Kalle Wallenius

Video Annotation for Studying the Brain in Naturalistic Settings

Faculty of Electronics, Communications and Automation

Thesis submitted for examination for the degree of Master of Science in Technology in

Espoo, 5 May 2010.

Instructor: Enrico Glerean, M. Sc.

Supervisor: Prof. Mikko Sams

Aalto University
School of Science
and Technology

AALTO UNIVERSITY
SCHOOL OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS
ABSTRACT

Author: Kalle Wallenius

Thesis Title: Video annotation for studying the brain in naturalistic settings

Date: 2010-05-05 Language: English Number of Pages: 8 + 57

Faculty of Electronics, Communications and Automation

Professorship: Cognitive technology Code: S-114

Supervisor: Prof. Mikko Sams

Instructor: Enrico Glerean, M.Sc.

Studying the brain in naturalistic settings is a recent trend in neuroscience. Traditional brain imaging experiments have relied on using highly simplified and artificial stimuli, but recently efforts have been put into studying the human brain in conditions closer to real-life. The methodology used in these studies involve imitating naturalistic stimuli with a movie.

Because of the complexity of the naturalistic stimulus, a simplified model of it is needed to handle it computationally. This model is obtained by making annotations; collecting information of salient features of the movie to form a data structure. This data is compared with the brain activity evolving in time to search for possible correlations. All the features of a movie cannot be reliably annotated automatically: semantic features of a movie require manual annotations, which is in some occasions problematic due to the various cinematic techniques adopted. Understanding these methods helps analyzing and annotating movies.

The movie Match Factory Girl (Aki Kaurismäki, 1990) was used as a stimulus in studying the brain in naturalistic settings. To help the analysis of the acquired data the salient visual features of the movie were annotated. In this work existing annotation approaches and available technologies for annotation were reviewed.

Annotations help organizing information, therefore they are nowadays found everywhere. Different tools and technologies are being developed constantly. Furthermore, development of automatic video analysis methods are going to provide more meaningful annotations in the future.

Keywords: Annotation, modeling, movie, fMRI, ICA

AALTO-YLIOPISTO TEKNILLINEN KORKEAKOULU	DIPLOMITYÖN TIIVISTELMÄ

Tekijä: Kalle Wallenius

Työn nimi: Videon annotointi aivojen tutkimiseksi luonnollisessa asetelmissa

Päivämäärä: 05.05.2010 Kieli: Englanti Sivumäärä: 8 + 57

Elektroniikan, tietoliikenteen ja automaation tiedekunta

Professuuri: Kognitiivinen teknologia Koodi: S-114

Valvoja: Prof. Mikko Sams

Ohjaaja: DI Enrico Glerean

Aivojen tutkiminen luonnollisissa asetelmissa on viimeaikainen suunta aivotutkimuksessa. Perinteisesti aivotutkimuksessa on käytetty hyvin yksinkertaistettuja ja keinotekoisia ärsykkeitä, mutta viime aikoina on alettu tutkia ihmisaivoja yhä luonnollisimmissa asetelmissa. Näissä kokeissa on käytetty elokuvaa luonnollisena ärsykkeenä.

Elokuvan monimutkaisuudesta johtuen tarvitaan siitä yksinkertaistettu malli laskennallisen käsittely mahdollistamiseksi. Tämä malli tuotetaan annotoimalla; keräämällä elokuvan keskeisistä ärsykepiirteistä dataa tietorakenteen muodostamiseksi. Tätä dataa verrataan aivojen aikariippuvaiseen aktivaatioon etsittäessä mahdollisia korrelaatiota.

Kaikkia elokuvan ominaisuuksia ei pystytä annotoimaan automaattisesti; ihmiselle merkitykselliset ominaisuudet on annotoitava käsin, joka on joissain tapauksissa ongelmallista johtuen elokuvan käyttämistä useista viestintämuodoista. Ymmärrys näistä viestinnän muodoista auttaa analysoimaan ja annotoimaan elokuvia.

Elokuvaa Tulitikkutehtaan Tyttö (Aki Kaurismäki, 1990) käytettiin ärsykkeenä aivojen tutkimiseksi luonnollisissa asetelmissa. Kokeista saadun datan analysoinnin helpottamiseksi annotoitiin elokuvan keskeiset visuaaliset ärsykepiirteet. Tässä työssä tutkittiin annotointiin käytettävissä olevia eri lähestymistapoja ja teknologioita.

Annotointi auttaa informaation organisoinnissa, mistä syystä annotointia ilmestyy nykyään kaikkialla. Erilaisia annotaatiotyökaluja ja -teknologioita kehitetään jatkuvasti. Lisäksi videoanalyysimenetelmät ovat alkaneet mahdollistaa yhä merkityksellisemmän informaation automaattisen annotoinnin tulevaisuudessa.

Avainsanat: Annotaatio, mallinnus, elokuva, fMRI, ICA

Foreword

This thesis work was done at the department of Biomedical Engineering and Computational Science in Aalto University School of Science and Technology. I would like to thank my supervisor Prof. Mikko Sams for this opportunity to participate in this unique project. I also want to express my gratitude to my instructor Enrico Glerean, M.Sc. for providing me with all the valuable advices and ideas.

Finally, huge thanks to my family and friends for all the support, and very special thanks to my beloved common-law wife Tiina for keeping me sane.

Helsinki, May 5, 2010

Kalle Wallenius

Table of contents

Symbols and abbreviations

Symbols

λ	Gyromagnetic constant
ω_0	Larmor frequency

Abbreviations

BECS	Biomedical Engineering and Computational Science
BOLD	Blood Oxygen-Level Dependent
BSS	Blind Source Separation
DC	Dublin Core
DCMI	Dublin Core Metadata Initiative
FMRI	Functional Magnetic Resonance Imaging
ICA	Independent Component Analysis
IC	Independent Component
ISC	Inter-Subject Correlation
MNI	Montreal Neurological Institute
MPEG	Motion Picture Experts Group
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
NMR	Nuclear Magnetic Resonance
OWL	Net Ontology Language
PICA	Probabilistic Independent Component Analysis
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RF	Radio Frequency
TR	Repetition Time

W3C	World Wide Web Consortium
WWW	World Wide Web
XML	Extensible Markup Language

1 Introduction

1.1 Background

Studying the brain in naturalistic settings is a recent trend in neuroscience. Its aim is to gain understanding about the brain activation in every day life by utilizing stimuli closer to real life experiences. Conventionally neuroscientific studies have concentrated on studying very specific phenomena by using highly restricted stimuli in strictly controlled conditions. However, stimuli in every day life are only seldom simple, thus this kind of experiments do not necessarily tell much about complex brain activity in real life. For this reason studying brain in naturalistic settings involves highly complex stimuli in less controlled experimental settings.

The methodology used in these studies involve a fairly new branch of science that connects the disciplines of film studies and neuroscience. Simply put, the basic method is to detect significant brain activation of the subjects while they are watching a movie. The brain activity during the viewing is recorded. Afterwards the data is compared to the events of the stimulus movie in order to make conclusions about the relations between stimuli and response. The goals are to find out if some event in the movie elicited activation in specific areas of the brain and to understand the dynamics of brain processes involved during the perception of real-life situations.

This thesis work is part of neuroscientific research project AivoAALTO conducted at Aalto University department of Biomedical Engineering and Computational Science (BECS). AivoAALTO is a project of the Aalto University, which aims to research social interaction, social decision making and the influence of a movie to the human mind.

1.2 The research problem

Because of the complexity of the natural stimuli, a simplified model is needed in order to gain understanding how it affects the brain activity. The model consists of descriptions of the stimulus movie and is obtained by making *annotations*. In this context annotating means that descriptions of salient events of the movie, with the times of their occurrences, are collected to form a data structure. This data is then compared with the brain activity evolving in time to search for possible correlations.

While at first glance the annotation of a movie might seem a trivial problem, we are going to see that it certainly is not. A movie is a very complex spatiotemporal audiovisual stimulus whose annotation is in many cases subjected to personal interpretation of the annotator. The quality of a given annotation depends also on the purpose of the annotation. There are many possible ways of performing the annotation of a movie and so far there is no standard procedure for the process. This thesis makes a contribution to this problem – the aim is to clarify essential viewpoints for annotations to study the brain in naturalistic conditions.

1.3 Goals of this thesis work

This thesis work is a contribution to the annotation part of the brain studies currently being carried. This thesis has two main goals:

1. Make a profound introduction to the field of annotations
2. Annotate visual features of the stimulus movie used in our neuroscientific studies.

In order to carry out the actual annotation, the field of annotations is studied in more depth. As the dive into the realm of annotations in general is quite shallow, a deeper one is taken into the field of movie annotation for our specific purposes.

1.4 Thesis outline

Chapter 2 discusses the background of this work. It introduces the benefits of studying the brain under naturalistic conditions and clarifies essential concepts. A brief description of the technology and the methods used in the studies is given.

Chapters 3 and 4 deal with the theoretical part of this thesis work. The 3rd chapter makes an extensive introduction to annotations – a clarification is given about what annotations are, what features they possess, how they can be used, and what kind of problems are related to annotating. As it covers a variety of annotation types, the possibilities and benefits of annotations, it also introduces state-of-the-art annotation technologies. The main focus of the chapter is in the visual feature annotation of a movie. Examination of the theory and possible ways of executing the actual annotation is also given. Chapter 4 describes a survey of the current state video annotation tools. This is done to find suitable framework to use in brain research. A short assessment of the tools is performed.

The practical part of this thesis work, annotation of the visual features of the movie is described in chapters 5. Chapter 6 discusses and concludes the thesis work.

2 Neurocinematics: Studying brain activation in naturalistic viewing conditions

2.1 Introduction

This chapter discusses the essential background and the methods used to study the brain under naturalistic viewing conditions. The chapter is based on the earlier work by Bartels and Zeki (2004a, 2004b, 2005), Hasson et al. (2004, 2005, 2008, 2009) and Jääskeläinen et al. (2008).

Functional neuroimaging attempts to localize different mental processes to different parts of the brain. The quest for an understanding of the functional organization of the human brain has been under way for a long time. Present- day researchers are still mapping the human brain with various methods.

Neurocinematics - studying brain activation during natural viewing conditions - is a fairly new way of assessing the effect of a given visual stimulus on viewers' brain. In its core essence the method is to record subjects' brain activations while freely viewing a movie. First experiments of this kind were conducted by Bartels and Zeki (2004b) and Hasson et al. (2004) The field of study was named later by Hasson et al. (2008) “Neurocinematics: the neuroscience of film”. As this implicates, neurocinematics can be used to study both film and neuroscience. In recent studies of neurocinematics, the state-of-the-art method for brain imaging has been the functional magnetic resonance imaging (fMRI).

The main advantage of neurocinematics is that it provides means to study the functioning of the human brain in close-to-every-day-life conditions. In fact this characteristic provides various advantages over conventional methods for functional neuroimaging. Traditionally brain imaging research has concentrated on studying only very specific areas of brain functioning at a time by using highly simplified and artificial stimuli. Commonly, static and isolated object images are briefly flashed on the screen during a carefully maintained fixation. Precisely parameterized stimuli usage has been critical for isolating relevant dimensions out of the various dimensions that natural stimuli possesses. However, this kind of stimuli differs drastically from natural vision (Hasson & Malach 2005).

In natural vision objects are not presented in spatial isolation, but are part of a complex scene which can include various object categories. Objects also move continuously within the scene. In addition, vision usually has interaction with other modalities, context and emotional valence of the stimuli. From this it is evident that controlled experiment settings are far away from a naturalistic viewing experiment (Hasson & Malach 2005).

Brain has evolved to deal with the dynamic complexity of the natural conditions, thus it is likely that studying brain in settings as natural as possible may reveal important principles of functional and structural brain organization. There is evidence that neural responses evoked with natural stimuli are more sparse and statistically independent than those evoked with artificial stimuli (Hasson et al. 2009). Studies by Bartels and Zeki

(2004a, 2004b) revealed that functional specificity of distinct brain regions maintain or potentially even enhance during natural viewing conditions. They showed that dynamic faces and movies activate brain regions that have not been activated in traditional studies with static pictures of faces. As specialization was shown to preserve when many features had to be processed simultaneously, they concluded that each specialized area is responsible of creation of feature-specific conscious experience, thus making this a method to study construction of consciousness as well.

More complex phenomena are only researchable within real-life natural contexts. For example phenomena that unveil only after relatively long period of time, play a major role in our everyday life. Traditionally used method of restricting the temporal dimension of a stimuli can impede this kind of research, that includes studying for example social interaction, engaging in dialogues and reacting to surrounding events. Thus, temporally restricting the stimuli used in experiments drastically narrows the scope of the phenomena to be studied. Also it has been shown that brain responses to complex stimuli are not always similar to combination of simplified stimuli (Hasson et al. 2009).

The use of natural-like stimuli breaks free from the traditional experiment restrictions and thus enables the researchers to study phenomena that have not been possible to study using conventional methods. Naturalistic settings may elicit broader set of connections, thus enabling to study phenomena, emotions for instance, as they occur naturally. As more naturalistic stimuli, a movie sequence is used with minimum amount of control over the viewing situation. A movie sequence consists of both visual and aural information with various levels of semantics.

There are of course drawbacks in this kind of method. The data analysis becomes much more complex compared to classical experimental settings. Because of the variety of the stimuli, conventional hypothesis-driven approach for analyzing the fMRI data is mostly of no use for the analysis (Hasson & Malach 2005). Instead, a data-driven approach is more convenient in this kind of studies. In data-driven analysis the researcher explores the structure of the data for task related activations. The data driven analysis introduces new possible methods to be used: Inter-subject correlation (ISC) and independent component analysis (ICA). With ISC it is possible to compare brain activation in one subject to the activation of another subject. It has been used to reveal, how similar brain activation a stimulus elicits across subjects. ICA is used to find brain activation patterns that are statistically independent from the the rest of the activity. These patterns are called independent components (IC). Each independent component (IC) forms a time series. A comparison between temporal elements of a movie with these ICs can reveal correlation between certain kind of stimuli and responses in brain activation. Thus, ICA can be used to reveal localization. It also might reveal the degree of similar brain activation between many subjects, which is also analyzed with intersubject correlation (ISC).

