1.4 Contributions

This research work provides the following contributions:

- The concept of pen-based video annotations, which can be defined as a set of ink strokes associated in time and space with video content. The main characteristics of these annotations are:

  - A familiar method for human-based video analysis and annotation.
  - The usage of motion tracking methods, in order to maintain the annotation context, i.e., the association between annotations and the annotated video features.
  - Augmentation of video content with digital ink and using with real-time motion tracking algorithms, in order to move the ink strokes according the movement of video objects.

- The concept of video as ink, in which digital ink is replaced by video content. This concept can be characterized by:
  - Applying the paint metaphor to video editing.
  - The canvas, in which the frames or segments are painted, works as 2D timeline and allows non-linear video editing.
  - The concept can also be applied to other video editing features, such as erasing frames and content selection.
  - Pen pressure is used for zooming features.

- A proof-of-concept prototype of pen-based video annotations for Tablet PCs, presenting the following features:
  - Sketches over a video stream (live or pre-recorded).
  - Annotations following motion video features detected by the difference of frames.
  - Replay of annotations in the same order that they were made.

- A multimodal video annotator for Tablet PCs and applied to contemporary dance, which exploits:
  - Bimanual pen and touch interaction, in which touch is mainly used for mode and tool switching, exploiting the non-preferred hand.
  - Video annotations that can be annotation marks, audio, text, ink strokes or hyperlinks.
  - Annotations that can be made in real-time, using a live video stream, or post-event, using a pre-recorded video stream.
  - Two annotations modes: (1) continuous, associating annotations to a video segment, and (2) suspended, which associates annotations to a video frame.
  - Two annotations methods: (1) hold and overlay, pausing the video while annotation task occurs and superimposing the live event in semi-transparent layer, and (2) hold and speed up, which pauses the video while annotation task occurs and plays it faster after, until the displayed video is synchronized with the live event.

  - Two video visualization modes: (1) real-time, showing the recorded video synchronized with the live event, and (2) delayed, in which the video recording is showed to the user with a time delay.
  - The integration with real-time state-of-the-art motion tracking methods, in order to maintain annotations context.

- A proof-of-concept prototype of video as ink for Tablet PCs, presenting the following features:
  - Video editing operations, such as adding, moving and removing video content, following the inking principle.
  - Transitions effects, like fading.
  - Two selection modes: one using pen gestures, called selection by inking, and the other using a lasso tool.
  - Zoom feature based on pen pressure.
- The integration of openFrameworks and Qt frameworks using Microsoft Visual Studio 2008 & 2010.
- The Microsoft Visual Studio 2010 update of the bbTablet library.

1.5 Publications

Different aspects of the work are included in the following publications:

- [CC09] Diogo Cabral and Nuno Correia. Pen-based video annotations: A proposal and a prototype for Tablet PCs. In Proceedings of the 12th IFIP TC13 Human-Computer Interaction International Conference, Part II, volume 5727 of Lecture Notes in Computer Science, INTERACT'09. Uppsala, Sweden: Springer, 2009, pages 17-20.
- [CCSVFC11] Diogo Cabral, Urândia Carvalho, João Silva, João Valente, Carla Fernandes, and Nuno Correia. Multimodal video annotation for contemporary dance creation. In Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems, CHI EA '11. Vancouver, BC, Canada: ACM, 2011, pages 2293-2298.
- [CV11] Diogo Cabral and João Valente. Programmer's Guide for QT Gui + openFrameWorks(of) in C++ (Visual Studio 2008 & 2010). Technical report, CITI and DI, FCT/UNL, 2011.
- [CVSAFC11] Diogo Cabral, João Valente, João Silva, Urândia Aragão, Carla Fernandes, and Nuno Correia. A creation-tool for contemporary dance using multimodal video annotation. In Proceedings of the 19th ACM international conference on Multimedia, MM'1. Scottsdale, AZ, USA: ACM, 2011, pages 905-908.

- [CVAFC12] Diogo Cabral, João G. Valente, Urândia Aragão, Carla Fernandes, and Nuno Correia. Evaluation of a multimodal video annotator for contemporary dance. In Proceedings of the 11th International Working Conference on Advanced Visual Interfaces, AVI'12. Capri, Italy: ACM, 2012 pages 572-579.

- [CC12] Diogo Cabral and Nuno Correia. VideoInk: A pen-based approach for video editing. In Adjunct proceedings of the 25th annual ACM symposium on User interface software and technology, UIST Adjunct Proceedings '12. Boston, MA, USA: ACM, 2012, pages 67-68.

- [SCFC12] João Silva, Diogo Cabral, Carla Fernandes, and Nuno Correia. Real-time annotation of video objects on tablet computers. In Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia, MUM '12. Ulm, Germany: ACM, 2012 pages 19:1-19:9.

1.6 Thesis Organization

The literature review of the different approaches regarding pen-based technology, video annotations and video manipulation is presented and discussed in Chapter 2.

In Chapter 3, it is presented the concept of pen-based video annotations and it is discussed how it can be implemented. In this chapter it is also described a case study of a multimodal video annotator applied to contemporary dance, which implements, compares and evaluates different input modalities and annotations methods.

Chapter 4 discusses how pen-based technology can be applied to video editing and presents the concept of video as ink as well as its implementation and application on video editing operations. In Chapter 4, it is also described the users' feedback on the implementation of the video as ink concept.

The conclusions and discussion regarding future work are reported in Chapter 5.

2

Background and Related Work

In the first section of this chapter, it is shown how important pen-based technology can be for human interactions as well as a summary of its different approaches. In the other two sections, two topics related to video interaction are discussed: video annotation and video editing. In each one, different approaches are described as well as how they led to pen-based interactions, i.e., pen-based video annotation and pen-based video editing. The first presents approaches that augment video content with annotations, particularly those made with pen strokes or digital ink. The second summarizes how video editing was changed by digital technology and its different perspectives, particularly those involving pen-based technology.

2.1 Pen-based Technology

The usage of hand tools for painting or writing is a natural human behavior, since prehistoric times. In this period, hand tools were used for bones engraving and cave painting. Among these tools are flint blades; chunks of charcoal and mangenese used as crayons; wooden sticks, bone splinters and animal-hairs brushes used with ink [Whi03]. These prehistoric tools can be considered the ancestors of the pen, brush and pencil tools.

Around the fourth millennium BC, the invention of writing deeply marked the evolution of hand tools [Fis04]. A reed stylus to write on clay or its combination with ink are first examples of pen tools [Fis04]. Since then, pen tools evolved through time, depending on the different tasks for which they were aimed [Mey95; Fis04].

With the development of electric and electronic devices, there has been the temptation to mimic pen and other hand tools in such devices. The first device like this was patented by Elisha Gray, in 1888, and was called the Telautograph [Gra88]. The device captured

the handwritten text using an electric pen and reproduced it on other device, working almost as a modern fax device (Figure 2.1). The Telautograph was part of Vannevar Bush's Memex [Bus45], as shown in section 1.1.

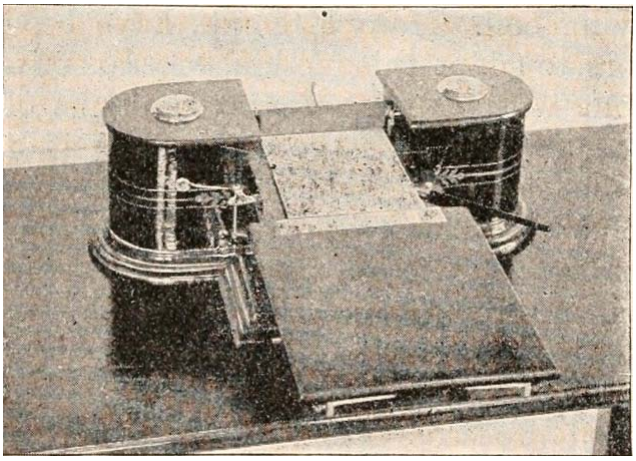


Figure 2.1: Telautograph transmitter.

It is possible to say that the first achievements of combining pen-based technology with computers were made in the 1950s, in the Whirlwind project, a collaboration between the American Navy and Air Force and the Massachusetts Institute of Technology (MIT), which included the development of the TX-0 and TX-2 computers [RS00; Wei08; Bux12]. The first approach was the light-gun, with the size and shape of a pistol, and developed by Robert Everett [RS00; Wei08; Bux12]. It sensed the light on the CRT screen and caused a computer interrupt [RS00; Wei08]. The second approach was the light pen, initially developed by Gurley and Woodward [GW59]. It was similar to the light-gun but with the size and shape of a pen [Wei08] and can be considered the first technology with a pen-shaped device used directly on a screen [Bux12]. The light pen was improved during the development of TX-0 and TX-2 computers [Wei08], and was used in the popular interactive system developed by Ivan Sutherland, the Sketchpad [Sut63] (Figure 2.2). The Sketchpad allowed a set of operations, like creation, deleting, merging and selection of geometric figures displayed in the CRT screen and using a light pen.

In the end of the 1950s and beginning of the 1960s, other type pen devices appeared: the digitizer tablet (also called graphics tablet or data tablet). The digitizer tablet is composed of a writing/drawing surface equipped with a stylus. It maps the two-dimensional coordinates of the stylus on the surface and sends them to other device. The idea of the digitizer tablet is similar to that already presented in Telautograph but with possibility to be attached to other devices, e.g., computers. The first approaches of the digitizer tablet were the Stylator [Dim57] and RAND Tablet [DE64]. Currently, digitizer tablets are commercialized by companies like Wacom[1] (Figure 2.3).

The superimposition of a display on a digitizer tablet surface was made in the end

[1] http://www.wacom.com/

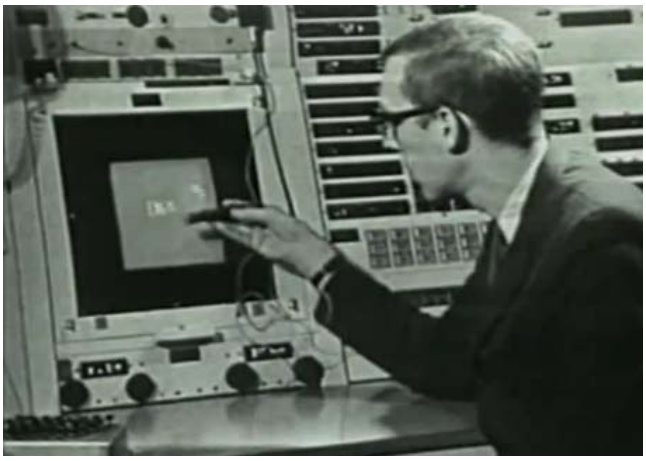


Figure 2.2: Light pen from Sketchpad.[2].

Figure 2.3: Wacom digitizer tablet connected to a laptop computer.

1960s by Gallenson [Gal67]. The idea was similar to the light pen, developed in the Whirlwind project, but with the technology used for digitizer tablets. In Gallenson's work, an upgrade of the RAND tablet, the CRT image was projected onto a rear-projection screen built in the tablet surface. The main advantage of this approach, compared with the traditional digitizer tablets, was the direct interaction of graphics elements displayed on the screen.

These developments inspired the idea of having a portable computer equipped with a pen or a stylus, which could be used directly on the computer screen for a more natural interaction. Alan Kay's Dynabook [Kay72] or the Apple Bashful [Gol10] are examples of this vision. Nonetheless, the first commercial pen-based computer only appeared in middle 1980s, the CASIO IF-8000, an electronic calculator with schedule and notebook functions [Bux01].

[2]From the Science Reporter TV series, episode "Computer Sketchpad", produced by WGBH, 1964, available at `http://www.youtube.com/watch?v=7fHFZcMD3-Ml`

In the end of the 1980s and beginning of the 1990s, the first Tablet Personal Computers (also known as Tablet PCs, Tablet Computers, Pen Computers or Slates) were developed and commercialized. The first commercialized Tablet was the GRiDpad [Mog06]. However, the GRiDPAD ran the same operating system and applications of regular PCs, like MS-DOS and Microsoft Windows [Bux01]. In the case of the GO Tablet, commercialized in the beginning of 1990s, it ran an operating system particularly developed for pen devices, called PenPoint [CS91; Kap96; Bux01]. Kaplan and Carr (both co-founders of GO Corp.) and Shafer (responsible by PenPoint development at GO Corp.) not only stressed the need of having a operating system particularly developed for a mobile pen computer but also the need of a specific designed interface for pen interactions [CS91; Kap96]. This interface, which they called *NUI: Notebook User Interface*, was inspired in the metaphor of paper notebooks [CS91]. Different tabs or sections of the PenPoint menu, were displayed as notebook bookmarks. The PenPoint also presented two main modalities for pen interfaces: indirect input through ink (using handwriting recognition) and direct input with graphical elements using the pen features, e.g., taping the screen with the pen tip [CS91]. Since the Penpoint, other research work has exploited these two modalities, like in XLibris [PGS98; SGP98], where ink was used for linking documents; in LEAN [RB03] and Zlider [RB05], where pen pressure was used for video browsing; in InkSeine [HZSBCST07] (Figure 2.4(a)), where handwritten notes were used for content searching and in Papiercraft [LGHH08], where pre-defined pen gestures were used for document editing and annotation tasks (Figure 2.4(b)).

(a) (b)

Figure 2.4: Pen interaction modes. (a) Inkseine: Handwritten notes to search information [HZSBCST07]. (b) Papiercraft: pen gesture to copy content [LGHH08].

These first developments of Tablet Computers led to other type of devices known as Personal Digital Assistants (PDAs) (Figure 2.5). PDAs are mobile devices focused on managing personal information. Among the first commercialized PDAs equipped with a stylus are the Apple Newton and the PalmPilot [Mog06]. In the last decade, the Tablets and the PDAs evolved and are in a continuous development. PDAs have been merged with mobile phones, changing their name for *smartphones*. The Samsung Galaxy Note[3] or the iPhone[4] are some of the current market examples. Different models of Tablet also appeared in the last years. The Tablets/Slates, which can use a pen or human touch

[3] http://www.samsung.com/global/microsite/galaxynote/note/index.html?type=find
[4] http://www.apple.com/iphone/

for input, or the so called *convertible tablets*, regular laptops that can be converted into Tablets/Slates and are more associated with the name Tablet PC. The Lenovo ThinkPad Tablets[5], the iPad[6], the Samsung ATIV Smart PC Pro[7] or the Microsoft Surface Pro[8] are some of the most recent examples of Tablet Computers.

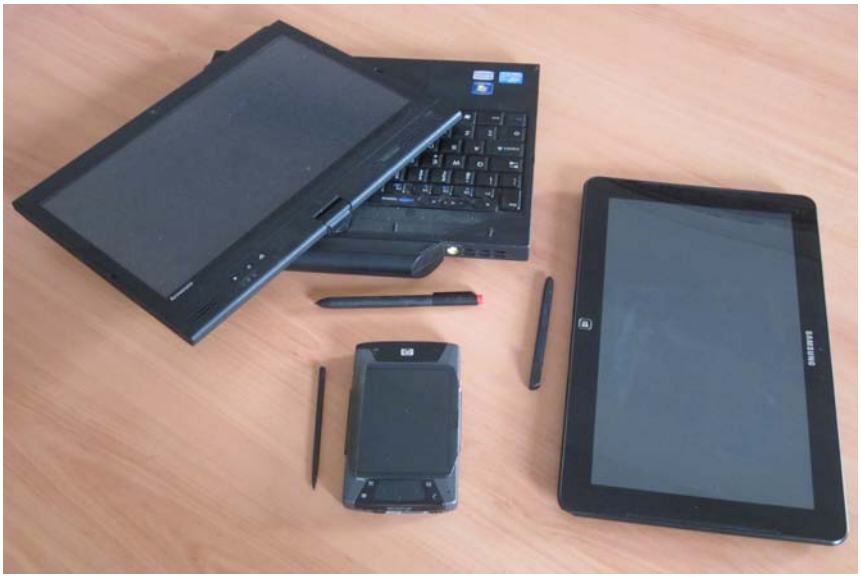


Figure 2.5: Tablet computers: a PDA on the left bottom, a convertible tablet on the left top and a tablet/slate computer on the right.

In the beginning of the 1990s, other two approaches of pen-based technology have appeared: the optical pen (or digital pen) and pen-based technology applied to large displays. The first uses optical hardware embedded in a digital pen, in order to digitize paper information and to record handwriting on paper. The first developments of optical pens were made by Bennett et al [BBDER91] and led to the current Anoto[9] (Figure 2.6) and Livescribe[10] digital pens.

Figure 2.6: Anoto Pen: Regular ink pen and camera[11].

[5] http://www.lenovo.com/products/us/tablet/thinkpad/
[6] http://www.apple.com/ipad/
[7] http://www.samsung.com/global/ativ/ativ_pc_pro.html
[8] http://www.microsoft.com/Surface/en-US/surface-with-windows-8-pro
[9] http://www.anoto.com
[10] http://www.livescribe.com/en-us/
[11] "Anoto digital pen", ©2007 by Anoto AB, at Flickr.com, used under a Creative Commons Attribution-NonCommercial-NoDerivs license: http://creativecommons.org/licenses/

In Liveboard[EBGGHJLMPPTW92], pen-based technology was applied to a vertical interactive large display transposing chalkboards and whiteboards interactions to digital technology. Currently, pen-based technology on large displays can be found on Smartboard[12] devices (Figure 2.7).

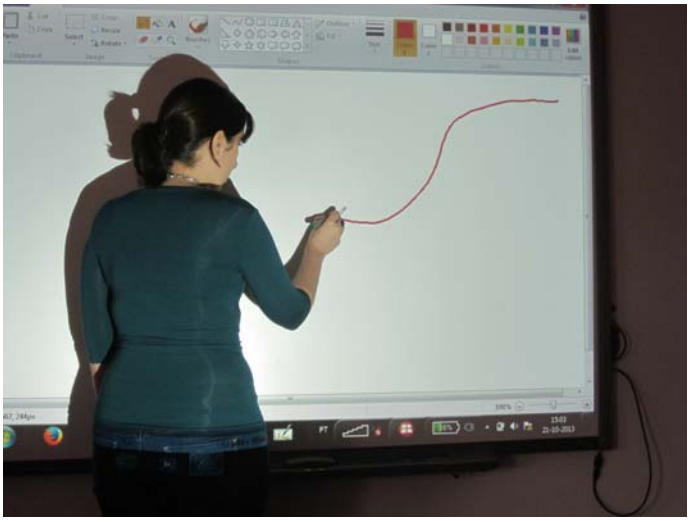


Figure 2.7: Smartboard.

An overview of the different hardware pen-based implementations were described by Meyer [Mey95], Foley et al [FDFH96] and Le et al [LMZ05].

2.1.1 Discussion

In this research, pen-based technology is used for video manipulation. Since one of the goals of this research was to exploit pen-based interactions, which could be applied to multimedia content on future electronic paper, Tablet PCs were chosen as the pen-based platform. The ability of direct input, important to control visual elements, their computational power, needed for video processing and their size, the same or less than regular laptops, made them the ideal technology for this research. As mentioned, both types of pen-based interaction, using ink or as direct input, were exploited. The first one was applied to video annotations and the second to video editing. Although they are exploited separately, they can be combined, as referred in section 1.1. These topics are discussed in the next sections, particularly in 2.2.1, about pen-based video annotations, and in 2.3.1, about pen-based video interaction and editing.

[12] http://smarttech.com/smartboard

2.2 Video Annotation

Marshal [Mar97], Sellen and Harper [SH03] reported the importance and useful role of annotations on paper documents, as mentioned in section 1.1. The first approaches for annotating video combined paper notes with video content, usually linked by a time stamp written on the beginning of each note [Mac89; Cha04; HHL10]. Bottoni et al [BCLOPT04] remarked the evolution from the traditional notion of annotation to digital annotations, which can be added to any multimedia document or some parts of it. Digital technology not only changed annotations but also video manipulation and visualization, as Mackay et al [MD89] and Yeo et al [YY97] pointed out. Among the affordances of digital technology, defined by Sellen and Harper [SH03], is the ability of storing and accessing a large amount of information and displaying multimedia documents, like video content. Both evolutions, of annotations and video, on digital platforms, allowed their combination, in a task called video annotation, which can be an important part of video data analysis, as previously shown in section 1.1.

Different approaches were tried in order to implement the concept of video annotations using digital platforms. One of the first digital systems for video annotation was EVA: Experimental Video Annotator [Mac89; MD89], which was developed and used for video protocol analysis (Figure 2.8). The EVA system allowed the previous creation of a set of buttons with tags, which were pressed down during a video recording session, in order to mark a video moment with that particular tag. During the recording session an user could also write additional text notes using a text editor and the keyboard. After the session, a user could also use the EVA system to transcribe text from the audio track, using the text editor. A few years later the VANNA [HB92] system presented similar features when compared to the EVA system. The key differences were: the possibility of grouping different types of annotations and the usage of different input devices including a touch screen, digital stylus, mouse, and keyboard.

Figure 2.8: The EVA system [Mac89; MD89].

Since the EVA and the VANNA development, other systems explored digital video annotations with different approaches. The MAD [BRFSC96] is a system for hierarchal structured multimedia authoring based on digital video, which also allowed voice annotations associated with time-based media. The MRAS: Microsoft Research Annotation System [BGGS99] is a Web-based client/server framework that supported the association of any segment of addressable media content with other segment of addressable media, allowing collaborative and asynchronous annotation of on-demand streaming video. In the MRAS system, the user annotated the video content with text and audio notes and shared these annotations with other users. In the VideoAnnEx[13] [NLSTB02], the MPEG-7 standard is used for annotation and an annotation is associated to a video region but in the interface they are shown separately. The AntV [CC99] went a little further by exploiting multimedia annotations, in the form of text, images or videos, over the main video window (Figure 2.9). In VAnnotator [CCGa02] the concept of video-lenses was introduced, allowing different perspectives of the same video stream. Video-lenses could take different forms, like an audio or visual playback window or a text window showing an annotation made by an user. For each video-lens there is a correspondent track in the global timeline, allowing video browsing based on the different video-lenses.

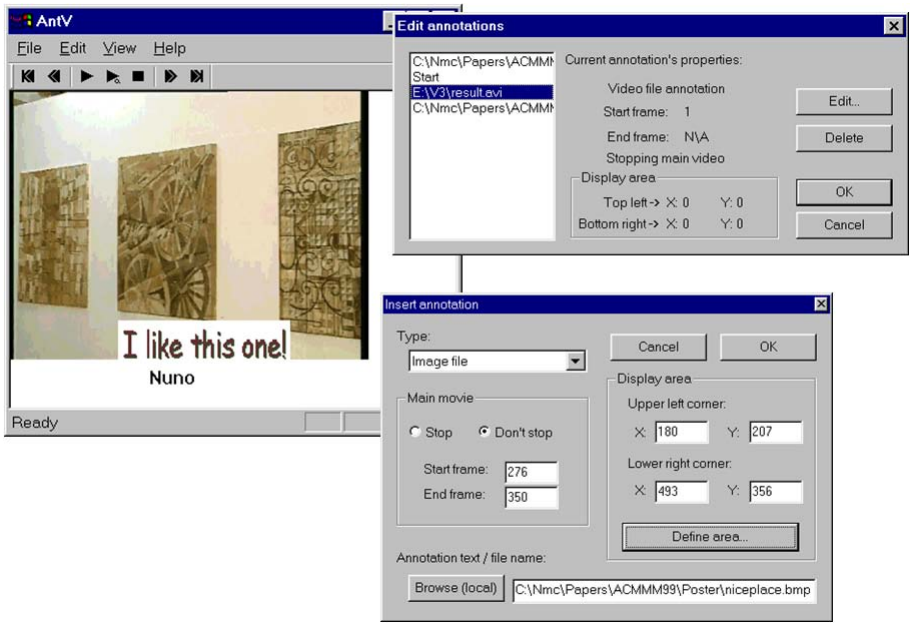


Figure 2.9: AntV annotator [CC99].

Other projects, like ANVIL[14] [Kip01] (Figure 2.10), Video Traces [CFS03], Family Video Archive [AGL03], ELAN[15] [BRN04], VCode & VData[16] [HHK08], VideoStore [CC06; CC07],

[13] http://www.research.ibm.com/VideoAnnEx/

[14] http://www.anvil-software.org/

[15] http://tla.mpi.nl/tools/tla-tools/elan/, Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands

[16] http://social.cs.uiuc.edu/projects/vcode.html

DanVideo [KAG10] or YouTube Video Annotations[17] implemented similar types of video annotation comparable with the systems mentioned above.

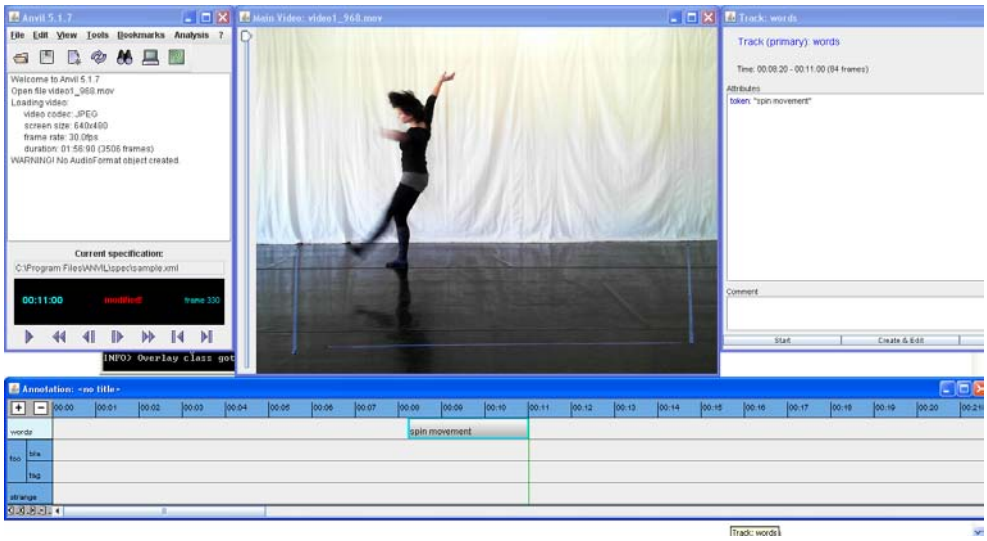


Figure 2.10: Anvil annotator [Kip01].

Although the projects mentioned above present valid approaches for video annotations, they missed the functionality to maintain the association between a particular note and a moving object. The annotations are linked to a video segment or to a video region, which should be (re)selected by the user, in each frame or second, in order to keep the annotation associated with a moving feature.

In DAVid [TH04] semi-automatic placement of annotations was exploited. In order to place annotations, global or associated with a moving object, on a low importance region, the system applied low-level visual features to detect these regions. The video content was analyzed based on elementary region properties: homogeneity, motion and clutter. In DAVid, tracking was achieved by a semi-automatic approach, during the preprocessing phase, where one has to mark the object in an arbitrary frame and after it, the system applies different low-level techniques, like active contours, feature tracking and optical flow, to object tracking.

The work presented by Rosten et al [RRD05] applied the concept of video annotations in an augmented reality environment. In Rosten's work, the ARToolkit [KB99] was used to detect and track physical marks placed in the environment. Once a mark becomes visible in the camera, a corresponding annotation was placed next to it, in an area free of any interesting features. Approaches that use augmented reality for video annotation, usually present two major limitations: the environment has to be physically modified, with a set of markers, and the annotations have to be previously made. Güven et al [GF03; GFO06] avoided to change physically the environment by using 3D models of space that surrounds the user, in which is possible to add different virtual/media objects. Since the user camera is synchronized with the 3D model, each virtual/media object is added not

[17] `http://www.youtube.com/t/annotations_about`

only to the 3D model but also to the scene viewed by the user. Nonetheless, the usage of a 3D model of the environment introduces other kind of limitations, like the size of the environment or the time needed to build it.

In the porTiVity Annotation Tool [NTM07] video annotations were pre-extracted using video content analysis, stored using the MPEG-7 standard and could be completed and edited by an user. Using the porTiVity Annotation Tool was possible to define the video region of a video object associated with an annotation. This object location could be used for searching similar regions in other shots of the video and for object tracking in each shot using a keypoint based generic tracker [TM07]. The Videolyzer [DGE09] presents a similar approach by analyzing the video and audio content for text transcription used as anchor for user made video annotations. Both of these systems present the need of video pre-processing.

The MediaDiver [MFAHIFFFJ11] applies annotations, text or hyperlinks, to a multi-camera context. This approach allows one to follow an object in different perspectives, e.g, following a player in a sports event. In MediaDiver, one has to manually identify the points of the video object on the keyframes, thus creating the base for motion interpolation. This system requires a multi-camera setup, only available in special events, and manual identification of video objects, which can be presented as strong limitations for regular usage.

Although these were interesting approaches to video annotation, they lack the interaction and creative fluidity usually found on paper annotation. As Backon [Bac06] pointed out, the keyboard fosters productivity, whereas the pen fosters creativity “as an extension of human hand and several regions of the brain”. This idea was reinforced by Buxton [Bux07] considering “sketching as an aid to thought”. Some years before Backon and Buxton, Marshall [Mar97] also mentioned “the informal and unconstrained pen-based sketching mechanisms [...] may be a far more appropriate model for annotating materials in the digital world” and Harrison and Baecker [HB92] pointed out that "Users access the various capabilities of the tool using interfaces which have low visual attentional demands. The kinds of mechanisms might include button presses, touch typing, the ability to point directly to the monitor using a touch screen or draw directly using a stylus. It may be desirable for interface mechanisms and graphic annotations to be overlaid on top of the video.". Bargeron and Moscovich [BM03] also mentioned this same idea, “Free-form document annotation is a crucial part of every knowledge worker’ s life. [...] One key advantage is the ease with which the reader may sketch unstructured notes and drawings in response to document content. There are definite advantages to emulating this annotation ability on a computer.”. In order to achieve this fluidity on video annotation, different approaches were made using pen-based technology, as described next.

2.2.1 Pen-based Video Annotations

One of the first approaches of annotating motion pictures with pen-based technology was patented by Leonard Reiffel [Rei71], in the end of the 1960s, and was called *Telestrator*. The system used two cameras: one to record the background scene and the other to record the linear images made with a light pen and shown in a display. A video mixer combined the two streams superimposing the sketches over a recorded video stream. Due the Telestrator invention, Reiffel was awarded with the Technology and Engineering EMMY Award 2003/2004, given by the National Academy of Television Arts and Sciences (NATAS) [Nat11].

Some few years later, the GALATEA system [PS76] integrated film and pen-based technology (Figure 2.11). The system was developed to analyze films produced in a biophysics laboratory. A film was projected and an user was able to show different features using an acoustic digitizing tablet. The pen movements were recorded by a computer, which was connected to a projector, reproducing these actions on the same screen of the film, combining both streams. The pen actions included graphical outputs, such as sketches or animations, and physical measurements of the film content, like the velocity of a motion feature. The system was later updated [BMR81] allowing the usage of video tapes.

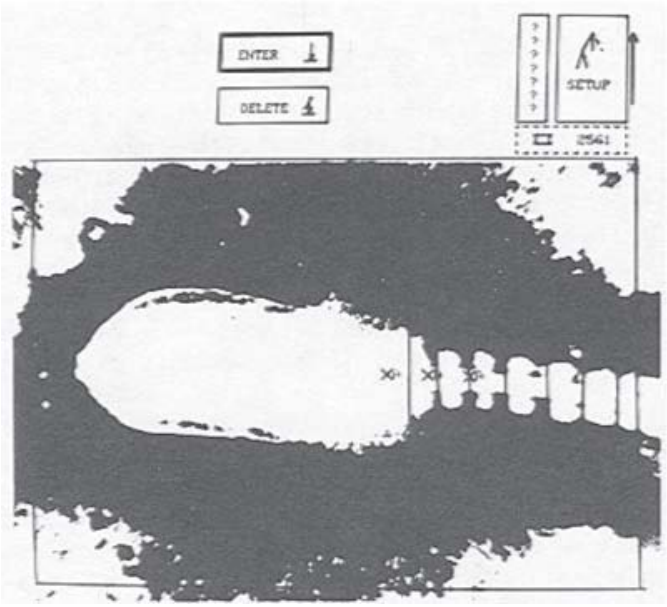


Figure 2.11: Galatea system: sketches combined with film projection [PS76].

Marquee [WP94], a pen-based video logging tool, allowed users to associate handwritten keywords and personal notes with a timezone of a videotape, during or after a particular event. The system used a flat tablet display and a Hi8 recording device, both attached to a personal computer, which allowed to create and review (after the event) annotations associated with video content but in separated devices. The system interface was composed of three main areas: a keyword area, a note taking area and an area with

the VCR controls. The note taking area was divided into different timezones, which were defined by drawing an horizontal line. Vertical pen-gestures were used to associated an keyword to a particular timezone.

NoteLook [CKRW99] was a client-server system for conference rooms. The client application ran on a wireless pen-based computer, which received video streams of the room activity and presentation material sent by the system server. The system allowed the ability to annotate, with pen strokes, still images from both streams (presentation slides and video frames), which were linked based on time stamps, and published on the Web.

Pen-based annotations, associated with video content, were also developed in the LEAN [RB03] and Videotater [DE06] projects. In the LEAN system, it was possible to annotate a particular frame (on top of it or aside), or a video segment, and browse video content based on the annotations, as shown in Figure 2.12. In Videotater, the annotations were associated to a video segment, on the global timeline.

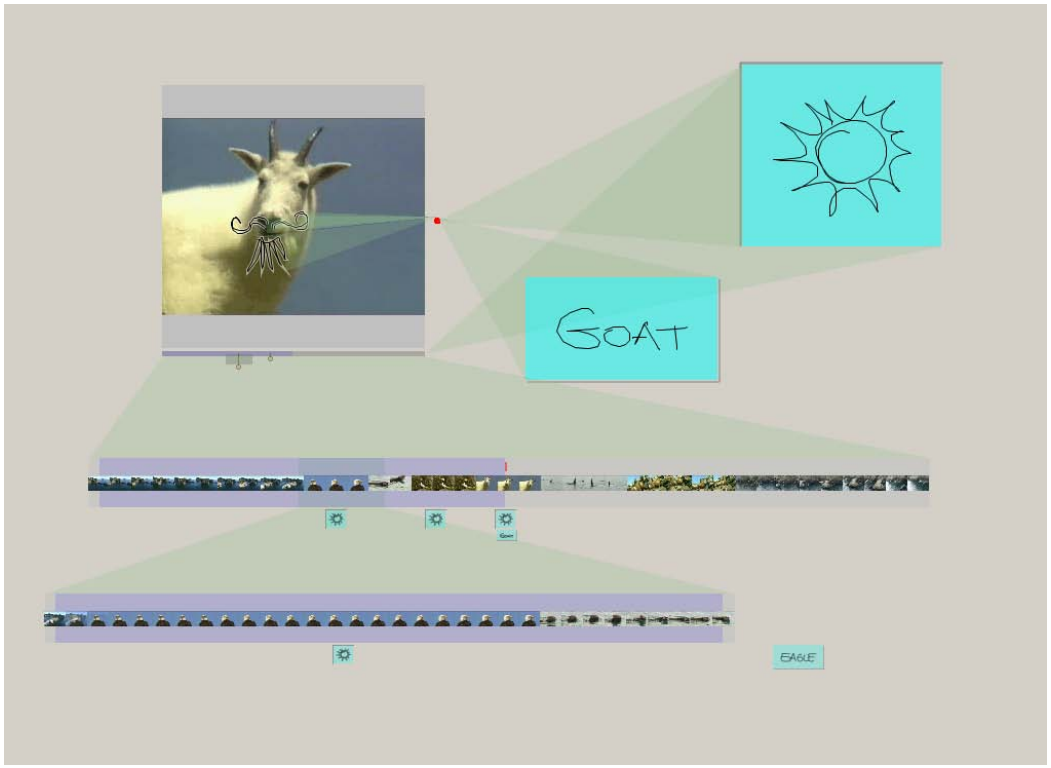


Figure 2.12: Pen-based annotations on the LEAN system [RB03]: associated to a video frame (on top of it or aside) or to video segment (at the bottom left).

The WaC tool [CTGP08] used video annotations in an interactive TV scenario. The annotations made by the user could be text, audio or ink strokes. The ink annotations could be made on a video frame previously selected by the user. The system produced a document, according to interactive TV specifications, which allowed the display of the annotated video content. The M4note [GCGIJCP04], a previous version of the WaC, also allowed the annotation of a live video stream. Nonetheless, the annotations were also

associated to a single frame.

The CAV [EG04; RLG08] tool, displayed the video frames separately frame-by-frame, allowing to sketch ink annotations in each frame. This mechanism enables to add ink annotations to a pre-recorded usability study video. The usability tests showed the need to improve the timeline interface and the navigation tool for browsing annotations.

A more recent collaborative Web-based video annotation system, the Choreographer's Notebook [SLCL11], allows video annotations composed by text, ink and video. A time stamp is used in order to associate the annotations to the pre-recorded video content. The annotations can be displayed on top of the video window or around it. Although it is possible to sketch with a stylus in the Choreographer's Notebook, the system was not specifically designed for Tablet PCs.

In the projects mentioned above, annotations have a temporal association but are spatially static and do not track the motion features of the frames. The Ambulant Player [Bul04], presented different types of annotations: text, audio, images or ink strokes. These annotations could be dynamic but their spatial path had to be defined by the user and not by tracking the video objects. Goldman's work [Gol07; GGCSS08] explored graphical video annotations, like graffiti, scribbles, speech balloons, path arrows and hyperlinks, combined with motion tracking but did not considered the particular case of pen-based annotations. In addition, the tracking algorithm based on point and group particles (Figure 2.13), needed a long time of video pre-processing and failed in some of the shots of the field study using a short film.

Figure 2.13: Selecting a region for tracking using point and group particles [GGCSS08].

2.2.2 Discussion

In order to provide a more natural interaction, this research uses pen-based technology applied to video annotations. Handwritten notes or sketches should be superimposed on video, keeping the visual connection between annotations and the annotated features,

like in the *Telestrator* [Rei71], and as recommended by Harrison and Baecker [HB92]. However, since video represents motion, this simple solution of overlaying graphical content on video segments or frames, as present on [CKRW99; RB03; GCGIJCP04; DE06; CTGP08; RLG08; SLCL11], lacks a dynamic connection between moving features and annotations. Therefore, annotations should also follow objects in motion. As previously mentioned, this idea was initially presented in the Ambulant Player [Bul04], but the user had to define the annotation path, requiring additional users' tasks and a previous knowledge of the video content. Goldman's work [Gol07; GGCSS08], went further by combining a motion tracking algorithm with graphical notes. The main advantage here was that the system automatically detected the moving features, without human intervention, focusing the user attention on the annotation task. However, the motion tracking algorithm required a long-time of video pre-processing and presented fails on tracking, making it useless for live video applications, which require a fluid user interaction.

This work exploits and defines the concept of pen-based video annotations: ink strokes associated with video changeable temporal and spatial positions, i.e., video motion. These annotations can be made on a recorded video or on a live video stream. Therefore, the presented approach uses motion tracking algorithms that can work on live streams and annotations methods to facilitate the annotation task on a visual dynamic medium. In addition, pen-based video annotations were compared with other type of annotations in a multimodal video annotator, called Creation-Tool.

	Video		Pen-based Annotations			
	Recorded	Live	Overlaid with Video	Associated with Video Frame	Associated with Video Segments	Associated with Moving Objects
Telestrator	Yes	Yes	Yes	No	Yes	No
Galatea	Yes	No	Yes	Yes	Yes	No
Marquee	Yes	Yes	No	No	Yes	No
Notelook	Yes	Yes	Yes	Yes	No	No
LEAN	Yes	No	Yes	Yes	Yes	No
Videotater	Yes	No	No	Yes	Yes	No
WaC/M4note	Yes	Yes	Yes	Yes	No	No
CAV	Yes	No	Yes	Yes	No	No
Choreographer's Notebook	Yes	No	Yes	Yes	No	No
Ambulant Player	Yes	No	Yes	No	Yes	Yes (Manually)
Goldman's work	Yes	No	Yes	No	Yes	Yes (Pre-processing)
Creation-Tool	Yes	Yes	Yes	Yes	Yes	Yes (Real-time)

Table 2.1: Pen-based Video Annotation Systems

2.3 Video Editing

As Dancyger [Dan11] refers in his book *The Technique of Film & Video Editing*, in the first motion pictures, dated from end of the XIX century, there was no editing. The camera recorded an event, an act, or an incident. Many of them were a single short shot. It was with Edwin S. Porter (e.g, *The Life of an American Fireman*[18], 1903, Figure 2.14), D.W. Griffith (e.g., *The Birth of a Nation*[19], 1915), Sergei Eisenstein (e.g., *Strike*, 1924) and Alfred Hitchcock (e.g., *Blackmail*, 1929 - initial sound film editing experiments) in the beginning of the XX century, that film editing took the form that is known today. Since these developments and achievements, film editing is a key process in cinematography.

Figure 2.14: First scene of Porter's film: *The Life of an American Fireman*.

In the *Dictionnaire mondial du Cinèma* [VBGLS+11] two types of editing are described: the traditional and the virtual. Traditional editing is described as the assembling of elements of film or tape (Figure 2.15(a)), which almost no one does anymore, whereas virtual editing can be described as the editing process using computers and digital technologies (Figure 2.15(b)). Virtual editing is the most common existing method. A consequence of this transformation was the removal of the two chapters about traditional editing from the first edition of Chandler's book *Cut by Cut* [Cha04] to the second edition, [Cha12]. In the first edition of *Cut by Cut* [Cha04], Chandler presents the different techniques of traditional editing, like cutting or splicing on film, as well as the technology used in this type of editing.

Davis [Dav03] and Chandler [Cha04; Cha12] describe three major technological phases of motion picture editing: physical film cutting; electronic videotape editing and digital nonlinear editing. Dancyger [Dan11] defines nonlinear editing as "random-access editing, sourcing shots, scenes, and sounds on an as needed basis". Although film and video editing suffered the revolution of passing from analog- to digital-driven technology [Dan11], still remain a time-consuming, frustrating and tedious tasks [GBCDFGUW00; CLMBSDYC02; RB03; GCSS06; DE06]. The difficulty of learning how to use current

[18] https://www.youtube.com/watch?v=p4C0gJ7BnLc
[19] http://www.youtube.com/watch?v=iEznh2JZvrI

Figure 2.15: Film editing equipment: (a) Old Moviola machine for editing on film[20]. (b) Modern Editing Equipment[21].

commercial editing software is discussed by Chandler [Cha12]. In addition, Jokela et al [JMK07], in their empirical study about video editing on a mobile context, present as primary users' motivations for not editing the lack of time and the perception that the editing task was not valuable enough for the trouble. In Jokela's study, the technical complexity of the editing process was a barrier for some users.

Chandler [Cha12] identifies three commercial systems for professional video editing: Avid Media Composer[22], Adobe Premiere Pro[23] and Final Cut Pro[24]. For non-professional use the most common commercial software are: Movie Maker[25] and iMovie[26]. Although non-professional applications try to facilitate the editing task, this is mostly achieved by reducing the number of features, when compared with professional systems, as shown in Figure 2.16.

The interface of the video editors are quite similar and the most relevant change, the transformation of a linear timeline in a stacked timeline, i.e., a timeline broken in rows (Figure 2.17), was made in recent versions of non-professional software (since iMovie '08, in 2007 and Windows Live Movie Maker 2009, in 2009). More recently, as Tablets with integrated cameras become more widely distributed, video editing tools particularly developed for these devices have also appeared. The Samsung S Camera[27] for Microsoft Windows 8 Tablets or the iMovie for iPad[28] are examples of this reality. Although, they can be considered adaptations of regular video editing applications for Tablet Computers,

[20] "J&R", ©2002 by User:Cutterette, at en.wikipedia, used under GFDL, http://www.gnu.org/copyleft/fdl.html, and Creative Commons Attribution-ShareAlike, http://creativecommons.org/licenses/by-sa/3.0/, licenses.

[21] "Edit Bay", ©2006 by Josh Miller, at Flickr.com, used under a Creative Commons Attribution-NonCommercial license: http://creativecommons.org/licenses/by-nc/2.0/deed.en.

[22] http://www.avid.com/US/products/media-composer

[23] http://www.adobe.com/products/premiere.html

[24] http://www.apple.com/finalcutpro/

[25] http://windows.microsoft.com/en-US/windows-live/movie-maker-get-started

[26] http://www.apple.com/ilife/imovie/

[27] http://www.samsung.com/uk/windows8apps/s_camera.html

[28] http://www.apple.com/apps/imovie/

(a)

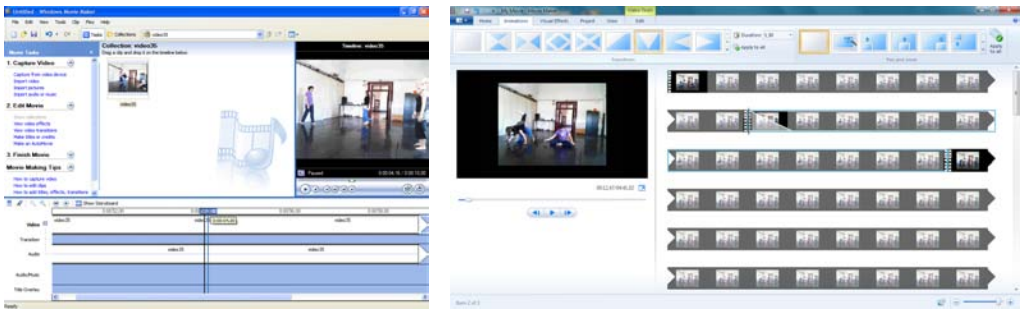
(b)

Figure 2.16: Editing Software: professional vs non-professional. (a) iMovie 11. (b) Final Cut Pro X.

their interface was mainly developed for touch input and missed pen-based interactions. It is also interesting to observe that, even though, Microsoft Windows 8 was developed considering Tablet computers [Win], the Windows Movie Maker interface remains the same as for regular desktops.

In addition, it is also possible to find Web-based platforms for video sharing, authoring and editing, like the Youtube Editor[29], One True Media[30] or the already closed

[29] http://www.youtube.com/editor
[30] http://www.onetruemedia.com/



(a) (b)

Figure 2.17: Editing software: linear timeline vs multiple rows.(a) Windows Movie Maker 5.1. (b) Windows Live Movie Maker 2012.

Jaycut[31]. While these platforms allow the editing of a video stream, they mimic the interfaces and interactions presented in the commercial applications mentioned above.

Most of the research improvements on digital editing were based on trying to automate the process. Projects like Lienhart's work [Lie99], which uses both metadata and sound analysis; Hitchcock [GBCDFGUW00], which uses a semi-automatic video processing approach; Silver [CLMBSDYC02], which uses metadata for automatic editing; Aner-Wolf and Wolf's work [AWW02], which generates a new video stream based on raw amateur videos and photos of the same event; AVE [HLZ03], which uses video, audio and music analysis for automatic editing; Wang and Hirakwa' work [WH06], which has a semi-automatic approach for composing a new video stream based on object movement and camera motion; and EWW [CWB07; WC07], which uses also a semi-automatic approach but using content analysis; are good examples of this philosophy defended and discussed by Davis [Dav03]. However, Dancyger [Dan11] points out that editing is a creative process, which cannot be made by machines, and most of the developments made on storytelling, interactivity and the relation between the storyteller and the audience were made in the fields of video games and education. By taking these facts in consideration and adding the fact of the increasing popularity of video sharing platforms, mentioned in section 1.1, it is possible to conclude that providing more familiar and powerful interfaces, in order to foster easiness and creativity, is a key issue for digital video editing systems.

The Video Mosaic [MP94] combined paper storyboards with video editing software. The system projected the computer output in a desk and recorded user gestures for input commands. In Video Mosaic, one could create a virtual storyboard, print it in paper sheets for sharing and adding notes, and digitize the modified paper sheets. This last step allowed to add the new notes to the virtual storyboards stored in the system. Later, Goldman [GCSS06; Gol07] also used storyboards for digital video visualization and editing. His work focused on creating schematic storyboards, a single static image composed from multiple input frames and annotated using the visual language of

[31] http://jaycut.com/

storyboards (outlines, arrows and text), describing the motion in the screen. An interactive interface was developed using this schematic storyboard for video visualization and editing. By traversing the paths defined by arrows and frames, one could visualize the content and generate a new video stream.

In the Hitchcock system [GBCDFGUW00; GBSBW01] the interface presented piles of clustered video clips from the raw data, based on color histogram analysis. The piles were displayed by temporal order of the first clip. In addition, the system hides to the user clips that do not present enough quality, like insufficient light or too much motion. The user could select the clip from each pile and drag them into the composition panel and generate the new video stream. In the users studies about the initial Hitchcock interface, the users missed some video control features. In order to solve this issue, a video clip length control based on the keyframe resizing was implemented (Figure 2.18).

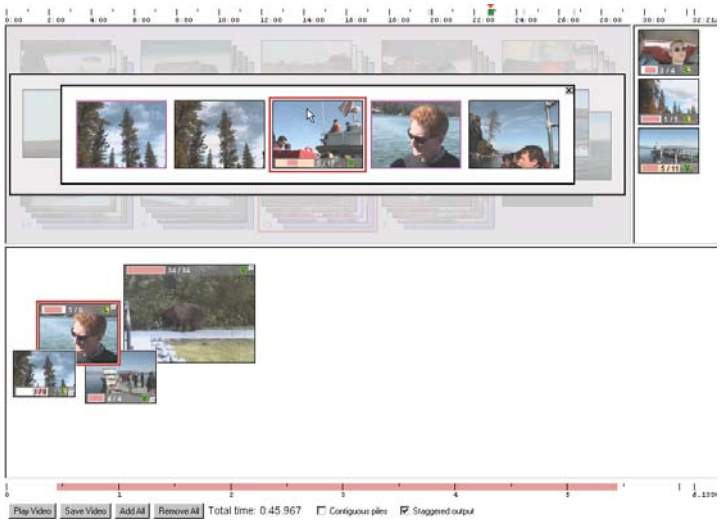


Figure 2.18: The Hitchcock system interface [GBSBW01].

The EnhancedMovie [NIKOS04] explored gesture commands for editing tasks. The system uses a camera for gesture detection and a video projection, in order to simulate a large-size display in a desk. The EnhancedMovie interface allowed one to select a video clip, from a set of clips, and define the start and the end points of a new video segment, using particular gestures.

The TextableMovie [VDJ03; VD04] constructs a new video stream while someone is composing a narrative, using text or voice. The video stream is composed by a set of recorded and annotated video clips. The system matched the narrative text with the keywords associated to each video clip and composed, in real-time, an equivalent video narrative. The TextableMovie was followed by a more tangible and collaborative approach, presented in Movie Pictures [VADWF05; VI07]. In Movie Pictures, the text narrative was replaced by RFID tokens, each associated with a video clip, and that could be reassembled composing a new video stream.

In the Tangible Video Editor (TVE) [ZHSJ07] a multi-user and tangible interface for

video editing based on active tokens was also exploited. The system was composed by a play-controller, a set of clip-holders (Pocket PCs inside plastic cases) and a set of transition connectors (three base types: minimize, rotate and fade). The clip-holders could be attached to each other, with or without a transition between them, on the left or right sides. A data stream flowed from the rightmost clip-holder to the leftmost clip-holder, until the play controller, placed in the beginning of the sequence. Each clip-holder received the information from its neighbor and appended information about its own clip and transition. The play-controller sends the information sequence to a desktop, which displayed the final movie on its monitor. In the users' study, some users missed more complete editing features in the TVE editor, like cut, merge, color correction, cropping or speed control.

The Mobile Video Editor [JKM07] exploited video editing operations in the context of mobile phones. The system used the timeline metaphor present in desktop video editors. The system was separated in different views: gallery view, showing the different video clips; play view, the regular video playback; edit view, allowing to edit a video clip; preview view, playback of the video being edited; insert object dialog, for inserting different objects in the clip; and cut view, in order to cut a video clip. Since the cut view could increase the complexity of the editing view, it was separated from it. The editing interface was composed by two tracks: the video track and the audio track. The transitions were represented by smaller boxes placed between the visual clips. The mobile phone keyboard was used for the different menu tasks, editing and cutting operations. In the two evaluation studies [JKM07; JMK07], the most problematic feature was the video cutting which presented a reduced rate of success.

In order to provide a more fluid video interaction, different research works explored pen-based technology for this task. However, most of them applied this type of technology to video browsing and navigation, as presented in the next section.

2.3.1 Pen-based Video Interaction and Editing

One of the first proposals that used pen computing to control video content was the Marquee [WP94], which uses a pen-based interface to control a VCR device. An user could control the direction and speed of regular VCR controls (play, pause, backward and forward) by drawing a horizontal line using the stylus. The direction and speed of the video content were controlled by the direction and length of the line. In addition, a stylus tap on the within the control area paused the tape.

The usage of pen-based technology to control or edit video content was also tried in other proposals, such as LEAN [RB03], Zlider [RB05], Videotater [DE06] and the MobileZoomSlider/ScrollWheel [HG08].

In the LEAN system [RB03], novel interfaces, the TLSlider and PVSlider, exploring the pen pressure feature, were developed to browse video content. The TLSlider departs from the fish-eye frame layout, which focus the user attention in a particular frame, to

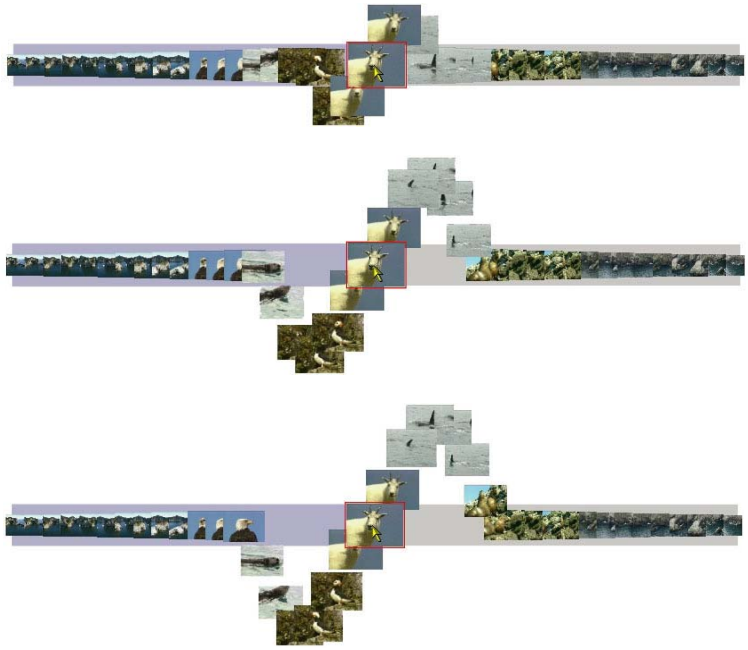


Figure 2.19: TLSlider interface on LEAN system [RB03]: sinusoidal size depends on the pen pressure.

a sinusoidal frame arrangement, that could be increased or decreased depending on the pen pressure and focusing the user attention in a set of frames, as shown in Figure 2.19. The PVSlider added an extra time bar to video navigation. By changing the PVSlider bar position relatively to the video window it was possible to change the video interval (clip) to watch. In addition, it was also possible to change the playback velocity by dragging the pen along the extra bar in the direction of its endpoints. In addition, LEAN used single-strokes gestures commands, which were parsed using Rubine's feature [Rub91]. Adding to the LEAN project, the authors have developed the Zlider [RB05] interface, a timeline with a variable scale which depends on the pen pressure made by the user.

The Videotater [DE06] used vertical and horizontal pen gestures, on a global timeline, in order to split and join different video segments, as shown in Figure 2.20. In addition, the Videotater presented a polyfocal visualization, allowing to watch the endpoints of a video segment and a pre-defined number of frames in its neighborhood.

The MobileZoomSlider/ScrollWheel [HG08] interface allowed one to browse video content using pen-based technology to change the timeline scale. The MobileZoomSlider scale is changed by moving the pen, up or down, in the vertical direction, whereas in the ScrollWheel a circular movement is used, in order to change the timeline scale.

The I/O Brush [RMI04], a drawing tool based on a physical brush, equipped with a small video camera, lights and touch sensors, enabled to record color, texture or movement of any physical object and to reproduce it in a digital canvas using the brush (Figure 2.21). The camera captured a real scene depending on the mode: one frame for the texture, RGB color for the color and 30-frames for movement. The capture was initialized

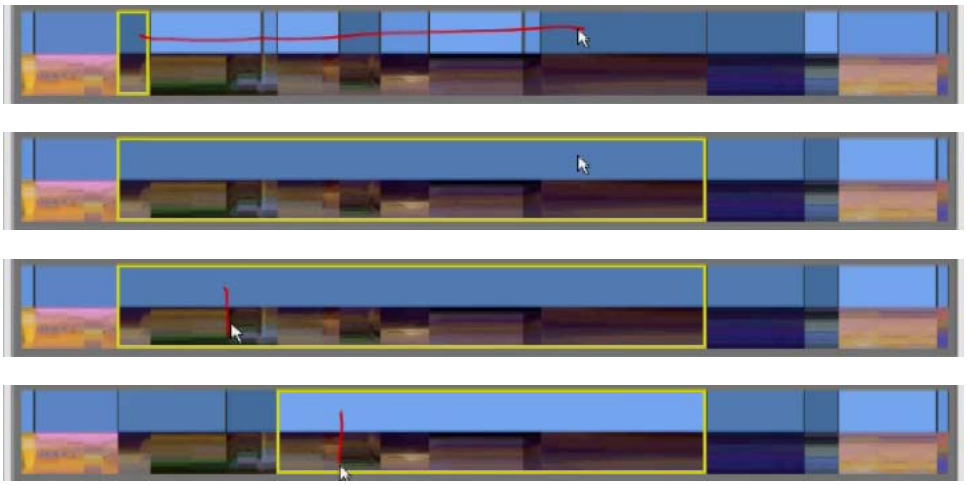


Figure 2.20: Videotater timeline [DE06]: horizontal and vertical pen gestures are used to join or split video segments.

by the touch sensors and illuminated by the lights of the brush. In addition, the coil of a pen tip was embedded in the brush tip with the aim to interact and control a pen-based display used as a digital canvas.

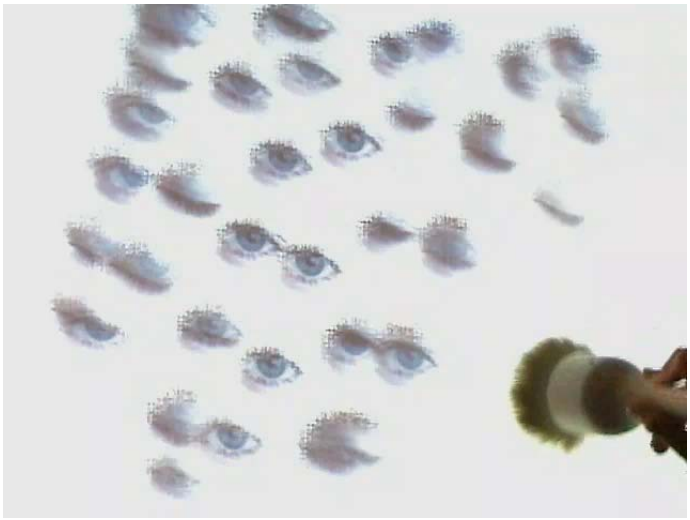


Figure 2.21: I/O Brush [RMI04]: Painting with video content.

2.3.2 Discussion

Marquee [WP94], LEAN [RB03], Zlider [RB05] and MobileZoomSlider/ScrollWheel [HG08] used pen-based interaction only for video browsing and visualization, without changing the content. In Videotater [DE06], pen-based technology was used to join and split video segments. Nevertheless, the Videotater interface follows the traditional scheme of the global timeline and a separated video window, without taking full advantage of the natural interaction provided by pen-based technology. Even though the I/O Brush [RMI04]

was not focused on pen-based video manipulation or control, it shows how digital ink can take different forms besides imitating regular physical ink. The work proposed in this research uses a pen as an input interface for video editing, by "inking" video frames in a canvas and using a video palette formed by video segments. In this canvas, video content can be painted, selected and manipulated, in different directions, fostering visual organization and creativity. In addition, pen pressure is used for zoom features. The different features were implemented in a Tablet PC prototype, called VideoInk.

	Video Browsing	Video Editing	Painting with Video Content
Marquee	Yes	No	No
LEAN	Yes	No	No
Zlider	Yes	No	No
MobileZoomSlide/ ScroolWheel	Yes	No	No
Videotater	Yes	Yes	No
I/O Brush	No	No	Yes
VideoInk	No	Yes	Yes

Table 2.2: Pen-based Video Interaction Systems

3

Pen-based Video Annotations

This chapter discusses the concept of pen-based video annotations as well as its implementation in a proof-of-concept prototype and in a multimodal video annotator applied to contemporary dance. The evaluation of the multimodal video annotator is also discussed in this chapter.

3.1 Pen-based Video Annotations: The Concept

In the previous chapter (section 2.2), the different input modalities used for video annotation were presented and it was discussed the importance of having a fluid input method like pen-based annotations.

Pen-based annotations generate ink-based annotations [CBP00]. Nevertheless, if a keyboard annotation is a set of characters [CBP00], then a pen-based annotation can be considered a set of ink strokes. Each ink stroke has its own attributes, like color or thickness, which can differ from one to another [Mar97; Mar98]. Therefore, a pen-based annotation can be considered a set of ink strokes with different attributes. In addition, a temporal dimension can be added to the ink strokes. Normally, a user makes these ink strokes sequentially [BM03]. This temporal order associated to each ink stroke can be crucial to understand the idea transmitted by the annotation [AHWA04].

On the other hand, video annotations should include spatial and temporal dimensions also associated with video content [CC99; Gol07]. Together, these dimensions generate the idea of motion, a spatial position that varies with time, which should be also associated to video annotations. Thus, video annotations should have the ability to follow specific motion features included in the video content. Goldman [Gol07] called this

type of annotations "Video Objects Annotations (VOA)", i.e., annotations that "are associated with specific objects or regions of the video".

The combination of both concepts, pen-based annotations and video annotations, generate a new kind of annotations, which can be called *pen-based video annotations* (Figure 3.1). It is composed by a set of time dependent ink strokes, each with specific attributes, and associated with video changeable temporal and spatial positions. The attributes of pen-based video annotations can include ink properties (e.g., color, thickness), private or public definitions, hypermedia properties (e.g., spatial anchors on the video content and links to other media types) and tracking features (e.g. tracking objects, colors and textures). Although private or public definitions can play an important role for annotations [Mar98], their study will diverge from the focus of this research.

Figure 3.1: Pen-based video annotations: proof-of-concept prototype - annotating on a recorded video.

Pen-based video annotations, like any other type of video annotations, can be made during a live event, using a video camera for live video recording, or after it, using a recorded video stream. When annotating on a recorded video, the user has full control of the video playback and the object tracking can be pre-processed, as shown in Goldman's work [Gol07; GGCSS08]. However, when the annotation is made on a video that is being recorded, additional challenges have to be faced: object tracking has to be efficient enough to work in real-time and the annotation has to be made in a context that is constantly changing. Both of these challenges are constrained by the fluidity of the user interaction. A slow motion tracker or a non-familiar annotation method can break the interaction fluidity, making hard to accomplish the task. The ability to associate annotations with objects and to create them in a live and mobile environment contributes to more spontaneous experiences, like the ones using a sketchbook. In addition, annotations

made after an event can also take advantage of the annotating methods developed for a live event, thereby improving the user experience, e.g., avoiding the pre-processing time required for motion tracking can improve the user experience on annotating a recorded video.

3.2 Pen-based Video Annotations: Proof-of-Concept Prototype

An initial prototype implementing pen-based video annotations was developed. The system runs on Tablet PCs and includes pen-based annotations displayed over video content, associated by time to video intervals, and with some changeable ink attributes, such as color and thickness. The prototype has two versions, one using a recorded video stream (Figure 3.1) and a second version using a camera for annotating during a live event (Figure 3.2) .

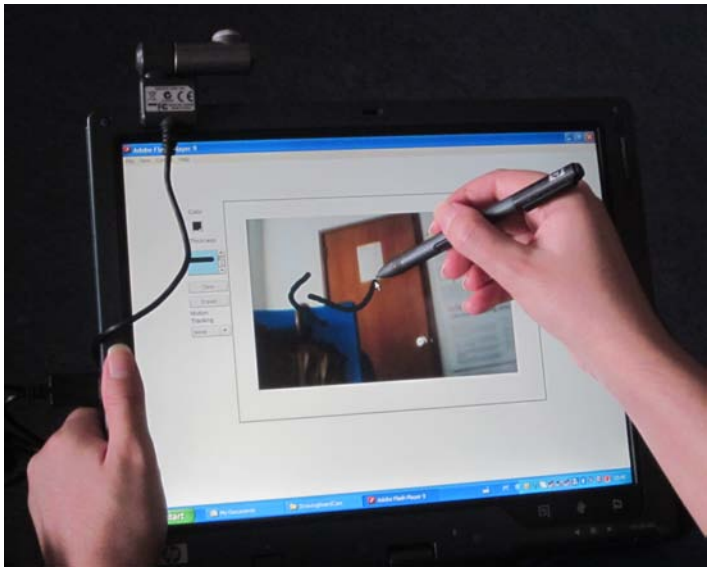


Figure 3.2: Pen-based video annotations: proof-of-concept prototype - annotating on a live stream, using a webcam.

In its first version, the system has a frame based timeline making possible to browse the video content, by selecting a particular frame, and to draw over a set of frames. A scrollbar was also developed, with the aim to navigate through the frame timeline, with the particular feature that the marker has fixed dimensions, sufficient to be selected using the pen, without missing the target, as can happen with regular scrollbars.

The prototype includes annotations motion tracking, which allows annotations to follow motion changes in the video content, as shown in Figure 3.3. The motion tracking is based on the difference of frames on the area defined by the annotation and computed in real-time. The annotation area, defined by a rectangle bounding box around the ink strokes, is equally divided in four smaller rectangles and the rectangle (of these four) that presents a larger change of pixels, indicates the direction that the annotation should

follow. This method does not require a lot of processing power, enabling its real-time usage. However, the frames difference algorithm does not recognize objects, only detects motion, making it incapable to work with motion tracking ambiguities, e.g., which object should be tracked if two objects are moving in the same region of interest. The annotations are drawn for every frame in a transparent layer over the video window, without the need to change the original video content. The system also redraws the ink strokes in the same order that they were made. Annotation motion stops when the object stops or when it is not visible in the video window.

Figure 3.3: Motion tracking based on frames difference.

In the prototype, annotations can also be moved, deleted, replayed and grouped/ungrouped. The annotation lifetime is by default equal to the video time length but the time interval can be changed using a dual slider, in the annotation properties dialog window (Figure 3.4). In this dialog window, it is also possible to change annotation color or thickness.

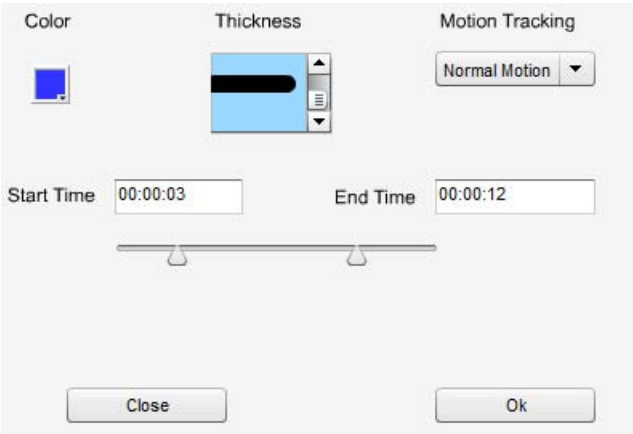


Figure 3.4: Pen-based video annotations: proof-of-concept prototype - annotation properties

The described prototype was implemented using Adobe Flash[1] and done as a proof-of-concept of pen-based annotations. This prototype has several limitations and would benefit with the implementation of different input modalities and annotation methods, as well as the implementation of more motion tracking algorithms. In order to provide a more complete study about pen-based video annotations, a multimodal video annotator was implemented as described next.

[1] http://www.adobe.com/products/flash.html

3.3 Creation-Tool using Multimodal Video Annotation - A Case Study

Creative processes of choreographers, or other authors in the performative arts, involve several rehearsal iterations where video based annotation can significantly enhance the process. The TKB (A Transmedia Knowledge-Base for Contemporary Dance) was a transdisciplinary project [FJ09; FJ13] that aimed at designing and constructing an open-ended multimodal knowledge-base to document, annotate and support the creation of contemporary dance pieces.

The TKB project included three main components: (1) the linguistic analysis and annotation of multimodal corpora; (2) the development of an original customizable software tool to support choreographic creation processes in real time, while allowing personal annotations of the respective authors; and (3) the design and development of a web-based collaborative archive for contemporary dance.

The second component is what it was called *Creation-Tool* [CCSVFC11; CVSAFC11; Val11; CVAFC12; SCFC12; Sil12] (Figure 3.5). The proposal was to design a creation-oriented tool deriving from the results of the annotated video corpora with the aim of feeding back to the choreographic creative process. It was conceived to assist the creative processes of choreographers and dance performers, working as a digital notebook for personal annotations. Therefore, it required a familiar interface and setup, so that the usage of such tool would not interfere with an already existing choreographic method. Choreographers can use the system to analyze and improve their work, by recording and annotating a rehearsal, or a live performance, for a later review or for sharing notes with the performers.

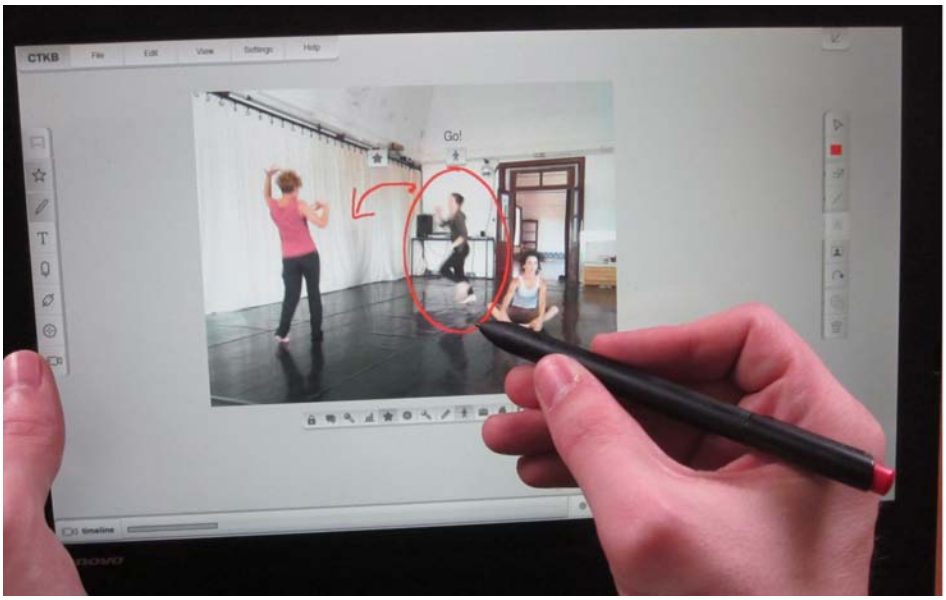


Figure 3.5: The Creation-Tool running on a Tablet PC.

Although the Creation-Tool was developed in a performance dance context, their possible usage on different work methodologies and on other fields, like those mentioned in section 1.1, was always taken in consideration during its development and design. Due to this open usage, the Creation-Tool was not limited to any particular choreographer or dance notation, being sufficiently flexible to be applied to any work environment.

The Creation-Tool was developed for Tablet PCs, using bimanual pen and touch interaction. Touch was mainly used for mode and tool switching, exploiting the non-preferred hand [LHGL05]. The tool supports the capture and multimodal annotation of video content. In addition, the system allows different modes and methods of annotation and video visualization, providing more familiar methods to add notes to dynamic content. In order to maintain the annotations context, motion tracking methods were applied to the annotations. The video content associated with the annotations is tracked and the annotations follow it along the video. The application is also able to capture up to two video input streams. Remote audio capture and control were implemented, thus allowing more freedom of movements and gestures, essential on dance performance environments.

3.3.1 Implementation Technologies

The prototype was implemented in C++, using OpenCV2.2[2], OpenNIv1.5.2.23[3], openFrameworksv0.062[4] and Qt GUIv4.7.3[5] frameworks [CV11]. openFrameworks, a framework for the development of multimedia and interactive applications, is the main platform used in the Creation-Tool. This framework is used for capture, storage and reproduction of video content and annotations but this framework does not include complete libraries for video processing and GUI elements. To address these two needs, we have used the OpenCV, OpenNI and QT frameworks. The first two were used for video processing and motion tracking, whereas the last adds GUI elements to the tool.

3.3.2 Video Annotation Modalities

The Creation-Tool supports the capture and multimodal annotation of video content. The annotations associated to video content can be ink strokes, annotation marks, audio, text and hyperlinks. The time interval of each annotation is defined by the time that user takes to make an annotation. However, this time interval can be changed, in the timeline, after the video is recorded, in a post-analysis context, as shown in Figure 3.6.