Recent studies

In their first study Hasson et al. (2004) let their subjects freely view a movie while the subjects' brain activity was being recorded with fMRI. They used the film *The Good, Bad and the Ugly* by Sergio Leone (1968) as a stimulus. They examined how the movie elicited similar brain responses across different subjects. For the data analysis they used inter-subject correlation (ISC), in which activation patterns of a given reference subject were used to predict activation in another subjects. They found that especially in primary visual and auditory cortices of the brain activity was very similar across subjects. They also studied stimulus specificity of certain brain regions: a well-known brain region that responds selectively to faces, *fusiform gyrus*, was activated during movie shots containing faces. They were also able to detect that highest activation in the *collateral sulcus* was elicited especially by shots containing sceneries or buildings

Another study was conducted by Bartels & Zeki (2004b) where they studied connectivity between different areas in the human brain during viewing a movie *Tomorrow Never Dies* directed by Roger Spottiswoode (1997). By using ICA they were able to map different functional subsystems of the brain into separate independent components, most of which were described as visual or auditory. The independent components showed also high degree of inter-subject correlation.

In a more recent study by Jääskeläinen et al. (2008) subjects were presented with the movie *Crash* directed by Paul Haggis (2005) – The first 72 minutes of the movie were viewed outside of the scanner in order to create the context for the subjects. The last 36 minutes of the movie were then shown in the fMRI scanner. The purpose was to study whether inter-subject correlation of the prefrontal cortex would occur when subjects would be emotionally engaged in the movie, while earlier studies lacked inter-subject synchronization in these areas (Hasson et al. 2004). They were able to detect significant inter-subject correlation especially in right hemisphere frontal cortical areas, in addition to temporal, occipital and parietal areas. Meaningful activation in individual subjects during viewing was also revealed by utilizing probabilistic ICA.

2.2 Methodology

2.2.1 Functional magnetic resonance imaging (fMRI)

In recent research of neurocinematics, the most applied method has been functional magnetic resonance imaging, fMRI. This short introduction on fMRI is based on the work of Huettel et al. (2004), unless otherwise stated.

Magnetic Resonance Imaging (MRI)

Magnetic Resonance Imaging (MRI) is a widely used non-invasive imaging method. It is based on detecting changes in magnetic properties of atomic nuclei elicited by oscillating magnetic field applied at the resonant frequency of the nucleus. Many atomic nuclei have an angular momentum which arises from their inherent property called *spin*. Because nucleus is electrically charged, its spin generates a small magnetic field. Spins contributed by protons and neutrons result in a net spin. Only nuclei with odd number of spins exhibit a net spin, because their protons and neutrons do not cancel each other out. In the absence of a strong magnetic field the spins are oriented randomly (Figure 1a) and tend to cancel each other out. In a external magnetic field nuclei with net spin align themselves to the direction of applied field (Figure 1b), either parallel or anti-parallel to it. The majority of nuclei align in parallel direction.

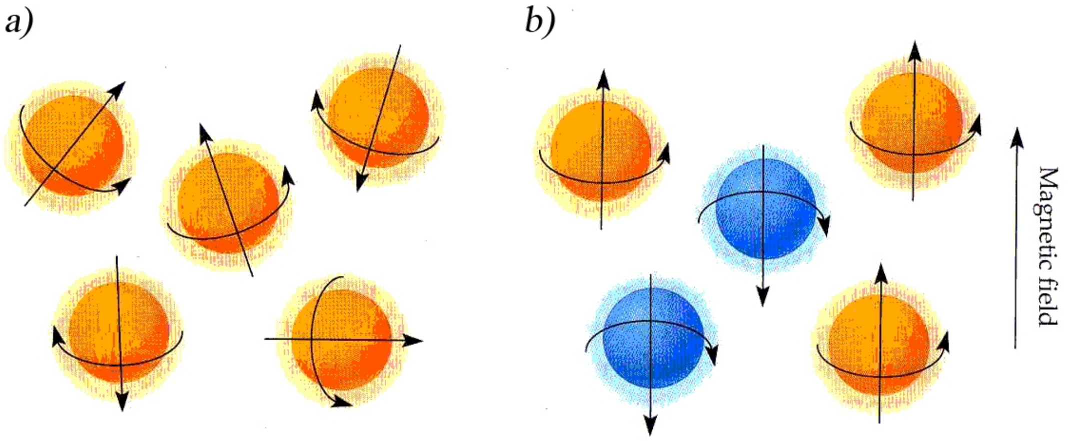


Figure 1. *Spins of a nuclei in free space with random orientation* (Huettel et al. 2004)

In the presence of the external magnetic field B_0 nuclei starts to precess around the direction of field with an angular frequency of ω_0 , called the Larmor frequency, which is defined by the Larmor equation

$$\omega_0 = B_0 \times \lambda \quad , \qquad (1)$$

where λ is a gyromagnetic ratio, constant to every atom in a particular field strength.

The static magnetic field does not produce any MR signal. The signal is produced by use of two type of electromagnetic coils also known as radio-frequency coils. A transmitter coil is used to excite nuclei by applying electromagnetic wave orthogonal to the field of B_0 that resonates at the Larmor frequency. This excitation causes the nuclei to absorb energy from the radio-frequency pulse. This phenomenon, nuclear magnetic resonance (NMR), forms the basis of MRI. Subsequently the nuclei moves back towards relaxation. While relaxing back, the nuclei emits signal back to environment, which can be measured by a detector coil. Because of the prevalence of the water in human body and hydrogen in water, MRI scanners are tuned to the frequency of hydrogen nuclei.

To create images from NMR, spatial information from raw MR signal must be gathered. For this, gradients coils are used to generate a magnetic field that increases in strength along one spatial direction, over the area of interest. This causes the rotation frequency of spins at different field locations also to vary. In case the RF pulse is introduced with a specific frequency range, only the protons with Larmor frequencies at that range are targeted. Measuring the amount of energy emitted at different frequencies enables to identify how much of that object was present at each spatial location. In order to image 3-dimensional objects, it is necessary to select a plane in which a slice to be imaged. A slice selection gradient is used for this purpose (Desmukh & Gadre 2008). Capability to detect between different tissue properties to distinguish between tissue types, is dependent on the pulse sequence used.

Functional magnetic resonance imaging (fMRI)

Functional magnetic resonance imaging is a non-invasive brain imaging method used to study brain functions *in vivo*. FMRI measures blood oxygenation level-dependent (BOLD) contrast, that is the changes in local brain haemodynamics. Thus, fMRI doesn't measure brain activity itself, but its metabolic correlate. Although its clear that neural activity and cerebral blood flow are linked, the precise nature of the link is still unknown and therefore causes debate (Rossier 2009).

The neural activity based blood flow causes changes to the homogeneity of the magnetic field. The oxygen is carried in blood stream by hemoglobin in red blood cells. The measurement of blood oxygenation level changes is based on the different magnetic properties between oxygenated and deoxygenated blood. Their susceptibilities are different; oxyhemoglobin is weakly diamagnetic whereas deoxyhemoglobin is paramagnetic. The paramagnetic deoxyhemoglobin causes a stronger inhomogeneity to the surrounding magnetic field, which results in a decrease in MR image density. The activation of a neuron causes its consumption of glucose to increase, which in turn causes flow of oxygenated blood to the neuron to increase. The increased blood flow starts with a delay; the first observable changes occur after 1 to 2 seconds after the actual neuronal activity; this is called the haemodynamic lag. The BOLD fMRI signal increases above baseline at about 2 s and peaks at about 5 s.

The basic sampling units of MRI are volume elements called voxels. The voxel size determines the spatial resolution of an image; as the size decreases, the spatial resolution increases. However, the total signal recovered from a voxel is proportional to the voxel size. Thus, too small a voxel may have an insufficient signal to create high-quality images. Typically in fMRI voxels are about 3 to 5 mm on a side. The basic sampling unit for temporal resolution in fMRI is repetition time, or TR. Decreasing TR improves temporal resolution of fMRI, but it may have disadvantageous effects as well – it may result in weaker MR signal and reduced spatial coverage. Because of the sluggish nature of haemodynamic response, reducing TR does not have a dramatic effect on estimating the absolute timing of the underlying neural response. Usually TR is about 1 to 3 seconds. The temporal resolution of FMRI is considered intermediate level, because it can discriminate events that are separated by intervals on the order of a few seconds.

2.2.2 Preprocessing the fMRI data

Before data analysis the image data is preprocessed to increase the quality of the signal. The goal is to remove or correct non-task related variability in data and to register each individual brain into a common template brain. Usual steps include head movement induced motion correction, slice timing correction, spatial and temporal filtering, removal of non-brain tissue and coregistration. Each step includes a risk of losing or distorting data.

Individual brains differ significantly in size, shape and functionality. In order to be able to compare the activation between individuals, which is done in almost every fMRI experiment, the brain areas must be made comparable. Traditionally this is done by registering the brain of each subject to a common template brain. The coregistration enables the use of all the data in standardized atlas so that the analysis becomes much more convenient. There are a couple of templates available to be used in the process. Talairach space, based on a brain of a 60-year-old woman, is maybe the most famous. Another template developed by Montreal Neurological Institute (MNI), known as MNI template, has also been used quite commonly. The template was generated by averaging combination of the scan information of multiple normal control subjects.

After the preprocessing, four dimensional volumes are obtained for each subject. The variation in time of the image density levels at a specific point in space (called *voxel*) contains the desired functional information.

2.2.3 Experiment setting

In these studies, participants lye on their back while in the MRI scanner. Digitized audio and video is supplied by a computer system. The video presentation is projected to a translucent screen behind subjects' head, and viewed via a angled mirror mounted over their eyes. Sound is delivered via MRI-compatible earphones. Figure 2 illustrates the experimental setting.

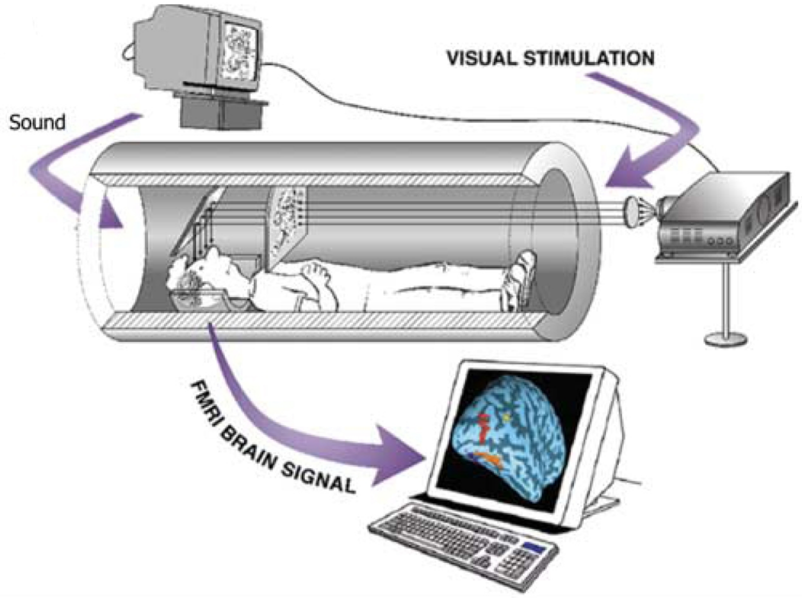


Figure 2. *Illustration of the experiment setting* (Hasson et al. 2008)

The fMRI experiment session includes acoustic noise, which often are very loud (<95dB). The primary source of acoustic noise is the gradient magnetic field. Rapid changes of current within the gradient coils, in the presence of a strong static magnetic field of the MR system, induce significant Lorenz forces causing physical displacement of the coil wires. This causes vibrations in the coils and their mountings, manifested as loud tapping or knocking.

The acoustic noise is reduced by using noise blocking earphones during the scanning (Nieminen 2010). Koskentalo (2007) also reported placing special kind of padding material between the coil and the scanner walls, as well as between the coil and the subject's head to reduce the noise level in the scanner and to help keep the head as immobilized as possible.

2.2.4 Independent Component Analysis (ICA)

Independent component analysis (ICA) is in general a method to identify a linear representation of mixed signals in which the separate components are statistically independent from the rest of the signals. BOLD FMRI data analysis is closely related to the blind source separation (BSS) problem – solving the original source signals using only the observed signals with unknown mixing and minimum information on the sources. Independent component analysis, ICA is a widely used method for performing BSS (Beckmann & Smith 2004). An simple illustration of BSS performed by with ICA (Figure 3) – Mixed signals from different sources, without any prior knowledge about the source signals, can be separated by using ICA.

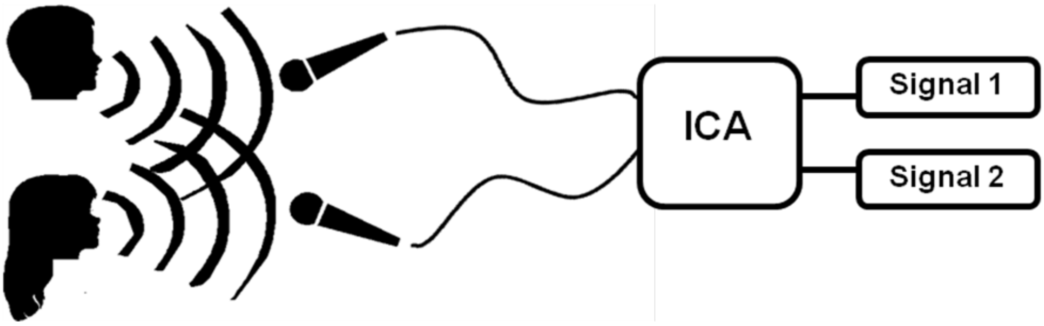


Figure 3. *Blind source separation performed by using ICA* (Eastaway et al. 2007)

The signal acquired from the BOLD fMRI suffers from a low signal to noise ratio deriving from different sources of variability, possibly including different kind of motion artefacts, physiological noise and with complex patterns. This sets a huge challenge for data analysis methods that seek to identify stimulus- or task- related activation (Tohka 2008),

ICA has proved useful studying brain in naturalistic viewing conditions (Jääskeläinen et al. 2008),(Bartels & Zeki 2005b) where it has an advantage that it doesn't need a previously constructed model of the stimulus sequences to which the activation patterns can be matched. In fact, no prior knowledge of the model is needed to identify differentially activated components (Beckmann & Smith 2004).

In fMRI data analysis ICA can be applied to localize statistically independent BOLD activation components from a voxel time series. Each independent component (IC) is composed of a one dimensional time series and a three dimensional image that represents the strength of the contribution of the time series to image voxels. Noise also forms independent components, which can be identified with ICA and removed from the data, thus denoising the fMRI signal by removing artifacts (Tohka 2008) .

A problem with traditional ICA is that the reliability of the acquired independent components is unknown. An ICA algorithm generates a specified number of components, but it is not known which ones to regard as important. Techniques to assess reliability exist; bootstrapping for instance is a computational method for computing the statistical reliability when simple mathematical formula cannot be found (Himberg et al. 2004).

Another solution to this is introduced by Beckmann and Smith (2004). They introduced an improved ICA method called probabilistic ICA (pICA). It reduces the problem of interpretation and selection of the significant ICs. By applying pICA, ICs are much more likely to derive from only one physical or physiological process. Another problem typical of ICA is that most of its algorithms have random elements that cause somewhat different results at different course of calculations, which cause additional work of estimating the variation of the variables(Himberg et al. 2004).

2.2.5 Intersubject Correlation (ISC)

Intersubject correlation (ISC) can be used to study how similarly a certain stimulus activates the brain across different subjects. Time series of each voxel from one subject's brain are compared with the corresponding time series of other subject's (Figure 4). As all the subjects are exposed to the same stimuli, computing ISC on a voxel-to-voxel basis reveals brain regions in which the response time series are similar across subjects.

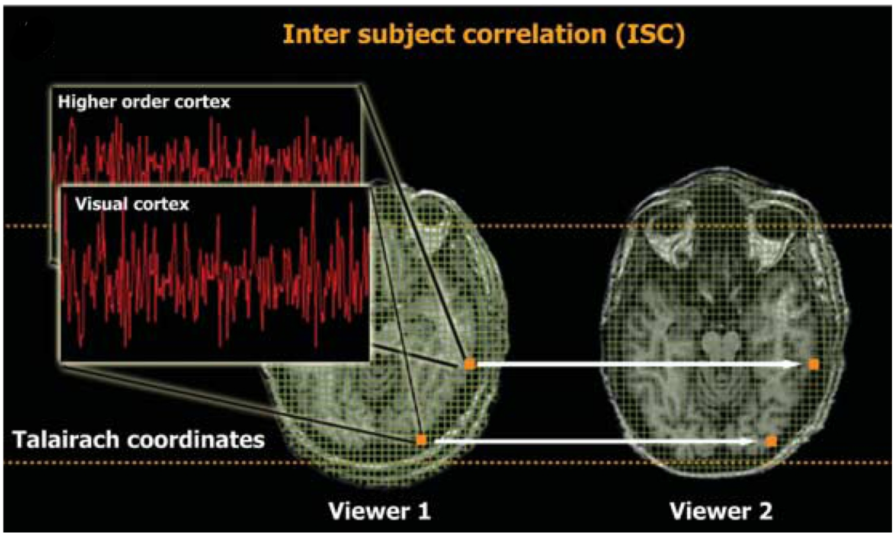


Figure 4. *Inter-Subject correlation* (Hasson et al. 2008)

Previously ISC analysis has been exploited in studies concerning the temporal scale of neural processing, the neural basis of inter-group differences, social cognition, memory, and learning. Recently Hasson et al. (2008) demonstrated how ISC can also provide a analytical tool for assessing and analyzing different aspects of films, film genres, and cinematic styles; directing and cutting for instance.

High ISC in visual or auditory areas is implication of high effectiveness of the visual content or the soundtrack of the stimulus movie. Likewise, the effectiveness of a movie in controlling viewers’ emotions and thoughts can be studied by computing ISC in brain areas to emotion processes or cognitive processes (Hasson et al. 2008). In addition to finding similar responses between subjects, ISC can also be used as a tool for tracing cultural and attentional differences among various populations. For example studying social dysfunction within autistic patients (Hasson & Malach 2005) .

2.3 Annotations in neurocinematics

The reason to annotate a stimulus movie is to build a model of it. The complexity of the movie affects also the workload of the annotation; the more completely the model is built by annotating the movie, the more can be concluded from the comparisons between the model and the obtained results. However, as the amount of annotated data grows the more complex the comparison becomes.

As the old cliché says, *one picture is worth a thousand words.* Because of the endless aspects that could be speculated, the complete and perfect annotation of a movie is nearly, if not completely, impossible task. For this reason annotations are also done in a data driven manner: not every aspect of the movie is needed to be annotated straight away, but the selection is done based on the research questions.

Various previous reports concerning studying the brain in naturalistic settings have not reported making annotations of the stimulus movie. Koskentalo (2007) reported that the reasoning was made in a data driven fashion, by utilizing peaks in the ICs. They inspected the film visually trying to identify a common factor from the events that had elicited strongest response in each IC. Only Lahnakoski (2010) described the features in a movie that were annotated, and methods for extracting features from the movie. However, the description part was not covered in detail (e.g. how the model was built).

For stimuli with higher levels of semantics the annotation becomes more challenging as the interpretation is involved in the stimuli assessment. Different interpretations of a single event can be done by different individuals. For this reason this kind of events are to be annotated by many people. For example annotation of emotional content is a challenging task because same event may elicit different emotions across individuals.

3 Introduction to annotation

3.1 Annotation is...

The word *annotation* is very ambiguous; its meaning is strongly related to the context of use. Generally in ordinary language, annotation is a comment or an explaination added to a document. Linguistics, speech and language engineers have used the term “annotation” to refer to a linguistic analysis of text or a speech signal. There has been a lot of other research regarding annotations over the past years in various fields, to gain understanding regarding how annotations could be useful in different contexts. The diverse starting points and interests have led to various views on what annotating is (Agosti et al. 2004). In most contexts the concept of annotations share many common factors.

Annotations are widely understood as *metadata*, additional data about data. The term meta origins from Greek word $\mu\epsilon\tau\alpha$, meaning after. The words “metadata” and “annotation” are used interchangeably in this case. The metadata describing document content can be roughly divided into two parts. First part, also referred to as *content-independent* metadata, is not directly concerned with content of the document, but is some way related to it; for example it concerns the concepts that give information about the creator of the content, the document format, the tools used in the process of creation, date, and other explicit properties of the image. The other part refers to the actual content of the document, and it can be divided into two layers: metadata may refer to low or intermediate-level features, like color, texture, shape, motion and spatial relationships in case of visual data. These are also referred to as *content-dependent* metadata. Metadata may also refer to content semantics. In this case the metadata are referred to as *content-descriptive*. They are concerned with the implicit properties of content that can be understood by perceiving the content itself. Under concern are relationships of content entities with real-world entities or temporal events, emotions and meaning associated with the content (Del Bimbo 1999).

In a broader sense, annotations can be seen as content. They possess similar properties to metadata, but according to Nagao (2003) it is reasonable to distinguish annotations from metadata. He defines annotations as:

Content about content.

This refers to metadata, which is traditionally known as *data about data.* The word metadata, however, has been adopted by computer science, where it is generally understood as information about a set of data with a restricted format. Nagao's definition emphasizes that annotations hold extra information about the meaning or context of content. The coverage of annotations are much broader than metadata. The annotations include any kind of content, for example text and voice commentaries or remarks, explanations, markups, hyperlinks or references (Nagao 2003).

Dong et al. (2008) view annotation as a process describing content and organizing the descriptions. This process can be understood as a process of *modeling* information, as the process of organizing information for communal use. Modeling helps to comprehend complex phenomena by using abstract descriptions, that hide certain details while illuminating others. Models help communication by describing information in an understandable way. They relate phenomena to one another, providing explanations and predictions about the world, and mediate among different viewpoints representing shared proven facts. (Allemang & Hendler 2008).

3.2 Characteristics of media annotations

3.2.1 Introduction to concepts

The function of medium in general is to serve means to communicate. Here a medium is defined to be a type of information representation, like alphanumeric data, audio, images, and video. Alphanumeric data is the kind of data that contains alphabetic and numeric data. Traditionally media types have analog and digital representations.

Divisions between digital media types can be made based on the ***modality*** they stimulate. Another kind of division of media types is obtained by classifying according to their relationship with time. There are two different classes of media types based on this division: *Static* and *temporal* media. The characteristic of temporal media is the time dimension, which static media do not possess. Examples of static types of media are alphanumeric data, structured text documents, images, and graphics. Temporal media can be audio, animations, plain video and audiovisual presentations (e.g. movies). Multimedia is a special type of data, which refers to a collection of media types used together. In this context we refer to multimedia presentations and multimedia objects, referring to a multimedia data to which a certain meaning has been attached. Annotating temporal media differs from annotating static media.

Media objects can be decomposed into smaller entities which characterize them. These derived characteristics are called *features* and can be described through annotations. Media content features shape the document, and define the modalities they activate. Features come in many forms, and they are usually divided between *high-level* and *low-level* features. This division is, however, only indicative. In some occasions it is difficult to label a feature as low- or high-level. Frequently in complex classification the term *mid-level* feature is used. Features are properties of the media object and annotation is its model. The derivation of the characteristics of media objects is called feature extraction (Blanken et al. 2007).

This process of feature extraction forms a basis for making annotations. Ultimately the annotations are organized to form a data structure (i.e. index), which is also referred to as *indexing*. There are alternative approaches to describing features and organizing annotations with numerous vocabularies and languages. These languages differ in their level of *expressiveness*, which affects the detail they are able to model. As an example the Figure 5 illustrates three different models of a water molecule with different levels of expressiveness.

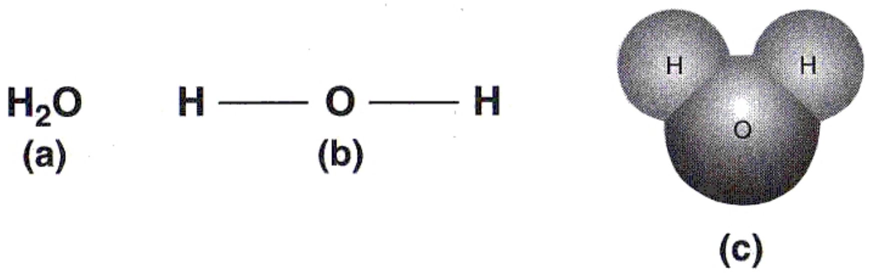


Figure 5. *Three different models of water molecule with different levels of expressiveness* (Allemang & Hendler 2008)

3.2.2 Features and the semantic gap

Low-level features include data patterns and statistics of media content and depend strongly on the content type. Low-level feature extraction can be done computationally by automated processes. With text documents the result of indexing could be a list of keywords with their corresponding frequencies, which are supposed to describe the content. Correspondingly audio signal might be represented by an amplitude time sequence. From amplitude time sequence other low-level features can be derived.

From images, for example, we can extract statistics on the pixel values, creating color histograms that can be used to classify images. Videos are sequences of images, thus they will share common features. Furthermore with video we can automatically classify the image sequences using also the time dimension. For example a shot - a sequence of images with no camera break - can be identified by computing the pixel difference between subsequent images. . A more extensive description on low-level video features is given in section 3.5.2. (Blanken et al. 2007).

High-level features bear more meaningful information than the low level ones. From a color histogram, it is hard to derive meaningful information on the image; for example a green image may indicate to a forest landscape, or to a golf course. High level features represent high level concepts that are meaningful only to humans. The gap between high and low level representations is called *the semantic gap*. Deriving meaningful concepts from low-level features of non-speech audio and video in general level is not possible, but focusing to a specific application domain improves possibilities to succeed (Blanken et al. 2007).

3.2.3 Temporal media annotations

A temporal media can be thought as a sequence of static media objects, thus one could think that its annotation would involve annotating each media object of the sequence one by one. Fortunately, this is not the usual case: changes in the high-level features of the content are relatively slow and thus make annotating necessary for only certain events of interest. With low-level features, annotating can be performed automatically, thus annotating each object in the sequence is not a problem.

There are two main approaches to temporal media annotation scheme: *segmented* and *stratified*. Segmented is the most simple traditional way of doing it:the idea is to partition the media object into consecutive temporal segments and describe each segment. Commonly this scheme has been extended to allow grouping related segments together, creating a hierarchical multilevel segmentation. Traditional structure of scenes and shots corresponds well to this kind of segmentation. Stratification is a context-based approach to modeling video content. It permits any subsequence of video frames to be modeled as rich multi-layered descriptions that can be easily parsed to support a wide range of applications (Chua et al. 2002)

3.3 Why to annotate

3.3.1 Introduction

Annotations are found everywhere. Libraries are kept in order by annotations, the book and film covers or their digital representatives can be thought as annotations. The Internet is full of material with comments and tags added to them. Annotations come in many forms, and the reasons to annotate are numerous.

One might already see that annotations can be used for many purposes. From shorter descriptions about content, with *overviews* to more extensive ones as the amount and complexity of data grows. The next sections give some examples of possible reasons to annotate.

3.3.2 Contexts of use

Research

For research purposes, for example in neurocinematics, annotations can be used to model the stimuli. This can be done to simplify the stimuli for machine accessibility to produce a simple model of it that is easier to compare with the data acquired from the experiments and thus ease *manageability* of the data. Researchers annotate data also by highlighting interesting parts of a document for future retrieval or *augment* data with personal observations. Although any kind of media are obviously meant to be self-sufficient, this kind of data may be very useful for research purposes. Annotations can be used in research also for *collaboration*, to share information or points of view in more efficient way.