3.3.2.1 Pen Annotations

Pen-based video annotations enable the user, in this case the choreographer, to sketch over a video stream, therefore allowing more freedom in the creative process (Figure

[2] http://opencv.willowgarage.com/wiki/
[3] http://www.openni.org/openni-sdk/
[4] http://www.openframeworks.cc/
[5] http://qt-project.org/

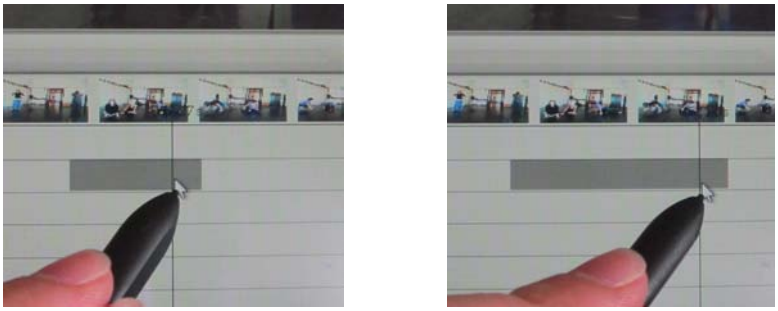


Figure 3.6: Changing annotations time interval using the pen.

3.7). Pen-based video annotations are composed by a set of ink strokes made by the user and, since the pen strokes cannot be made continuously, i.e., without maintaining the pen tip in contact with the display, a timeout mechanism was defined, deciding if two consecutive strokes belong to the same annotation. If the time interval between two strokes passes a defined threshold, then they belong to two different annotations, otherwise they are parts of the same annotation. This threshold was experimentally set to two seconds.

Figure 3.7: Creation-Tool: Pen annotations.

Pen-based annotations can also be used for sketching the icons of annotation marks (Figure 3.8). The sketched icons can be saved as PNG images and included in the icons' library or used as image annotations. This transformation allows one to reuse and search pen-based annotations without requiring handwriting recognition.

3.3.2.2 Annotation Marks

Annotation marks correspond to concepts defined by the user, e.g. These marks are represented by a keyword and an icon (Figure 3.8), e.g., the keyword "like" with a star icon, in contrast to regular annotations that do not have a pre-defined structure. The user has to define the keyword associated to each icon and a set of default icons are available in the tool, although the user can add more icons to it. In order to add a mark to the video, the user should press the corresponding button in the annotation marks bar and press it

again for its removal.

(a) (b)

Figure 3.8: Creation-Tool: Annotations marks. (a) Defining a new mark. (b) Adding annotation marks to a video, red mark on the right was made using pen annotations.

3.3.2.3 Audio Annotations

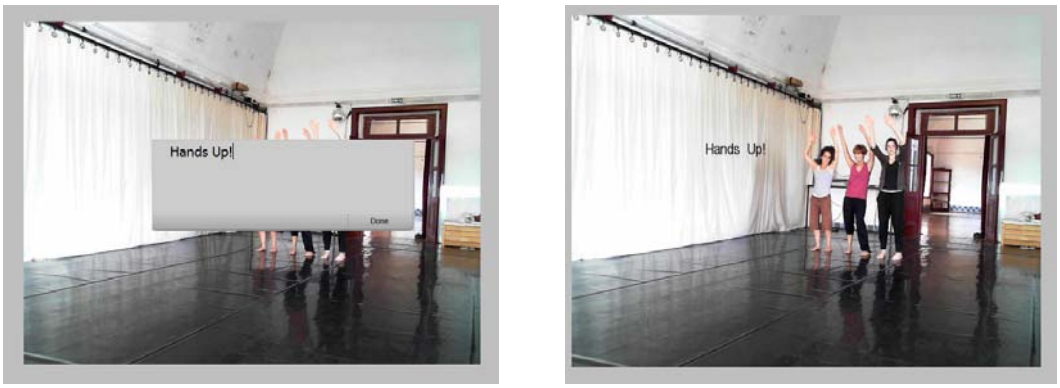
The Creation-Tool enables the recording of voice annotations. When the user wants to record an audio annotation, he/she has to press the correspondent button, an on/off button, which will start (on) or stop (off) the audio recording and define the time interval of the annotation. A waveform of sound the annotation is displayed, providing a visual feedback to the user (Figure 3.9). The system produces a sound file for each annotation made and a main sound file for the background sound of the video. A wireless microphone and a remote control, with a start/stop audio recording button implemented in a mobile device (e.g., smartphone), allows the remote recording of audio annotations, as detailed in section 3.3.6.1.

Figure 3.9: Creation-Tool: Audio annotations with visual feedback (waveform).

3.3.2.4 Text Annotations

Text annotations can be made using a physical keyboard or a virtual keyboard. After pressing the text annotation button, the user has to press in the point of screen where he/she wants to add the note and a text box will appear (Figure 3.10(a)). Since the "enter"

button could work ambiguously for adding a new paragraph to the note or finishing the note, a "Done" button was added to the text box. When the user presses on the "Done" button, the text box disappears and is replaced by the written text (Figure 3.10(b)). Moreover, the Tablet stylus, combined with a handwriting recognizer, can also be used for text annotations, thus replacing the usage of a keyboard.



(a) (b)

Figure 3.10: Creation-Tool: Text annotations. (a) Writing a text annotation. (b) Text annotation added to the video.

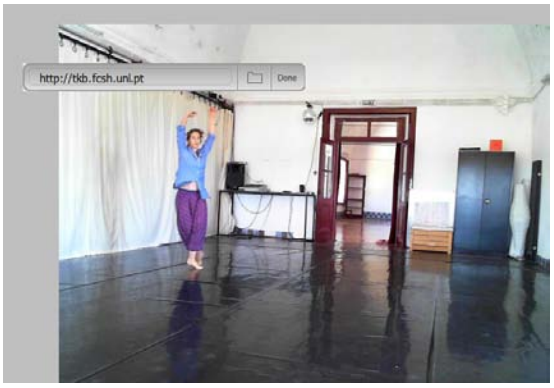
3.3.2.5 Hyperlinks

The system offers two types of hyperlinks: local and external. The local links are composed of other documents owned by the user, such as text, images or other videos. The external links are Websites defined by a URL. The input mechanism of the hyperlinks is similar to the text notes but the text box, is replaced by a URL box and a button that will open a folder dialog, in order to search and select a local file (Figure 3.11(a)). After the user has defined which document or Website he/she wants to add to the video content, the system will display a thumbnail of the hyperlink (Figures 3.11(b) and 3.11(c)).

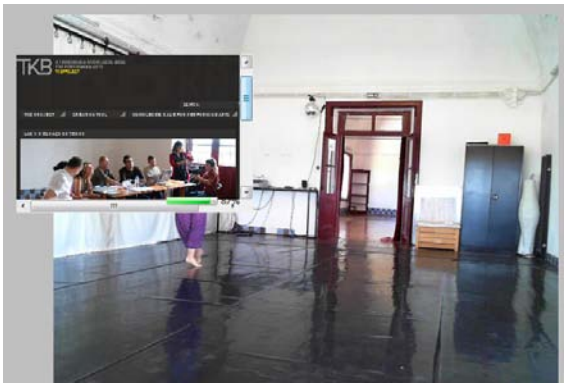
3.3.2.6 Grouping/Ungrouping Annotations

The annotations can be grouped, behaving as one single annotation. This allows the control of different annotations at the same time and is particularly useful when two or more input modalities are used, e.g., a sketched arrow pointing at a hyperlink. The group/ungroup operation can be achieved by selecting each annotation or an area. The original time intervals associated to each annotation are maintained.

[6] "Tornado & Lightning", ©by *Unknown photographer*, uploaded in 2009 by Evonne Heyning, at Flickr.com, used under a Creative Commons Attribution-NonCommercial license: `http://creativecommons.org/licenses/by-nc/2.0/deed.en`



(a)



(b)

(c)

Figure 3.11: Creation-Tool: Annotations as hyperlinks. (a) Adding a hyperlink: website or local file. (b) Website preview. (c) Image preview[6].

3.3.3 Motion Tracking

In order to maintain the association between pen-based video annotations and the annotated objects, object tracking methods are used. In this context, "an object can be defined as anything of interest for further analysis" [YJS06]. Although, the focus of this research were pen-based video annotations, motion tracking methods can be applied to any graphical video annotation, as shown in Goldman's work [Gol07; GGCSS08].

In the annotations associated with moving objects, the time interval of each annotation is calculated in a different form. If an annotation is associated with a particular object, it is more acceptable that its time interval is defined by the period in which this object appears in the scene than with the time that the user takes to make the note. Therefore, it was defined that annotations combined with motion tracking methods will disappear from screen, fifteen seconds after the object is no longer appearing in the scene. These fifteen seconds were experimentally defined as the time interval given to the system to detect the object and continue its tracking. After this period, the system considers that the object will no longer appear in the scene, stops the tracker and sets the time interval of annotations associated with that particular object.

A *naif* approach, would track the video area defined by the annotation. However,

this approach would present a major limitation: annotation would have to be made over its associated (and tracked) object. Marshall's studies [Mar97; Mar98] show that this is not always true, i.e., annotations can be made next to its associated object or even on documents' margins. In this research, it was defined a concept of anchors, specific annotations made by the user or the system, associated with objects of interest and onto which further annotations can be attached. The usage of anchors allows one to define the region or object to be tracked and associate it with other annotations.

Object tracking has been a research topic for a long time and a wide set of solutions was presented by different authors [YJS06]. Nonetheless, object tracking methods that require a long period of video pre-processing could break the interaction fluidity needed in a system that should work as a digital notebook. Therefore, the possibility of real-time performance was an important choice factor for choosing motion tracking algorithms used on the Creation-Tool.

With the goal of tracking objects in the scene, an object tracking library[7] was developed [Sil12]. This framework was first tested and refined by adapting OpenCV algorithms and features. The CAMSHIFT tracker [Bra98], which uses image color segmentation, allowed for a quick way to test the object tracking framework. This tracker has proven rather limited for use in actual dance videos, especially whenever the dancers were being tracked and parts of the background have similar color histograms (Figure 3.12). In order to improve the system, an attempt of background subtraction [ZH06] was tried but it was too sensitive to illumination changes to be useful for this application.

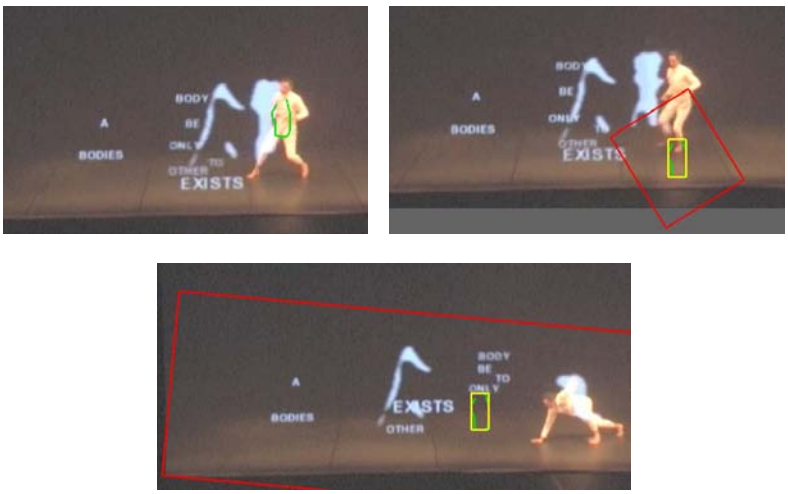


Figure 3.12: Tracking with CAMSHIFT: green ink strokes - annotation; yellow small rectangle - anchor; and the large red rectangle - CAMSHIFT's output[8].

A custom tracker was developed by extracting interest points from each image and matching descriptors between frames. This process was tested with SIFT [Low04] and SURF [BETVG08] point detectors. However, this approach was too slow for real-time

[7] https://github.com/jmfs/libobtrack
[8] Video of *.txt* choreography, authors: Nabais, F., Jürgens, S, Galrito, F., 2009.

tracking. A third OpenCV tracker, also based on point matching between successive frames, was tested. It detected the interest points using the FAST [RD06] corner detector, build BRIEF [CLSF10] descriptors out of those interest points, and removed the features which had irregular movement, compared to the others. Problems still occur, however, when the background's and the performer's movement were similar(Figure 3.13) .

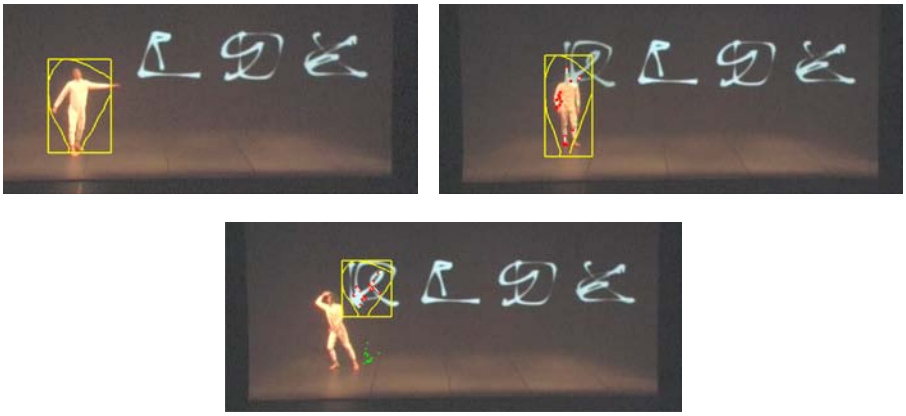


Figure 3.13: Tracking with FAST and BRIEF: ink strokes - annotation; rectangle - anchor; and red dots - interest points[9]

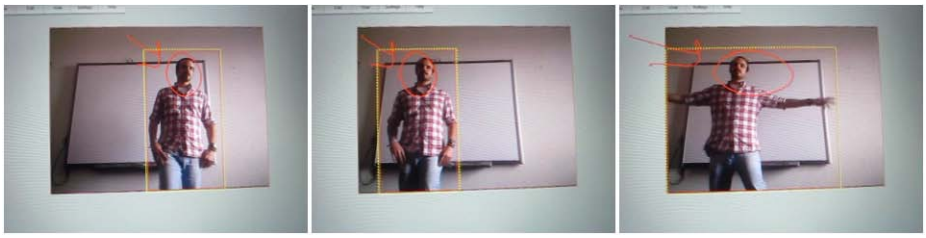
Since, the results of the previous approaches were not accurate enough for the Creation-Tool purposes, other two real-time motion tracking methods were integrated in the developed framework, Microsoft Kinect[10] [SFCSFMKB11] and the Tracking-Learning-Detection (TLD) algorithm [KMM12].

The Kinect [SFCSFMKB11] detects from a single depth image a small set of 3D position candidates for each skeletal joint, presenting a reasonable accuracy and speed for real-time people tracking. It uses a depth camera to capture the depth images, which it then feeds to a previously trained randomized decision forest classifier. This classifier consists of a set of decision trees, each containing split nodes and leaf nodes. Each split node consists of a feature and a threshold. If a given feature evaluates below the threshold, the left branch is followed, otherwise the right one is followed. Leaf nodes then store a learned probability distribution of body parts for a given pixel, based on the features. The final classification for the pixel is reached by averaging the probabilities of all body parts across all trees, and retaining the highest one. Since Kinect recognizes human bodies, the system can automatically draw bounding boxes around them, which work as anchors (Figure 3.14). However, Microsoft Kinect is limited to performing people tracking on live video and was not designed for mobile devices. Due to its size, geometry, power consumption and lack of a battery, attaching the Kinect to a mobile device is no easy task.

TLD [KMM12] is a more generic tracker algorithm, which can be used with any camera and by any type of computer device (including mobile devices). TLD presents a

[9] Video of *.txt* choreography, authors: Nabais, F., Jürgens, S, Galrito, F., 2009.
[10] `http://www.xbox.com/kinect`

(a)



(b)

Figure 3.14: Tracking with Kinect: (a) Kinect Camera. (b) Tracking using Kinect. Ink strokes - annotation; and rectangle - anchor.

learning component, which observes the tracker and detector performances, estimating detector's errors and generating training examples to avoid these errors in the future. In TLD, an initial bounding box made by the user, around the tracking object, is required. This user selection is used as TLD anchors (Figure 3.15).

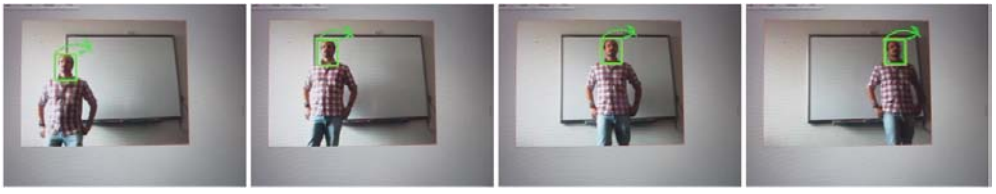


Figure 3.15: Tracking with TLD.
Tracking with TLD: ink strokes - annotation; and rectangle - anchor.

The TLD detector works by generating a large set of bounding boxes (patches) where the object might be and then running a three-stage cascaded classifier on each of the image patches. A patch is rejected if it does not pass any of the classifiers. The tracker computes the optical flow of several points within the bounding box, estimates the reliability of that flow, and uses the most reliable displacements to estimate the object's movement. The estimate of the object's current position (or its absence) is then calculated

by combining the estimates of the tracker and the detector. After this process, the learning component is engaged to estimate and correct errors in the detector and the tracker. It consists of two "experts", P-expert and N-expert, which estimate errors of the other components. The P-expert exploits the temporal structure in the video and assumes that the object moves along a trajectory, estimating false negatives and adding them to the training set as a positive. The N-expert exploits the spatial structure in the video and assumes that the object can appear at a single location only, estimating false positives and adding them to the training set but as a negative. The training set is updated at each frame, feeding the ensemble classifier. This work used the C++ Nebehay's implementation[11] of TLD [Neb12].

In preliminary tests it was observed that these two last motion tracking approaches, Microsoft Kinect and TLD, were suitable for the Creation-Tool and can be used together (Figure 3.16), using the Kinect camera for both methods, or separately, using a regular camera for TLD.

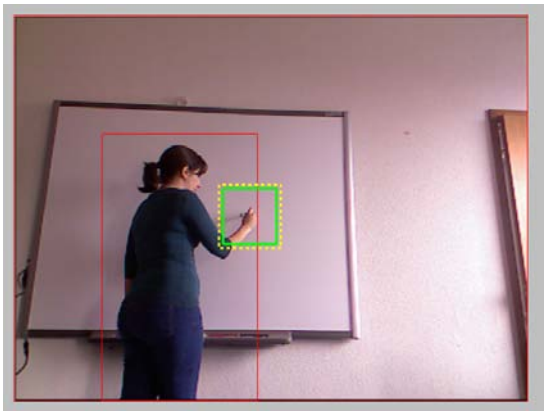


Figure 3.16: Tracking with Kinect and TLD: large red rectangle - Kinect-based anchor; and small green rectangle -TLD-based anchor.

3.3.4 Annotations Storage

The annotations are stored in an XML file [Val11; Sil12]. Each annotation type is stored in a different node. Each node contains the in and out points, i.e., the first frame and the last frame to which the annotation is associated. Additional stored data includes the format attributes (e.g., font size, line thickness and color), for ink and text notes, the recognized text, for audio and ink notes, and the reference to the sound file, for audio notes. The XML structure for ink annotations also contains the ink path, a sequence of 2D points of the drawn lines, and the identification of a parent note, if it is associated with a motion tracking anchor. These anchors are also stored in the XML file as annotations and include the path made by the tracked object through the different frames.

[11] http://gnebehay.github.io/OpenTLD/

3.3.5 Annotation and Video Modes

Annotating a live event creates a set of interaction challenges: should the user's attention be focused on the tool or in the event? And what if the user wishes to annotate a moment which has already occurred? In order to answer these questions, different annotation and video visualization modes and methods were developed.

The system offers two modes for video annotation, continuous and suspended, and two modes for video visualization were added to the system, real-time and delayed. In addition, two annotation methods were also developed: hold and overlay and hold and speed up. These modes and methods are explained next.

3.3.5.1 Continuous vs Suspended

In the continuous mode, the annotations are made as the video is being captured and directly in that same video window. The annotations are saved along the video segments. This mode allows the user to watch the video being captured continuously and annotate it at the same time. In this mode, the annotations gradually disappear, since their time span has already ended (Figure 3.17). The amount of time that an annotation takes to disappear can be defined the user in an interval between one and five seconds. After the capture, the system shows the annotations in the same order as they were made.

Figure 3.17: Continuous mode: annotation gradually disappears.

The suspended mode copies a particular frame to an area where the user has the opportunity to annotate it. The video stream is displayed on the left side of the interface and when the user presses onto this video window, the corresponding frame is copied to the right side, as shown in Figure 3.18. The annotations can be made on top of this particular frame and remain associated to it. The capture itself does not stop or pause while the user is annotating. In the suspended mode, the annotations remain visible, until a mode change occurs or a new frame is picked. In order to visualize the annotations after the capture, the video playback pauses in the annotated frame during a period.

3.3.5.2 Real-Time vs Delayed

Two modes for video visualization were developed: real-time and delayed. The real-time video mode records an event and displays it simultaneously, enabling a straightforward method for video annotation. However, this mode directs the user's attention to the tool

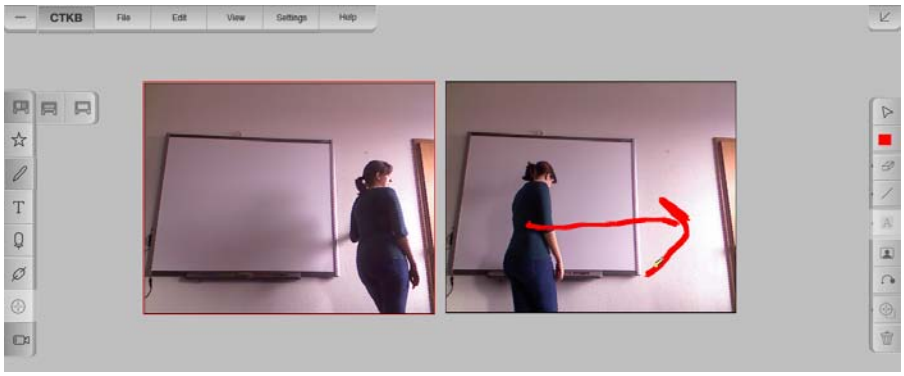


Figure 3.18: Suspended mode: video window on the left side and annotated frame on the right side

and not to the event, working as a "see through" application. In the delayed video mode, the event is displayed with a time delay, thus allowing later annotation. In this mode, the user' attention can be focused on the event and not on the tool. The delay definition time is configurable by the user, allowing a different delay timing for each user. The annotation and video visualization modes can be combined. For example, one can use the continuous mode in the real-time visualization mode or in the delayed mode.

3.3.5.3 Hold and Overlay vs Hold and Speed Up

In order to help the task of selecting and annotating a moving object in a live stream, two annotation methods were developed, Hold and Overlay and Hold Speed Up. Both of them were based in the technique developed by Hajri et al [AHFMI11], called Hold. In the Hold method (also referred as Click-to-Pause), when a user clicks the mouse button down, the moving targets temporarily pause while the user interacts with those targets. When one releases the button, the target starts moving again. Although the user study presents good results, the method cannot be directly applied to the annotation task. The main problem of using the Hold method on the annotation process is that this technique was mainly developed for target selection, whereas the annotation task requires more than just selecting an object. In the annotation process, the user adds information (sketches or handwriting, in the case of pen-based annotations) to a moving object and in a particular context. In addition, annotating a live video stream also presents other challenge: the user does not have full playback control of the video, i.e., the user cannot perform playback operations, like stop, rewind, go fast forward or play frame-by-frame. The event is happening and being recorded, independently of the fact that the user wishes to stop it or rewind it to make a note.

Aiming to adapt the Hold method for the annotation on a live stream, the Hold and Overlay and the Hold and Speed Up methods were added to the Creation Tool. The first one, the Hold and Overlay, is a video adaptation of the graphics Target Ghost selection technique, presented by Hasan et al [HGI11]. It consists in freezing a frame when the

user starts the note and it overlays a translucent live video feed, until the user ends the annotation task (Figure 3.19). This method provides a cue of what's currently on the scene while the user makes a note. The live video's opacity has experimentally been set at 20%.

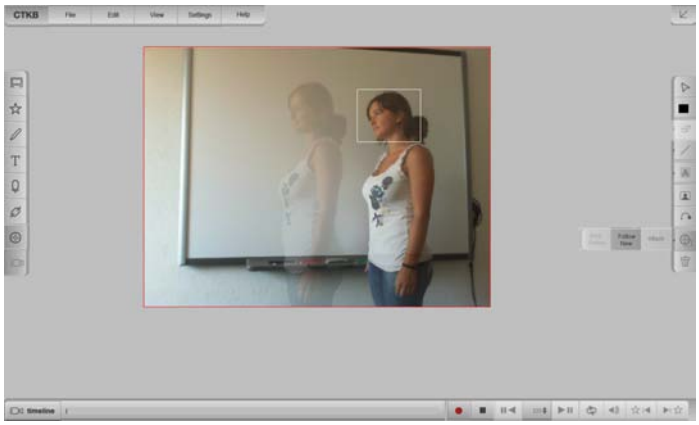


Figure 3.19: Hold and Overlay method.

The Hold and Speed Up is similar to the *flashback* concept for windows management, presented by Bezerianos et al [BDB06]. It also works by freezing the frame when the user starts the note but, once this is done, it speeds up the video, displaying the missed and buffered frames until it returns to the live video feed. However, rather than making the transition occur at a constant speed, the system smooths it by gradually speeding it up at the beginning, continuing at a nearly constant speed and then slowing it down, until the live video is reached. In both methods, the recording of the live video continues, even if it is in the Hold phase.

3.3.6 Interface Design & Development Process

Usability, user-friendly design and the ability to categorize the information into a coherent structure were the main concerns during the interface design and development process. The Creation-Tool was designed following an iterative process, in which two choreographers with different working habits and needs were involved from the beginning. Additional input was also obtained from dancers and dance technology experts, in residence-lab workshops and in two preliminary tests made during the development process. The first preliminary test was focused on annotation modalities and modes of annotation and video visualization, whereas the second was focused on the two motion tracking methods, using Microsoft Kinect and TLD, and the two annotation methods, Hold and Overlay and Hold Speed Up.

3.3.6.1 Creation-Tool Main Interface

The video annotator interface (Figure 3.20) is composed of a video display area, presenting a live or pre-recorded stream, in which it is possible to augment the content with

Figure 3.20: The Creation-Tool main interface.

annotations. It is also possible to annotate in the area around the video window, avoiding the occlusion of video elements by the user notes. All annotations are shown in a timeline, as well as the corresponding video frames. The menus and toolbars are organized in five sections: on the upper left side, there is the system menu and, on the left side, the tools bar. On the upper right side, the project menu manages all the content and files, like videos, annotations or icons' marks, and on the right side are the toolsets (related with toolbars, e.g., line color, line width, eraser). In order to retrieve a wider view of the annotations, a navigator menu was added to the project menu. This navigator shows all annotations associated to the video content and allows its navigation, as shown in Figure 3.21. The video navigation can be made by pressing one annotation, causing a change in the video time position.

At the bottom, the timeline is used for video and notes navigation and allows to change annotations temporal positions and length (Figure 3.22). The timeline is divided in different tracks, one per each annotation modality, and it can be dragged down and partially hidden, avoiding the occlusion of tool buttons, video content and annotations. Sizes and spacing between buttons are related with the touch screen specifications in order to have a better performance [Iph].

In order to allow more freedom of movements and gestures, essential on dance performance environments, some application features can be controlled remotely with a mobile device via Open Sound Control (OSC)[12] [Val11]. The device (e.g., an iOS or Android based device) uses the TouchOSC[13] application to send messages to the system, via Wi-Fi

[12] http://opensoundcontrol.org/
[13] http://hexler.net/software/touchosc

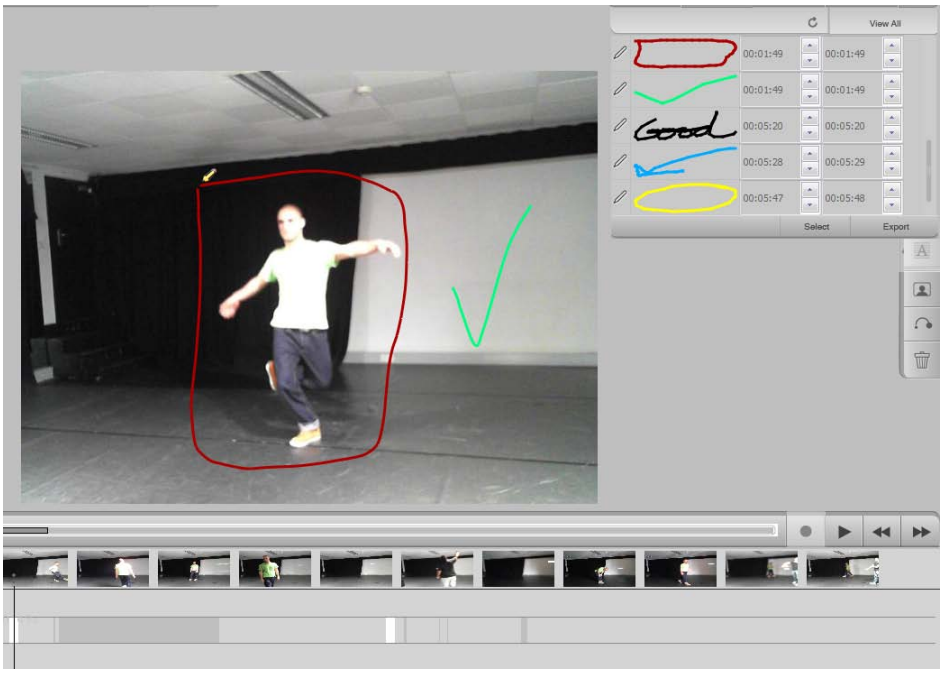


Figure 3.21: Creation-Tool navigator: time intervals of the different pen-based video annotations.

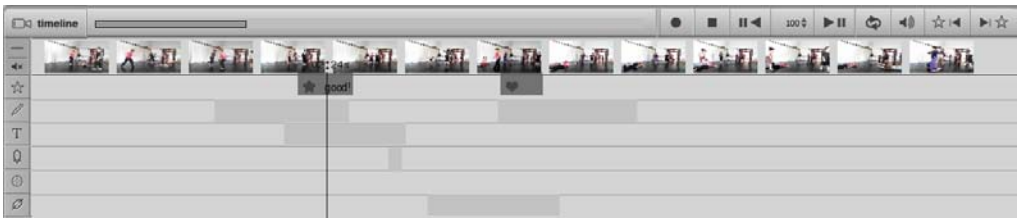


Figure 3.22: Creation-Tool timeline.

(UDP/IP Protocol), using a graphical interface: one switch button for starting or finishing a voice annotation and other switch button for starting or stopping a video capture (Figure 3.23).

3.3.6.2 Preliminary Test - I

This preliminary test was done with a choreographer, simulating a dance rehearsal. For this test, regular and HD webcams, connected to a Windows Tablet PC, were used to capture the scene to be annotated. In addition, a Bluetooth headset was used for audio annotations and an iPod Touch as a remote control. (Figure 3.24)

The main interface was well perceived but some interaction problems were observed. The major interaction difficulties were related with the ability to use touch and pen simultaneously in the Tablet PC. The pen usage blocks the touch mode, which makes a true pen+touch interaction difficult to achieve. It was also found that it was important to improve the recording buttons, as well as the timeline's hiding mechanism, so their actions could be performed with more fluid gestures (Figure 3.25). The annotation marks

(a) (b)

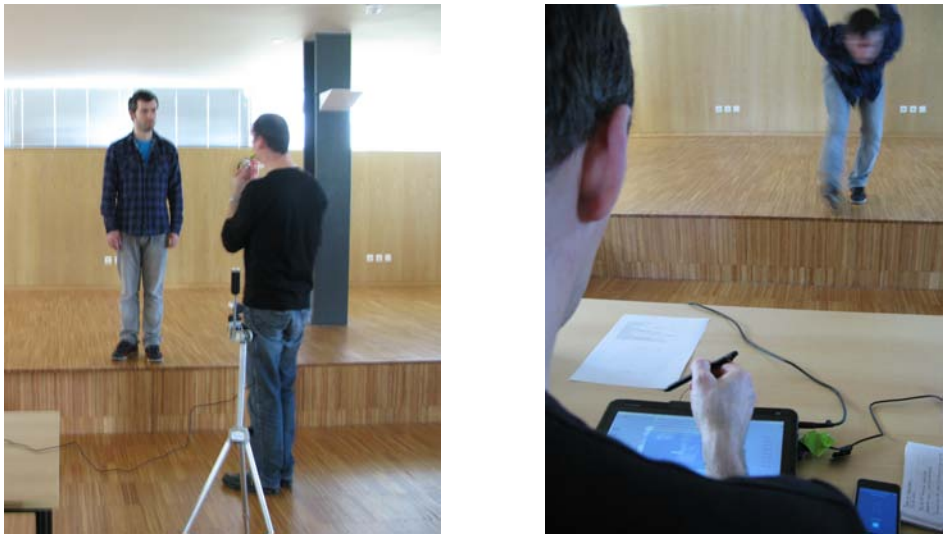
Figure 3.23: Creation-Tool Hardware. (a) Tablet PC, iPod Touch and Bluetooth headset. (b) TouchOSC interface.

interaction had to be improved, since pressing the used mark to define its endpoint distracted the choreographer. The input for text and audio annotations did not present major difficulties.

3.3.6.3 Preliminary Test - II

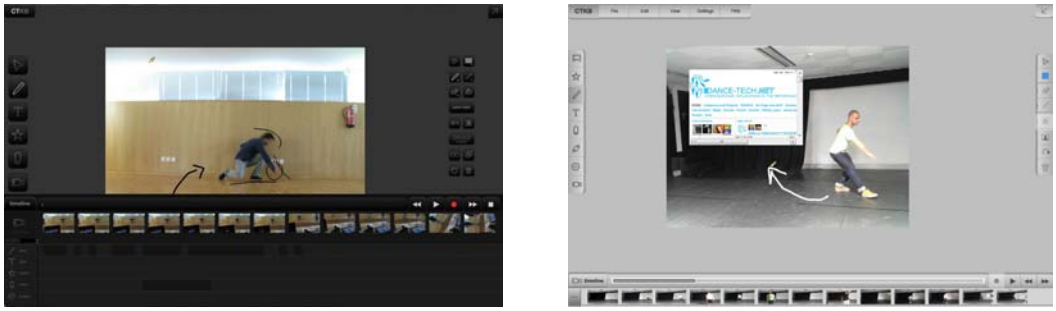
This preliminary test occurred during a workshop with choreographers and dancers and additional outdoor experiments were made (Figure 3.26). In this preliminary test, a Windows Tablet equipped with Microsoft Kinect and a HD webcam, was used. The informal discussions revealed that the trackers have worked satisfactorily for users' purposes. During the outdoor tests it was observed that the Kinect cannot be reliably used outdoors, even if a power outlet is available. A possible cause of this is that the Kinect's infrared camera cannot find the projected infrared dots in the presence of strong ambient light. TLD's outdoor use was satisfactory, even if somewhat sensitive to sudden illumination changes. The algorithm's learning component can, however, deal with gradual changes and still continue following the object.

Although the main conclusion was that the motion tracking methods, Kinect and TLD, were sufficient accurate to be used in the Creation-Tool, two changes were needed in the application. The first one was the need of annotation methods for live video stream, already described in section 3.3.5.3. The second was the need of improving the original mechanism of associating a regular annotation with an anchor. The initial interface used explicit anchor selection and note attachment, which required too many steps during the user interaction. When using Kinect, the anchors are always displayed, as long as a person is detected by the system. Therefore, a previous anchor selection, during the association process, is always needed. Nonetheless, with TLD, the anchor is made by the user. In this case the anchor is automatically selected as soon as it is created, avoiding



(a) (b)

Figure 3.24: Choreographer testing the tool. (a) Using a bluetooth microphone for audio annotations. (b) Testing the Creation-Tool using the tablet's pen.



(a) (b)

Figure 3.25: Creation-Tool Design Process. (a) First interface (b) New interface

the users' need to make an additional step, in order to select the anchor. The "attach" button used for associating a regular annotation and an anchor was also removed (Figure 3.27). New annotations are automatically attached to the currently selected anchor. These two improvements have reduced the number of steps needed to select an anchor and to associate it to a set of annotations.