World Wide Web

The digital annotations are often used in projects which aim to make information *retrieval* easier. Usually in this kind of projects there is a lot of information which needs to be organized and kept in order by dynamic information dissemination to ease the retrieval process. A huge framework of this kind is the Semantic Web, a set of tools to dramatically improve the current World Wide Web. (Antoniou & van Harmelen 2008).

Education and training

For educational and training purposes annotations can be used to code observational data. In case of education annotations could mean annotating videotaped behavior of pupils in a school, with a video annotation tool. Thus, allowing teachers to analyze their classroom events (Edyburn 2008). Sports and any other kind of training with movements might also benefit from annotations.

Enactive Cinema

The Enactive Cinema is a novel kind of interactive cinema genre, in which spectator's experiences of the emotional dynamics between the characters affect on the way the narrative unfolds. Each scene is annotated into various dimensions and by using the viewer's feelings a new montage is achieved with adaptive retrieval. A set of bio-sensors continuously measure each spectator's heart rate, breathing and movements while following the narrative. Thus the spectator's experience is connected to the installation's real-time emotional dynamics even without conscious attention on interaction. (Tikka 2008).

3.3.3 Reasons to annotate

From the examples above, one could see that there are many reasons to annotate. The most important ones are summarized here:

- **Retrieval**. The most important use of annotations is retrieval, the process of finding and accessing relevant information. Finding relevant data is a common problem and annotations provide a way of filtering unessential data, thus making the search more efficient. Usually either too much unessential information or too few relevant information is included in the search results. With additional semantics embedded in the annotation of a documents, the amount of essential hits can increase notably.

- **Manageability.** Annotations are used to describe the content for easier manageability. Simplified information is easier to handle. Models are built to ease the management of the information. Transforming information one form to another can be done to ease communication, for example. Manageability can be improved also by simply ordering information.

- **Overviews.** Overviews are needed everywhere, in order to survive this era of information. Overviews are means to decide about relevancy, usefulness and suitability of a piece information for a specific need. Overviews also setup a context. Keywords extracted from text documents can be thought as overviews.

- **Augmentation.** The data can be augmented with additional personal observations and comments. Comments provide a means to share points of view about media content.

- **Highlighting.** Visual search can be enhanced greatly with highlighting. Areas of interest are annotated to point out. This can be done for example as a social act or for own purposes to easier retrieval of relevant parts of a document in the future.

3.4 *Annotation* approaches

3.4.1 Introduction

Annotation process varies according to the context. As stated earlier, it can also be thought as modeling. This process consists of three phases, the order of whose can vary:

1. Feature extraction
2. Describing the feature with annotations
3. Organizing annotations - indexing

3.4.2 Feature extraction and annotation

Feature extraction is the most fundamental step of the whole annotation process; if no information is is retrieved from a media object there is no reason to annotate it. Features are extracted manually, automatically or semi-automatically.

Manual annotations are performed by subjectively extracting features from a media object. The feature can be any kind of piece of information that the user has chosen to annotate. There is plenty of tools with graphical user interface, which can be used to assist the annotator in the process. Because of the cost of the manual annotations**,** they presumably would not be done in first place if they were not necessary. Amongst the disadvantages manual annotations are time consuming, they are affected by subjectivity and incompleteness and in many contexts they are not practical.(Blanken et al. 2007) Subjectivity especially can destroy the consistency of the annotations . For this reason it is recommended to adopt *rule-based annotations*, which use expert knowledge to classify the annotations. (Xu et al. 2009) .

Automatic annotation techniques can be applied to content features that can be interpreted by a computer system. Traditionally these features have been low in semantics, but the development of pattern recognition and computer vision techniques have started to show good results reducing the semantic gap. There is nowadays a great scope of systems which provide basic annotations automatically. This requires relevant software that can track interesting features from the content (Ranguelova & Huiskes 2007).

Semi-automatic annotation is a hybrid model of the preceding ones. It is a promising approach to sort out the problem of automation by a combination of pattern classification methods and carefully designed concept hierarchies to provide standardized content descriptions. Many concepts ontologies already exist and they can be used to describe content. Given a well defined set of concepts is in use, the approach is to design pattern classifiers for each of the concepts in the set. In practice this requires annotating a small feature set manually to obtain a set of examples, to function as ground-truth annotations, that can be used to train the classifiers. After the classifiers have been trained, the rest of the collection is then annotated automatically (Ranguelova & Huiskes 2007).

The demand for reliability is the main criteria affecting the choice between feature extraction techniques. It depends mostly on the context of use of the annotations. The choice between manual or automatic annotations depends also on a level of semantics related to the content and available technologies. Manual approaches are used to discover most relevant annotations to describe content especially when the level of semantics is relatively high.

3.4.3 Describing the features – Vocabularies

Describing the features of a real-world phenomenon is in principle about building a model of it. Feature descriptions determine the level of expressiveness of the model acquired. The features can be expressed in a wide range of languages and with a wide range of vocabularies. A brief insight to the formal, computer accessible, vocabulary types is given in this section.

Controlled Vocabularies

Controlled vocabulary consists of a closed list of named subjects, which can be used for classification. The units of the controlled vocabularies are referred to as terms and each term is disambiguated to refer to a single subject. A controlled vocabulary prevents authors from defining meaningless, too broad or narrow terms. It also prevents misspellings and different authors from choosing slightly different terms(Lassila & Mcguinness 2001) .

Taxonomies

Taxonomies provide means to classify a set of terms of a controlled vocabulary. Taxonomy is defined in the following manner:

The classification of information entities in the form of a hierarchy, according to the presumed relationships of the real-world entities that they represent.

A hierarchy has a treelike structure with a *root* and *branches*. Each branching point is called a *node*. Each node of the taxonomy tree is an information entity that represents a real-world entity. Each link between nodes represents a relation called *is subclassification of* or i*s superclassification of* (depending on the direction of the link's direction). Sometimes this relation is defined to be *is subclass of or is superclass of,* when the information entities are classes of objects. A *class* is a generic entity. A class is included in its superclass. Figure 6 illustrates this idea with the class academic staff member. It has subclasses professor, associate professor, and assistant professor and the superclass staff member.

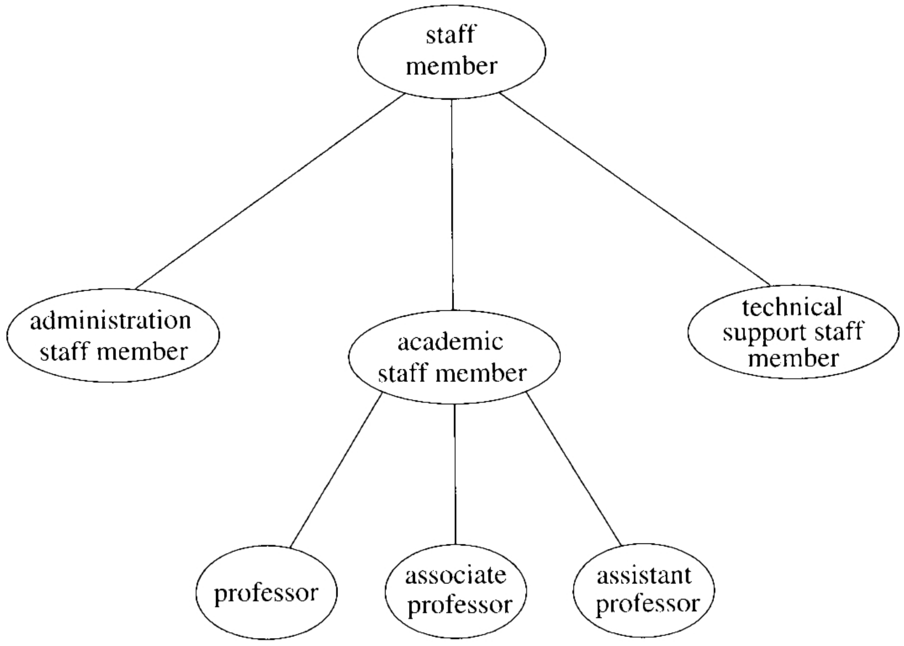


Figure 6. *A class hierarchy (Antoniou & van Harmelen 2008)*

The most general information entity is the root node and as going towards down the taxonomy towards leaves the entities become more specialized. Taxonomies are good for classifying information entities. They are able to express a little amount of semantics to distinguish among the objects in the information space. They are simple models of the distinguishable items under interest (Daconta et al. 2003).

Thesauri

Thesaurus is a controlled vocabulary, which provide some additional semantics in their relations between terms. There are various standards defining thesauri, but here it only a general properties of thesauri are discussed. In general thesauri is a extended taxonomy with additional relations:

- equivalence (a term X has nearly the same meaning as a term Y)
- homographic (terms X and Y are spelled same way, but their meaning differ)
- hierarchical (a term X is broader in meaning than a term Y)
- hierarchical (a term X is broader in meaning than a term Y)
- associative (a term X is associated in unknown way with a term Y).

With these properties a thesaurus ensures that concepts are described in a consistent manner. As a summary, thesauri provide a much richer vocabulary for describing the terms than taxonomies thus, they are much more expressive and provide better way to classifying objects and searching for them. WordNet thesaurus is a widely used large database of English. (Daconta et al. 2003).

Ontologies

The term *ontology* has its origins in philosophy, where it is used as a name of the field that studies the nature of existence. In computer science ontologies define words and concepts used to describe an area of knowledge. They use taxonomies as their foundation. Taxonomies and thesauri are also considered lightweight ontologies. In addition to the properties of taxonomies, ontologies have also additional relations, properties and property values. Taxonomies provide the basic structure of information space for ontologies (Antoniou & van Harmelen 2008). A widely accepted T.R Gruber's definition of an ontology that was later refined by R. Studer is used here:

An ontology is an explicit and formal specification of a conceptualization.

A “conceptualization” is an abstract simplified view of a phenomena. “Explicit” means that the type of the used concepts and constraints suit for describing also other phenomena of the same kind. “Formal” refers to the fact that the ontology is machine readable. “Shared” reflects the notion that the ontology is not restricted to some individual, but accepted by a larger group.

In general ontologies are used to describe formally a certain domain of discourse. Typically they consist of finite set of *terms*. Terms denote important concepts (classes of objects). For example in neurocinematics, the terms could denote concepts like movie event, camera event, social event, dance event, brain area, character or IC. The relationships include typically a taxonomy-based hierarchy. For example a camera event is a subclass of movie event and dance event could be a subclass of social event. Apart form subclass relationships, ontologies may also include other kind of information such as *properties* (stimuli X elicits activation in brain area Y) and *value restrictions* (social event must include more than one character).Ontology languages have been used among others in the fields of artificial intelligence and Semantic Web .

3.4.4 Organizing the annotations

The approach used to describe a feature determines how the annotation data is organized and related to the content and it affects how the data can be used afterwards. Organizing annotations is an often overlooked problem. Frequently annotations are structured on default as flat attribute-value pairs. More refined structures may have various advantages.(Geurts et al. 2005), (Dong et al. 2008). In the previous section a brief glimpse into different kind of vocabularies was taken. This section is based on work by Hyvönen et al. (2002) and Styrman (2008), unless otherwise stated. It discusses the common ways to describe features with metadata:

- free text annotations
- keyword annotations
- field-based annotations
- ontology-based annotations.

Annotation using free text

This is the most simple way of describing features. For this type annotator can use any combination of words or sentences, since there is no predefined lexicon the annotations should follow. Thus annotating is easy with this approach, but annotations are harder to utilize when there is no structure for the vocabulary: everything can be described with natural language, but it is hard to solve the intended meaning of the textual descriptions automatically with computers.

Annotation using keywords

In this approach the content is described with keywords, which the annotator can choose arbitrarily or from a controlled vocabulary. Controlled vocabularies are a finite list of *terms* representing real-world phenomena. Catalogs are examples of this category. They can provide an unambiguous interpretation of terms for example with identifiers linked to each term. Annotations can be specified separately for the content-independent, content-dependent and content-descriptive metadata. Keywords make the vocabulary more consistent compared to the free-text approach, but they are not sufficient for automatic processing.

Field-based annotations

Field-based annotations can be seen as an extension to text base annotations. Here the structure describes and retrieves an item by one or more attribute-value pairs. Typically this is implemented with a metadata schema that describes a set of fields, and some indication is given about the type of values that can be assigned to a particular field. In the usual case the annotator has a number of fields with required data types like integer, boolean, date, etc. Field based annotations allow validating the annotations, and refining syntactic relations between annotations these could also enhance semantics by requiring values to come from a particular ontology

The use of an agreed vocabulary helps handling the data computationally and thus improves the usability of the data. When the vocabulary used in the annotation process has not been decided beforehand, fields with predefined value specifications can be very helpful. In the case of multiple attribute-value pairs for an annotation, deciding the vocabulary and addressing values to all the possible fields would be very time-consuming and unnecessary. Field-based search can be performed for example by first giving values to only a few fields and then start the search.

Ontology-based annotations

Ontology-based annotation approach is an improved version of the field-based approach. Whereas field-based annotations include typing descriptions of a certain data type with a keyboard, the ontology approach allows selecting them from an ontology. The ontology approach allows more complex descriptions including relations between descriptions as discussed in section. For example the description of a human may include descriptions of the body parts (e.g. hand or leg). It depends on the available structure type, what kind of descriptions are supported. The ontology could be a taxonomy, which is basically a collection of keywords with a hierarchical structure. Ontologies with higher level of semantics provide more relationships and other properties (Lassila & Mcguinness 2001). Ontology-based retrieval can be done in many ways, which depends on the way the annotations are implemented. By semantically refining queries based on relationships between concepts and their descriptors, relevant concepts can be extracted.

Ontologies-based annotations have many advantages: Formal ontology languages can be used to build ontologies, which means that can it is possible use computer programs to process information and infer knowledge. Ontologies use controlled vocabularies, which alleviates the problem of inconsistency in annotation metadata and thus enables information sharing and exchange among different parties. Also multiple ontologies can be integrated. Thus descriptions of the same content from different perspectives can be integrated (Dong et al. 2008). A brief introduction to ontology languages and related techniques is given in next section.