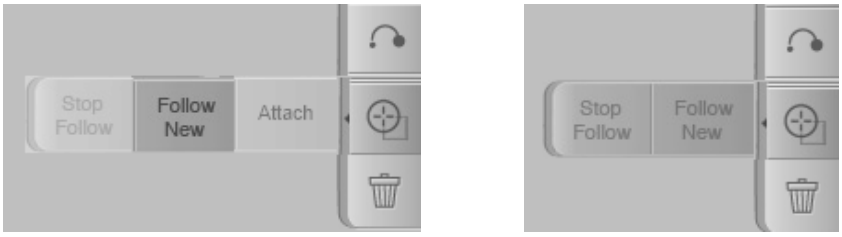
3.3.7 Evaluation

In order to evaluate and have users' feedback about the different tool features two usability tests were carried out. They follow the same generic structure of the preliminary tests: the first was focused on annotation modalities and modes of annotation and video visualization, whereas the second one was focused on the two motion tracking methods, using Microsoft Kinect and TLD, and the two annotation methods, Hold and Overlay and Hold Speed Up. In both usability studies, the users answered a questionnaire with numerical

(a) (b)

(c)

Figure 3.26: Motion tracking: indoor and outdoor tests. (a) Dance performers testing indoor tracking. (b) Tracking a person. (c) Tracking a car.



(a) (b)

Figure 3.27: Motion tracking interface improvements. (a) First interface with the "attach" button. (b) New interface.

semantical differential scales. Although this type of scale can be considered an approximation of an interval scale [Gd82; SA04; TA08], allowing the application of parametric tests that are more powerful, this approach it is not consensual [Kna90; WYWH99; Jam04; CP07; Nor10]. Thus, both parametric (paired-samples t-tests and one-way ANOVA) and non-parametric (Wilcoxon Matched-Pairs Signed-Ranks and Friedman tests) tests were made, in order to confirm possible significant differences between the different answers. Due the small number of participants in both studies, 12 in the first and 9 in the second, the results of them should be considered as user tendencies. However, the number of participants in each study respects the 5-participants rule, which claims that with five participants most of the usability issues (around 80%) are observed [TA08].

3.3.7.1 Usability Tests - I

The first usability test was conducted, with 12 international dance performers from a contemporary dance "residence-workshop", aiming the evaluation of the prototype and the different modalities. The participants were divided into 3 separate groups, 4 participants per group. While one participant was doing the usability test, the other three were improvising a performance for the test, as shown in Figure 3.28. A HD webcam, the Microsoft LiveCam Studio, connected to a HP TM2-2150 Tablet PC were used in this usability test.

The usability test was composed of 10-15 minutes of tool experimentation, assisted by a member of the development team, followed by a questionnaire. Two main scenarios were sequentially tested by each user: during a live event, with a camera recording the other users; and after the event, using the video stream previously recorded. The questionnaire had three major parts: participants' information, participants' working habits and tool evaluation.

Participants and Working Habits The subjects were primarily females (83.33%), the mean of ages was $\bar{x} = 25$ ($\sigma = 2.56$) and they were all right handed. Most of the subjects had a Bachelors degree (58.33%), a quarter (25.00%) had studied until high school, 8.33% held a Master's degree and the remaining 8.33% had attended elementary school education. Most of the subjects did not have previous experience with pen-based technology

Figure 3.28: User experimenting the Creation-Tool during the test.

(83.33%) and did not have experience with touch technology either (91.67%).

All participants were used to recording their work and usually make use of video recordings. Almost all of them use regular paper notebooks (91.67%) and a quarter (25.00%) uses audio recordings (Figure 3.29).

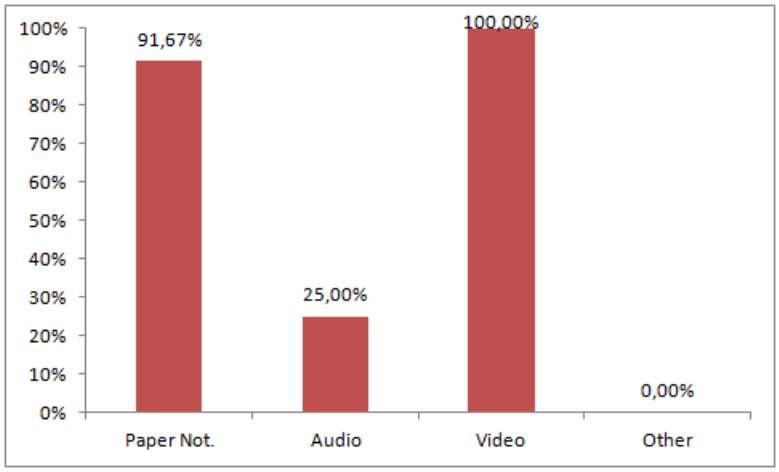


Figure 3.29: Work Recording

Almost all, 91.67%, of the subjects annotate their work in some way. From that percentage, almost all of them use regular paper notebooks (90.91%), more than half use a laptop (54.55%) and only 9.09% use mobile phones (Figure 3.30).

Most of the users share their working documents (75.00%), mostly by e-mail (88.89%) or by sharing hardware (77.78%), such as pen drives, CDs or DVDs. Some of them use Web sites (33.33%), instant messaging (33.33%), post mail (11,11%) or other media (22.22%), such as sharing printed books or network file sharing (Figure 3.31).

Questionnaire The tool evaluation was composed of eight questions with semantic differential numerical scale answers, five (Q1, Q2, Q3, Q5, Q7) about mode usage rate (1 for

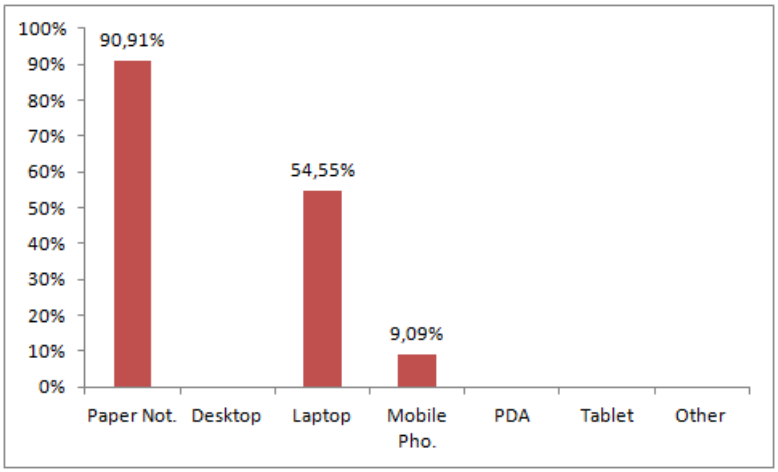


Figure 3.30: Work Annotation

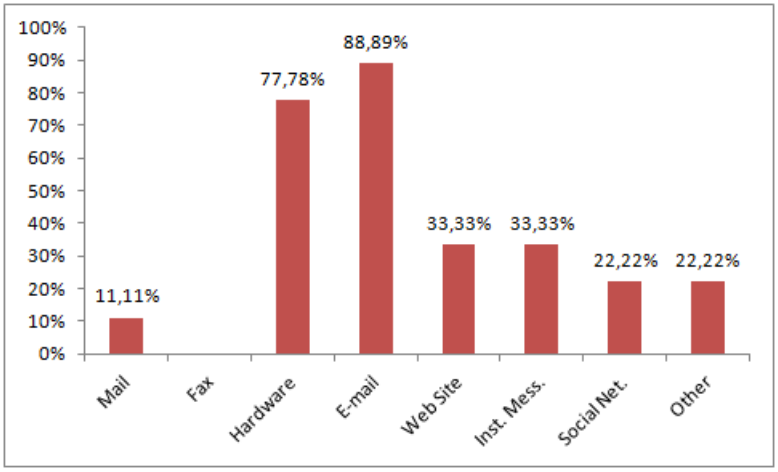
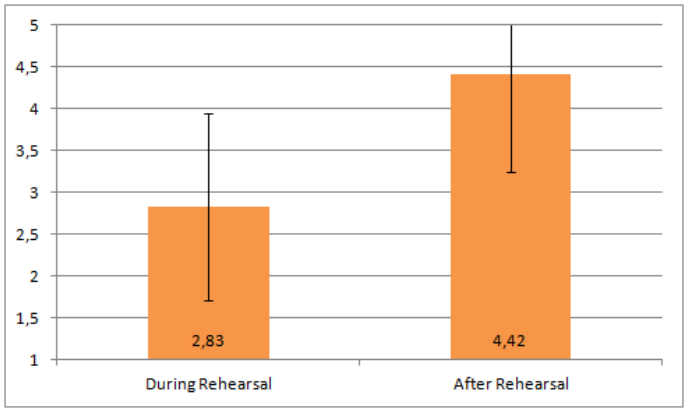


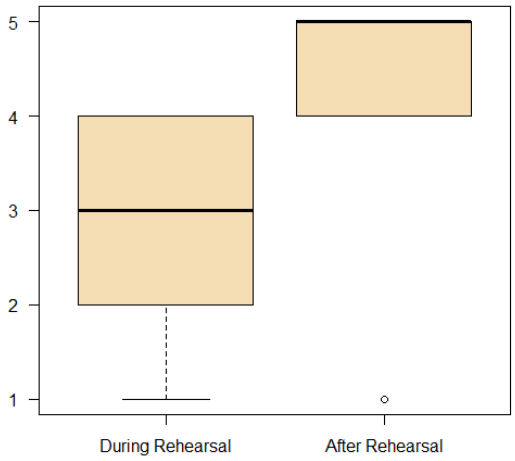
Figure 3.31: Work Sharing

Rarely - 5 for Frequently) and three (Q4, Q6, Q8) about perceived difficulty (1 for Difficult - 5 for Easy); one (Q9) based on Microsoft "Product Reaction Cards" classification [BM02] and one (Q10) open question for comments and suggestions (see Appendix A). In order to compare the mode preferences and the perceived difficulty, paired-samples t-tests and Wilcoxon Matched-Pairs Signed-Ranks tests (Q1, Q5, Q6 and Q7) as well as one-way ANOVA and Friedman tests (Q2, Q3 and Q4), were conducted based on the null hypothesis (H0), i.e, there was not a significant difference between answers. In all tests the alpha level was 0.05, in order to achieve an interval of confidence of 95%.

In Q1, the participants were asked to rate the tool usage in two different scenarios: during a rehearsal ($\bar{x} = 2.83, \sigma = 1.11, \tilde{x} = 3.00$) and after a rehearsal ($\bar{x} = 4.42, \sigma = 1.16, \tilde{x} = 5.00$) (Figure 3.32). The t-test presented a significant difference ($t(11) = -3.98, p < 0.05$, Cohen's $d = 1.15$) between the two scenarios, showing that the participants prefer to use the tool after a rehearsal. The Wilcoxon Matched-Pairs Signed-Ranks test also showed a significant difference ($W = 7, Z = -2.54, p < 0.05, r = 0.52$).



(a)

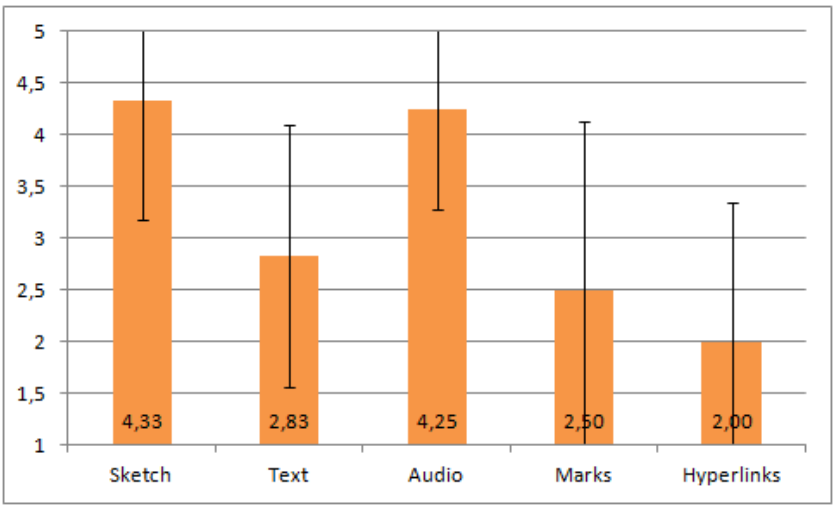


(b)

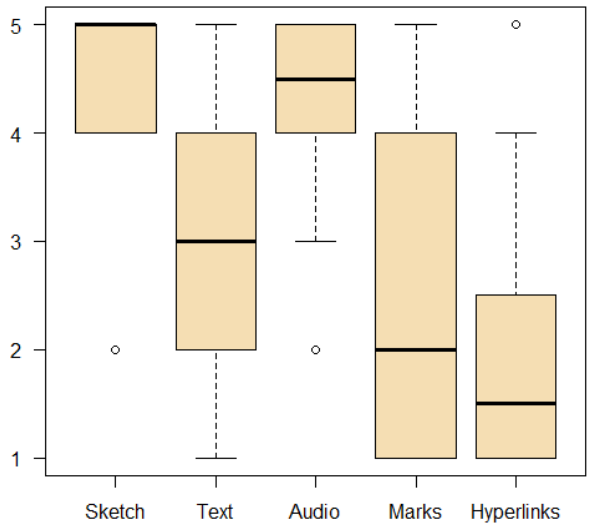
Figure 3.32: Results the two different usage scenarios: during a rehearsal and after a rehearsal. (a) Mean scores. Error bars represent the standard deviation.(b) Median scores

Afterward they were asked to rate the usage of the different annotation types, considering the two scenarios: during a rehearsal (Q2) and after a rehearsal (Q3). In Q2 (Figure 3.33), the ANOVA test showed a significant difference ($F_{4,55} = 8.02, p < 0.0001$) between the different annotation types during rehearsal scenario. Post hoc comparisons, using the Tukey HSD test, indicated that the mean score for the sketch ($\bar{x} = 4.33, \sigma = 1.15, \tilde{x} = 5.00$) was significantly different from text ($\bar{x} = 2.83, \sigma = 1.27, \tilde{x} = 3.00, p < 0.05$), marks ($\bar{x} = 2.50, \sigma = 1.62, \tilde{x} = 2.00, p < 0.01$) and hyperlinks ($\bar{x} = 2.00, \sigma = 1.35, \tilde{x} = 1.50, p < 0.001$) but not from audio ($\bar{x} = 4.25, \sigma = 1.27, \tilde{x} = 4.50, p > 0.05$). However, the mean score for audio was only significantly different from marks ($p < 0.05$) and hyperlinks ($p < 0.001$). The other means did not present any significant differences. The Friedman test also showed a significant difference between the different modalities ($\chi^2 = 20.39, df = 4, p < 0.05$). Wilcoxon Matched-Pairs Signed-Ranks tests presented the same significant differences: sketch from text ($W = 10, Z = -2.13, p < 0.05, r = 0.43$), marks ($W = 7.5, Z = -2.23, p < 0.05, r = 0.47$) and hyperlinks ($W = 3, Z = -2.73, p <$

$0.05, r = 0.56$); and audio from marks ($W = 9.5, Z = -2.34, p < 0.05, r = 0.48$) and ($W = 7.5, Z = -2.51, p < 0.05, r = 0.51$). These results show a usage preference for sketching, during a rehearsal, since sketch presents more significant differences when compared with the other annotation modalities.



(a)



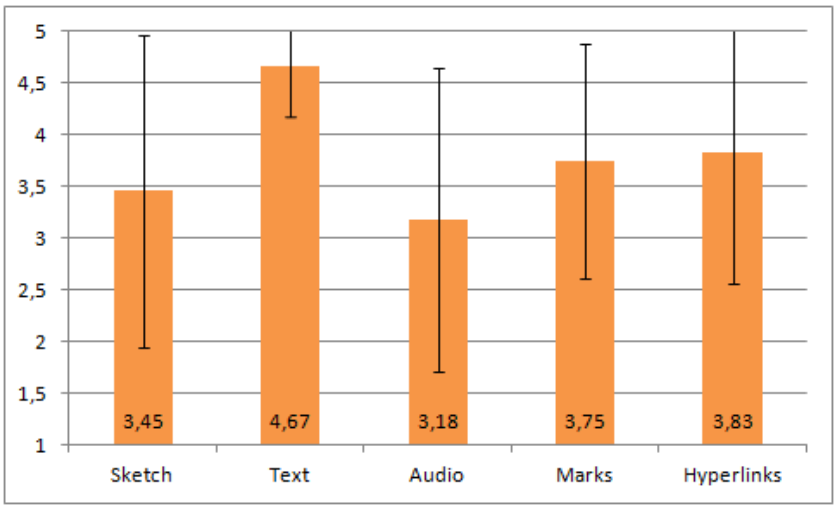
(b)

Figure 3.33: Results for the usage of the different annotation types, during a rehearsal. (a) Mean scores. Error bars represent the standard deviation. (b) Median scores.

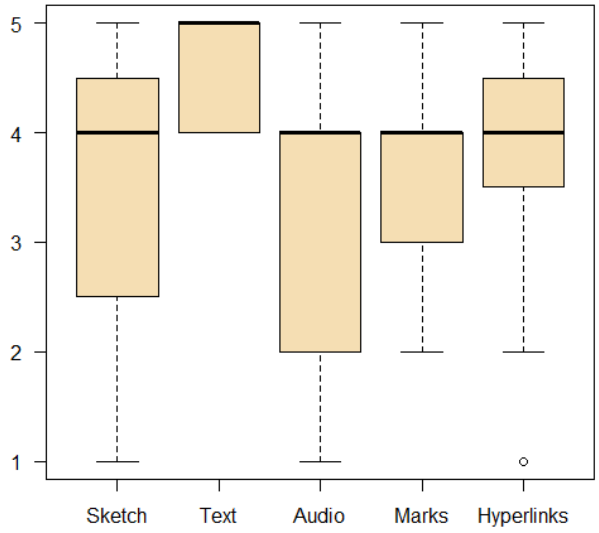
In Q3 (Figure 3.34), the ANOVA test did not present a significant difference ($F_{4,53} = 2.45, p > 0.05$)[14] between the different annotation types in an after-rehearsal scenario. On the other hand, the Friedman test showed a significant difference between the different

[14]Due to missing data, we considered only 11 participants for Sketch and Audio and 12 for other annotation types. A second ANOVA was conducted, but not considering the answers of the participant that caused the missing data. The conclusion of this test was the same: there was not a significant difference ($F_{4,50} = 2.19, p > 0.05$).

modalities ($\chi^2 = 10.92, df = 4, p < 0.05$)[15]. Nonetheless, the Wilcoxon Matched-Pairs Signed-Ranks tests only showed a significant different between text ($\tilde{x} = 5.00$) and the other modalities: sketch ($\tilde{x} = 4.00, W = 0, Z = -2.58, p < 0.05, r = 0.55$), marks ($\tilde{x} = 4.00, W = 3, Z = -2.19, p < 0.05, r = 0.47$), hyperlinks ($\tilde{x} = 4.00, W = 3, Z = -2.19, p < 0.05, r = 0.47$) and audio ($\tilde{x} = 4.00, W = 0, Z = -2.57, p < 0.05, r = 0.55$).



(a)



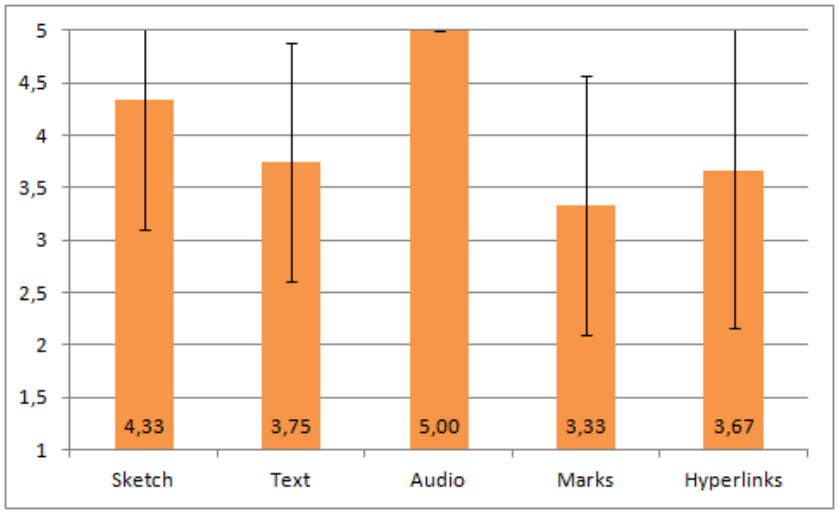
(b)

Figure 3.34: Results for the usage of the different annotation type after a rehearsal. (a) Mean scores. Error bars represent the standard deviation. (b) Median scores.

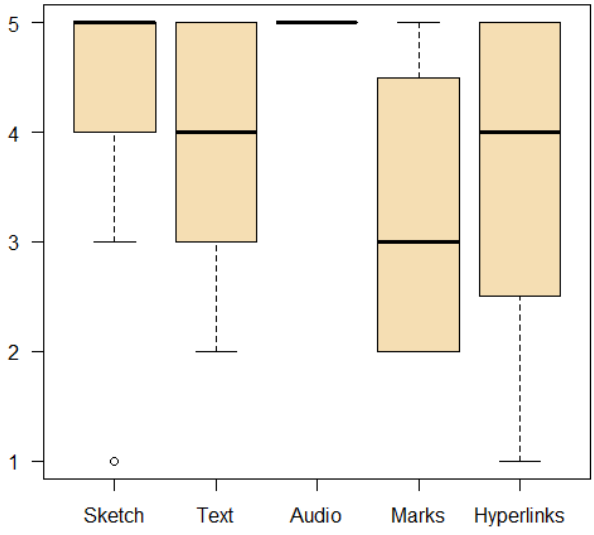
The perceived difficulty of the different annotation types was studied in Q4 (Figure 3.35) and an initial ANOVA test presented a significant difference ($F_{4,55} = 3.46, p < 0.01$). However, the audio annotations were rated, by all participants, with the maximum of 5 points, resulting in $\sigma = 0$ and $Var(X) = 0$. A second ANOVA test was conducted

[15]For the Friedman test and the Wilcoxon Matched-Pairs Signed-Ranks tests the participant that caused the missing data was not considered.

without the audio annotations. This second test showed that there was not a significant difference ($F_{3,44} = 1.27, p > 0.05$) between sketch, text, marks and hyperlinks. The Friedman test presented a significant difference between the different modalities ($\chi^2 = 14.55, df = 4, p < 0.05$). The Wilcoxon Matched-Pairs Signed-Ranks tests confirmed a significant difference audio ($\tilde{x} = 5,$) when compared with text ($\tilde{x} = 4.00, W = 0, Z = -2.74, p < 0.05, r = 0.56$), marks ($\tilde{x} = 3.00, W = 0, Z = -2.87, p < 0.05, r = 0.59$) and hyperlinks ($\tilde{x} = 4.00, W = 0, Z = -2.42, p < 0.05, r = 0.49$). However, the Wilcoxon test did not show a significant difference between audio and sketches ($\tilde{x} = 5.00, W = 0, p > 0.05$), neither between the other annotation modalities.



(a)

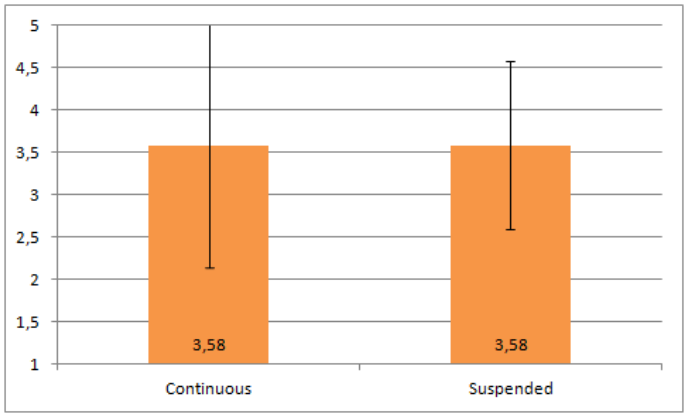


(b)

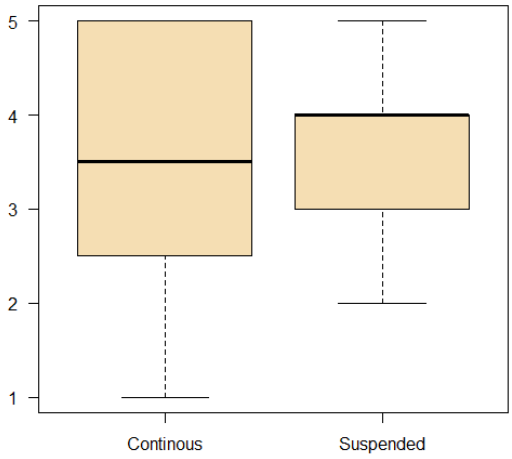
Figure 3.35: Results for the perceived difficulty of the different annotation type during a rehearsal. (a) Mean scores. Error bars represent the standard deviation. (b) Median scores.

In Q5, participants were asked to rate the usage of two annotation modes: continuous

($\bar{x} = 3.58, \sigma = 1.44, \tilde{x} = 3.50$) and suspended ($\bar{x} = 3.58, \sigma = 1.00, \tilde{x} = 4.00$) (Figure 3.36). The t-test did not present a significant difference, ($t(11) = 0, p > 0.05$) between the means of the two annotation modes. A Wilcoxon Matched-Pairs Signed-Ranks test did not also showed a significant difference ($W = 32.5, Z = 0, p > 0.05$).



(a)

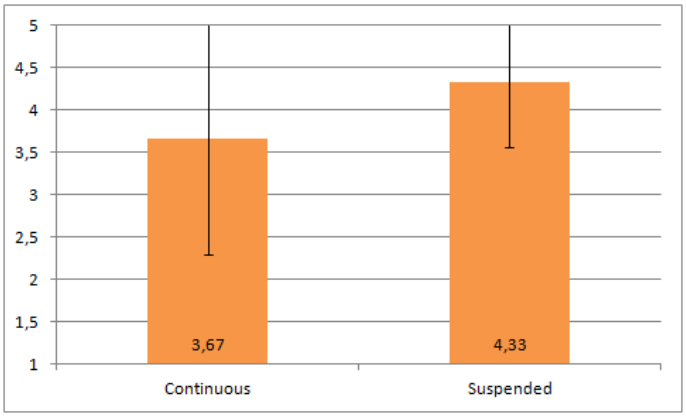


(b)

Figure 3.36: Results for the usage of the different annotation modes: continuous and suspended. (a) Mean scores. Error bars represent the standard deviation. (b) Median scores.

The perceived difficulty of the continuous ($\bar{x} = 3.67, \sigma = 1.37, \tilde{x} = 3.50$) and suspended ($\bar{x} = 4.33, \sigma = 0.78, \tilde{x} = 4.50$) annotation modes was studied in Q6 (Figure 3.37) and the t-test did not also present a significant difference ($t(11) = -1.61, p > 0.05$) between the means of the two annotation modes. A Wilcoxon Matched-Pairs Signed-Ranks test presented the same result ($W = 5, Z = -1,36, p > 0.05$). The answers from Q5 and Q6 show that there is not a preference between the continuous and the suspended mode.

In Q7, the participants were asked to rate the usage of two video visualization modes: real-time ($\bar{x} = 2.83, \sigma = 1.45, \tilde{x} = 3.00$) and delayed ($\bar{x} = 4.5, \sigma = 0.90, \tilde{x} = 5.00$), during a rehearsal (Figure 3.38). The t-test presented a significant difference ($t(11) = -3.46, p <$



(a)

(b)

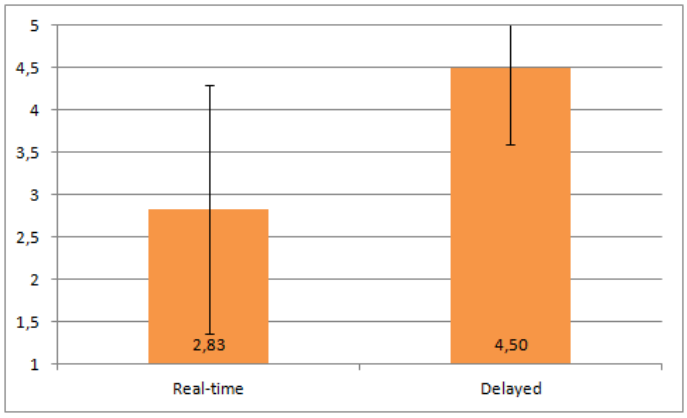
Figure 3.37: Results for the perceived difficulty the different annotation modes: continuous and suspended. (a) Mean scores. Error bars represent the standard deviation. (b) Median scores.

0.05, Cohen's $d = 0.998$) between the means of the two video modes, showing the preference for the delayed mode. A Wilcoxon Matched-Pairs Signed-Ranks test confirmed this result ($W = 2.5, Z = -2.62, p < 0.05, r = 0.53$).

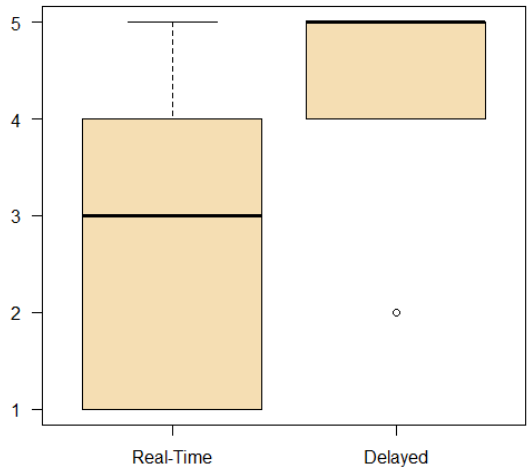
The perceived difficulty of the tool interaction was rated in Q8, with a mean of $\bar{x} = 4.08$ ($\sigma = 0.90$). In Q9, they were asked to classify the tool with 48 words of the Microsoft "Product Reaction Cards" [BM02]. Figure 3.39 presents the percentage for each word. The most selected (>=50%) words were: "attractive", "useful", "clear", "helpful", "time-saving", "innovative" and "organized". In Q10, we have asked for open comments and suggestions. There were a few open comments, reporting a positive feedback on the tool.

3.3.7.2 Usability Tests - II

The second usability test was focused on the two motion tracking methods, using Microsoft Kinect and TLD, and the two annotation methods, Hold and Overlay and Hold



(a)



(b)

Figure 3.38: Results for the usage of the different video visualization modes, real-time and delayed. (a) Mean scores. Error bars represent the standard deviation. (b) Median scores.

and Speed Up. People tracking is the only common and comparable feature between TLD and Kinect and it is a crucial issue on dance performances, thus, it was the focus of this test. The test was made on a Windows Tablet PC, a Lenovo X220, connected to the Microsoft Kinect, obtaining a live video stream. For this test, a person was moving in front of the Kinect camera and ready to make any particular movement asked by the subjects. The Tablet stylus was used as input interface in the usability test, exploiting pen-based video annotations.

The users were able to experiment the tool a few minutes before the test, in order to be familiarized with the interface. The features related with motion tracking, the focus of this study, were not experimented, avoiding a previous learning of these features. After this initial tool experimentation the users have been asked to perform a set of tasks related to motion tracking features. During the test a member of the development team was available in the case the users needed assistance.

In order to start the test, the users were asked to select the anchor, generated with

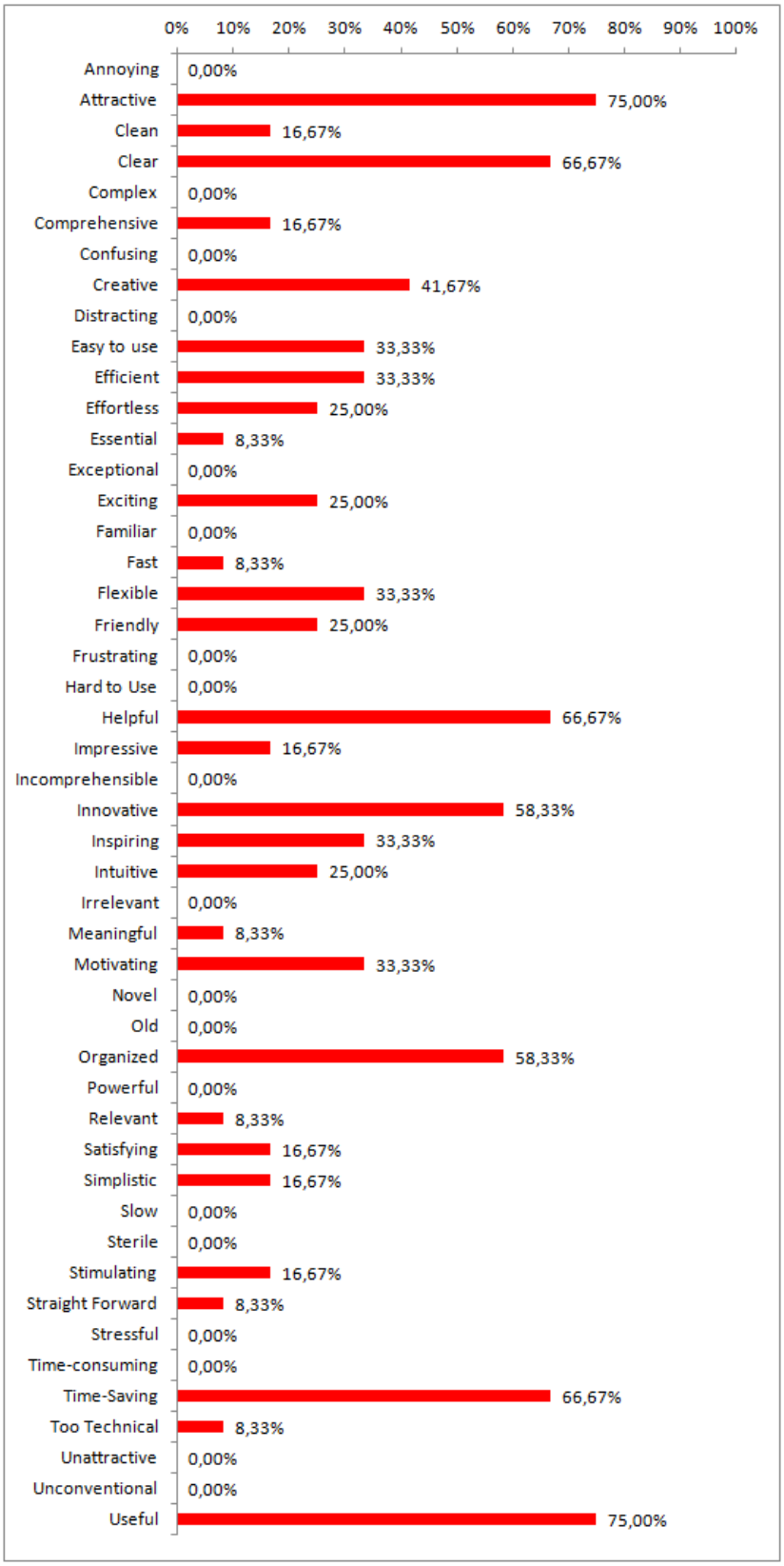


Figure 3.39: Classification with Microsoft "Product Reaction Cards".

Kinect and associated with the moving person, and to make an annotation composed by a circle around the tracked person (Figure 3.40). Then they were asked to create an anchor, in order to track the person using the TLD algorithm, and finally to make a second circle associated with this second anchor. In this initial task, none of the annotation methods have been used. Since the Kinect does not require an anchor creation by the user, no learning bias was introduced. After this, the same steps were repeated, firstly using the Hold and Overlay annotation method and then the Hold and Speed Up method.

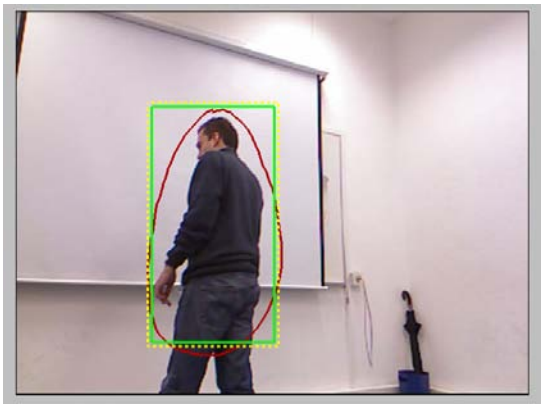


Figure 3.40: User task: select (Kinect) and make (TLD) the anchor around the person and make an annotation - the circle

Participants and Working Habits The subjects were mostly male (66.67%), the mean of ages was $\bar{x} = 31.67$ ($\sigma = 6.40$). Most of the subjects had a Master degree (66.67%), 22.22% had a Bachelor degree and 11.11% held a PhD. Most of the subjects had previous experience with pen-based technology (88.89%) and usually annotate their work (88.89%). From those who annotate their work, 87.50% use a paper notebook, 75.00% use a mobile phone and 50.00% use a laptop (Figure 3.41).