3.4.5 Ontologies and technologies

In this section a brief introduction to different ontology-based annotation technologies is given. A lot of work has been done recently in the field of ontology languages and ontology based video annotation. The most of the recent work is building upon the Semantic Web technologies. A fundamental technology is Extensible Markup Language (XML) by World Wide Web Consortium (W3C). XML is a simple, very flexible text format, which forms the foundations for various other technologies. It is a metalanguage for defining markup (Antoniou & van Harmelen 2008).

Metadata ontologies

Dublin Core (DC) is a framework for descriptive metadata, that standardizes the way how descriptive data is presented. It is designed by a broadly representative group from the library, research, and academic communities, as well as from the industry. The Dublin Core Metadata Initiative (DCMI) is an open forum engaged in the development of inter-operable on-line metadata standards. Dublin core can be seen as a language for making particular class of statements about resources. This language has two classes of terms: elements (nouns) and qualifiers (adjectives), which can be arranged into statements.

DCMI has defined the Dublin Core standard to include to levels: Simple and Qualified element sets, which are standards for the cross-domain information resource description. The Simple set consists of the following 15 elements: Title, Subject, Description, Type, Source, Relation, Coverage, Creator, Publisher, Contributor, Rights, Date, Format, Identifier, Language. The Qualified level includes three additional elements Audience, Provenance and RightsHolder, as well as a group of element qualifiers that refine the semantics of the elements. Metadata content creation is assisted with guidelines for each element. (Hillmann 2007).

MPEG-7 is a standard for describing the multimedia content. It supports some degree of interpretation of the semantics. MPEG-7 is not aimed to any particular application, but attempts to keep the application domain as broad as possible. This results in an extensive, but complex standard; MPEG-7 supports multiple ways of structuring annotations with number of fields ranging from low level encoding scheme descriptors to high level content descriptors. In addition, domain dependent descriptors can be added if the default descriptors do not provide enough detail.

MPEG-7 is a XML - based standard with Dublin Core-like metadata elements (i.e. title, creator and rights); semantic(high-level); and structural (low-level) features of the content.

The MPEG-7 standard has four main elements:

1. A set of descriptors: A descriptor is a representation of a feature that defines the syntax and semantics of the feature representation.
2. A set of description schemes (DSs): A DS specifies the structure and semantics of the relationships between its components, which may be both descriptors and description schemes.
3. A Description Definition Language (DDL), which also allows for the extension and modification of existing DSs.
4. System tools, to support binary coded representation for efficient storage and transmission, synchronization of descriptions with content, management and protection of intellectual properties.

The tools allow a broad scale of descriptions: Audiovisual content can be segmented into its consecutive elements and into relationships between those elements. A moving region withing a video segment can be extracted into subset of moving regions. Objects within a moving region can be further annotated. An object can also be annotated by its low-level features and by its trajectory feature within the motion region. The spatiotemporal relationships of objects within frames of a stream can also be indexed (Martínez 2004).

Ontology Technologies

Ontology languages have been developed and used in Artificial Intelligence (AI) for a long time. The use of ontologies allows exploring semantic relations between concepts through reasoning - Information that is not explicit in the data may be automatically obtained by reasoning. The following list of technologies are recommendations of W3C, and this section is based on work by Antoniou & van Harmelen (2008); Allemang & Hendler (2008); and Daconta et al. (2003).

Resource Description Framework (RDF**)** is a data model for objects (resources) and relations between them; it provides simple semantics for this data model. RDF is based on the widely adopted XML syntax.

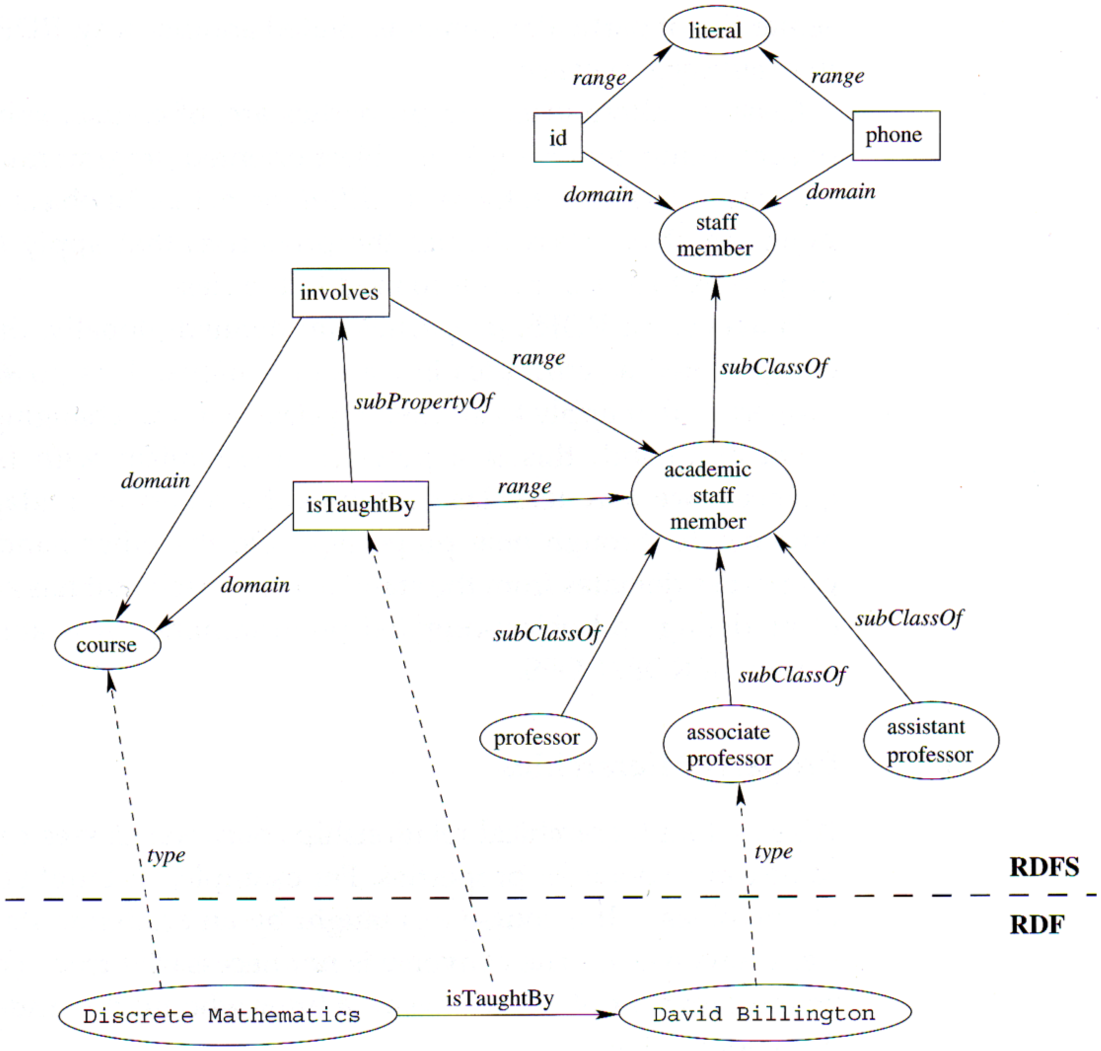
The RDF has three fundamental concepts:

1. *Resource* is an object under discussion. It could be anything; a media object, a book, an author, a hotel and so on. Resources are identified with Uniform Resource Identifiers, URIs. Each resource has an URI.
2. *Properties* are resources that describe relations between resources (e.g. owner).
3. *Statements* are object-attribute-value triples, which state something about a resource. They consist of resource, property and a value. Values can be resources or literals (i.e. strings).

RDF provides a model for metadata, it does not define vocabularies. It lets users to define their own vocabularies. The languages are defined with RDF Schema (RDFS).

RDF Schema is a vocabulary description language. Its function is to describe properties and classes of RDF resources, including semantics for generalization hierarchies of such properties and classes. Classes are generic objects. Individual objects that belong to a class are *instances* of that class. To allow properties to have only certain type of values, the *range* of that property is restricted. Similarly, to allow only certain objects to have some property, the *domain* of that property can be restricted. RDFS provides simple modeling primitives that can be used in constructing very simple ontologies, but more specific ontologies should be created with other languages.

RDF can be represented with graphs. Figure 6 illustrates, how different layers of RDF and RDFS are linked. In this example a schema defined with RDFS is linked with a RDF statement “Discrete Mathematics is taught by David Billington”. In the figure blocks are properties, ellipses above the dashed line are classes and ellipses below the dashed line are instances.



Web Ontology Language (OWL) is a semantically richer vocabulary description language. It supports describing properties and classes such as relations between classes. These include disjointness, cardinality (e.g. “exactly one”), equality, characteristics of properties (e.g. symmetry) and enumerated classes. The fundamental property of OWL is its special language construct called a *Restriction* with which it is possible to describe classes in terms of other things that have been already modeled. OWL builds on RDF and RDFS and uses the syntax of RDF.

Relevant tools and technologies

Protégé is a Java-based ontology editor and knowledge-base framework. It is a free, open-source platform with a suite of tools to construct domain models and knowledge-based applications with ontologies. It offers two main ways of modeling ontologies: via the Protégé-Frames and Protégé-OWL editors. Protégé supports exporting ontologies into a variety of formats including RDF(S), OWL, and XML Schema.

Protégé provides with a rich set of knowledge-modeling structures and actions that support the creation, visualization and manipulation of ontologies in various representation formats. Protégé supports also queries to RDFS or OWL model with SPARQL, which can be used to filter out individuals with specific characteristic. SPARQL is a query language and a protocol for accessing RDF. Protégé can be customized to provide domain-friendly support for creating knowledge models and entering data. Protégé also offers Java-based Application Programming Interface (API) for building knowledge-based tools and applications (Musen 2010).

Jena is a open-source Java framework that provides programming environment for RDF, RDFS, OWL and SPARQL. Jena includes RDF API, an OWL API, SPARQL query engine, in memory and persistent storage. Jena can be used to navigate and query models (ontologies), and supports three operations for manipulating models: union, intersection and difference (McBride 2009).

3.5 Annotation of a movie

3.5.1 Introduction

A movie is a very complex stimulus. from which it is computationally very hard to derive high- or even mid-level features and concepts. Since movies are used as a stimuli in neurocinematics, it is desirable to have a computational and easier handling of the data. The aim is to be able to link the stimuli computationally to the changes in the brain activity for example relating the stimuli with the ICs componets or with the ISC varying in time. For this reason a simplified model of the stimulus is built by annotating the stimulus movie.

The next section gives a short introduction to movie theory. It is based on the work by Pirilä & Kivi (2005, 2008), Nordberg (1986) and Prunes et al. (2002). The discussion is not all-embracing, but introduces to essential concepts considered relevant in this context. Section 3.5.3 gives an brief introduction to video annotation. It is based on work by Del Bimbo (1999), unless otherwise stated.

3.5.2 Movie structure

Movie consists of single images (e.g. frames) played at fixed rate (typically at rate of 24 to 30 frames / s) to create illusion of movement. Also one or more synchronized audio tracks are included. Thus it has two fundamental building blocks:

1. frames (images)
2. audio samples.

A single static image of a movie is from now on defined as a frame. Consecutive frames and audio samples form two temporally continuous structures, audio and video. Movie is recorded during the filming and constructed through *editing,* which consists of cutting the rearranging the audiovisual material into a single continuous entity. Audio can be mixed with other sounds, the result of which is then aligned to the video structure.

Video consists of clear segments shots, which are made of smallest units – frames. The shots are composed into scenes and sequences. In the shots there are also *images.* The following four elements can be considered a basic structural elements of a movie:

- *a shot* is a set of frames between camera turn-on and a camera turn-off
- *a scene* is a collections of shots that share the properties of simultaneity in space, time and action
- *a sequence* is an independent structural entity identified by similar action, atmosphere or theme
- an *image is* a basic building element of a movie; a set of frames characterized by uniform shot size, cropping and stationarity. The time and space composition during an image remains the same and no significant changes in the images storytelling occur.

All these elements are joined together with *transitions*. Types of visual edits include both sharp shot transitions, cuts, and gradual shot transitions, like fades, wipes, mattes and dissolves. S*mall transitions* are changes in a shot between images. These include camera movement and also A structure of a movie is presented in Figure 7. Empty spaces between structural units are transitions.

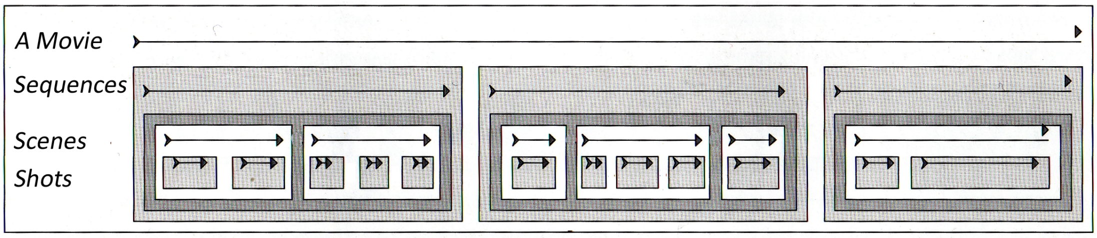


Figure 7. *A Structure of a movie (Pirilä & Kivi 2005)*

3.5.3 Language of the cinema

Storytelling

The central part of a movie is the story. The *diegesis* is the world in which the narrated story occurs. It is constructed by audience from the material presented and consists of all the objects, events, spaces and the characters of the story. It also includes things, actions, and attitudes not explicitly presented in the film but inferred by the audience.

Audio can be divided into *actual* (diegetic) and *commentary* (non-diegetig) sound. Actual sound includes sounds whose source is visible or whose source is implied to be present by the action on the film:

- characters' voice
- sounds made by objects in the story
- music represented as coming from instruments in the story space.