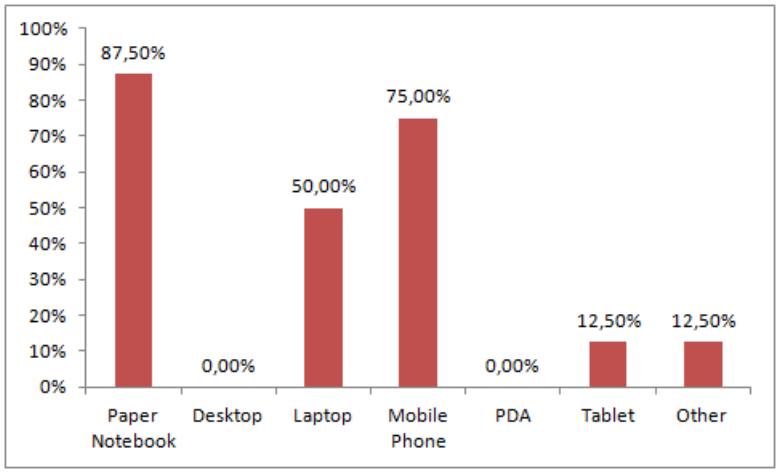


Figure 3.41: Work Annotation

Questionnaire The questionnaire was composed of five questions with semantic differential numerical scale answers, one (Q1) trackers performance (1 for Poor - 5 for Excellent), one (Q2) about mode usage rate (1 for Rarely - 5 for Frequently) and three (Q4, Q5, Q6) about perceived difficulty (1 for Difficult - 5 for Easy); one (Q7) based on Microsoft "Product Reaction Cards" classification [BM02] and one (Q8) open question for comments and suggestions (see Appendix B). The first three questions (Q1, Q2 and Q3) were directly related with the tasks and were answered by the users immediately after each task. An informal discussion with each user took place after the test. In order to compare the mode preferences and the perceived difficulty paired-samples t-tests and Wilcoxon Matched-Pairs Signed-Ranks tests were conducted based on the null hypothesis (H0), i.e., there was not a significant difference between answers. In all tests the alpha level was set to 0.05, aiming to achieve an interval of confidence of 95%.

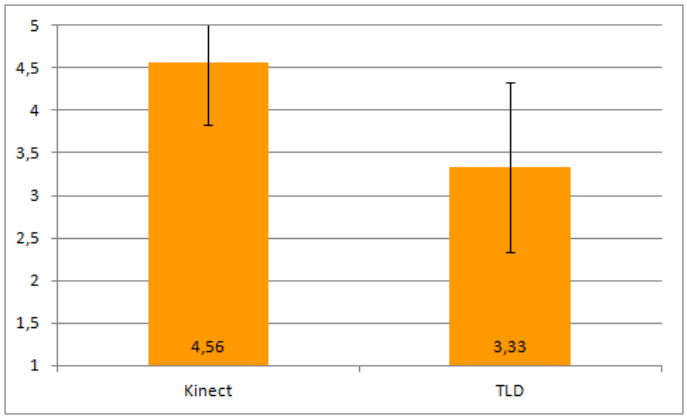
In Q1, the participants were asked to rate the performance of the two trackers regarding people tracking, i.e., how successfully the trackers performed on tracking the person in front of the camera. The Kinect performance was rated with $\bar{x} = 4.56, \sigma = 0.73, \tilde{x} = 5.00$ and the TLD was rated with $\bar{x} = 3.33, \sigma = 1.00, \tilde{x} = 3.00$ (3.42) (Figure 3.42). The t-test presented a significant difference ($t(8) = 3.05, p < 0.05$, Cohen's $d = 1.02$) between the performance of the two trackers and showing a better performance of the Kinect, in the users' opinion. A Wilcoxon Matched-Pairs Signed-Ranks test confirmed this result ($W = 2.5, Z = -2.22, p < 0.05, r = 0.52$).

Afterward they have been asked to rate the usage of the two annotation methods: Hold and Overlay ($\bar{x} = 4.22, \sigma = 0.83, \tilde{x} = 4.00$) and Hold and Speed Up ($\bar{x} = 3.78, \sigma = 0.83, \tilde{x} = 4.00$), as shown in Figure 3.43 . The t-test did not present a significant difference ($t(8) = 0.94, p > 0.05$) between the two annotation methods. A Wilcoxon Matched-Pairs Signed-Ranks test confirmed this result ($W = 15, Z = -0.92, p > 0.05$).

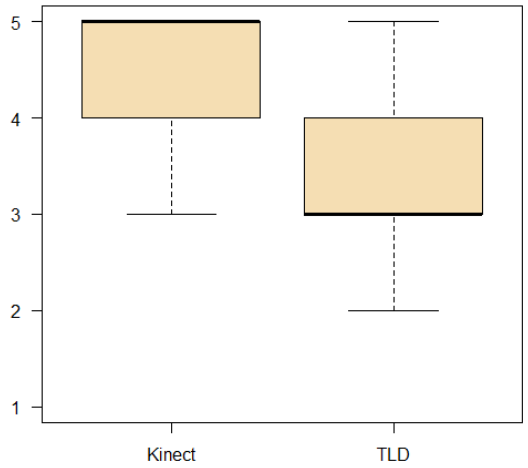
In Q3 (Figure 3.44), the users were asked to rate perceived difficulty of the annotation task, using each tracker and considering each annotation method (3.44). Considering the Hold and Overlay, the Kinect ($\bar{x} = 4.56, \sigma = 0.53, \tilde{x} = 5.00$) and TLD ($\bar{x} = 3.89, \sigma = 0.60, \tilde{x} = 4.00$), the t-test showed a significant difference ($t(8) = 4, p < 0.05$, Cohen's $d = 1.33$) between the two trackers. A Wilcoxon Matched-Pairs Signed-Ranks test confirmed this result ($W = 0, Z = -2.45, p < 0.05, r = 0.58$).

Considering the Hold and Speed Up annotation method, the Kinect ($\bar{x} = 4.33, \sigma = 0.87, \tilde{x} = 5.00$) and TLD ($\bar{x} = 3.89, \sigma = 0.78, \tilde{x} = 4.00$), the t-test also showed a significant difference ($t(8) = 2.53, p < 0.05$ Cohen's $d = 0.84$) between the two tracker but a Wilcoxon Matched-Pairs Signed-Ranks test did not confirmed this result ($W = 0, Z = -2, p > 0.05$). This contradiction should be further studied with a larger sample of users. However, the significant difference shown in these tests can be considered natural, since TLD requires the manual creation of an anchor (a rectangle bounding box) on the tracked objects, whereas when using the Kinect its selection is sufficient, since the anchors are automatically created by the system.

On the other hand, considering the Kinect, the Hold and Overlay ($\bar{x} = 4.56, \sigma =$



(a)



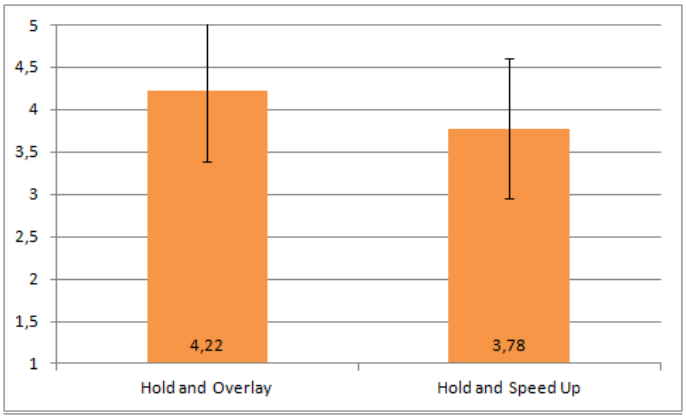
(b)

Figure 3.42: Results for trackers' performance: (a) Mean scores. Error bars represent the standard deviation. (b) Median scores.

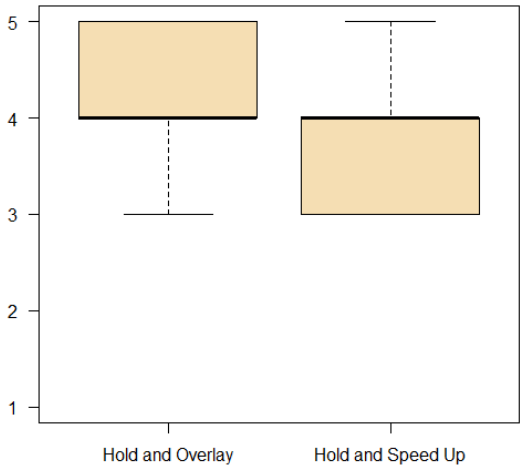
$0.53, \tilde{x} = 5.00$) and Hold and Speed Up ($\bar{x} = 4.33, \sigma = 0.87, \tilde{x} = 5.00$) methods, the t-test did not present a significant difference ($t(8) = 0.69, p > 0.05$). A Wilcoxon Matched-Pairs Signed-Ranks test confirmed this result ($W = 5, Z = -0.57, p > 0.05$).

Considering TLD, Hold and Overlay ($\bar{x} = 3.89, \sigma = 0.60, \tilde{x} = 4.00$) and Hold and Speed Up ($\bar{x} = 3.89, \sigma = 0.78, \tilde{x} = 4.00$), the t-test did not present a significant difference either ($t(8) = 0, p > 0.05$) between the two annotation methods. A Wilcoxon Matched-Pairs Signed-Ranks test also confirmed this result ($W = 10.5, Z = 0, p > 0.05$). These last tests show that the two annotation methods did not influence the task of creating anchors.

The perceived difficulty of attaching annotations with a tracked object was rated in Q4, with $\bar{x} = 4.33$ ($\sigma = 0.50$) $\tilde{x} = 4.00$). In Q5, the perceived difficulty of the tool interaction was rated with $\bar{x} = 4.33$ ($\sigma = 0.87$) $\tilde{x} = 5.00$).



(a)



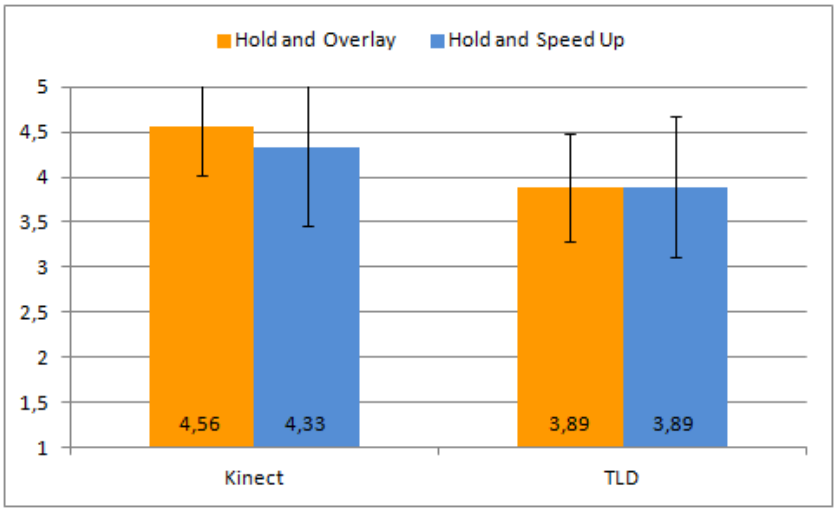
(b)

Figure 3.43: Results for annotation method preference: (a) Mean scores. Error bars represent the standard deviation. (b) Median scores.

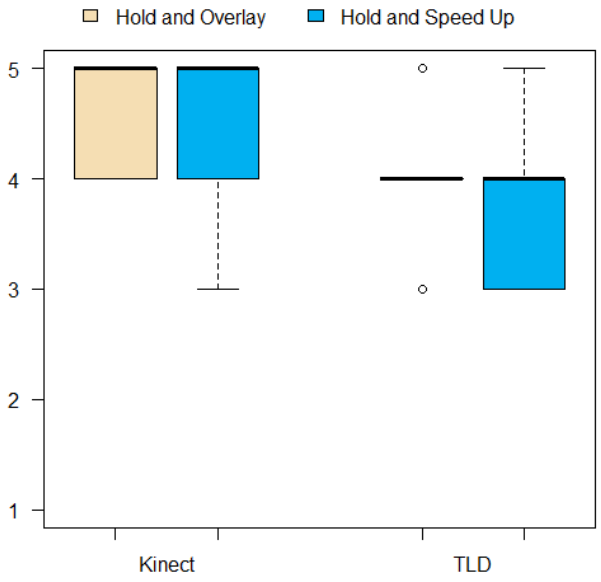
In Q6, the users were asked to classify the tool using 36 words of the Microsoft "Product Reaction Cards" [BM02]. The set of words was properly adapted for non-expert subjects. Figure 3.45 presents the percentage for each word. The most selected (>=50%) words were "attractive" and "useful". The open comments, answered in Q7, were mainly related with two topics: the need of having more control features in the Hold and Overlay method and the difficulty of understanding some of the icons on the interface. The informal discussion taking place after the questionnaire was related to the same topics of the open comments.

3.4 Discussion

Regarding the evaluation, in general, there are two major results: 1) the interface was well perceived by the testing participants and 2) the participants have recognized the positive contribution a tool like this can have on creative processes.



(a)



(b)

Figure 3.44: Results for the perceived difficulty for using each tracker and considering each annotation method.: (a) Mean scores. Error bars represent the standard deviation. (b) Median scores.

From the first usability test, it is possible to conclude that there was a preference for the tool usage on a post-rehearsal scenario. However, from iterative design process with the choreographers results that this preference truly depends on the task to perform. Dance performers, during a rehearsal, need to concentrate on their gestures and movements, thus preferring to use the tool after a rehearsal. In contrast, choreographers need to be focused on the different details of the performance. Therefore, they can use it during or after rehearsal. Regarding the annotation types, there was a preference for sketching during a rehearsal and for text after it. However, if it is also considered the high

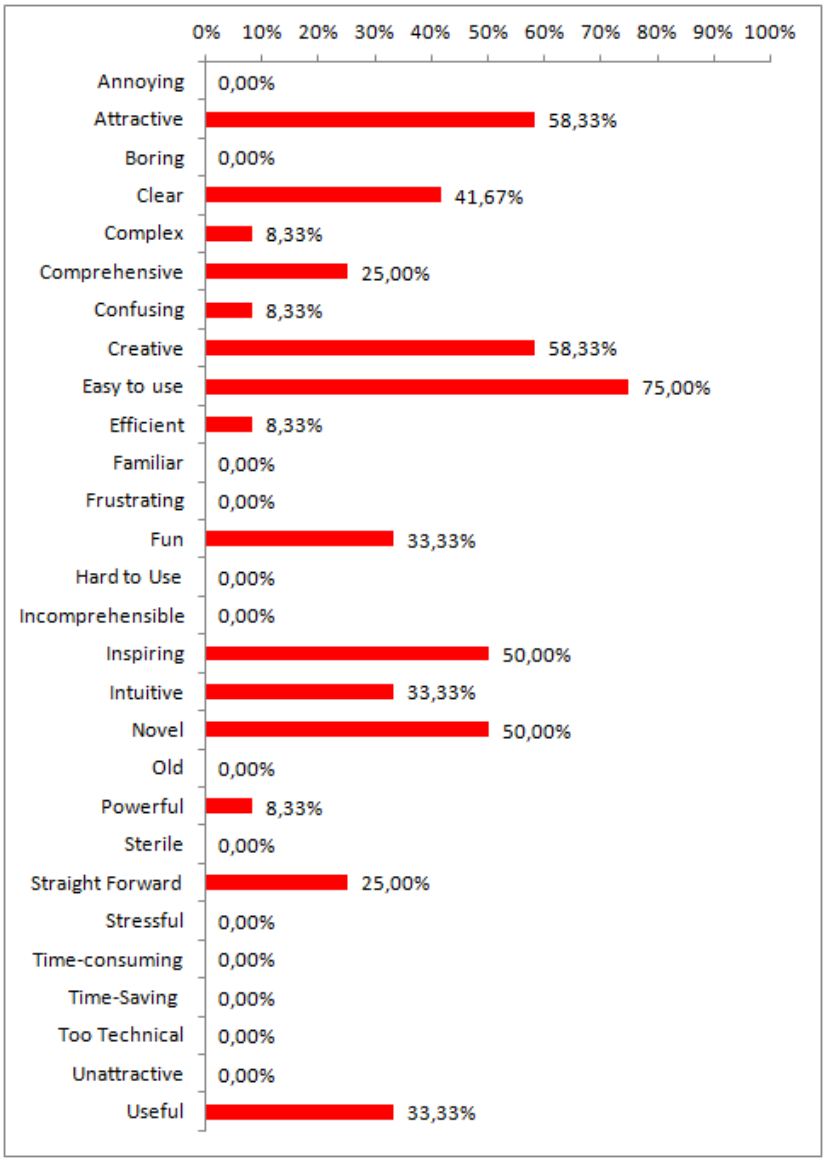


Figure 3.45: Classification with Microsoft "Product Reaction Cards".

percentages of work documentation using video and work annotation using paper notebooks (on both studies), it is possible to conclude that sketching is an important modality for video annotation. Initially, the study was not conclusive regarding the perceived difficulty of the different annotation types, but the non-parametric tests showed that the easiest annotation type was the audio. This choice can be considered natural, since audio only required an on/off button. However, it is important to refer that there was not a significant difference of perceived difficulty between audio and sketches. The study did not show a clear preference between the different annotation modes, but there was a preference for the delayed video visualization mode, when compared with the real-time mode. This result shows the importance of a mode which directs the user's attention to the real event rather than to the application.

Although the users have classified the Kinect with a better performance when compared with TLD and regarding people tracking, the TLD algorithm allows the tracking of any object [KMM12], works with any camera and can be used in outdoor environments. The tests did not show a preference for an annotation method, Hold and Overlay or Hold and Speed Up, but both methods were rated above the median value of the scale, which represents an improvement for annotating moving objects. It is possible to say that it is better to have one (or both) of the annotation methods than to have none. This idea was confirmed during the informal discussions with the users, in the end of each test. In addition, there is no difference between the two methods for the task of creating an anchor.

The results about perceived difficulty of attaching annotations and tool interaction, as well as the results of the tool classification, show that the improvements made on the interface have been quite well received by the users, and once again that a tool such as this one can make a positive contribution for both professional and daily life activities. From the open comments and the informal discussion, two main topics were discussed. The first compares the two annotation methods. Although Hold and Overlay provides the context of the live event, when compared with the Hold and Speed Up, it can also be more confusing, particularly in the cases where the event presents a large quantity of moving features. The usage of each mode will depend on the type of the recorded event. A third approach, combining the advantages of both developed methods, i.e., displaying live events but without increasing the visual noise of the main video window, has also been suggested by the users. The second topic was related to the difficulty of understanding some of the "icons" functions. There is space for improvement in this issue, especially due to the lack of a standard iconography for annotating moving objects.

4

VideoInk

The concept of video as ink, its implementation in a proof-of-concept prototype and how it can be applied to video editing tasks using pen-based technology are discussed in this chapter. A usability study of the prototype is also described in this chapter.

4.1 Video as Ink: The Concept

One advantage of using digital pens, when compared with regular pens, is that they can be used to perform different tasks and digital ink can be replaced by other forms. This idea can be found in the research work developed by Ryokay et al [RMI04], in which the pen (embedded in a physical brush) is used to paint different types of media in a digital canvas, and by Hinkley et al [HYPCRWBB10], where the pen takes the form of an x-acto that cuts digital images.

Considering this principle, that digital ink is not limited to imitate physical ink, the concept of *videoink* explores a painting metaphor where ink is composed of video content. In the same way that a painter places a brush in an ink bucket or a palette and paints in a canvas with the selected ink, one can select a video clip and use a pen to paint on a screen the video frames that belong to that clip. Therefore, instead of imitating regular ink, the trail left by the pen is replaced by video content. This change cannot only give the idea that the user is directly manipulating the video content, using a familiar interaction, but can also reduce the number of widgets usually necessary for video manipulation and editing. An implementation of this concept is described next.

4.2 Video as Ink: Proof-of-Concept Prototype

A proof-of-concept prototype of the *videoink* concept was implemented as part of this research. The prototype was developed for Tablets and exploited pen gestures and pressure for video editing.

In the implementation, the timeline is represented in two dimensions, instead of the more usual 1D horizontal timeline. The two dimensional timeline allows one to the paint the video content horizontally, vertically or diagonally. The direct use of pen coordinates to add a frame in the canvas will cause successive frame occlusions, causing interesting but not very useful visual effects. In order to avoid these occlusions, the canvas is mapped to a 2D matrix, which is invisible for the user, but where the frames have a pre-reserved space. Therefore, when the user drags the pen in screen, the video frames are placed in the correspondent place of the matrix (Figure 4.1).

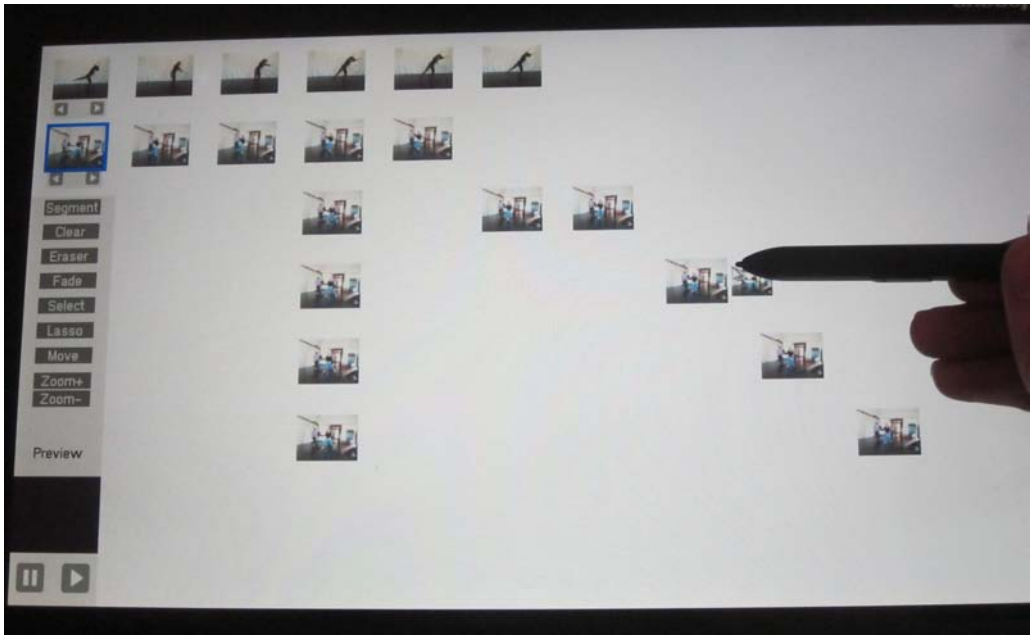


Figure 4.1: Inking with video frames.

The prototype interface is composed of a menu bar, on the left, and canvas area, on the right. Figure 4.1, shows two video clips, used as examples, on the top left corner, which work as video buckets, and a trail of painted frames is displayed on the right. The two buttons below each clip allow to move the current frame of the clip to the next or previous one. Different editing features can be chosen in the menu below the clips and, at the bottom of the menu bar, a video window shows a preview of the new video stream. Two buttons, in the corners of the right side of the canvas, will move up or down content displayed in the canvas.

The pen's cursor shows a thumbnail of the next video frame to be "inked". A final video stream can be composed by all the video content displayed in the canvas, from top left corner to bottom right, or by selecting a particular set of frames or segments from

the canvas. Since selection can be done horizontally, vertically or diagonally, this second method can be used for non-linear video editing.

4.2.1 Implementation Technologies

The prototype was implemented in C++, using openFrameworksv7[1], OpenCV2.3.1[2] and bbTablet[3] [4] platforms. The openFrameworks toolkit is the main platform for graphics and video display, whereas OpenCV is used for video and image processing, particularly for transitions effects. bbTablet is used for pen data access, e.g., pen pressure values, from Wacom tablets (digitizer tablets or Tablet PCs).

4.2.2 The Canvas

The prototype includes a canvas where the video content can be painted, selected, moved or erased. This canvas works as a timeline, like in other regular video editing software, but with the difference that video content can be displayed horizontally, vertically or diagonally. The canvas was defined as horizontally limited, i.e., there is a maximum of frames that can be painted in each row of the canvas, but vertically unlimited (Figure 4.2). A bidirectionally limited canvas will reduce the working space, whereas an unlimited canvas in both directions could be too confusing. The implemented approach breaks the traditional horizontal timeline, providing a better visual organization and allowing to explore different alternatives more easier. Thus, it was decided that rows should be limited, forcing the user to change to the row below.

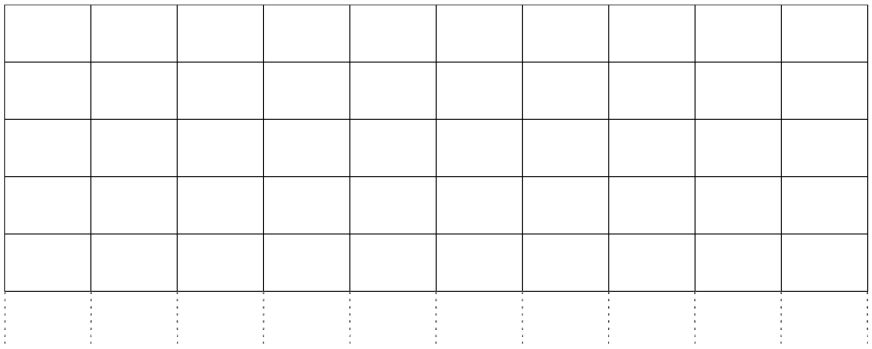


Figure 4.2: The matrix that composes the canvas.

[1] http://www.openframeworks.cc/
[2] http://opencv.willowgarage.com/wiki/
[3] http://www.billbaxter.com/projects/bbtablet/
[4] MSCV2010 version: http://img.di.fct.unl.pt/~diogocabral/bbTablet_MSVC2010.zip

4.2.3 Painting Video: Video Frames vs Video Segments

In the proof-of-concept prototype two basic modes were defined: "Frame" and "Segment". In the "Frame" mode, the user paints a single frame in the canvas, while in the "Segment" mode it is possible to paint a video segment, i.e., a set of consecutive frames, using one single gesture. In the "Frame" mode, all painted frames are displayed (Figure 4.3), whereas in the "Segment" mode, only the start and end frames of a segment are shown (Figure 4.4). Each segment is represented, horizontally in the canvas, by its start (on the left) and end (on the right) frames connected by a gray box. Even if the frames were painted one by one, the system considers that they compose a video segment if they are horizontally consecutive, i.e., with no holes between them. In this situation the start and end frames are automatically defined by the end points of the frame set. An isolated frame is considered a special segment represented by a single frame, which is simultaneously its start and end frames. Transitions frames placed between two different clips were considered separated segments.

The visual change of the canvas is synchronized with transformation on the menu. Each mode is represented by a visual change on the selected clip: in the "Frame" mode the clip only shows the current frame whereas in the "Segment" mode part of the last frame it is displayed behind the current frame (Figures 4.3 and 4.4).

Figure 4.3: Frame mode.

The selection of each mode follows the principle that when a brush is sinked in an ink bucked, more ink will be attached to it and more of it will paint in the canvas. This idea was implemented using the pressure made by the pen tip against the screen. A pressure threshold was defined and if it is passed, the "Segment" mode is triggered. In order to return to the "Frame" mode, is it sufficient to tap on top of one clip with a pressure below this threshold. The threshold was experimentally defined at 99% of the maximum level of pressure represented by the pen. Nonetheless, a second (and more traditional) mechanism, based on a switch button (Figures 4.3 and 4.4), was implemented with the goal of comparing both techniques.

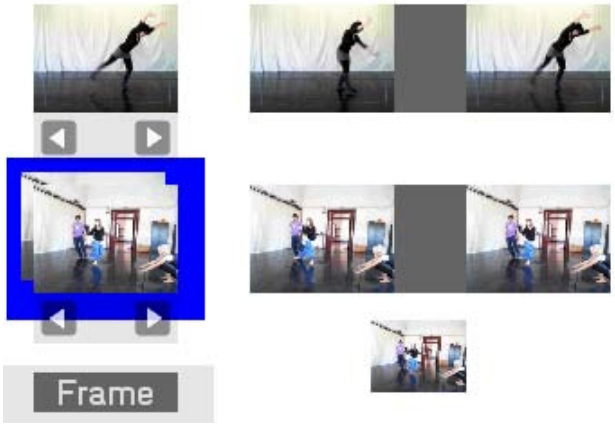


Figure 4.4: Segment mode.

After selecting the mode, the user has to tap or drag in the canvas, in order to paint the video content of the selected video clip (Figure 4.5). In the case of the "Frame" mode a single frame is painted each time the user passes with the pen tip over a rectangle of the matrix. After a frame is painted, the clip moves to the next one, following the ink metaphor, i.e., the ink attached to a brush or inside a pen moves into a physical surface, when it is in contact with this surface. If the user selects the "Segment" mode, a video segment, defined by the current (start frame) and last frames (end frame) of the selected clip, is painted in the canvas.

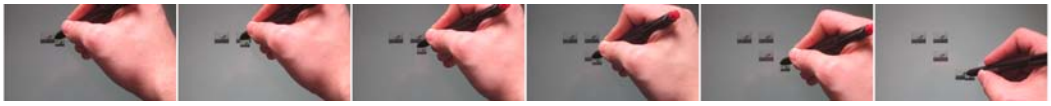


Figure 4.5: Painting frames by dragging the pen.

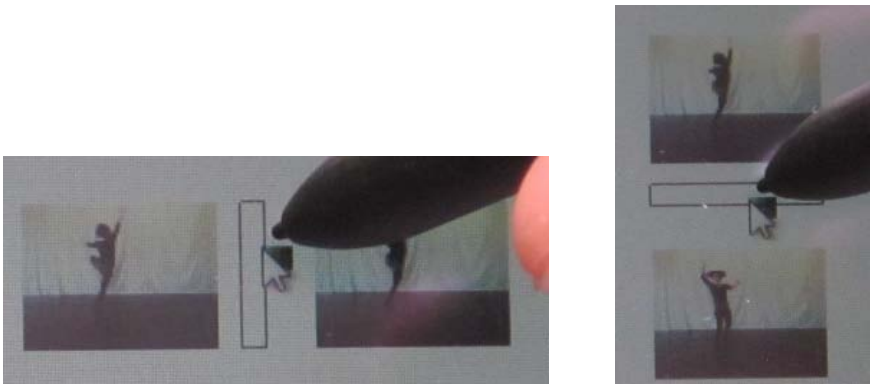
In the "Segment" mode, the implementation presents two major limitations: 1) the video segments are only represented horizontally, left to right, in the canvas and 2) the dragging gesture is not used for painting segments, i.e., it is sufficient to tap with the pen tip somewhere in the canvas, in order to add a video segment. The vertical representation of video segments can perfectly fit in the concept of video as ink. However, it would introduce an additional level of complexity that was too early to implement, without having the users' feedback about simpler situations, e.g., painting or selecting frame-by-frame in any direction or sequence. Regarding the dragging gesture, since the video segment is defined by the current and last frames of each clip, there is no video content left to be painted. Nonetheless, combining these two limitations it is possible to observe that the dragging gesture could be used to indicate the direction of the video segment to be painted.

4.2.4 Video Editing Features

In this implementation, the video editing main features are divided in two main categories: operation features and transitions effects. Operation features such as adding, moving or erasing content follow the "ink principle". The selection of each operation in the menu changes the behavior of the pen in the canvas. Adding and erasing content work as previously explained but with opposite functions, i.e., by selecting the eraser mode, pen gestures will erase the elements displayed in the canvas. If the user inks in a place where a frame already exists, the older frame will be removed and replaced by the new one. Moving frames can be achieved by dragging each frame from its original position to the new place, as shown in Figure 4.6. The implementation was limited to move single frames for the same issues presented above, i.e., it would introduce an additional level of complexity without a previous feedback from the users.

Figure 4.6: Move a frame by dragging

When the user hovers the pen over the space between two painted frames, an empty box is displayed (Figure 4.7), indicating that the user can add or move other frame between those two. In this situation, the frames on the right or below are shifted and the new frame is added to the canvas.



(a) (b)

Figure 4.7: Hit area to add content (hovering the pen). (a) On a row. (b) On a column.

Regarding the transitions effects, the same input method can be used to add several of them. Therefore, in this implementation only the fade effect was developed, as an example of such features. The fade effect can be selected from the menu and made by tapping the pen in the middle of the two frames used for the transition effect. A new transition frame, resulting from the composition of the adjacent frames (50 % of each

one), can be added using the same process of adding a new frame between two already painted. The frames on the right or below are shifted and the new frame is added to the canvas. The fade effect can be completed by adding successive blend frames, as shown in Figure 4.8. As already mentioned, other transition effects could be developed using the same input method.

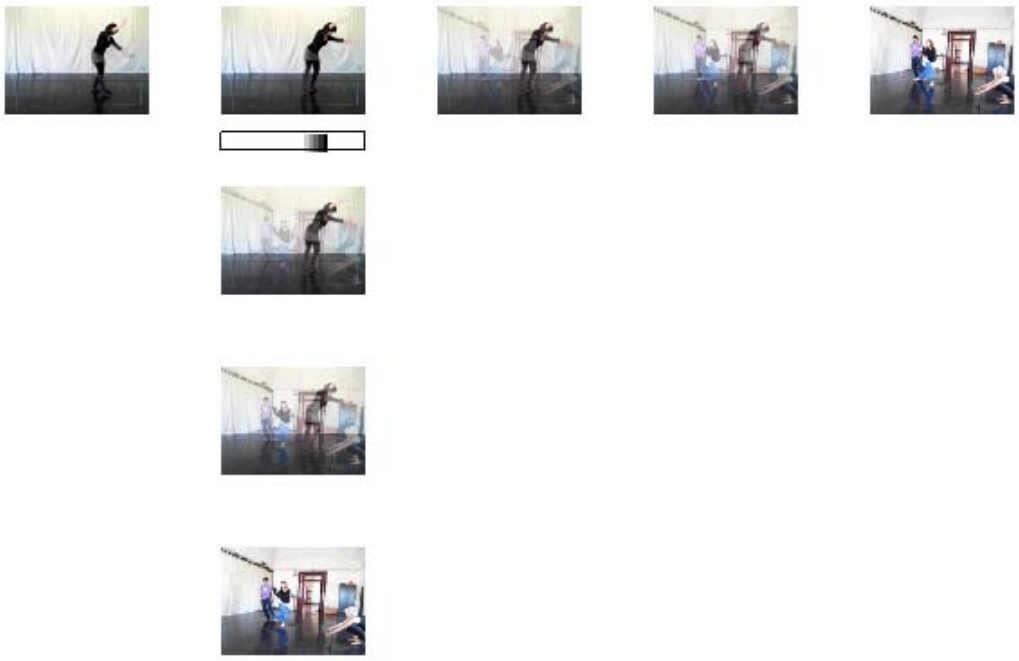


Figure 4.8: Transition: fade effect (horizontal and vertical).

4.2.5 Selecting Elements

Two selection methods were developed: paint selection and lasso selection. The paint selection also follows the "inking principle". In this mode, the selection is made by pressing or dragging the pen on screen (Figures 4.9 and 4.11(a)). A frame or segment is selected when the pen passes on top of it. Following the same principle of adding or removing content, paint selection can be made horizontally, vertically or diagonally.

In the lasso selection, the user has to draw a lasso around the frames or segments that should be selected (Figures 4.10 and 4.11(b)). The selection is made by traversing the frames or segments displayed in the canvas from left to right, top to bottom. The point that represents the center of mass of each frame, in the case of the "Frame" mode, or the point in middle of start and end frames of a segment, in the case of the "Segment" mode, is verified if it is inside or outside the lasso. This verification is achieved by using a ray-casting algorithm, i.e., counting how many times a ray, starting from that point and going to any fixed direction, intersects the edges of the polygon (the lasso). If the number of intersections is even, the point is outside, if it is odd is inside. The prototype

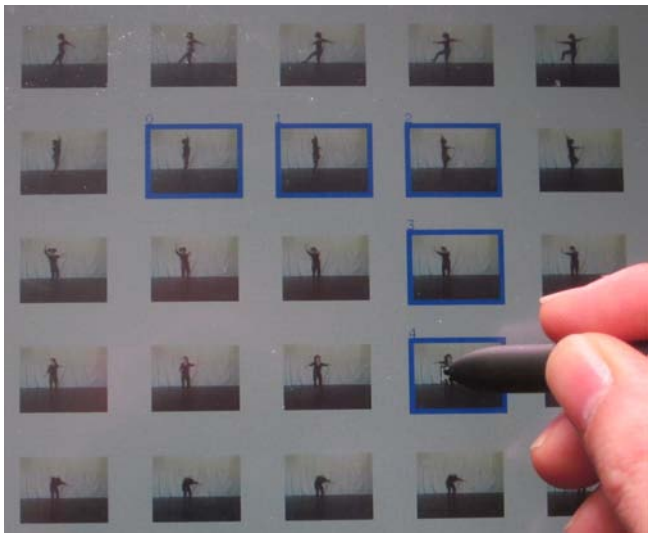


Figure 4.9: Paint selection on frames.

uses an implementation of the ray-casting algorithm made by Alexander Motrichuk[5]. In this implementation the rays are calculated using a horizontal left cross over direction approach and the programmer has to choose if the points that are in the boundary of the polygon can be considered inside or outside. Since, it could be awkward to draw a lasso line over a point and this point (and the corresponding frame or segment) will not be included in the selection, it was considered that boundary points are inside the lasso. In both methods, paint and lasso selection, the selected frames present a number on top of each one, defining the order of the frame selection. This ordering is considered for creating a new video stream.

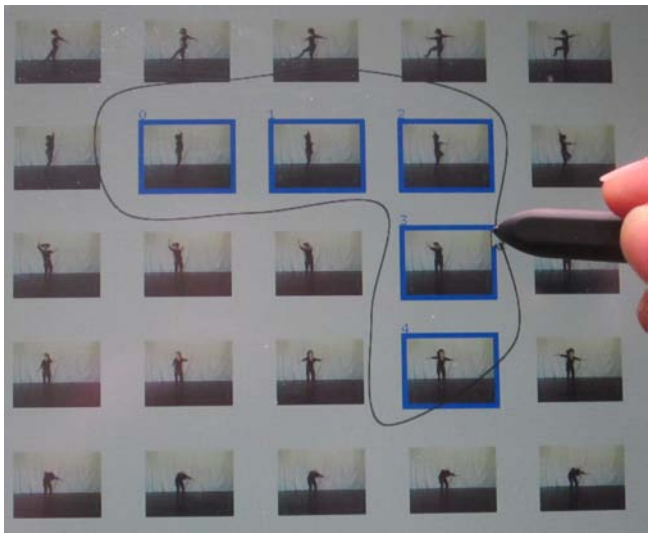
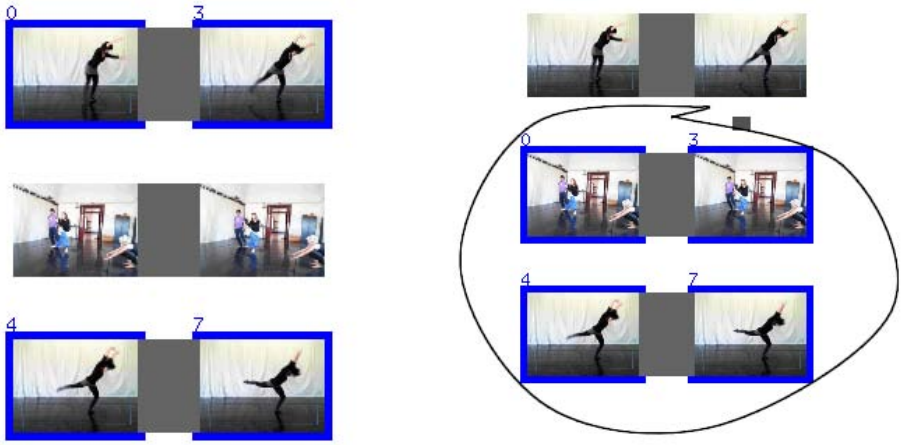


Figure 4.10: Lasso selection on frames.

In this implementation, the two selection methods are limited to create a new video

[5] http://paulbourke.net/geometry/polygonmesh/InsidePolygonWithBounds.cpp

stream. However, they can be combined with other video editing features, in order to apply an action to a set of frames, e.g., move a selected set of frames or segments.

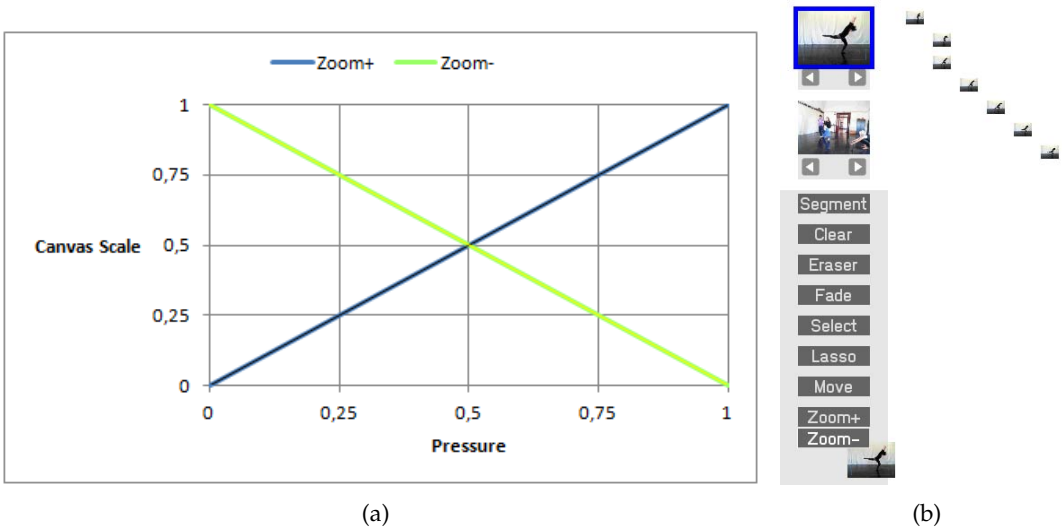


(a) (b)

Figure 4.11: Selecting segments. (a) Paint Selection. (b) Lasso Selection.

4.2.6 Pressure-based Zoom

A pressure-based zoom mechanism was developed as part of the proof-of-concept prototype. It is composed of two buttons, "Zoom+" and "Zoom-", and the pressure made by the pen tip on top of each button causes a proportional scale of the canvas (Figure 4.12). The pressure levels are proportionally scaled into an interval of values between 0 (lowest pressure level) and 1 (maximum pressure level). These values are used directly for scaling the content displayed in the canvas.



(a) (b)

Figure 4.12: Pressure-based zoom. (a) Relation between pressure and scale. (b) Canvas zoomed out (scaled at 0.25).

4.2.7 Evaluation

The evaluation of the proof-of-concept prototype was made by conducting a usability study with 12 subjects, 8 non-experts and 4 experts. The study was composed of a set of basic tasks, a questionnaire and an informal discussion about the tool at the end of the test, particularly in the tests that involved expert users. Since numerical semantical differential scales were included in the questionnaire, it was decided to follow the same approach described in section 3.3.7, i.e., to make parametric (paired-samples t-tests and one-way ANOVA) and non-parametric (Wilcoxon Matched-Pairs Signed-Ranks and Friedman tests) tests, verifying the significant difference between the answers. The considerations mentioned in section 3.3.7 about the number of participants in the tests are also valid in this study.

Before the start of each test, the video as ink concept was briefly introduced to the users. After, they were asked to experiment the different tool features (Figure 4.13), which took around 15 minutes, and to answer the questionnaire. Questions related to particular features were answered immediately after each feature was experimented and more generic questions were left to the end of the test.

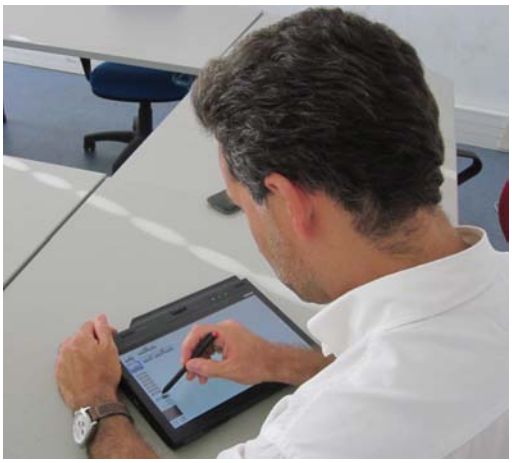


Figure 4.13: A user experimenting videoink prototype.

The test was made on a Windows Tablet PC, a Lenovo X220, with the rotated screen blocking the use of physical keyboard and touchpad. It was asked to the users to only use the tablet's pen, during the test. The tablet's pen detects 1024 pressures levels.

4.2.7.1 Participants

As previously mentioned, the study included 12 participants, 8 non-experts and 4 experts. All the experts users worked (now or in the past) in video or film productions, one accumulates his professional work with video-jockey (VJ) activities and other is a professional designer. The non-expert users record video content for fun or for work purposes. Those that record video in work contexts also use professional software.

The subjects were mostly male (66,67%) and the mean of ages was $\bar{x} = 33.67$ ($\sigma = 6.89$). Most of the subjects had a Master degree (58.33%), 16.67% had a Bachelor degree, 16.67% had a Bachelor degree complemented with post-graduated studies and 8.33% held a PhD. All of them usually record video content, most (83,3%) use a video or photo camera, 66,67% use their mobile phones and 8.33%, corresponding to one user, record video using a webcam (Figure 4.14). 75% of the users edit their videos, from those that do not edit one reported the difficulty and boredom of the task, a second reported the lack of time to do it (the same issue was reported by one of the users that edit) and the third does not simple care about editing his own videos. Most of the users (75 %) had occasionally experimented pen-based technology, 16,67% never had experimented this technology before and 8,33% (one of the experts) frequently use it.

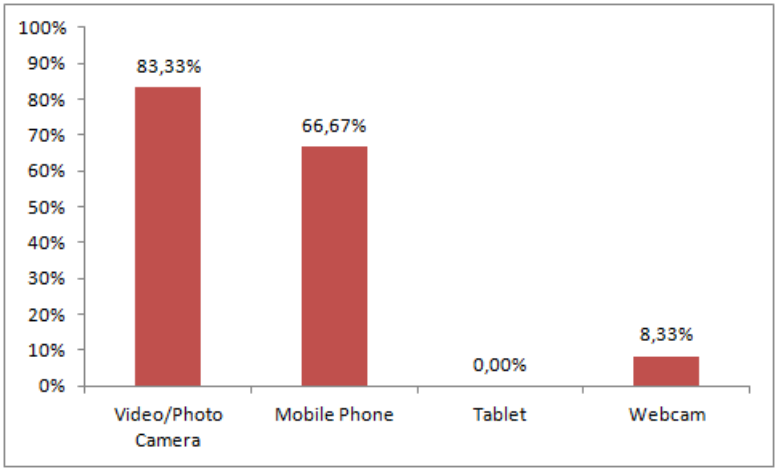


Figure 4.14: Device that participants usually use to record videos.

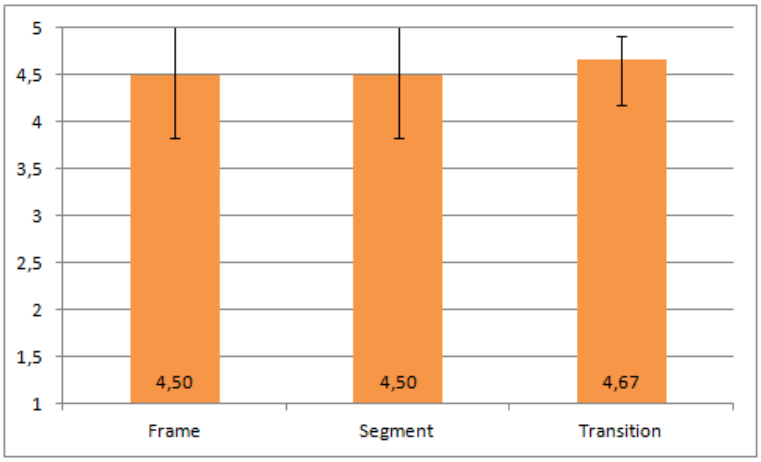
4.2.7.2 Questionnaire

The questionnaire was composed of seven questions with semantic differential numerical scale answers, four (Q1, Q3, Q4, Q6) about perceived difficulty (1 for Difficult - 5 for Easy) and three (Q2, Q5, Q7) about mode usage rate (1 for Rarely - 5 for Frequently); a set of questions (Q8) and pairwise factor rankings (Q9) that define the Creative Support Index (CSI) [CLFT09]; one (Q10) based on Microsoft "Product Reaction Cards" classification [BM02] and one (Q11) open question for comments and suggestions (see Appendix C).

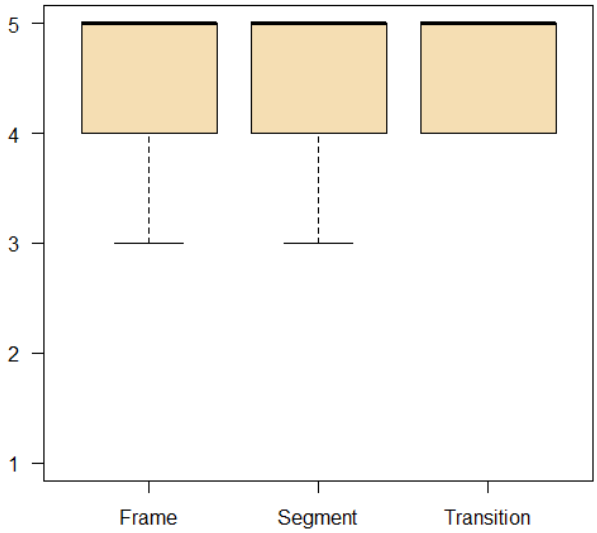
In order to compare the mode preferences and the perceived difficulty, one-way ANOVA and Friedman tests (Q1, Q4 and Q7) as well as paired-samples t-tests and Wilcoxon Matched-Pairs Signed-Ranks tests (Q2, Q3, Q5 and Q6), were conducted based on the null hypothesis (H0), i.e, there was not a significant difference between answers. In all tests the alpha level was 0.05, in order to achieve an interval of confidence of 95%.

In Q1, the users were asked to rate the perceived difficulty for adding a video frame ($\bar{x} = 4.50, \sigma = 0.67, \tilde{x} = 5.00$), a video segment ($\bar{x} = 4.50, \sigma = 0.67, \tilde{x} = 5.00$) and a transition (the fade effect) ($\bar{x} = 4.67, \sigma = 0.49, \tilde{x} = 5.00$) (Figure 4.15). The ANOVA

test did not present a significant difference ($F_{2,33} = 0.29, p > 0.05$) between the different features and a Friedman test confirmed this result ($\chi^2 = 0.7, df = 2, p > 0.05$).



(a)

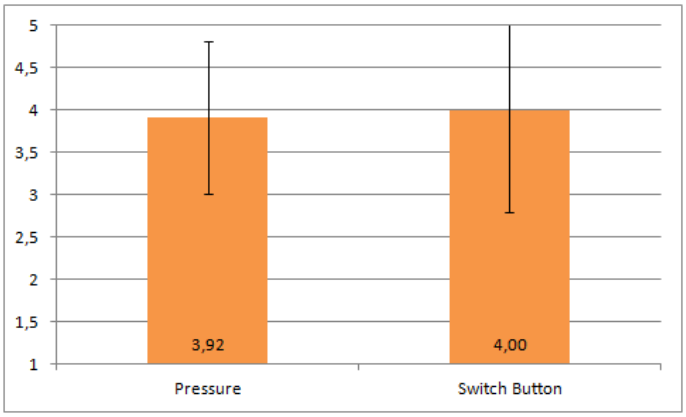


(b)

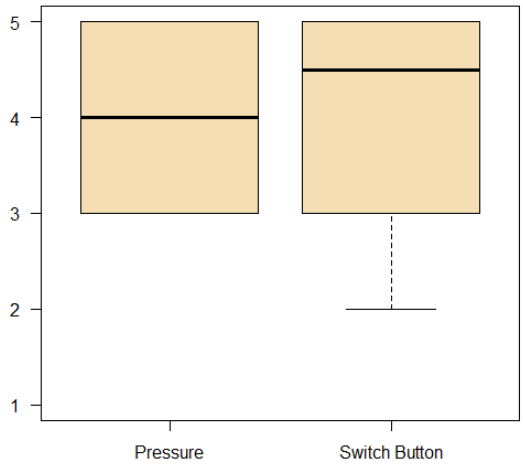
Figure 4.15: Results for the perceived difficulty for adding video frame, a video segment and a transition (the fade effect). (a) Mean scores. Error bars represent the standard deviation. (b) Median scores.

Afterward they were asked to rate the usage (Q2) and the perceived difficulty (Q3) of the two methods for switching between "Frame" and "Segment" modes: by pressing on top of the selected clip or tapping on the switch button. Regarding the usage rating, the t-test ($t(11) = -0.16, p > 0.05$) as well as the Wilcoxon Matched-Pairs Signed-Ranks test ($W = 31, Z = -0.20, p > 0.05$) did not present a significant difference between using the pressure mechanism ($\bar{x} = 3.92, \sigma = 0.90, \tilde{x} = 4.00$) and the switch button ($\bar{x} = 4.00, \sigma = 1.21, \tilde{x} = 4.50$) (Figure 4.16).

However, regarding the perceived difficulty, both the t-test ($t(11) = -2.97, p < 0.05$, Cohen's $d = 0.856$) and the Wilcoxon Matched-Pairs Signed-Ranks test ($W = 4.5, Z =$



(a)

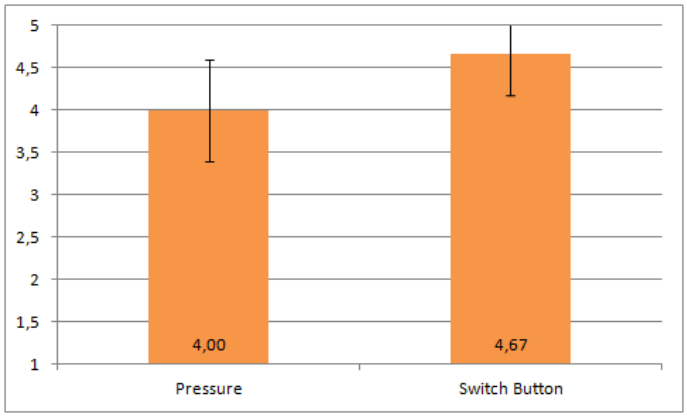


(b)

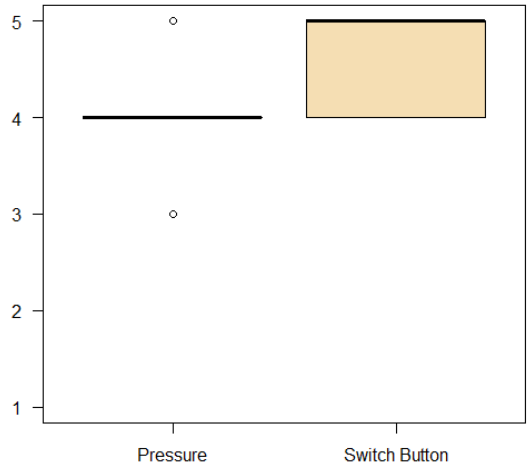
Figure 4.16: Results for the usage of the two methods for mode switching: pressing on top of the selected clip or tapping on the switch button. (a) Mean scores. Error bars represent the standard deviation. (b) Median scores.

$-2.34, p < 0.05, r = 0.48$) showed a significant difference between using the pressure mechanism ($\bar{x} = 4.00, \sigma = 0.49, \tilde{x} = 4.00$) and the switch button ($\bar{x} = 4.67, \sigma = 0.60, \tilde{x} = 5.00$) (Figure 4.17). The results of Q2 and Q3 show that there is not a usage preference between pressing on top of the selected clip or tapping on the switch button but the switch button is perceived as the easiest method.

The perceived difficulty of the different ways of generating a new video stream, with no selection ($\bar{x} = 4.75, \sigma = 0.87, \tilde{x} = 5.00$), using the paint selection ($\bar{x} = 4.67, \sigma = 0.65, \tilde{x} = 5.00$) and using the lasso selection ($\bar{x} = 4.17, \sigma = 0.94, \tilde{x} = 4.00$) was studied in Q4 (Figure 4.18). The ANOVA test did not present a significant difference ($F_{2,33} = 1.74, p > 0.05$) whereas a Friedman test showed a significant difference ($\chi^2 = 6.41, df = 2, p < 0.05$). However, the pairwise comparison, using Wilcoxon Matched-Pairs Signed-Ranks tests, did not show a significant difference between the different possible comparisons, i.e., between no selection and lasso selection ($W = 7.5, Z = -1.82, p > 0.05$), no selection and paint selection ($W = 2.5, Z = -0.52, p > 0.05$) and paint selection and



(a)



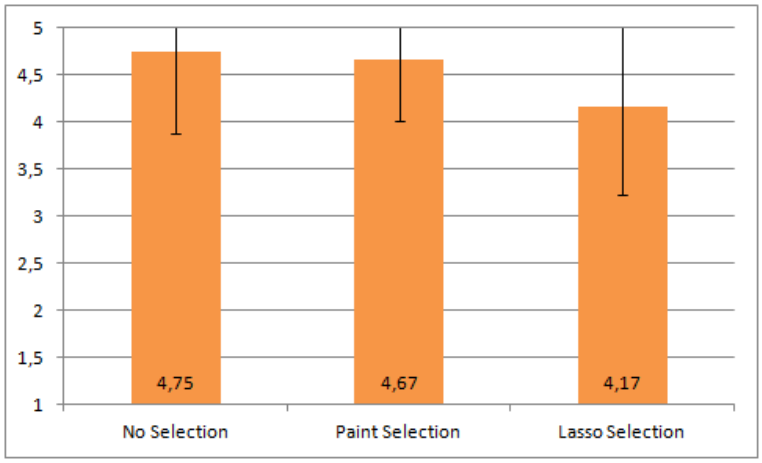
(b)

Figure 4.17: Results for the perceived difficulty for two methods for mode switching. (a) Mean scores. Error bars represent the standard deviation. (b) Median scores.

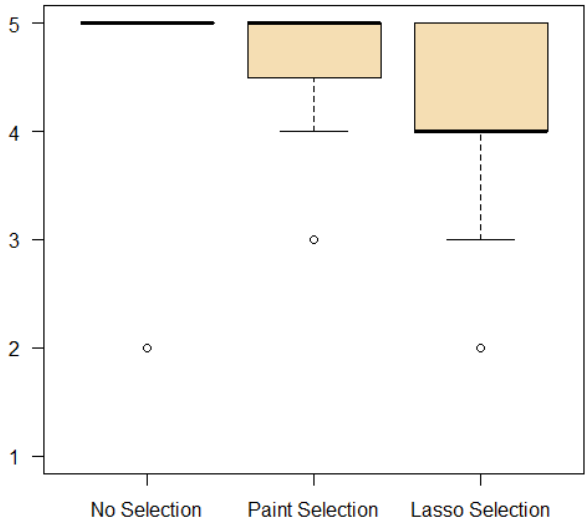
lasso selection ($W = 11.5, Z = -1.49, p > 0.05$). Therefore, it is possible to conclude that there is no significant difference between the perceived difficulty of the different ways of generating a new video stream.

In Q5, it was asked to the participants to rate the usage of the selection modes: paint selection and lasso selection. The t-test ($t(11) = -2.88, p < 0.05$, Cohen's $d = 0.831$) as well as the Wilcoxon Matched-Pairs Signed-Ranks test ($W = 11.5, Z = -2.20, p < 0.05, r = 0.45$) showed a significant difference between paint selection ($\bar{x} = 4.58, \sigma = 0.67, \tilde{x} = 5.00$) and lasso selection ($\bar{x} = 3.42, \sigma = 1.00, \tilde{x} = 3.00$) (Figure 4.19). Q5 shows a preference for using the paint selection mode.

The perceived difficulty of the pressure-based zoom mechanism was rated in Q6. A t-test ($t(11) = -0.52, p > 0.05$) and a Wilcoxon Matched-Pairs Signed-Ranks test ($W = 14, Z = -0.33, p > 0.05$) did not show a significant difference between pressure-based zoom mechanism ($\bar{x} = 3.83, \sigma = 0.83, \tilde{x} = 4.00$) and the pressure mode switch "Frame"/"Segment" mechanism ($\bar{x} = 3.92, \sigma = 0.90, \tilde{x} = 4.00$) classified in Q3 (Figure



(a)

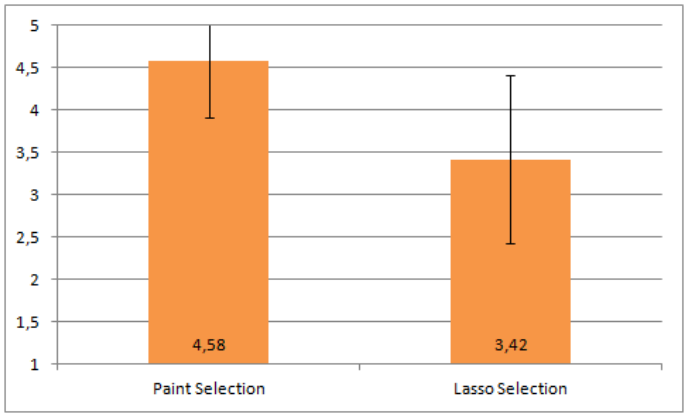


(b)

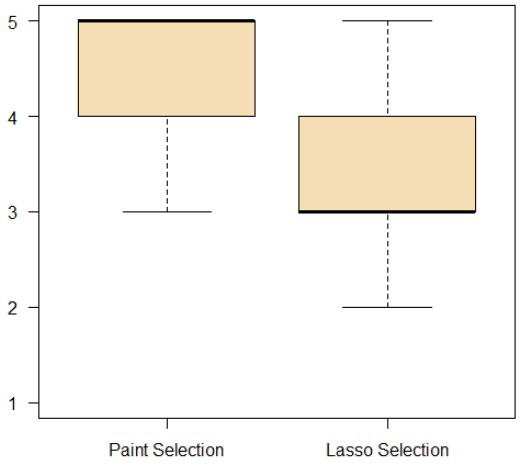
Figure 4.18: Results for the perceived difficulty for different ways of generating a new video stream: no selection, paint selection and lasso selection. (a) Mean scores. Error bars represent the standard deviation. (b) Median scores

4.20).

In Q7 (Figure 4.21), it was asked to the users to rate the usage between the pressure-based zoom ($\bar{x} = 4.25, \sigma = 0.75, \tilde{x} = 4.00$) and other two traditional zooming GUI mechanisms: a slider that goes to the right or left ($\bar{x} = 4.25, \sigma = 0.75, \tilde{x} = 4.00$), increasing or decreasing the zoom, and two simple buttons ($\bar{x} = 3.50, \sigma = 0.80, \tilde{x} = 3.50$), one that zooms in and other that zooms out. Although the ANOVA test showed a significant difference ($F_{2,33} = 3.81, p < 0.05$), Post hoc comparisons, using the Tukey HSD test, did not show a significant difference between any of the comparisons ($p > 0.05$). The Friedman test did also not showed a significant difference ($\chi^2 = 3.80, df = 2, p > 0.05$) between the usage of the different zooming mechanisms. It is possible to say there is not a clear preference between the different zooming methods but more studies are needed to confirm



(a)



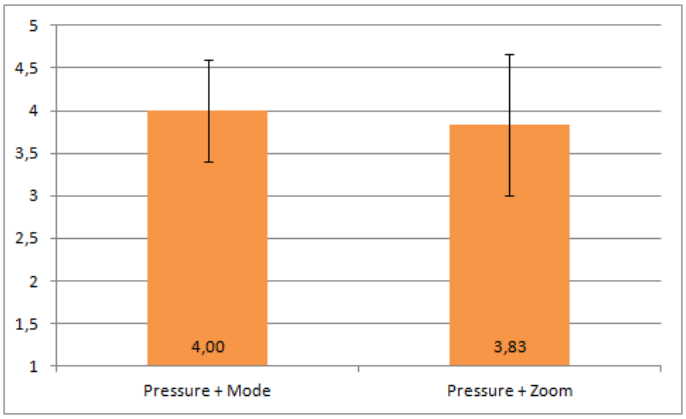
(b)

Figure 4.19: Results for the usage of the selection modes: paint selection and lasso selection: (a) Mean scores. Error bars represent the standard deviation. (b) Median scores.

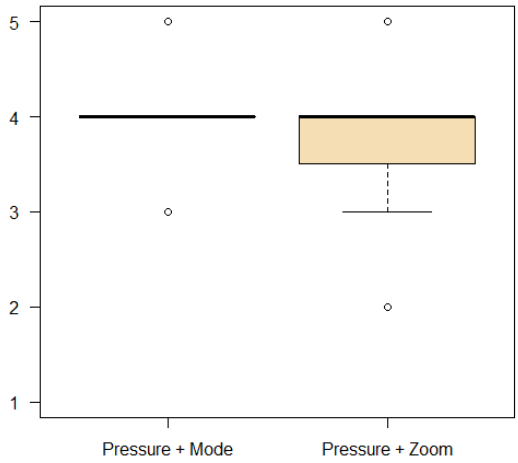
this hypothesis. However, the results presented in Q7, show lower preference for the two buttons interface for zooming.

The Creative Support Index (CSI) [CLFT09] was studied in questions Q8 and Q9. The CSI is a measurement tool for evaluating creativity support. The CSI is composed of six factors: *exploration*, *expressiveness*, *enjoyment*, *immersion*, *collaboration* and *results worth effort*. The survey metric generates an index between 0 and 100 of the creativity support afforded by a system, tool or interface.

In CSI, users have to answer a set of questions, each related to one of the factors, and compare each factor against the other five, assessing the relative importance of these factors (Table 4.1). The questions are scored in a scale from 0 (Highly Disagree) to 10 (Highly Agree) and the factors are pairwise ranked from 0 to 5. Each answer is multiplied by its associated ranked factor and the overall score for the CSI is calculated by summing all the weighted answers and the result of the sum is divided by 1.5, resulting in a value between 0 and 100.



(a)



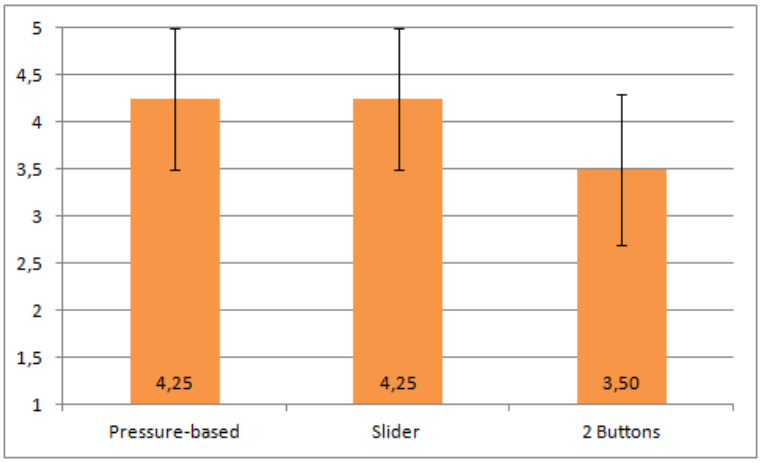
(b)

Figure 4.20: Results for the perceived difficulty of the two pressure-based mechanisms: mode switch and zoom. (a) Mean scores. Error bars represent the standard deviation. (b) Median scores.

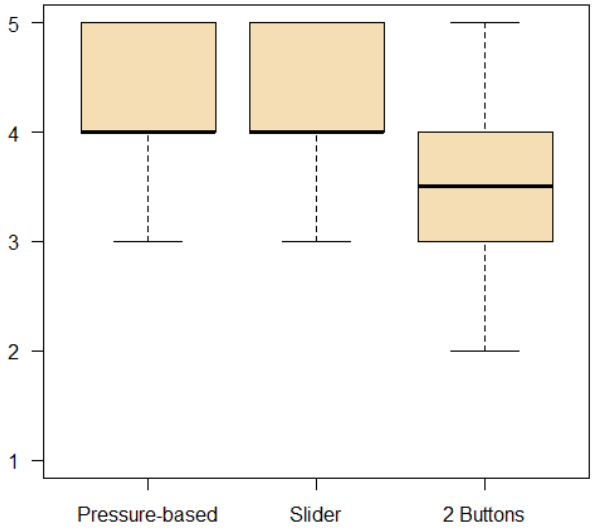
The original work [CLFT09] includes a set of six questions but the software[6] developed by the authors considers two sets of six questions. The calculus of the CSI for these two sets considers the average between the two answers associated to each factor.

As reported in Carrol et al work [CLFT09], for non-collaborative tools (as the one being evaluated), questions on the collaboration factor can confuse the users. The authors argument that, in these situations, the users tend to give low scores in the questions and pairwise comparisons related to collaboration. In addition, the authors of this metric also report that the 15 pairwise factor comparisons can be too tedious for the participants. Considering these aspects, it was decided to remove the two questions about collaboration (one per set), and to eliminate the collaboration factor from the pairwise comparisons. This decision reduced the numbers of questions to five per set, ten in total, and the pairwise comparisons to ten. In order to compute the final calculus of the overall CSI score, the lack of answers related to the collaboration factor was compensated by adding

[6] http://www.erincarroll.net/csi.html



(a)



(b)

Figure 4.21: Results for the usage of the different zoom interfaces: pressure-based, slider and two buttons. (a) Mean scores. Error bars represent the standard deviation. (b) Median scores.

one unit to the other factors and setting the collaboration factor to zero. This approach assumes that if the users had to choose between collaboration and other factor in the pairwise comparisons, they would always choose the other factor.

Figure 4.22 presents the means CSI factors used for the index calculation and Figure 4.23 shows the different result of the CSI index. It is possible to observe that the means for non-expert users and for expert users are very close. The overall CSI mean is 81.53, which can be considered a high value in the CSI scale. The comparison with other tools is difficult due the lack of other studies using this metric. The only comparable study [Mar12] reports CSI scores between 70 and 81, for an interactive quadruped animation tool.

Set 1	
Questions	**Factor**
I was satisfied with what I got out the system or tool.	Results Worth Effort
It was easy for me to explore many different ideas, outcomes, and possibilities.	Exploration
I would be happy to use this system or tool on a regular basis.	Enjoyment
I was able to be very creative while doing this activity.	Expressiveness
My attention was fully tuned to the activity, and I forgot about the system or tool that I was using.	Immersion
It was really easy to share ideas and designs with other people inside this tool.	Collaboration (not used)
Set 2	
Questions	**Factor**
The system was helpful in allowing me to track different ideas, outcomes, or possibilities.	Exploration
I enjoyed this system or tool.	Enjoyment
What I was able to produce was worth the effort I had to exert to produce it.	Results Worth Effort
The system or tool allowed me to be very expressive.	Expressiveness
I became so absorbed in the activity that I forgot about the system or tool that I was using.	Immersion
The system or tool offered support for multiple users.	Collaboration (not used)

Table 4.1: CSI Questions and Factors.

In Q10, the users were asked to classify the tool with 28 words of the Microsoft "Product Reaction Cards" [BM02]. Figure 4.24 presents the percentage for each word. The most selected (>=50%) words were: "easy to use", attractive", "creative", inspiring", and "novel".

The comments and suggestions, asked in Q11, were sometimes replaced by an informal discussion. Most of the written comments were about GUI features, like having tooltips or having some visual feedback in the canvas synchronized with the preview window.

4.2.7.3 Informal Discussions with Participants

During or after each test, there was an informal discussion with the participants, both expert and non-expert. However, as previously reported, there was an intentional focus on this type of feedback in the tests with expert users. The general feedback from the expert users was positive. They particularly enjoy the possibility to explore different

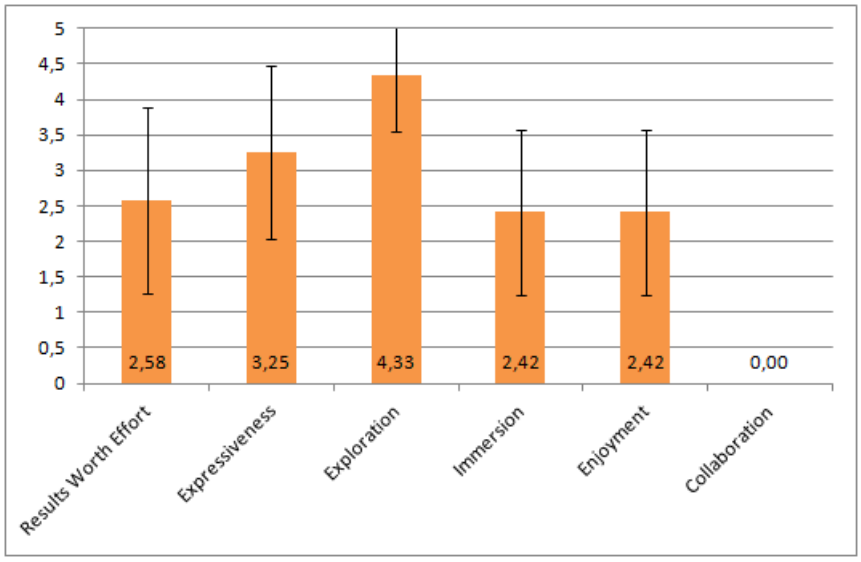


Figure 4.22: CSI factor means used for index calculation. Error bars represent the standard deviation.