Commentary sound is represented as coming from a source that is outside the story space. It includes sounds whose source is neither visible on the screen nor has been implied to be present in the action:

- narrator's commentary
- sound effects
- mood music.

Wide variety of modern movies lean on *Poetics* by Aristotle(384 BC – 322 BC). Movie mimics the reality and thus creates a strong illusion of it. Objects and characters in the movie represent real-world objects and meanings are conveyed through the use of narrative elements like characters, objects, symbols, metaphors, etc - different kind of *signs*. For example an image of a decelerating train wheel is an indicator of an incoming train.

Although movies are presented in two spatial dimensions, they form a strong experience of space and movement in the spectator. Thus, movie as a form of expression is considered four dimensional; in addition to the width and the height of the image also time and space are included. A movie is a unity comprising of an image space and a sound space. The messages are conveyed through these spaces using multiple planes of communication.

Juxtaposition of two images affects their shared experience. Directors Vsevolod Pudovkin (1893-1953) and Lev Kulesov (1899-1970) demonstrated this effect in their infamous experiment. They connected an image of a famous actor to a plate soup, a coffin and a playing child. To subjects these different compositions of the same picture of the actor appeared differently: His face was considered to express hunger, sadness, and joy. The movie that is based on *montage (* image juxtaposition) activates the viewer to interact and viewer empathizes with the presented emotions. Spectator empathizes emotions and identifies to the emotions of the movie. According to Iacoboni (2008) this is caused by mirror neurons that elicit emotions similar to ones seen in the movie.

3.5.4 Staging – Mise-en-scene

Staging (mise-en-scene) refers to the visible characteristics of the movie. It comprises of *set* (decór), *costuming*, *lightning* and *acting*. Staging determines the image composition.

The set covers the setting; scenery, and a set of articles. The set can be used to bring viewer to the right context and to adjust the mood or to amplify character's emotions. The expressive meaning of lighting is extensive: eventually everything that is visible in an image is light. The intensity, direction and quality of lightning provide a effective means to the way an image is perceived. Low lighting creates strong contrasts between the brightest and darkest parts of an image.

The representation of space is a mean to affect the reading of the film. Spatial relations of the image like proximity, depth, size and proportions of the objects and surroundings in an image can be manipulated through camera placement and lenses. Also lighting and set can be used to manipulate the representation of space. Offscreen space exists in the diegesis, but is not visible in the image. Costuming refers to the clothes worn by the characters. Costuming is a way of characterizing and signifying a character. It is another way to authenticate the illusion of reality. Acting defines the behavior of characters in the movie. The behavior or actions reveal the nature of the characters.

Cinematographic elements

Cinematographic elements strongly contribute the storytelling and affect the meaning. These elements are produced by the way the image is shot. The viewer can be manipulated to pay attention to certain events using various techniques: for example *focusing*, *framing*, and relevant use of *shot sizes*. The use of *color* and *lighting* especially pre-attentive color (e.g. red) can be used very effectively to direct the attention of the viewer. The duration of a scene can be used to advise the viewer to draw conclusions about its meaningfulness.

The impression of image is strongly affected by its fundamental features like color, contrast and areas of focus. Colors enable powerful creation of aesthetic patterns and establishment of mood. Contrast can be used also as a symbol; clichéd example would be to distinguish between good and evil. The areas in focus reveal the significance of elements in the image. The depth of field determines the distance through which the elements in an image are in sharp focus. Varying the depth of field and focus is used as a means to show, which planes in the image space have significance in a given image. Deep focus causes the image composition to be such that significant elements occupy widely separated planes in the image. Shallow focus us used to direct the viewers attention to one element of a scene. Very common usage of this is close-up images.

Framing determines what are the included and excluded aspects of the events that occur in front of the camera. The expressive qualities of framing include the angle of the camera to the object, the relationship between camera and object, and the association of camera with character. Camera angle can be used to indicate the relation between a character and the camera's point of view. It is frequently used to depict either superiority or inferiority. *Level of framing* means varying the height of the camera. The technique can be used to to signify sympathy for characters with faces at similar height with the camera.

Size of shot (also referred to as scaling) determines the the significant elements in an image. There are conventional sizes of shots used in the film production. The eight shot size system is represented in Figure 8. The shot sizes are determined with relation to size of a characters in the image. In broader images the significance of the surroundings is high, whereas with closer shots signify the role and expressions of the character.

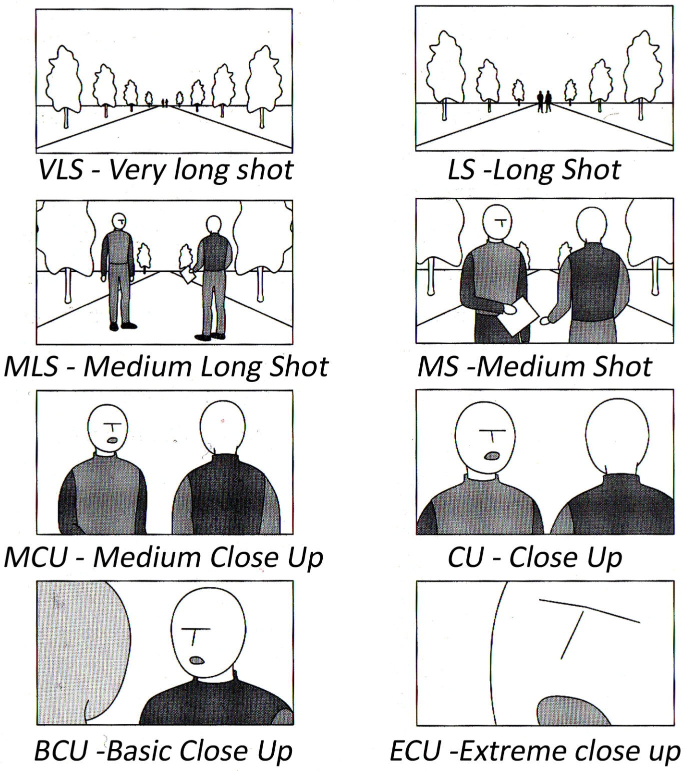


Figure 8. Shot size formats (Pirilä & Kivi 2005)

Camera movements include *zooming*, *dollying*, *panning* and *tilting*. Pan and tilt produce mobile framings; pan scans the space horizontally, whereas tilt scans the space vertically. Pan and tilt mimic head movements. Dollying is a tracking shot that usually follows moving character or a object along the screen.

3.5.5 Video annotation

Structural features

There are many possible ways to annotate features associated with video production data such as *shot boundaries* and *key-frames*. Segmentation breaks down the video into its constituent shots and their higher-level combinations – scenes – respectively. One approach is to utilize an edit decision list created by producers that would provide edit type and their corresponding frame number.

Automatic segmentation is generally based on using combination of image analysis methods and a rule set, which models conditions under which certain situation occurs. The methods concentrate on detecting shot boundaries, both sharp shot transitions and gradual shot transitions Automatic cut detection is based on extracting the information that contributes to the cut. The idea is to exploit similarity of consecutive images in a same shot - Cuts with little motion and constant illumination are easily detected, but as the changes between consecutive frames reproduce the detection becomes harder. Detecting fades and dissolves is based on the fact that they make boundary between shots spread across multiple frames.

Perceptual features

Color perception is related to several factors including color attributes (brightness, saturation and chromaticity), surrounding colors, spatial organization of colors, viewing surroundings, and so on. Geometric color models are used to represent colors. They represent color stimuli as points in a color space, they allow discriminating between color stimuli and also permit similarity judgment and identification. The usual way to describe color is with a color histograms, but also symbolically by color names. *Texture* is basically characterized by differences in brightness, that can be analyzed from the high-frequencies of the image spectrum. Image patch size used in pattern recognition task, number of distinguishable grey-level primitives and the spatial distribution of these primitives are interrelated factors which all characterize the texture. Textures are usually described with numerical vectors and are modeled as a point in a suitable feature space.

The form of an object is identified by its *shape*, which is traditionally described through a set of features. Features can characterize either the global form of the shape(e.g. area or major axis orientation) or local elements of its boundary like corners or characteristic points. Shapes also bear semantical meaning, which automatic analysis cannot extract. For this reason describing shapes involves manual annotation.

Spatial relationships between spatial entities (points, lines, regions, objects) can be described according to their directional and topological relationships. Directional relationships concern relative directions (right/left of, above, below). To describe directions, orientation must be established, which is taken with respect to the reference entity or to an external reference frame. *Metric relationships* are usually described with distance. Other metric relationships regard angles. Distances are usually defined according to Euclidean metric. Topological relationships capture set-theoretical concepts (e.g.like adjacency, containment, disjunction and overlapping) between close entities.

Spatial relationships between spatial entities (points, lines, regions, objects) can be described according to their directional and topological relationships. Directional relationships concern relative directions (right/left of, above, below). To describe directions, orientation must be established, which is taken with respect to the reference entity or to an external reference frame. *Metric relationships* are usually described with distance. Other metric relationships regard angles. Distances are usually defined according to Euclidean metric. Topological relationships capture set-theoretical concepts (e.g. adjacency, containment, disjunction and overlapping) between close entities.

Motion is directly related to a change in the relative position of spatial entities or to a camera movement. There is two main approaches to compute an approximate estimation of motion. First involves analyzing temporal changes of grey-level primitives of consecutive frames, and computing a dense flow field usually at every pixel of the frame. Under suitable restrictions the two unknown components of the flow field are obtained by solving a system of non-linear equations. The other approach is based on extraction of set of sparse characteristic features of the object, such as corners or salient points, and tracking them in consecutive frames. Motion can be described through a set of parameter values or through symbolic strings which indicate how spatial relationships change through time.

Camera motion can be extracted from either uncompressed of compressed video streams. The former approach is based on evaluation and analysis of flow field approximating the motion field of pixels between two consecutive frames. The latter is concerned with the analysis of frames of the compressed streams which encode motion vectors. Region or object motion requires tracking regions with coherent motion over consecutive image and object segmentation. Different possible approaches include detection of brightness changes, motion vectors, and motion and color information.

Semantic features

Semantic feature extraction from a video can be done automatically in restricted domains, while broad domain semantic content requires manual annotations. Some features can be extracted with use of knowledge on conventional methods to build scenes. For example conversations scenes are built by using the shot-reverse-shot format, which can be detected to some level of accuracy.

However, to gain reliable feature extraction of higher-level features manual annotations are needed. Tthere is a huge amount of meaningful content features in a movie. As discussed in in the previous chapters movies use multiple high-level techniques for storytelling. Annotating a movie requires analyzing, interpreting and explaining it. When analyzing film, it is important to focus on how a specific film constructs a certain reality. Annotating high-level features of a movie requires analyzing the means of movie communication through signs that represent reality.

The signs and meaning in the cinema have been studied to a large extent. According to Wollen (1972) studies are based on studies by Fernand de Saussure on linguistics in which he predicted a new science of semiology – A study of signs and laws that govern them. Wollen also describes the work of C. S. Pierce - The classification of signs into *icons, indices* and *symbols.* An icon represents its object by similarity to it. For example an portrait of a man resembles him. “An index is a sign by virtue of an existential bond between itself and its object.” For example a clock indicates the time of the day. A symbol has neither resemblance with its object nor any existential bond with it. The symbols (e.g. written words) are understood and used for conventional reasons.

3.5.6 Viewpoints on neuroscience of film

Using movies as naturalistic stimuli has good justification. Video makes a good replication of the real life, allowing understanding of the content with minimum background knowledge used to make interpretations. Usually making movies includes paying extra attention for creating credible – similar to real life – representation of events.

For the annotations it is recommended to use an unambiguous and controlled vocabulary which is defined by an expert. The terms used for the annotation should be defined beforehand. To support collaboration, the terms should be general. Sharing the annotation should be easy, and connecting annotations of different individuals should be supported.

The model's expressiveness should describe temporal, spatial and semantical features of the movie. For the annotations results in better computability and the use of organized, ontology-based annotations and existing vocabularies is recommended in (Antoniou & van Harmelen 2008) and (Geurts et al. 2005).

The model could be used to reason about correlations between ICs and the stimuli and to explain certain brain activity changing in time. For this purpose, ontology-based annotations could be used, by embedding the IC components to the model. To ease the overall process of comparing the model of the stimulus to the acquired fMRI data, also MRI images of each independent component could be annotated. Ontology based MRI image annotation has been suggested by Mechouche et al. (2008, 2009) With a mechanisms to relate these annotations to the movie annotations, the reasoning process could be assisted computationally.

4 Survey on video annotation tools

4.1 Introduction

A lot of research has been done in the field of video annotation tools recently, many projects are somehow driven by the strive for Semantic Web. This chapter describes the state-of-the-art available video annotation tools considering their availability and terms of use as criteria for their selection. At the moment of this research, the discussed tools are available for download. The

The purpose of a video annotation tool is to describe, index and augment the content of a video with machine readable annotations. As visual information retrieval and the Semantic Web are hot topics nowadays, new image and video annotation technologies are being developed all the time. Previous reviews on video annotation tools by Chaudhary (2008) and Heggland (2005a), have already started to outdate and do not provide comprehensive information on tools useful for studying brain in naturalistic settings. This part of the thesis is a contribution either to find suitable tool or acquire viewpoints for building a tool for future annotations.

The next section describes a survey on annotation tools, starting from the features that define a tool. The section is based on the work of Heggland (2005a), in which he defined proper dimensions of a content model to evaluate the properties of a video annotation tool.

4.2 Video annotation tool properties

4.2.1 Introduction

Video annotation tools are used to describe the content of a video. The possibilities to make descriptions of the video define the *expressiveness* of each tool. To discuss and compare video annotation tools, a clarification of their measurable properties is needed. The properties of annotation tools define the model they are able to build called a *content model*. To assess content models and their properties, a set of measurable dimensions is needed.

A video is a spatiotemporal medium - it holds information in both *spatial* and *temporal* dimension. Information in these dimensions frequently convey a message that carries also a *semantic* dimension – in other words the spatio-temporal information also mediates a meaning. In order to express content in profound way, all these dimensions have to be addressed through annotation.

4.2.2 Temporal expressiveness

Video is a temporal media, in which events occur along the time. The time is frequently the most important part of the video content model, because events change in time and video events often occur simultaneously, either dependent or independent of each other. Different level of event features form a complex structure. A video content model has to be able to handle all these levels. This can be done by marking time intervals of each feature of interest. There are two possible approaches to this as described in section 3.4.2, segmentation of stratification. With stratified option, during certain time intervals only the relevant descriptions are valid. The segmentation approach refers to a segmentation of video based on interesting events and attach descriptions to these fragments. A content model may put restrictions on these fragments. The level of restrictions define the temporal expressiveness of a tool. The different levels are discussed next. Figure 9 illustrates these different levels.

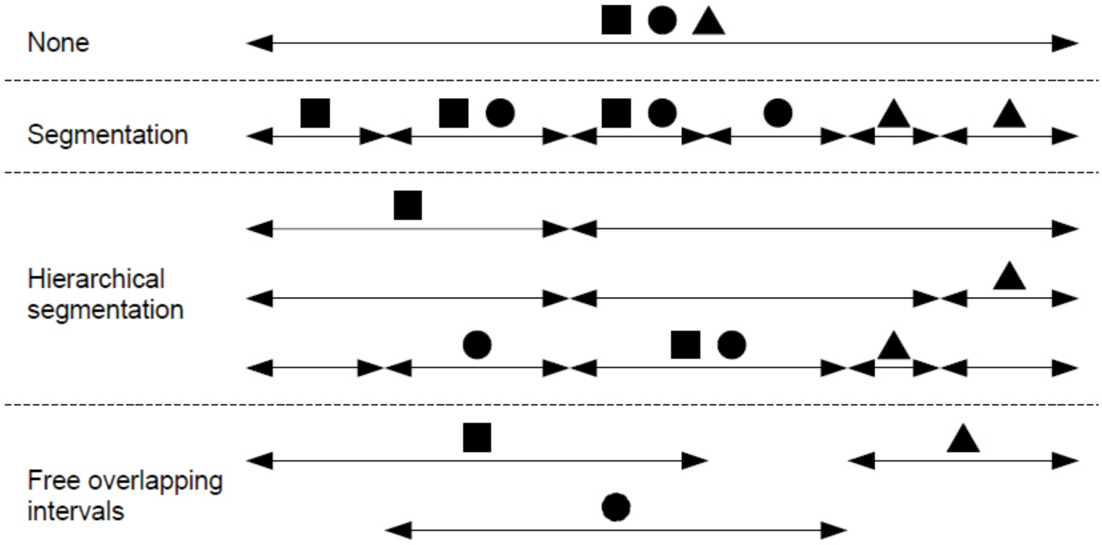


Figure 9. *Temporal expressiveness* (Heggland 2005a)

None - No temporal expressiveness means that there is no representation of time. Temporal media is described as a unity. Example of this kind of description is a general genre classification that is used for general description of movies. Content model with no temporal expressiveness allows only descriptions including content-independent metadata like directors, producers and actors – the kind of data that is valid for the entire film or at least somehow describes the whole movie.

Segmentation means here logically dividing the video into non-overlapping parts. Content descriptions are connected to segments. There are few possible segmentation approaches to determine the segment boundaries. *Fixed segmentation* produces fixed, content-independent sized segments. In this case segmentation presumably does not correspond to the content, but its implementation is simple and straightforward. *Structure-based* segmentation is based on detecting shot boundaries,segmenting the scenes and providing a generally logical way to segment a film. Shots often correlate to the semantic structure of the video. For videos with no shot boundaries or with very fast shots only partially correct descriptions can be achieved. This model can be developed further by introducing *user-defined segmentation*, for which the structure based segmentation can provide a basis. This feature makes it possible for user to define the segmentation according to interesting video events. The disability of annotations to describe events that overlap segment boundaries result in redundant annotations. Also audio and video often do not go together, so a conflict between them would often occur.

Hierarchical segmentation offers increased expressiveness by having several layers of segments arranged hierarchically. Levels in the hierarchy may increase the amount of segments defined on the lower levels, thus overlapping is still not possible. Layer containing a scene segmented film would be at one of the highest levels, while shorter segments containing some specific events would be lower in the hierarchy. Layered architecture of course results in increased complexity at implementation of the system.

Free overlapping intervals - The greatest expressiveness and flexibility is obtained by allowing time intervals to be independent of each other, which makes it possible to annotate independent aspects of the video accurately. Drawback to this approach is the complexity: The start- and ending point of each interval has to be stored, compared to just the position of boundaries of segments. However this results in more accurate descriptions too. Common technique of implementing this is to to arrange each interesting entity or concept to a layer called *strata.* This approach requires some predefined notion of what is represented in the video.

4.2.3 Spatial expressiveness

The spatial dimension of a video bears information that can be attended in multiple ways. Eye tracking technologies have been used to study the gazing patterns of subjects. Especially in subjective video research it is very important to be able to know where subjects are looking at. Accordingly, the video data can be annotated spatially on frame by frame basis or focusing on a smaller region identified from eye movements. Spatial expressiveness of the models represent the the available precision to address spatial content. Expressiveness of different levels can be characterized as none, static and dynamic.

None - If there is no spatial expressiveness the descriptions are expected to belong to the whole picture. No matter how significant object or event is on video, all descriptions have similar value. Also in the case of multiple similar objects in the scene simultaneously it is impossible to refer to a particular one. This is disadvantageous for research purposes, with which eye tracking is often used as a means to follow which parts of the image the subject is looking at.

Static regions of video are 2D shapes that do not change in time. Two-dimensional spatial component may contain different levels of complexity. The form, size and location of it may be defined to be anything. Figure 10 illustrates a static rectangular region. With static regions static objects are easy to model, but for moving objects it is cumbersome task. Static regions are relatively simple to handle, and suit pretty well to situations where the level of movement is low. Movement can be modeled with temporally consecutive regions with slightly different spatial regions.

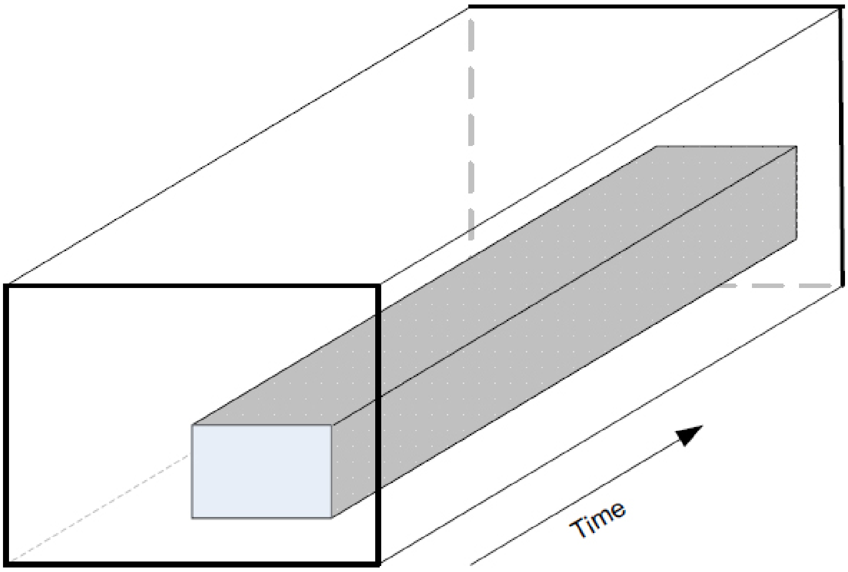


Figure 10. *Static rectangular region (Heggland 2005a)*

Dynamic regions – In this case the regions to be annotated can change in time. Two different levels of expressiveness can be identified. A simple kind of dynamic region could be specified by its form and its location at the beginning and at the end of the temporal interval. The position of the region in the frames in between would be interpolated linearly. Advanced model of this could be achieved by allowing more points between the beginning and the end, where the direction of movement could change. This kind of region is illustrated in Figure 11. Adding these points would ultimately result in fully dynamic interpolation. This kind of system could be produced using pattern recognition techniques.

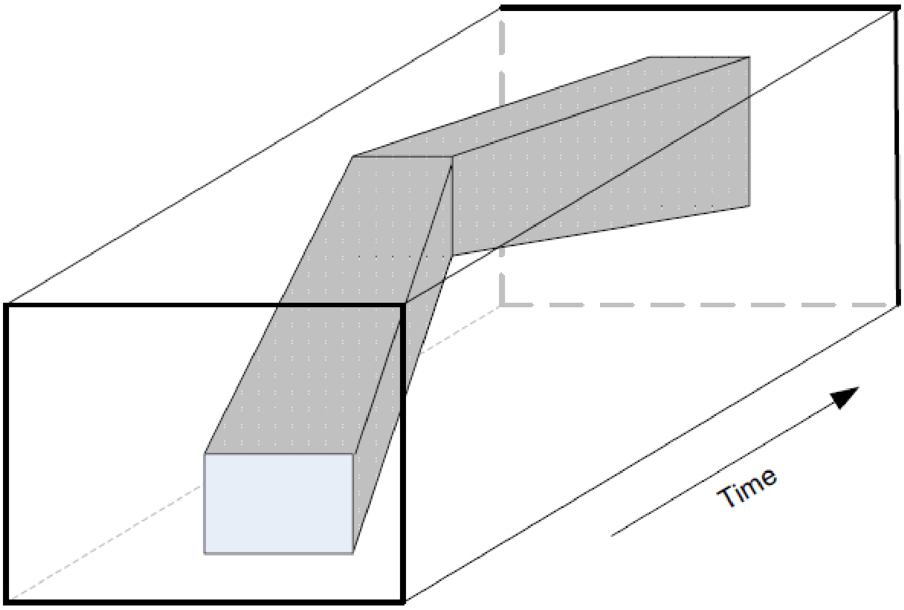


Figure 11. *Dynamic rectangular region with linear interpolation (Heggland 2005a)*

4.2.4 Semantic expressiveness

Video annotation should also describe meaningful phenomena in the video. But how to represent them - what kind of lexicon is used and what is the relationships between the descriptions? What kind of names, properties and data values are supported? Extremely simple content models may support only a single unstructured description field, whereas a more advanced model may support much more refined type of system. These properties of a model are here referred to as semantic expressiveness of the content model. This dimension is harder to describe with defined and discrete steps as the two previous ones, but it still has clear properties that can be used to describe the content model:

Single, free-text descriptor - This is the most simple mechanism to describe a temporal interval. As the descriptor can be anything, indexing and retrieval of descriptions can become imprecise. This model however gives to the user full freedom of choice – the descriptor can be used for anything.

Keyword descriptors - Formality increases notably with a priori named keyword descriptors to describe temporal elements. These descriptors allow more precise annotation and are easier to handle computationally. The efficiency of this mechanism depends to large extent of the decided vocabulary. Also classification implemented by descriptors makes the search of a content of certain type more efficient.

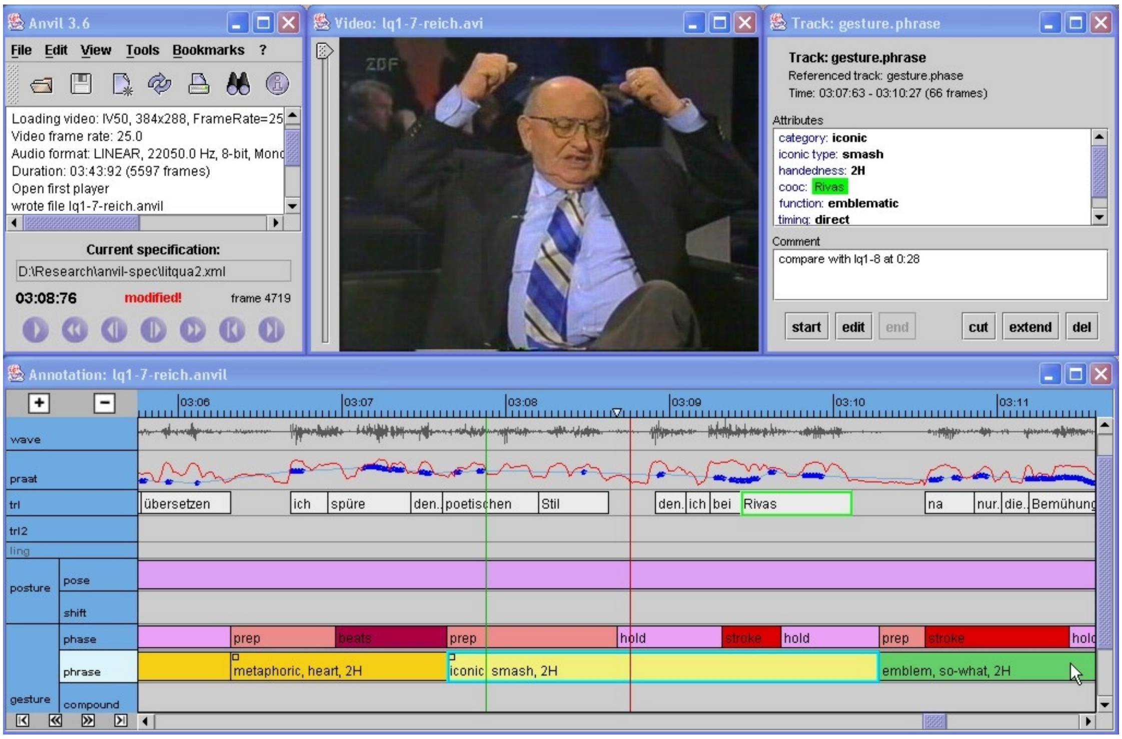
Field-based descriptors - The content is described with one or more attribute-value pairs. Typically the annotator has a number of fields implemented with a metadata schema with required data types like integer, boolean, date, etc. The schema that describes a set of fields By this way constraining the values of the descriptors, the computability and integrity in the descriptions can be increased notably. This could mean demanding specific type of format for certain descriptions (e.g. time).