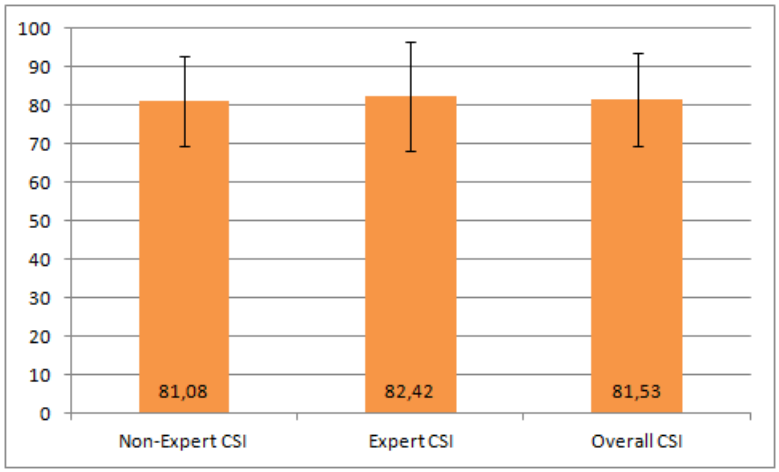


Figure 4.23: CSI means: non-expert, expert and overall. Error bars represent the standard deviation.

outcomes using the two dimensional canvas as well as the use of the pen, when compared with the mouse. Three of them pointed out the visual organization of the video material in the canvas as an advantage of the tool. When they were asked if they considered the canvas and paint mechanism confusing or creative, all them answered as being more creative. Two of them reported the advantage of having pen and touch interactions in such approach.

Three of them also would like to have the possibility to use multiple lasso selections, in order to do multiple operations, i.e., select different video blocks for preview or move them around the canvas, and one of them would like to have the possibility to change the selection order using this tool, e.g., use the path of lasso to order them. One user would like to have the ability to zoom in a frame and edit it, as in an animation tool, and have

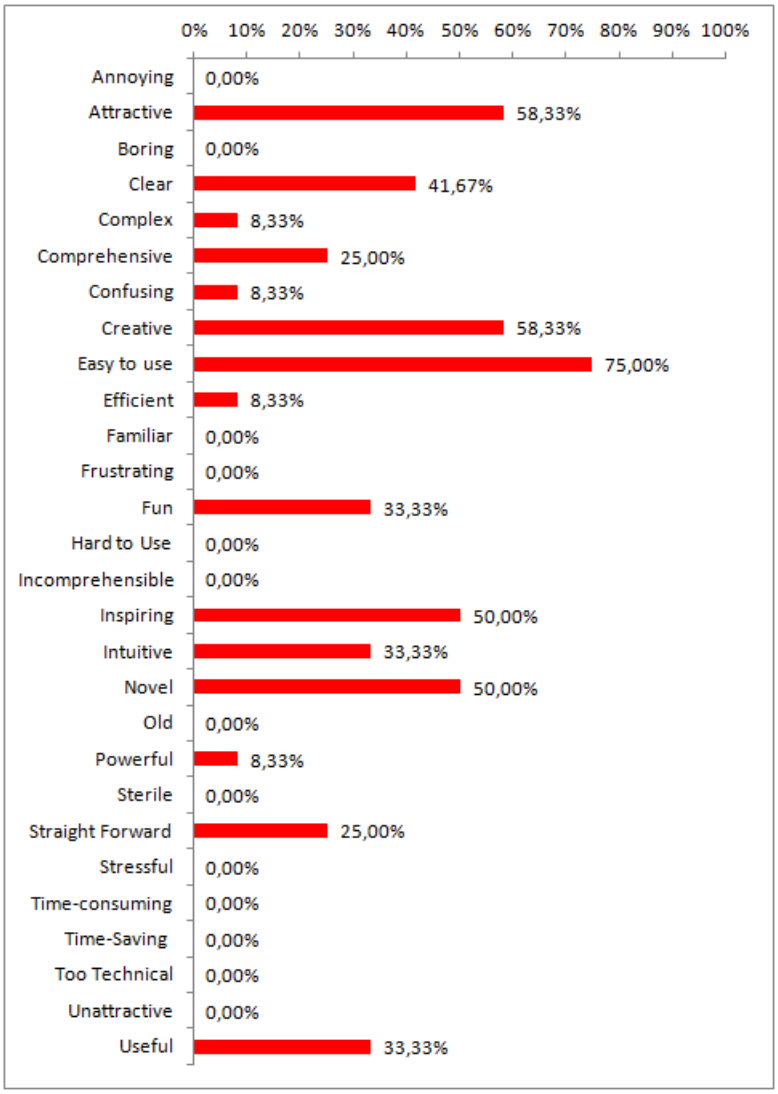


Figure 4.24: Classification with Microsoft "Product Reaction Cards".

some visual feedback on the frames about the segment that they belong to. In addition, other user referred that using this tool, the hand gestures easily followed his thoughts but would like to have more time to experiment the tool.

Some expert and non-expert users referred the need of moving the menu bar to other places of the canvas, particularly to the right side of the screen. The reason of this observation was that the right-handed users had to pass their hand over the canvas, in order to reach the menu bar, causing uncomfortable gestures. In addition, one non-expert user referred, in the context of the two "Frame"/"Segment" modes, that he would like to have the ability to define the start and end frames of a video segment, independently of how they are placed in a frame set.

4.3 Discussion

The general aspects of the video as ink concept were well received by the participants. Adding content to the canvas was generally well perceived by the users, although, it was observed that the majority of them did not realize at first, the rectangles between the two frames, indicating that additional content can be added between them. Regarding these rectangles, it was also observed that sometimes it was difficult to hit them with the pen. Larger hit areas with additional visual tips could help on the task. It was also observed, during the tests, that the usage of a small frame (the next frame to be painted) as a cursor could be confused and some users, during the moving tasks, tried to pick and drop the frame, instead of dragging it on canvas.

The participants could not perceive which mechanism for mode switching, pressing on the clip or tapping on the switch button, would use more. However, the switch button was perceived as the easiest one. In addition, the user that frequently uses the pen tested the two modes a couple of times. It was observed that the pressing technique reduces one step, i.e., using a single gesture one selects the clip and the segment mode, but the amount of time needed to achieve the pressure threshold was longer than just tapping on the switch button.

Regarding the construction of a new video stream, there was not a significant difference between the different methods: no selection, paint selection and lasso selection. Nonetheless, when comparing the two modes of selection, the participants preferred the paint selection. Considering the comments about the usage of multiple lasso selection, this improvement could change the users feedback on the usage of this selection tool.

In some tests the pressure-based zoom mechanism was confused by a time-based mechanism, i.e., the zoom was proportional to the time that the user pressed on one of the two buttons. After the users discover that the zoom was dependent on the pressure made on the buttons, the users tried to control it. Some users considered hard to control the pressure, due to the high sensibility of the pen, specially with low pressure levels. The same control problem was reported in Ramo's et al study [RBB04] about pressure widgets. In addition, some users would like to have a scale, e.g., 0 % to 100 %, in order to perceive the maximum and minimum zoom levels. Even tough the tests did not show a significant different between the pressure-based mechanism presented in the prototype and the traditional zoom methods, slider and two buttons, this last one was the less considered for usage by the users. Therefore, it could be interesting to use an interface similar to the Zlider [RB05], developed by Ramos et al, in the *videoink* prototype.

The CSI score presents a high value in the scale but the lack of other studies using this metric difficult the comparison with other tools. Regarding the usage of this metric and due to the short experience with the tool, some users reported some difficulty to answer questions related with *results worth effort* and *immersion* factors. It is important to note that the most ranked factors of CSI metric were *expressiveness* and *exploration*. With these results combined with words chosen by the majority of the users to classify their

experience with the prototype ("easy to use", attractive", "creative", inspiring","novel") and with the feedback given during the informal discussions, it is possible to conclude that the *videoink* concept fosters creativity using natural interactions.

5

Conclusions and Research Status

This chapter presents a summary of the developed work and its findings as well as a discussion of its limitations, future research directions and implications.

5.1 Research Summary and Findings

The main motivation of this work was the improvement of human-media interaction. The human familiarity with pen-based devices makes them a potential candidate for video interaction improvement. However, their usage for such tasks was not totally exploited before.

The projects like those developed before by Ramos et al [RB03; RB05], Bulterman et al [Bul04], Diakopoulos et al [DE06] or Hurst et al [HG08] have exploited the usage of pen-based technology on different types video interactions: video browsing, video annotation and video editing. Although pen-based video browsing has been fairly studied, as shown in Chapter 2, pen-based video annotation and pen-based video editing can still benefit with more natural interactions. By studying the usage of pen-based technology on these two types of video manipulation, it would be also possible to study the two interaction modes enabled by this technology: indirectly, through digital ink, or directly, trough pen gestures or pressure.

In this research, digital ink was applied to video annotations, developing the concept of *pen-based video annotations*, and pen gestures and pressure was applied to video editing in a concept called *video as ink*. The first one is focused on adding pen-based annotations to video content, which can also be used for video navigation or for the generation of a new video stream. The second approach replaces the digital ink by video frames or segments, which are used for video editing.

A proof-of-concept prototype of pen-based video annotations was implemented using Tablet PCs. In order to study further this concept, a multimodal video annotator for Tablet PCs, called Creation-Tool, was also developed. The system supports annotations composed of ink, annotation marks, text, audio and hyperlinks. The prototype presents different annotation modes and methods, allowing the annotation of a live or a recorded event. In addition, pen-based video annotation were combined with real-time motion tracking algorithms, in order to maintain the association between the moving features and its annotations. The system evaluation aimed to study users' opinion about the different annotation types, modes, methods and scenarios as well as their judgment about trackers' performance. From users' feedback it is possible to conclude pen-based video annotations will be used more often during a live event and while annotating it there is preference for annotation modes that help to focus the user attention on the event and not on the application. The tests also showed that Kinect [SFCSFMKB11] and TLD [KMM12] were sufficiently accurate to work as real-time trackers. Kinect presented a better performance in the users view, but shows more limitations when compared with TLD. It also was stressed by the users the need to have annotations methods that allow pausing the video, while an annotation is being made, but somehow showing the live event without much visual noise. Since annotating directly on video is a task only made possible with recent technology, standard iconography and guidelines are needed, in order to help users with this task.

The main principles for a pen-based approach to video editing were also presented in this work. The described approach uses video content as digital ink, which can be painted in a canvas that works as a two dimensional timeline. In the context of this research, a Tablet PC implementation of the concept combined with different video editing features was carried out. The prototype includes operation features such as add, erase and move video content. A fade effect and two selection modes, one using pen gestures and the other using a lasso tool, as well as pressure-based features, like mode switching or zooming, were also implemented. In order to know the users' opinion about the concept and the different prototype features a usability study was conducted. The concept was well received by the users, allowing them to explore different outcomes of the final video content. In this study, users pointed out the visual organization and creativity fostered by the *videoink* concept. The easiness to use the pen to manipulate directly video content, when compared with regular software tools that are mainly developed for the mouse or touchpad, was also noted by the users. The tests showed a preference for the usage of paint selection when compared with lasso selection. However, an improvement of the lasso tool by allowing multiple "lassos" in the canvas, could change the users' opinion. In addition, the pressure for zooming tasks had a positive feedback from the users but it should be improved with a visual scale and by making it less sensitive with low levels.

5.2 Limitations and Future Work

This dissertation aimed to study interaction with video using pen-based technology, particularly on two main topics: video annotation and video editing. Despite the findings reported in the previous section, a set of limitations and future research directions were detected, as reported next.

5.2.1 Collaborative and Shared Pen-based Video Annotations

Research questions about sharing pen-based video annotations would not fit in the scope of this dissertation. However, they are important, particularly for collaborative work. They require solutions that involve issues like private vs public or individual vs group. How pen-based video annotations can be adapted to these concepts is still an open question. Systems, like the Ambulant Annotator [CBJ06; CBGJKS08], ARA [GCD11] or CoVida [ZWLS12], started to exploit the idea of sharing pen-based annotations when doing collaborative work but the topic require further study, especially when applied to dynamic content like video.

5.2.2 Pen-based Hyperlinks

During the development of the multimodal video annotator, presented in 3.3, pen-based annotations for hyperlinks were tested. However, the representation of websites by URLs limits the pen usage to handwriting. Therefore, this interaction can also be achieved by handwriting recognition algorithms, transforming the handwritten text to a regular URL. This solution not only can break the interaction fluidity but also requires that the user should know the URL from memory, particularly in the case of long and complex URL strings. Simplified representations of URLs and websites are needed, in order to use them as pen-based annotations.

5.2.3 Attaching Anchors and Annotations: Selection vs Line Connectors

The Creation-Tool used a mechanism of anchor selection, in order to attach annotations to a particular anchor. However, an alternative method based on line connectors could also be used. These lines could be sketched by the user using the pen and would connect an annotation to an anchor. The usage of line connectors could pin annotations in a particular point, whereas lines would move or scale according to anchor's motion (Figure 5.1). However, this solution not only requires more engagement when doing the task of attaching a note to an anchor, i.e., the user would need to draw a line connector instead of just pointing to an anchor, but also requires the creation of a particular annotation category, the line connector, that could be explicit selected by the user or could be recognized somehow by the system.

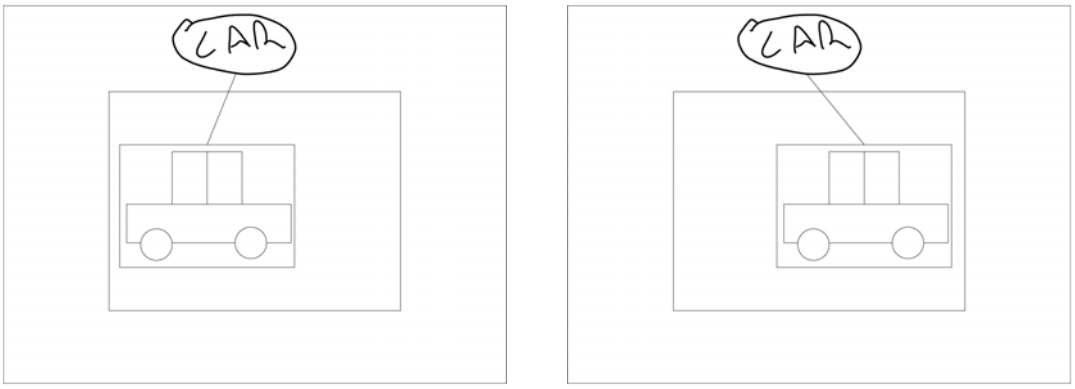


Figure 5.1: Line connector links a pinned note and an anchor of a moving object

5.2.4 Annotations Methods for Live Annotation

In section 3.4, it is discussed a hybrid method of annotation, which combined the advantages of both "Hold and Overlay" and "Hold and Speed Up" methods, i.e., displaying the live event without increasing the visual noise of the main video window. Which kind of method would be and how it can be achieved is an open question. The annotations and video visualization modes as well as the annotations methods developed in the context of this dissertation, can serve as background work for the study of such method.

5.2.5 Bimanual Pen + Touch Video Interactions

In Chapter 1, it was mentioned the importance of having pen and touch bimanual interactions. The focus of this dissertation was on pen-based video interaction but with the understanding that this was a step to reach full pen and touch interaction. The attempts of using simultaneously pen and touch, during this research, failed due the limitations of Tablet PCs, as reported in 3.3.6.2. Research works, like those developed Hinckley et al [HYPCRWBB10], Matulic et al [MNAKS13] and Santosa et al [SCBS13], applied pen and touch interaction to image, text and animation editing. However, how this bimanual interaction can be applied to video annotation or video editing should be addressed in future work.

5.2.6 More Pen-based Gestures

The recent developments of multi-touch pens unfold new modes of video interaction using pen-based technology. The work developed by Song et al [SBGICH11] (Figure 5.2) and by Liu and Guimbretière [LG12] use specific sensors on the pen's barrel, in order to detect grips and gestures made by users' fingers. The pen prototype developed by Hinkley et al [HCB13], uses accelerometer, gyro and magnetometer sensors, in order to capture motion and context sensing techniques. How these finger grips and gestures can be used on video interaction still needs an answer.

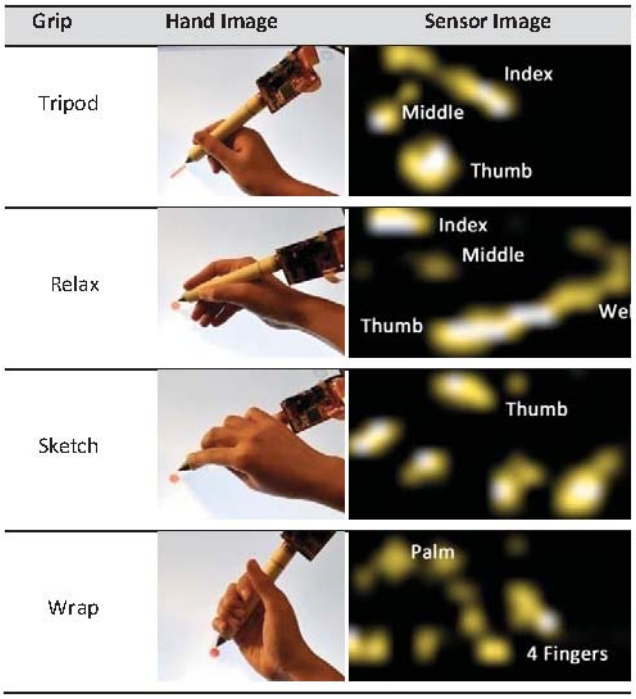


Figure 5.2: Multi-touch Pen [SBGICH11]

5.2.7 Bidirectional Video Segments

As mentioned in section 4.2.3, the proof-of-concept of the *videoink* prototype presents two major limitations regarding video segments: 1) the video segments are only be represented horizontally, left to right, in the canvas and 2) the dragging gesture is not used for painting segments. The positive feedback from the users on painting and selecting video content on a canvas as a way to explore different outcomes of the final video stream, indicates that the development of bidirectional video segments could worth the effort of implementing such complex feature. The direction of a video segment could be calculated based on the direction of the pen dragging on the screen (Figure 5.3).

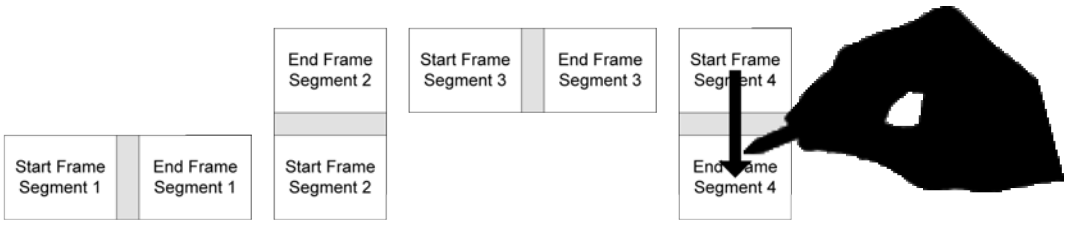


Figure 5.3: Bidirectional video segments using pen dragging

5.2.8 Pen-based Sound Editing

This research only considered the visual dimension of video editing. However, sound is also an important factor on video and film editing. Pen-based interactions for sound

editing should also be considered on future research work, e.g., moving sound tracks around the canvas.

5.3 Final Remarks

This work aimed to study how pen-based technology could be applied to video interaction, making it more natural and fostering user's creativity. As previously mentioned, two main topics and usage contexts were studied: video annotation and video editing. In this research, each topic was studied by exploiting one of the main modes interaction of using pen-based technology, through digital ink and using gestures and pressure. Two concepts resulted from this study: pen-based video annotations, used for video data analysis, and video as ink, used for video editing. Each concept not only provides a more natural video interaction but fosters creativity, crucial in both usage contexts. Pen-based video annotations improve the interaction fluidity of note taking and sketching and, at the same time, maintain the dynamic context associated to video content, and video as ink provides more familiar interactions for video editing and fosters the creation and exploration of new video content. From these findings results that pen-based technology should not be discarded as a method of interaction, particularly when applied to multimedia content. In addition, it is important to report that, during the development of this research, it was observed that computer interfaces should be properly adapted to the affordances and features of pen-based technology and better designed, in order to improve video interaction. Regarding pen-based technology, there are familiar gestures and interactions, learned from physical pens and hand tools, that should also be used on digital technology, whereas on video interaction, interfaces should be carefully designed, in order to facilitate and simplify the manipulation of such complex media. If these issues are considered during the design and development of future human-media interactions, then, less barriers will exist in the creation and sharing of new media content, ideas and knowledge.

Bibliography

[AGL03] G. D. Abowd, M. Gauger, and A. Lachenmann. "The Family Video Archive: an annotation and browsing environment for home movies". In: *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*. MIR '03. Berkeley, CA, USA: ACM, 2003, pp. 1–8.

[AHFMI11] A. Al Hajri, S. Fels, G. Miller, and M. Ilich. "Moving target selection in 2D graphical user interfaces". In: *Proceedings of the 13th IFIP TC 13 international conference on Human-computer interaction - Volume Part II*. Vol. 6947. Lecture Notes in Computer Science, INTERACT'11. Lisbon, Portugal: Springer-Verlag, 2011, pp. 141–161.

[AHWA04] R. J. Anderson, C. Hoyer, S. A. Wolfman, and R. Anderson. "A study of digital ink in lecture presentation". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '04. Vienna, Austria: ACM, 2004, pp. 567–574.

[AWW02] A. Aner-Wolf and L. Wolf. "Video de-Abstraction or How to save money on your wedding video". In: *Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision*. WACV '02. Orlando, FL, USA: IEEE Computer Society, 2002, pp. 264–268.

[Bac06] J. Backon. "Student Minds and Pen Technologies: A Wonderful Pedagogical Marriage". In: *The Impact of Tablet PCs and Pen-based Technology on Education: Vignettes, Evaluations and Future Directions.* West Lafayette, IN, USA: Purdue University Press, 2006, pp. 1–11.

[BMR81] R. Baecker, D. Miller, and W. Reeves. "Towards a laboratory instrument for motion analysis". In: *SIGGRAPH Comput. Graph.* 15.3 (1981), pp. 191–197.

[BRFSC96] R. Baecker, A. J. Rosenthal, N. Friedlander, E. Smith, and A. Cohen. "A multimedia system for authoring motion pictures". In: *Proceedings of the fourth ACM international conference on Multimedia*. MULTIMEDIA '96. Boston, MA, USA: ACM, 1996, pp. 31–42.

[BGGS99] D. Bargeron, A. Gupta, J. Grudin, and E. Sanocki. "Annotations for streaming video on the Web: system design and usage studies". In: *Comput. Netw.* 31.11-16 (May 1999), pp. 1139–1153.

[BM03] D. Bargeron and T. Moscovich. "Reflowing digital ink annotations". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. CHI '03. Ft. Lauderdale, FL, USA: ACM, 2003, pp. 385–393.

[BETVG08] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. "Speeded-Up Robust Features (SURF)". In: *Computer Vision and Image Understanding* 110.3 (June 2008), pp. 346–359.

[BM02] J. Benedek and T. Miner. "Measuring Desirability: New methods for evaluating desirability in a usability lab setting". In: *Proceedings of UPA Usability Professional Association Conference*. Orlando, FL, USA: Microsoft Corporation, 2002.

[BBDER91] W. E. Bennett, S. J. Boies, A. R. Davies, K.-F. Etzold, and T. K. Rodgers. "Optical stylus and passive digitizing tablet data input system". Pat. US 5051736. 1991.

[BDB06] A. Bezerianos, P. Dragicevic, and R. Balakrishnan. "Mnemonic rendering: an image-based approach for exposing hidden changes in dynamic displays". In: *Proceedings of the 19th annual ACM symposium on User interface software and technology*. UIST '06. Montreux, Switzerland: ACM, 2006, pp. 159–168.

[BCLOPT04] P. Bottoni, R. Civica, S. Levialdi, L. Orso, E. Panizzi, and R. Trinchese. "MADCOW: a multimedia digital annotation system". In: *Proceedings of the working conference on Advanced visual interfaces*. AVI '04. Gallipoli, Italy: ACM, 2004, pp. 55–62.

[Bra98] G. R. Bradski. "Computer Vision Face Tracking For Use in a Perceptual User Interface". In: *Intel Technology Journal* 2.2 (1998), pp. 1–15.

[BFWHS08] P. Brandl, C. Forlines, D. Wigdor, M. Haller, and C. Shen. "Combining and measuring the benefits of bimanual pen and direct-touch interaction on horizontal interfaces". In: *Proceedings of the working conference on Advanced visual interfaces*. AVI '08. Napoli, Italy: ACM, 2008, pp. 154–161.

[BRN04] H. Brugman, A. Russel, and X. Nijmegen. "Annotating multimedia / multimodal resources with ELAN". In: *Proceedings of Fourth International Conference on Language Resources and Evaluation*. LREC 2004. Lisbon, Portugal, 2004, pp. 2065–2068.

[Bul04] D. Bulterman. "Animating Peer-Level Annotations Within Web-Based Multimedia". In: *EuroGraphics Multimedia Workshop*. Nanjing, China: Eurographics Association, 2004, pp. 49–57.

[Bur06] B. Burr. "VACA: a tool for qualitative video analysis". In: *CHI '06 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '06. Montréal, Québec, Canada: ACM, 2006, pp. 622–627.

[Bus45] V. Bush. "As We May Think". In: *The Atlantic Monthly* 176.1 (1945), pp. 101–108.

[Bux01] B. Buxton. *Buxton Collection* `http://research.microsoft.com/en-us/um/people/bibuxton/buxtoncollection/`. ONLINE. Last access: 30-10-2013. 2001.

[Bux07] B. Buxton. *Sketching User Experiences: Getting the Design Right and the Right Design*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007.

[Bux12] B. Buxton. *Some Milestones in Computer Input Devices:An Informal Timeline* `http://www.billbuxton.com/inputTimeline.html`. ONLINE. Last access: 30-10-2013. 2012.

[CCSVFC11] D. Cabral, U. Carvalho, J. Silva, J. Valente, C. Fernandes, and N. Correia. "Multimodal video annotation for contemporary dance creation". In: *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*. CHI EA '11. Vancouver, BC, Canada: ACM, 2011, pp. 2293–2298.

[CC09] D. Cabral and N. Correia. "Pen-Based Video Annotations: A Proposal and a Prototype for Tablet PCs". In: *Proceedings of the 12th IFIP TC13 Human-Computer Interaction International Conference, Part II*. Vol. 5727. Lecture Notes in Computer Science, INTERACT'09. Uppsala, Sweden: Springer, 2009, pp. 17–20.

[CC12] D. Cabral and N. Correia. "Videoink: a pen-based approach for video editing". In: *Adjunct proceedings of the 25th annual ACM symposium on User interface software and technology*. UIST Adjunct Proceedings '12. Cambridge, MA, USA: ACM, 2012, pp. 67–68.

[CC07] D. Cabral and N. Corrreia. "Mobile and Web Tools for Participative Learning". In: *Proceedings of International Conference e-Learning*. Lisbon, Portugal: IADIS, 2007, pp. 483–490.

[CV11] D. Cabral and J. Valente. *Programmer's Guide for QT Gui + openFrameworks (OF) in C++ (Visual Studio 2008 & 2010)*. Technical Report. CITI and DI, FCT/UNL, 2011.

[CVSAFC11] D. Cabral, J. Valente, J. Silva, U. Aragão, C. Fernandes, and N. Correia. "A creation-tool for contemporary dance using multimodal video annotation". In: *Proceedings of the 19th ACM international conference on Multimedia*. MM'11. Scottsdale, AZ, USA: ACM, 2011, pp. 905–908.

[CVAFC12] D. Cabral, J. G. Valente, U. Aragão, C. Fernandes, and N. Correia. "Evaluation of a Multimodal Video Annotator for Contemporary Dance". In: *Proceedings of the 11th International Working Conference on Advanced Visual Interfaces*. AVI'12. Capri, Italy: ACM, 2012, pp. 572–579.

[CLSF10] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. "BRIEF: binary robust independent elementary features". In: *Proceedings of the 11th European conference on Computer vision: Part IV*. ECCV'10. Heraklion, Crete, Greece: Springer-Verlag, 2010, pp. 778–792.

[CWB07] M. Campanella, H. Weda, and M. Barbieri. "Edit while watching: home video editing made easy". In: *Proceedings of the SPIE Electronic Imaging 2007 - Multimedia Content Access: Algorithms and Systems*. Vol. 6506. EI 2007. San Jose, CA, USA: SPIE, 2007, 65060L:1–65060L:10.

[CP07] J. Carifio and R. J. Perla. "Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes". In: *Journal of the Social Sciences* 3.3 (2007), pp. 106–116.

[CS91] R. Carr and D. Shafer. *The Power of Penpoint*. USA: Addison-Wesley, 1991.

[CLFT09] E. A. Carroll, C. Latulipe, R. Fung, and M. Terry. "Creativity factor evaluation: towards a standardized survey metric for creativity support". In: *Proceedings of the seventh ACM conference on Creativity and cognition*. C&C '09. Berkeley, CA, USA: ACM, 2009, pp. 127–136.

[CLMBSDYC02] J. Casares, A. C. Long, B. A. Myers, R. Bhatnagar, S. M. Stevens, L. Dabbish, D. Yocum, and A. Corbett. "Simplifying video editing using metadata". In: *Proceedings of the 4th conference on Designing interactive systems*. DIS '02. London, England: ACM, 2002, pp. 157–166.

[CTGP08] R. G. Cattelan, C. Teixeira, R. Goularte, and M. D. G. C. Pimentel. "Watch-and-comment as a paradigm toward ubiquitous interactive video editing". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)* 4.4 (Nov. 2008), 28:1–28:24.

[CBJ06] P. Cesar, D. C. A. Bulterman, and A. J. Jansen. "The ambulant annotator: empowering viewer-side enrichment of multimedia content". In: *Proceedings of the 2006 ACM symposium on Document engineering*. DocEng '06. Amsterdam, The Netherlands: ACM, 2006, pp. 186–187.

[CBGJKS08] P. Cesar, D. C. Bulterman, D. Geerts, J. Jansen, H. Knoche, and W. Seager. "Enhancing social sharing of videos: fragment, annotate, enrich, and share". In: *Proceedings of the 16th ACM international conference on Multimedia*. MM '08. Vancouver, BC, Canada: ACM, 2008, pp. 11–20.

[Cha04] G. Chandler. *Cut by Cut: Editing Your Film or Video*. Studio City, CA, USA: Michael Wiese Produtions, 2004.

[Cha12] G. Chandler. *Cut by Cut: Editing Your Film or Video (2nd Edition)*. Studio City, CA, USA: Michael Wiese Produtions, 2012.

[CFS03] G Cherry, J Fournier, and R Stevens. "Using a digital video annotation tool to teach dance composition". In: *Interactive Multimedia Electronic Journal of ComputerEnhanced Learning* 5.1 (2003).

[CKRW99] P. Chiu, A. Kapuskar, S. Reitmeier, and L. Wilcox. "NoteLook: taking notes in meetings with digital video and ink". In: *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. MULTIMEDIA '99. Orlando, FL, USA: ACM, 1999, pp. 149–158.

[CC06] N. Correia and D. Cabral. "Interfaces for Video Based Web Lectures". In: *Proceedings of the 6th IEEE International Conference on Advanced Learning Technologies*. ICALT '06. Kerkrade, The Netherlands: IEEE Computer Society, 2006, pp. 634–638.

[CC99] N. Correia and T. Chambel. "Active video watching using annotation". In: *Proceedings of the seventh ACM international conference on Multimedia (Part 2)*. MULTIMEDIA '99. Orlando, FL, USA: ACM, 1999, pp. 151–154.

[CCGa02] M. Costa, N. Correia, and N. Guimarães. "Annotations as multiple perspectives of video content". In: *Proceedings of the tenth ACM international conference on Multimedia*. MULTIMEDIA '02. Juan-les-Pins, France: ACM, 2002, pp. 283–286.

[CBP00] S. B. Cousins, M. Baldonado, and A. Paepcke. *A Systems View of Annotations*. Technical Report P9910022. Xerox PARC, 2000.

[Dan11] K. Dancyger. *The Technique of Film and Video Editing: History, Theory, and Practice*. 5th. Focal Press, 2011.

[Dav03] M. Davis. "Editing out video editing". In: *MultiMedia, IEEE* 10.2 (2003), pp. 54 –64.

[DE64] M. R. Davis and T. O. Ellis. "The RAND tablet: a man-machine graphical communication device". In: *Proceedings of the October 27-29, 1964, fall joint computer conference, part I*. AFIPS '64 (Fall, part I). San Francisco, CA, USA: ACM, 1964, pp. 325–331.

[DE06] N. Diakopoulos and I. Essa. "Videotater: an approach for pen-based digital video segmentation and tagging". In: *Proceedings of the 19th annual ACM symposium on User interface software and technology*. UIST '06. Montreux, Switzerland: ACM, 2006, pp. 221–224.

[DGE09] N. Diakopoulos, S. Goldenberg, and I. Essa. "Videolyzer: quality analysis of online informational video for bloggers and journalists". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. Boston, MA, USA: ACM, 2009, pp. 799–808.

[Dim57] T. L. Dimond. "Devices for reading handwritten characters". In: *Proceedings of the eastern joint computer conference: Computers with deadlines to meet*. IRE-ACM-AIEE '57 (Eastern). Washington, D.C., USA: ACM, 1957, pp. 232–237.

[DFAB04] A. Dix, J. Finlay, G. D. Abowd, and R. Beale. *Human-Computer Interaction*. 3rd. Prentice Hall, 2004.

[DRBNBS08] P. Dragicevic, G. Ramos, J. Bibliowitcz, D. Nowrouzezahrai, R. Balakrishnan, and K. Singh. "Video browsing by direct manipulation". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. Florence, Italy: ACM, 2008, pp. 237–246.

[ED94] E. Elliott and G. Davenport. "Video streamer". In: *Conference companion on Human factors in computing systems*. CHI '94. Boston, MA, USA: ACM, 1994, pp. 65–68.

[EG04] S. E. Ellis and D. P. Groth. "A collaborative annotation system for data visualization". In: *Proceedings of the working conference on Advanced visual interfaces*. AVI '04. Gallipoli, Italy: ACM, 2004, pp. 411–414.

[EBGGHJLMPPTW92] S. Elrod, R. Bruce, R. Gold, D. Goldberg, F. Halasz, W. Janssen, D. Lee, K. McCall, E. Pedersen, K. Pier, J. Tang, and B. Welch. "Liveboard: a large interactive display supporting group meetings, presentations, and remote collaboration". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '92. Monterey, CA, USA: ACM, 1992, pp. 599–607.

[FJ13] C. Fernandes and S. Jurgens. "Video annotation in the TKB project: Linguistics meets choreography meets technology". In: *International Journal of Performance Arts & Digital Media* 9.1 (2013), pp. 115–134.

[FJ09] C. M. Fernandes and S. Jürgens. "Transdisciplinary research bridging cognitive linguistics and digital performance: from multimodal corpora to choreographic knowledge-bases". In: *Performing Technology: User Content and the New Digital Media*. Newcastle upon Tyne, UK: Cambridge Scholars Publishing, 2009. Chap. 2, pp. 19–34.

[Fis04] S. R. Fischer. *A History of writing*. London, UK.: Steven R. Fischer, 2004.

[FDFH96] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes. *Computer graphics: principles and practice (2nd ed. in C)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1996.

[Gal67] L. Gallenson. "A graphic tablet display console for use under time-sharing". In: *Proceedings of the November 14-16, 1967, fall joint computer conference*. AFIPS '67 (Fall): Anaheim, CA, USA: ACM, 1967, pp. 689–695.

[GBSBW01] A. Girgensohn, S. Bly, F. Shipman, J. Boreczky, and L. Wilcox. "Home Video Editing Made Easy - Balancing Automation and User Control". In: *Proceedings of the 8th IFIP TC13 Human-Computer Interaction International Conference*. INTERACT '01. Tokyo, Japan: IOS Press, 2001, pp. 464–471.

[GBCDFGUW00] A. Girgensohn, J. Boreczky, P. Chiu, J. Doherty, J. Foote, G. Golovchinsky, S. Uchihashi, and L. Wilcox. "A semi-automatic approach to home video editing". In: *Proceedings of the 13th annual ACM symposium on User interface software and technology*. .Girgensohn00. UIST '00. San Diego, CA, USA: ACM, 2000, pp. 81–89.

[GCSS06] D. B. Goldman, B. Curless, D. Salesin, and S. M. Seitz. "Schematic storyboarding for video visualization and editing". In: *ACM Transactions on Graphics* 25.3 (July 2006), pp. 862–871.

[GGCSS08] D. B. Goldman, C. Gonterman, B. Curless, D. Salesin, and S. M. Seitz. "Video object annotation, navigation, and composition". In: *Proceedings of the 21st annual ACM symposium on User interface software and technology*. UIST '08. Monterey, CA, USA: ACM, 2008, pp. 3–12.

[Gol07] D. R. Goldman. "A framework for video annotation, visualization, and interaction". PhD thesis. University of Washington, 2007.

[Gol10] R. Golijan. *GIzmodo: A 27-Year-Old Apple Tablet Prototype* `http://gizmodo.com/5455130/a-27+year+old-apple-tablet-prototype`. ONLINE. Last access: 30-10-2013. 2010.

[GCD11] G. Golovchinksy, S. Carter, and A. Dunnigan. "ARA: the active reading application". In: *Proceedings of the 19th ACM international conference on Multimedia*. MM '11. Scottsdale, AZ, USA: ACM, 2011, pp. 799–800.

[GCGIJCP04] R. Goularte, J. A. Camacho-Guerrero, V. R. Inacio Jr., R. G. Cattelan, and M. d. G. C. Pimentel. "M4Note: A Multimodal Tool for Multimedia Annotations". In: *Proceedings of the WebMedia & LA-Web 2004 Joint Conference 10th Brazilian Symposium on Multimedia and the Web 2nd Latin American Web Congress*. LA-WEBMEDIA '04 & LA-Web'04. Ribeirao Preto-SP, Brazil: IEEE Computer Society, 2004, pp. 142–149.

[Gra88] E. Gray. "Telautograph". Pat. US 386815. 1888.

[Gd82] J. Greene and M. d'Oliveira. *Learning to Use Statistical Tests in Psychology: A Student's Guide*. Milton Keynes, UK: Open University Press, 1982.

[GW59] B. Gurley and C. Woodward. "Light-pen links computer to operator". In: *Electronics* 32.47 (1959), pp. 85–87.

[GF03] S. Güven and S. Feiner. "Authoring 3D Hypermedia for Wearable Augmented and Virtual Reality". In: *Proceedings of the 7th IEEE International Symposium on Wearable Computers*. ISWC '03. Sanibel Island, FL, USA: IEEE Computer Society, 2003, pp. 118–126.

[GFO06] S. Güven, S. Feiner, and O. Oda. "Mobile augmented reality interaction techniques for authoring situated media on-site". In: *Proceedings of the 5th IEEE and ACM International Symposium on Mixed and Augmented Reality*. ISMAR '06. Santa Barbara, CA, USA: IEEE Computer Society, 2006, pp. 235–236.

[HHK08] J. Hagedorn, J. Hailpern, and K. G. Karahalios. "VCode and VData: illustrating a new framework for supporting the video annotation workflow". In: *Proceedings of the working conference on Advanced visual interfaces*. AVI '08. Napoli, Italy: ACM, 2008, pp. 317–321.

[HB92] B. L. Harrison and R. M. Baecker. "Designing video annotation and analysis systems". In: *Proceedings of the conference on Graphics interface '92*. Vancouver, BC, Canada: Morgan Kaufmann Publishers Inc., 1992, pp. 157–166.

[HGI11] K. Hasan, T. Grossman, and P. Irani. "Comet and target ghost: techniques for selecting moving targets". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. Vancouver, BC, Canada: ACM, 2011, pp. 839–848.

[HHL10] C. Heath, J. Hindmarsh, and P. Luff. *Video in Qualitative Research: Analysing Social Interaction in Everyday Life*. Introducing Qualitative Methods series. London, UK: SAGE Publications, 2010.

[HDYK11] J. Heikenfeld, P. Drzaic, J.-S. Yeo, and T. Koch. "Review Paper: A critical review of the present and future prospects for electronic paper". In: *Journal of the Society for Information Display* 19.2 (2011), pp. 129–156.

[HCB13] K. Hinckley, X. A. Chen, and H. Benko. "Motion and context sensing techniques for pen computing". In: *Proceedings of the 2013 Graphics Interface Conference*. GI '13. Regina, Sascatchewan, Canada: Canadian Information Processing Society, 2013, pp. 71–78.

[HYPCRWBB10] K. Hinckley, K. Yatani, M. Pahud, N. Coddington, J. Rodenhouse, A. Wilson, H. Benko, and B. Buxton. "Pen + touch = new tools". In: *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. UIST '10. New York, NY, USA: ACM, 2010, pp. 27–36.

[HZSBCST07] K. Hinckley, S. Zhao, R. Sarin, P. Baudisch, E. Cutrell, M. Shilman, and D. Tan. "InkSeine: In Situ search for active note taking". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. San Jose, CA, USA: ACM, 2007, pp. 251–260.

[HLZ03] X.-S. Hua, L. Li S., and H.-J. Zhang. "AVE: automated home video editing". In: *Proceedings of the eleventh ACM international conference on Multimedia*. MULTIMEDIA '03. Berkeley, CA, USA: ACM, 2003, pp. 490–497.

[HG08] W. Hürst and G. Götz. "Interface designs for pen-based mobile video browsing". In: *Proceedings of the 7th ACM conference on Designing interactive systems*. DIS '08. Cape Town, South Africa: ACM, 2008, pp. 395–404.

[Iph] *iPhone Human Interface Guidelines for Web Applications: User Experience.* Technical Report. Apple, 2010.

[Jam04] S. Jamieson. "Likert scales: how to (ab)use them". In: *Medical Education* 38.12 (2004), pp. 1217–1218.

[JKM07] T. Jokela, M. Karukka, and K. Mäkelä. "Mobile video editor: design and evaluation". In: *Proceedings of the 12th international conference on Human-computer interaction: interaction platforms and techniques*. HCI'07. Beijing, China: Springer-Verlag, 2007, pp. 344–353.

[JMK07] T. Jokela, K. Mäkelä, and M. Karukka. "Empirical observations on video editing in the mobile context". In: *Proceedings of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology*. Mobility '07. Singapore: ACM, 2007, pp. 482–489.

[KMM12] Z. Kalal, K. Mikolajczyk, and J. Matas. "Tracking - Learning - Detection". In: *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* 34.7 (2012), pp. 1409–1422.

[KAG10] R. Kannan, F. Andres, and C. Guetl. "DanVideo: an MPEG-7 authoring and retrieval system for dance videos". In: *Multimedia Tools and Applications* 46 (2 2010). 10.1007/s11042-009-0388-3, pp. 545–572.

[Kap96] J. Kaplan. *Startup: A Silicon Valley Adventure*. New York, NY, USA.: Penguin Books, 1996.

[KWLB08] T. Karrer, M. Weiss, E. Lee, and J. Borchers. "DRAGON: a direct manipulation interface for frame-accurate in-scene video navigation". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. Florence, Italy: ACM, 2008, pp. 247–250.

[KB99] H. Kato and M. Billinghurst. "Marker Tracking and HMD Calibration for a Video-Based Augmented Reality Conferencing System". In: *Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality*. IWAR '99. San Francisco, CA, USA: IEEE Computer Society, 1999, pp. 85–94.

[Kay72] A. C. Kay. "A Personal Computer for Children of All Ages". In: *Proceedings of the ACM annual conference - Volume 1*. ACM '72. Boston, MA, USA: ACM, 1972.

[Kip01] M. Kipp. "ANVIL - A Generic Annotation Tool for Multimodal Dialogue". In: *Proceedings of the 7th European Conference on Speech Communication and Technology*. Eurospeech'01. Aalborg, Danmark, 2001, pp. 1367–1370.

[Kna90] T. R. Knapp. "Treating ordinal scales as interval scales: an attempt to resolve the controversy." In: *Nursing Research* 39.2 (1990), pp. 121–123.

[LMZ05] K. Le, W. Mingxu, and W. Zhongcheng. "An overview of pen computing". In: *Proceedings of the 2005 IEEE International Conference on Information Acquisition*. ICIA 2005. Hong Kong and Macau, China: IEEE Computer Society, 2005, pp. 576–583.

[LHGL05] Y. Li, K. Hinckley, Z. Guan, and J. A. Landay. "Experimental analysis of mode switching techniques in pen-based user interfaces". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. CHI' 05. Portland, OR, USA: ACM, 2005, pp. 461–470.

[LGHH08] C. Liao, F. Guimbretière, K. Hinckley, and J. Hollan. "Papiercraft: A gesture-based command system for interactive paper". In: *ACM Transactions on Computer-Human Interaction* 14.4 (Jan. 2008), 18:1–18:27.

[Lie99] R. Lienhart. "Abstracting home video automatically". In: *Proceedings of the seventh ACM international conference on Multimedia (Part 2)*. MULTIMEDIA '99. Orlando, FL, USA: ACM, 1999, pp. 37–40.

[LG12] S. Liu and F. Guimbretière. "FlexAura: a flexible near-surface range sensor". In: *Proceedings of the 25th annual ACM symposium on User interface software and technology*. UIST '12. Cambridge, MA, USA: ACM, 2012, pp. 327–330.

[Low04] D. G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". In: *International Journal of Computer Vision* 60.2 (Nov. 2004), pp. 91–110.

[MP94] W. Mackay and D. Pagani. "Video mosaic: laying out time in a physical space". In: *Proceedings of the second ACM international conference on Multimedia*. MULTIMEDIA '94. San Francisco, CA, USA: ACM, 1994, pp. 165–172.

[Mac89] W. E. Mackay. "EVA: an experimental video annotator for symbolic analysis of video data". In: *SIGCHI Bulletin* 21.2 (1989), pp. 68–71.

[MD89] W. E. Mackay and G. Davenport. "Virtual video editing in interactive multimedia applications". In: *Communications of the ACM* 32.7 (1989), pp. 802–810.

[Mar10] C. C. Marshall. *Reading and writing the electronic book*. Morgan and Claypool Publishers, 2010.

[Mar97] C. C. Marshall. "Annotation: from paper books to the digital library". In: *Proceedings of the second ACM international conference on Digital libraries*. DL '97. Philadelphia, PA, United States: ACM, 1997, pp. 131–140.

[Mar98] C. C. Marshall. "Toward an ecology of hypertext annotation". In: *Proceedings of the ninth ACM conference on Hypertext and hypermedia*. HYPERTEXT '98. Pittsburgh, PA, USA: ACM, 1998, pp. 40–49.

[MPGS99] C. C. Marshall, M. N. Price, G. Golovchinsky, and B. N. Schilit. "Introducing a digital library reading appliance into a reading group". In: *Proceedings of the fourth ACM conference on Digital libraries*. DL '99. Berkeley, CA, USA: ACM, 1999, pp. 77–84.

[Mar12] T. Martin. "Interactive Quadruped Animation". MA thesis. University of California, Santa Barbara, 2012.

[MNAKS13] F. Matulic, M. C. Norrie, I. Al Kabary, and H. Schuldt. "Gesture-supported document creation on pen and touch tabletops". In: *CHI '13 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '13. Paris, France: ACM, 2013, pp. 1191–1196.

[Mey95] A. Meyer. "Pen computing: a technology overview and a vision". In: *ACM SIGCHI Bulletin* 27.3 (1995), pp. 46–90.

[Win] *Microsoft Windows 8* `http://windows.microsoft.com/en-us/windows-8/meet`. Online. Last access: 30-10-2013. 2013.

[MFAHIFFFJ11] G. Miller, S. Fels, A. Al Hajri, M. Ilich, Z. Foley-Fisher, M. Fernandez, and D. Jang. "MediaDiver: viewing and annotating multi-view video". In: *CHI '11 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '11. Vancouver, BC, Canada: ACM, 2011, pp. 1141–1146.

[Mog06] B. Moggridge. *Designing Interactions*. The MIT Press, 2006.

[NIKOS04] Y. Nakanishi, Y. Ishii, H. Koike, K. Oka, and Y. Sato. "Enhanced-Movie: Movie Editing on an Augmented Desk as a Large-Sized Display". In: *Adjunct Proceedings of the 17th annual ACM Symposium on User Interface Software and Technology*. UIST Adjunct Proceedings '04. UIST ARCHIVE. Santa Fe, NM, USA: ACM, 2004.

[NLSTB02] M. R. Naphade, C.-Y. Lin, J. R. Smith, B. L. Tseng, and S. Basu. "Learning to annotate video databases". In: *Proceedings of the SPIE Electronic Imaging'2002 - Storage and Retrieval for Media Databases Conference*. Vol. 4676. EI 2002. San Jose, CA, USA: SPIE, 2002, pp. 264–275.

[Nat11] National Academy of Television Arts and Sciences. *Outstanding Achievement in Technical/Engineering Development Awards:* `http://www.emmyonline.org/tech/applications/engineering_award_winners_rev8.pdf`. Online. Last access: 30-10-2013. 2011.

[Neb12] G. Nebehay. "Robust Object Tracking Based on Tracking-Learning-Detection". MA thesis. Faculty of Informatics, TU Vienna, 2012.

[NTM07] H. Neuschmied, R. Trichet, and B. Merialdo. "Fast annotation of video objects for interactive TV". In: *Proceedings of the 15th international conference on Multimedia*. MULTIMEDIA '07. Augsburg, Germany: ACM, 2007, pp. 158–159.

[Nor10] G. Norman. "Likert scales, levels of measurement and the "laws" of statistics". English. In: *Advances in Health Sciences Education* 15.5 (2010), pp. 625–632.

[PS76] M. J. Potel and R. E. Sayre. "Interacting with the GALATEA film analysis system". In: *ACM SIGGRAPH Computer Graphics* 10.2 (1976), pp. 52–59.

[PGS98] M. N. Price, G. Golovchinsky, and B. N. Schilit. "Linking by inking: trailblazing in a paper-like hypertext". In: *Proceedings of the ninth ACM conference on Hypertext and hypermedia : links, objects, time and space—structure in hypermedia systems: links, objects, time and space—structure in hypermedia systems*. HYPERTEXT '98. Pittsburgh, PA, USA: ACM, 1998, pp. 30–39.

[RLG08] N. C. Rahn, Y.-k. Lim, and D. P. Groth. "Redesigning video analysis: an interactive ink annotation tool". In: *CHI '08 extended abstracts on Human factors in computing systems*. CHI EA '08. Florence, Italy: ACM, 2008, pp. 3339–3344.

[RB03] G. Ramos and R. Balakrishnan. "Fluid interaction techniques for the control and annotation of digital video". In: *Proceedings of the 16th annual ACM symposium on User interface software and technology*. UIST '03. Vancouver, BC, Canada: ACM, 2003, pp. 105–114.

[RB05] G. Ramos and R. Balakrishnan. "Zliding: fluid zooming and sliding for high precision parameter manipulation". In: *Proceedings of the 18th annual ACM symposium on User interface software and technology*. UIST '05. Seattle, WA, USA: ACM, 2005, pp. 143–152.

[RBB04] G. Ramos, M. Boulos, and R. Balakrishnan. "Pressure widgets". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '04. Vienna, Austria: ACM, 2004, pp. 487–494.

[Rat03] D. Ratcliff. "Video methods in qualitative research." In: *Qualitative research in psychology: Expanding perspectives in methodology and design.* Washington, DC, US: American Psychological Association, 2003. Chap. 7, pp. 113–129.

[RS00] K. C. Redmond and T. M. Smith. *From Whirlwind to Mitre: The R&D Story of the Sage Air Defense Computer*. USA: MIT Press, 2000.

[Rei71] L. Reiffel. "Superimposed dynamic television display system". Pat. US 3617630. Nov. 1971.

[RD06] E. Rosten and T. Drummond. "Machine learning for high-speed corner detection". In: *Proceedings of the 9th European conference on Computer Vision - Volume Part I*. ECCV'06. Graz, Austria: Springer-Verlag, 2006, pp. 430–443.

[RRD05] E. Rosten, G. Reitmayr, and T. Drummond. "Real-Time video annotations for augmented reality". In: *Proceedings of the First international conference on Advances in Visual Computing*. ISVC'05. Lake Tahoe, NV, USA: Springer-Verlag, 2005, pp. 294–302.

[Rub91] D. Rubine. "Specifying gestures by example". In: *ACM SIGGRAPH Computer Graphics* 25.4 (1991), pp. 329–337.

[RMI04] K. Ryokai, S. Marti, and H. Ishii. "I/O brush: drawing with everyday objects as ink". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. CHI '04. Vienna, Austria: ACM, 2004, pp. 303–310.

[SCBS13] S. Santosa, F. Chevalier, R. Balakrishnan, and K. Singh. "Direct space-time trajectory control for visual media editing". In: *Proceedings of the 2013 ACM annual conference on Human factors in computing systems*. CHI '13. Paris, France: ACM, 2013, pp. 1149–1158.

[SGP98] B. N. Schilit, G. Golovchinsky, and M. N. Price. "Beyond paper: supporting active reading with free form digital ink annotations". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '98. Los Angeles, CA, USA: ACM Press/Addison-Wesley Publishing Co., 1998, pp. 249–256.

[SHMBJ10] K. Schoeffmann, F. Hopfgartner, O. Marques, L. Boeszoermenyi, and J. M. Jose. "Video browsing interfaces and applications: a review". In: *SPIE Reviews* 1.1 (2010), pp. 018004–1–018004–35.

[SH03] A. J. Sellen and R. H. Harper. *The Myth of the Paperless Office*. Cambridge, MA, USA: MIT Press, 2003.

[SW04] M. Shilman and Z. Wei. "Recognizing Freeform Digital Ink Annotations". In: *Document Analysis Systems VI*. Vol. 3163. LNCS, DAS 2004. Florence, Italy: Springer, 2004, pp. 322–331.

[SFCSFMKB11] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. "Real-time human pose recognition in parts from single depth images". In: *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR '11. Colorado Springs, CO, USA: IEEE Computer Society, 2011, pp. 1297–1304.

[Sil12] J. M. F. da Silva. "People and Object Tracking for Video Annotation". MA thesis. FCT/UNL, 2012.

[SCFC12] J. Silva, D. Cabral, C. Fernandes, and N. Correia. "Real-time annotation of video objects on tablet computers". In: *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia*. MUM '12. Ulm, Germany: ACM, 2012, 19:1–19:9.

[SLCL11] V. Singh, C. Latulipe, E. Carroll, and D. Lottridge. "The choreographer's notebook: a video annotation system for dancers and choreographers". In: *Proceedings of the 8th ACM conference on Creativity and cognition*. C&C '11. Atlanta, GA, USA: ACM, 2011, pp. 197–206.

[SA04] S. M. Smith and G. Albaum. *Fundamentals of Marketing Research*. Thousand Oaks, CA, USA: SAGE Publications, Inc, 2004.

[SBGICH11] H. Song, H. Benko, F. Guimbretiere, S. Izadi, X. Cao, and K. Hinckley. "Grips and gestures on a multi-touch pen". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. Vancouver, BC, Canada: ACM, 2011, pp. 1323–1332.

[Sut63] I. E. Sutherland. "Sketchpad: a man-machine graphical communication system". In: *Proceedings of the May 21-23, 1963, spring joint computer conference*. AFIPS '63 (Spring). Detroit, MI, USA: ACM, 1963, pp. 329–346.

[TH04] V. Thanedar and T. Höllerer. *Semi-automated placement of annotations in videos*. Technical Report. UCSB, 2004.

[TB09] R. Thompson and C. J. Bowen. *Grammar of the Edit (Second Edition)*. Second. Burlington, MA, USA: Focal Press, 2009.

[TM07] R. Trichet and B. Merialdo. "Generic object tracking for fast video annotation". In: *2nd International Conference on Computer Vision Theory and Applications*. Vol. 2. VISAPP 2007. Barcelona, Spain: INSTICC - Institute for Systems, Technologies of Information, Control, and Communication, 2007, pp. 419–426.

[TMDSK98] T. Tse, G. Marchionini, W. Ding, L. Slaughter, and A. Komlodi. "Dynamic key frame presentation techniques for augmenting video browsing". In: *Proceedings of the working conference on Advanced visual interfaces*. AVI '98. L'Aquila, Italy: ACM, 1998, pp. 185–194.

[TA08] T. Tullis and B. Albert. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Burlington, MA, USA: Morgan Kaufmann, 2008.

[Val11] J. G. Valente. "Sistema Multimodal para Captura e Anotação de Vídeo". MA thesis. FCT/UNL, 2011.

[VD04] C. Vaucelle and G. Davenport. "A System to Compose Movies for Cross-Cultural Storytelling: Textable Movie". In: *Proceedings of the Second International Conference on Technologies for Interactive Digital Storytelling and Entertainment*. Lecture Notes in Computer Science, TIDSE 2004. Darmstadt, Germany: Springer-Verlag, 2004, pp. 126–131.

[VDJ03] C. Vaucelle, G. Davenport, and T. Jehan. "Textable movie: improvising with a personal movie database". In: *ACM SIGGRAPH 2003 Sketches & Applications*. SIGGRAPH '03. San Diego, CA, USA: ACM, 2003, pp. 1–1.

[VADWF05] C. Vaucelle, D. Africano, G. Davenport, M. Wiberg, and O. Fjellstrom. "Moving pictures: looking out/looking in". In: *ACM SIGGRAPH 2005 Educators program*. SIGGRAPH '05. Los Angeles, CA, USA: ACM, 2005, 27:1–27:7.

[VI07] C. Vaucelle and H. Ishii. "Interfacing video capture, editing and publication in a tangible environment". In: *Proceedings of the 11th IFIP TC 13 international conference on Human-computer interaction - Volume Part II*. INTERACT'07. Rio de Janeiro, Brazil: Springer-Verlag, 2007, pp. 1–14.

[VBGLS+11] C. Viviani, M. Baptiste, J. A. Gili, L. Logette, D. Sauvaget, et al. *Dictionnaire mondial du Cinèma*. Paris, France: Larousse, 2011.

[VB10] D. Vogel and R. Balakrishnan. "Direct Pen Interaction With a Conventional Graphical User Interface". In: *Human-Computer Interaction* 25.4 (2010), pp. 324–388.

[WYWH99] S.-T. Wang, M.-L. Yu, C.-J. Wang, and C.-C. Huang. "Bridging the gap between the pros and cons in treating ordinal scales as interval scales from an analysis point of view". In: *Nursing research* 48.4 (1999), pp. 226–229.

[WH06] Y. Wang and M. Hirakawa. "Video editing based on object movement and camera motion". In: *Proceedings of the working conference on Advanced visual interfaces*. AVI '06. Venezia, Italy: ACM, 2006, pp. 108–111.

[WC07] H. Weda and M. Campanella. "Use study on a home video editing system". In: *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...but not as we know it - Volume 2*. BCS-HCI '07. Lancaster, United Kingdom: British Computer Society, 2007, pp. 123–126.

[WP94] K. Weher and A. Poon. "Marquee: a tool for real-time video logging". In: *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*. CHI '94. Boston, MA, USA: ACM, 1994, pp. 58–64.

[Wei08] D. E. Weisberg. *The Engineering Design Revolution: The People, Companies and Computer Systems That Changed Forever the Practice of Engineering* `http://www.cadhistory.net/`. ONLINE Book, Last Access: 30-10-2013. 2008.

[Whi03] R. White. *Prehistoric Art: The Symbolic Journey of Humankind*. New York, NY, USA: Harry N. Abrams, Inc, 2003.

[WMYL12] Y. Wu, T. Mei, N. Yu, and S. Li. "Accelerometer-based single-handed video browsing on mobile devices: design and user studies". In: *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service*. ICIMCS '12. Wuhan, China: ACM, 2012, pp. 157–160.

[YY97] B.-L. Yeo and M. M. Yeung. "Retrieving and visualizing video". In: *Communications of the ACM* 40.12 (1997), pp. 43–52.

[YJS06] A. Yilmaz, O. Javed, and M. Shah. "Object tracking: A survey". In: *ACM Computing Surveys* 38.4 (Dec. 2006), 13:1–13:45.

[You] *Youtube Statistics:* `http://www.youtube.com/t/press_statistics`. Onlne. Last access: 11-11-2012. 2012.

[ZHSJ07] J. Zigelbaum, M. S. Horn, O. Shaer, and R. J. K. Jacob. "The tangible video editor: collaborative video editing with active tokens". In: *Proceedings of the 1st international conference on Tangible and embedded interaction*. TEI '07. Baton Rouge, LA, USA: ACM, 2007, pp. 43–46.

[ZWLS12] T. Zimmermann, M. Weber, M. Liwicki, and D. Stricker. "CoVidA: pen-based collaborative video annotation". In: *Proceedings of the 1st International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications*. VIGTA '12. Capri, Italy: ACM, 2012, 10:1–10:6.

[ZH06] Z. Zivkovic and F. van der Heijden. "Efficient adaptive density estimation per image pixel for the task of background subtraction". In: *Pattern Recognition Letters* 27.7 (May 2006), pp. 773–780.

Appendix: Creation-Tool Questionnaire

1. **Consider two types of scenarios, during a rehearsal (in video capture) and after a rehearsal (after video capture). Please, compare and classify how often are you willing to use the Creation-Tool, in each scenario:**

	Rarely				Frequently
During a rehearsal	1	2	3	4	5
After a rehearsal	1	2	3	4	5

2. **Consider the different annotation types of the Creation-Tool. Please, compare and classify how often are you willing to use each type, during a rehearsal.**

	Rarely				Frequently
Sketch	1	2	3	4	5
Text	1	2	3	4	5
Audio	1	2	3	4	5
Marks	1	2	3	4	5
Hyperlinks	1	2	3	4	5

3. **Consider the different annotation types of the Creation-Tool. Please, compare and classify how often are you willing to use each type, after a rehearsal.**

	Rarely				Frequently
Sketch	1	2	3	4	5
Text	1	2	3	4	5
Audio	1	2	3	4	5
Marks	1	2	3	4	5
Hyperlinks	1	2	3	4	5

4. **Consider the insertion of different annotation types of the Creation-Tool. Please, compare and classify them.**

	Difficult				Easy
Sketch	1	2	3	4	5
Text	1	2	3	4	5
Audio	1	2	3	4	5
Marks	1	2	3	4	5
Hyperlinks	1	2	3	4	5

5. **Consider the continuous and suspended annotation modes of the Creation-Tool. Please, compare and classify how often are you willing to use each mode.**

	Rarely				Frequently
Continuous Mode	1	2	3	4	5
Suspended Mode	1	2	3	4	5

6. **Consider the continuous and suspended annotation modes of the Creation-Tool. Please, compare and classify them.**

	Difficult				Easy
Continuous Mode	1	2	3	4	5
Suspended Mode	1	2	3	4	5

7. **Consider the real-time and delayed video modes of the Creation-Tool. Please, compare and classify how often are you willing to use each mode, during a rehearsal.**

	Rarely				Frequently
Real-Time Mode	1	2	3	4	5
Delayed Mode	1	2	3	4	5

8. **Consider your interaction with the Creation-Tool. Please, classify it.**

Difficult				Easy
1	2	3	4	5

9. Choose the expression(s) that better classifies your experience with the Creation-Tool.

☐ Annoying	☐ Essential	☐ Innovative	☐ Simplistic
☐ Attractive	☐ Exceptional	☐ Inspiring	☐ Slow
☐ Clean	☐ Exciting	☐ Intuitive	☐ Sterile
☐ Clear	☐ Familiar	☐ Irrelevant	☐ Stimulating
☐ Complex	☐ Fast	☐ Meaningful	☐ Straight Forward
☐ Comprehensive	☐ Flexible	☐ Motivating	☐ Stressful
☐ Confusing	☐ Friendly	☐ Novel	☐ Time-consuming
☐ Creative	☐ Frustrating	☐ Old	☐ Time-Saving
☐ Distracting	☐ Hard to Use	☐ Organized	☐ Too Technical
☐ Easy to use	☐ Helpful	☐ Powerful	☐ Unattractive
☐ Efficient	☐ Impressive	☐ Relevant	☐ Unconventional
☐ Effortless	☐ Incomprehensible	☐ Satisfying	☐ Useful

10. Comments and Suggestions:

_____

_____

_____

_____

_____

_____

_____

11. **Do you usually record your work?** ☐ Yes ☐ No

If yes, in which media?
- ☐ Paper Notebook
- ☐ Audio
- ☐ Video
- ☐ Other: _____

12. **Do you usually annotate during your work process?** ☐ Yes ☐ No

If yes, in which technology?
- ☐ Paper Notebook
- ☐ Desktop Computer
- ☐ Laptop Computer
- ☐ Mobile Phone
- ☐ PDA
- ☐ Tablet
- ☐ Other: _____

13. **Do you usually share your work documents?** ☐ Yes ☐ No

If yes, in which platform?
- ☐ Mail (Post)
- ☐ Fax
- ☐ Hardware (Sharing CDs, DVDs, Pen Drives, etc..)
- ☐ E-mail
- ☐ Web Site (personal site, blog, etc..)
- ☐ Instant Messaging (Skype, MSN, etc..)
- ☐ Social Network (Facebook, Myspace, etc..)
- ☐ Other: _____

14. **Were you familiar with pen-based technology (Tablets or PDAs) before:** ☐ Yes ☐ No

15. **Were you familiar with touch-based technology (Tablets or Touch Phones) before:** ☐ Yes ☐ No

16. **Age:**_____

17. **Genre:** ☐ Masculine ☐ Feminine

18. **Education:** ☐ Elementary School ☐ High School ☐ Bachelor ☐ Master ☐ PhD

19. **Handedness:** ☐ Left ☐ Right

B

Appendix: Creation-Tool and Motion Tracking Questionnaire

1. **Considering the two trackers, A (Kinect) and B (TLD), and their performance on tracking people. Please classify the performance of each tracker:**

	Poor				Excellent
A	1	2	3	4	5
B	1	2	3	4	5

2. **Consider the two techniques of video annotation with motion tracking, Hold & Overlay and Hold and Speed Up, recording a live event:**
   Hold & Overlay – the video is paused during the task of the annotation and a half-transparent video window is overlaid, showing the live event and until the task ends.
   Hold & Speed Up – the video is paused during the task of the annotation and played it after, in fastfoward, until the video is synchronized with the live event.
   Please, classify how often are you willing to use each type.

	Rarely				Frequently
Hold & Overlay	1	2	3	4	5
Hold & Speed Up	1	2	3	4	5

3. **Consider the two techniques of video annotation with motion tracking, Hold & Overlay and Hold and Speed Up, as well as the tracking methods A (Kinect) and B (TLD). Please, classify them.**

Annot. Method	Tracker	Difficult				Easy
Hold & Overlay	A	1	2	3	4	5
	B	1	2	3	4	5
Hold & Speed Up	A	1	2	3	4	5
	B	1	2	3	4	5

4. **Consider the task of associating an annotation with a tracked object or person. Please, classify it.**

Difficult				Easy
1	2	3	4	5

5. **Consider the annotation with motion tracking. Please, classify it.**

Difficult				Easy
1	2	3	4	5

6. **Choose the expression(s) that better classifies your experience with the Creation-Tool & Motion Tracking.**

☐ Annoying	☐ Exceptional	☐ Inspiring	☐ Stimulating
☐ Attractive	☐ Exciting	☐ Intuitive	☐ Straight Forward
☐ Complex	☐ Familiar	☐ Irrelevant	☐ Stressful
☐ Comprehensive	☐ Frustrating	☐ Meaningful	☐ Time-consuming
☐ Confusing	☐ Hard to Use	☐ Motivating	☐ Time-Saving
☐ Creative	☐ Helpful	☐ Novel	☐ Too Technical
☐ Distracting	☐ Impressive	☐ Old	☐ Unattractive
☐ Easy to use	☐ Incomprehensible	☐ Powerful	☐ Unconventional
☐ Essential	☐ Innovative	☐ Relevant	☐ Useful

7. **Comments and Suggestions:**

_____

_____

_____

_____

_____

_____

_____

8. **Do you usually record your work?** ☐ Yes ☐ No

   If yes, in which media?
   - ☐ Paper Notebook
   - ☐ Audio
   - ☐ Video
   - ☐ Other: _____

9. **Do you usually annotate during your work process?** ☐ Yes ☐ No

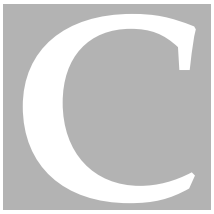
   If yes, in which technology?
   - ☐ Paper Notebook
   - ☐ Desktop Computer
   - ☐ Laptop Computer
   - ☐ Mobile Phone
   - ☐ PDA
   - ☐ Tablet
   - ☐ Other: _____

10. **Were you familiar with pen-based technology (Tablets or PDAs) before:** ☐ Yes ☐ No

11. **Age:**_____

12. **Gender:** ☐ Masculine ☐ Feminine

13. **Education:** ☐ Elementary School ☐ High School ☐ Bachelor ☐ Master ☐ PhD



Appendix: VideoInk Questionnaire

1. **Please, consider the different types of adding video content and classify them.**

	Difficult				Easy
Video Frame	1	2	3	4	5
Video Clip/Segment	1	2	3	4	5
Transition (Fade)	1	2	3	4	5

2. **Consider the two techniques for switching between video segments and video frames: pressure-based and button-based. Please, classify how often are you willing to use each type.**

	Rarely				Frequently
Pressure	1	2	3	4	5
Buttons	1	2	3	4	5

3. **Consider the two techniques for switching between video segments and video frames: pressure-based and button-based. Please, classify them.**

	Difficult				Easy
Pressure	1	2	3	4	5
Buttons	1	2	3	4	5

4. **Please, consider the different techniques for creating the new video (No Selection, Paint Selection and Lasso Selection) and classify them.**

	Difficult				Easy
No Selection	1	2	3	4	5
Paint Selection	1	2	3	4	5
Lasso Selection	1	2	3	4	5

5. Consider the two techniques of selecting video content: Paint Selection and Lasso. Please, classify how often are you willing to use each type.

	Rarely				Frequently
Paint Selection	1	2	3	4	5
Lasso Selection	1	2	3	4	5

6. Please, consider the zoom pressure-based mechanism and classify it.

Difficult				Easy
1	2	3	4	5

7. Considering the zoom pressure-based mechanism and other traditional zoom methods, please, classify how often are you willing to use each one.

	Rarely				Frequently
Pressure	1	2	3	4	5
Slider (-)-----|-----(+)	1	2	3	4	5
Two Buttons (-) (+)	1	2	3	4	5

8. Please rate your agreement with the following statements:

I. I was satisfied with what I got out the system or tool.

Highly Disagree	0	1	2	3	4	5	6	7	8	9	10	Highly Agree

II. It was easy for me to explore many different ideas, outcomes, and possibilities.

Highly Disagree	0	1	2	3	4	5	6	7	8	9	10	Highly Agree

III. I would be happy to use this system or tool on a regular basis.

Highly Disagree	0	1	2	3	4	5	6	7	8	9	10	Highly Agree

IV. I was able to be very creative while doing this activity.

Highly Disagree	0	1	2	3	4	5	6	7	8	9	10	Highly Agree

V. My attention was fully tuned to the activity, and I forgot about the system or tool that I was using.

Highly Disagree	0	1	2	3	4	5	6	7	8	9	10	Highly Agree

VI. The system was helpful in allowing me to track different ideas, outcomes, or possibilities.

Highly Disagree	0	1	2	3	4	5	6	7	8	9	10	Highly Agree

VII. I enjoyed this system or tool.

Highly Disagree	0	1	2	3	4	5	6	7	8	9	10	Highly Agree

VIII. What I was able to produce was worth the effort I had to exert to produce it.

Highly Disagree	0	1	2	3	4	5	6	7	8	9	10	Highly Agree

IX. The system or tool allowed me to be very expressive.

Highly Disagree	0	1	2	3	4	5	6	7	8	9	10	Highly Agree

X. I became so absorbed in the activity that I forgot about the system or tool that I was using.

Highly Disagree	0	1	2	3	4	5	6	7	8	9	10	Highly Agree

9. When using this system or tool, it's most important that I'm able to...

Be creative and expressive		or		Produce results that are worth the effort I put in
Enjoy using the system or tool		or		Become immersed in the activity
Become immersed in the activity		or		Produce results that are worth the effort I put in
Produce results that are worth the effort I put in		or		Explore many different ideas, outcomes, or possibilities
Become immersed in the activity		or		Be creative and expressive
Be creative and expressive		or		Enjoy using the system or tool
Explore many different ideas, outcomes, or possibilities		or		Become immersed in the activity
Produce results that are worth the effort I put in		or		Enjoy using the system or tool
Explore many different ideas, outcomes, or possibilities		or		Be creative and expressive
Enjoy using the system or tool		or		Explore many different ideas, outcomes, or possibilities

10. Choose the expression(s) that better classifies your experience with the Video Ink Tool.

	Annoying		Creative		Incomprehensible		Straight Forward
	Attractive		Easy to use		Inspiring		Stressful
	Boring		Efficient		Intuitive		Time-consuming
	Clear		Familiar		Novel		Time-Saving
	Complex		Frustrating		Old		Too Technical
	Comprehensive		Fun		Powerful		Unattractive
	Confusing		Hard to Use		Sterile		Useful

11. Comments and Suggestions:

_____

_____

_____

_____

_____

_____

_____

12. Do you usually record video? ☐ Yes ☐ No

If yes, using which device?
- ☐ Video/Photo Camera
- ☐ Mobile Phone
- ☐ Tablet
- ☐ Other: _____

13. If you usually record videos, do you usually edit them? ☐ Yes ☐ No

If no, why?
- ☐ It's a boring task.
- ☐ I don't care about editing.
- ☐ It's a difficult task.
- ☐ Other: _____

14. Ever used pen-based technology before? ☐ Never ☐ Occasionally ☐ Frequently

15. Age:_____

16. Gender: ☐ Masculine ☐ Feminine

17. Education: ☐ Elementary School ☐ High School ☐ Bachelor ☐ Post-Grad ☐ Master ☐ PhD