Ontology-based descriptors provide formal methods to define different type of data objects, their properties and relations to each other. They allow mechanisms to define objects with their relations to other objects or object classes. For example object could be defined to belong to a subclass of another object. As the properties of relations are formally defined, it is possible to computationally reason about the knowledge represented in the model.

4.3 Video annotation tools

4.3.1 Anvil

Anvil is a free Java-based research tool by Kipp (2008) for annotating digital video. It is platform-independent and was designed to be used in research areas like linguistics, human-computer interaction, or film studies. With anvil annotation elements are created with graphical user interface (Figure 12). The elements are constructed on temporal, hierarchical and completely user-defined layers.



Loading video: IV50, 384x288, FrameRate=25
Video frame rate: 25.0
Audio format: LINEAR, 22050.0 Hz, 8-bit, Mono
Duration: 03:43:92 (5597 frames)
Open first player
wrote file lq1-7-reich.anvil

Current specification: D:\Research\anvil-spec\litqua2.xml

03:08:76 modified! frame 4719

Track: gesture.phrase
Referenced track: gesture.phase
Time: 03:07:63 - 03:10:27 (66 frames)

Attributes
category: iconic
iconic type: smash
handedness: 2H
cooc: Rivas
function: emblematic
timing: direct

Comment
compare with lq1-8 at 0:28

Figure 12. *The user interface of Anvil*

Anvil's design is object-oriented. The largest unit of the object model is an annotation, which hold a number of entities called tracks. These tracks contain any arbitrary number of user-defined attribute-value pairs. Different attribute types are supported. Tracks can be grouped together to form a structural entity, that can be used for ergonomic reasons. Anvil is capable displaying waveform and pitch contour of audio.

Expressiveness

- Temporal - Free overlapping intervals
- Spatial - Dynamic regions (points)
- Semantic - Field-based descriptors (XML Schema)

4.3.2 OntoLog

OntoLog is an annotation system for temporal media. It is a Java application built by Heggland (2005a), build to "explore the issues and benefits of using ontologies, application profiles and RDF to temporal annotations." It uses annotation theme based on hierarchical ontologies, and an RDF-based data model that may be adapted and extended through the use of RDF Schema. As a default description scheme Dublin Core is used.

OntoLog is a tool for annotating temporal media – video and audio using ontologies. It is designed to be flexible and adaptable; user may decide the form and content of annotations. It contains default description schemes which can be used to describe content, but also creating new ontologies or importing existing ones is easy. OntoLog uses stratified approach to annotate video. The annotations themselves are thus organized into layers, which is intuitively visualized in the user interface (Figure 13).

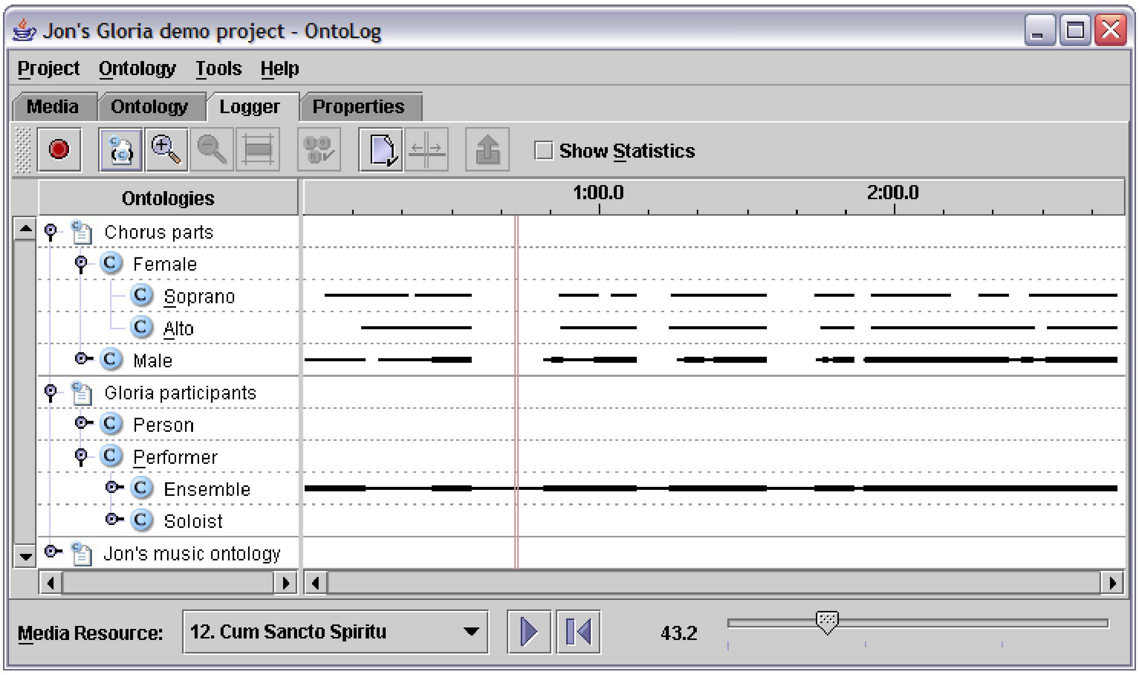


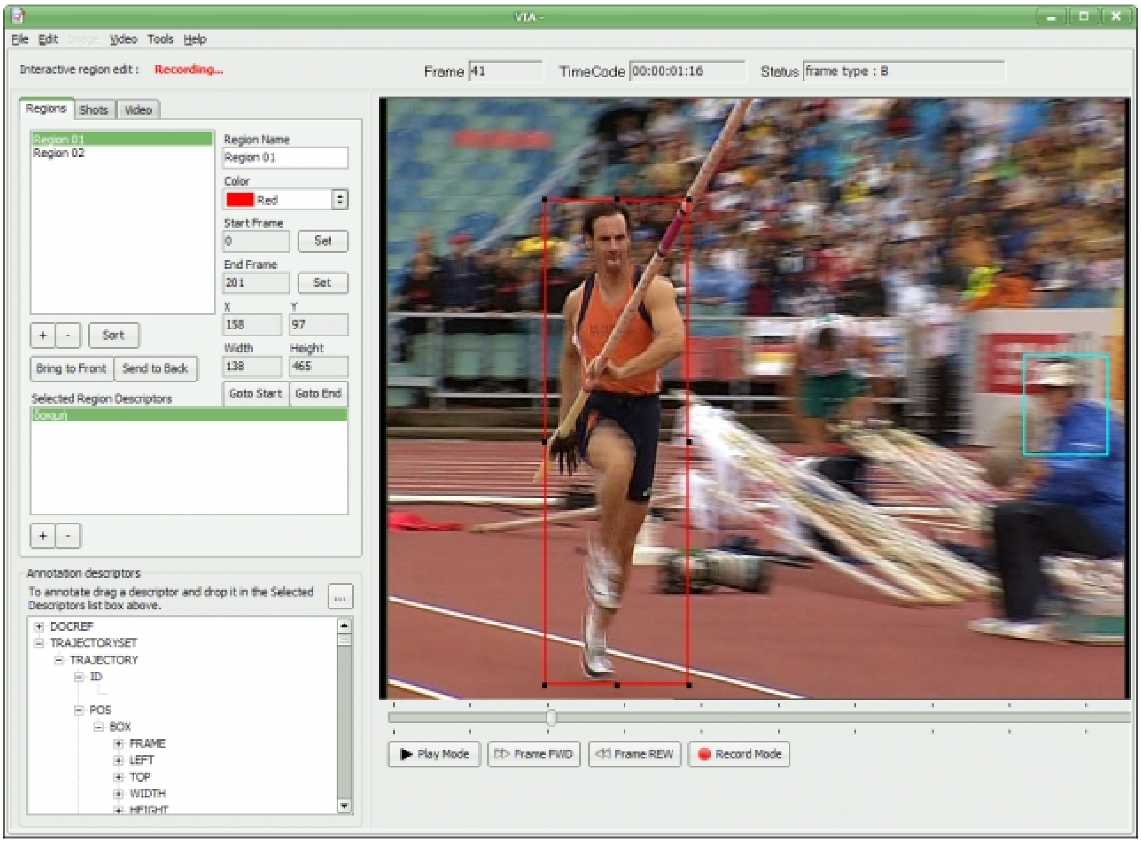
Figure 13. *The user interface of OntoLog* (Heggland 2005b)

Expressiveness

- Temporal - Free overlapping intervals
- Spatial - None
- Semantic – Ontology-based descriptors

4.3.3 Video Image Annotation tool (VIA)

Video Image annotation tool is an open source Windows application by Informatics and Telematics Institute, Centre for Research and Technology Hellas(CERTH-ITI). It is designed to be used to manually annotate video and still images. Video annotations support frame based annotations and segmentation. It also supports spatial annotations – region to be annotated is selected with a bounding box, the size and position of which is to be adjusted per each frame. The annotation methods support live annotations – annotating while playing and frame by frame annotations. The type of annotations is free text, however the annotations can be saved to a structured format XML for instance, which allows a level of semantics to the annotations. The tool supports MPEG-1 and MPEG-2 video playback .



Interactive region edit : Recording...

Frame 41

TimeCode 00:00:01:16

Status frame type : B

Region 01
Region 02

Region Name: Region 01

Color: 🟥 Red

Start Frame: 0

End Frame: 201

X: 158 Y: 97

Width: 138 Height: 465

Selected Region Descriptors: Boxquh

Annotation descriptors

To annotate drag a descriptor and drop it in the Selected Descriptors list box above.

Figure 14. *The user interface of VIA*

Expressiveness

- Temporal - Free overlapping intervals
- Spatial - Static
- Semantic – Keyword descriptors

4.3.4 M-OntoMat-Annotizer

M-OntoMat-Annotizer is an Java-based annotation tool, or software environment, developed inside the acemedia project. The tool is an extension of CREAM (CREAting Metadata for the Semantic Web) framework. OntoMat-Annotizer tool is a CREAM's concrete implementation, which support methods for manual annotations, authoring the documents, semi-automatic annotations and annotation of dynamic content. Visual Descriptor Extraction Tool (VDE) is a plug-in to OntoMat-Annotizer that allows initialization and linking of RDF(S) domain ontologies with low-level MPEG-7 visual descriptors. For describing content a Visual Descriptor Ontology (VDO) has been developed. It contains the expression of visual descriptors in RDFS, tightly coupled with the specification MPEG-7. Tool can be used to automatically extract features (descriptors included in the VDO) of a specific object inside a frame. Annotations of video are done with an graphical user interface (Figure 15) frame-by-frame manner, but there is no support for visualizing the annotations performed so far (Petridis et al. 2006) .

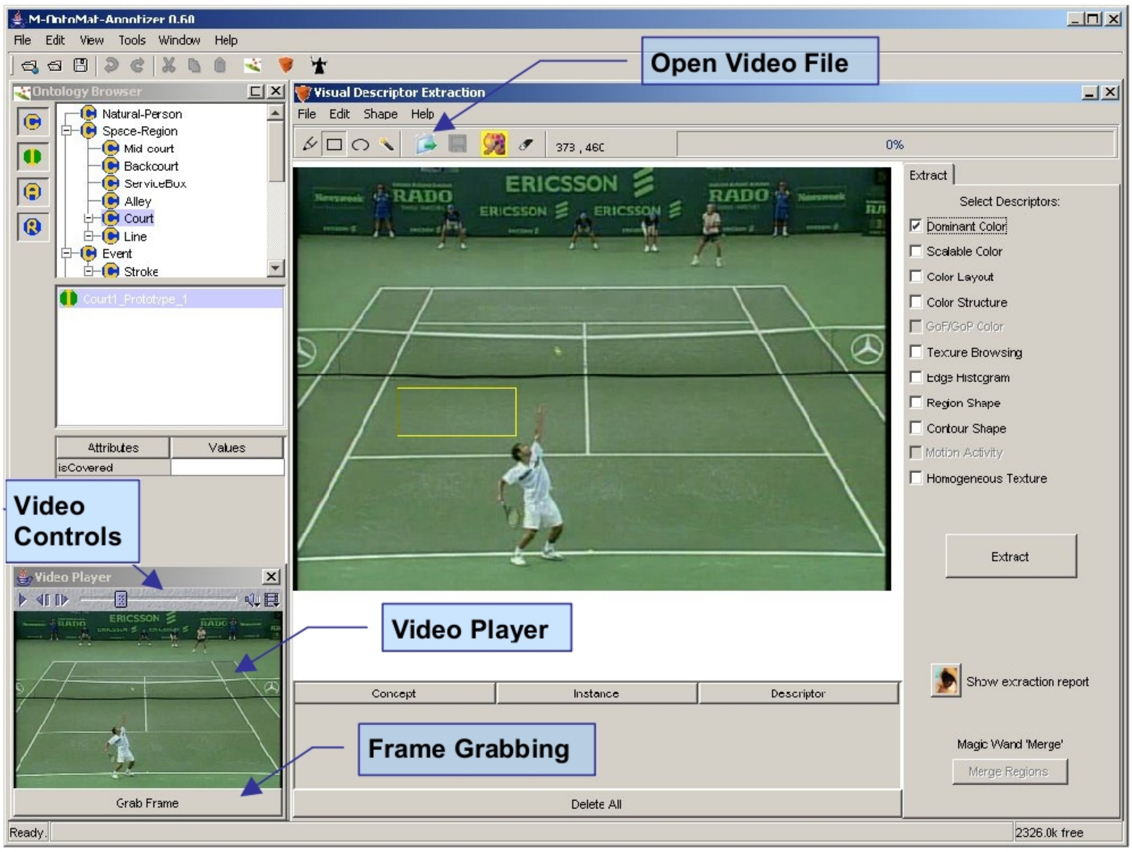


Figure 15. *The user interface of m-Ontomat-Annotizer* (Petridis et al. 2006)

Expressiveness

- Temporal – Fixed segmentation (frame)
- Spatial - Static regions (frame-by-frame)
- Semantic – Ontology-based descriptors (RDFS)

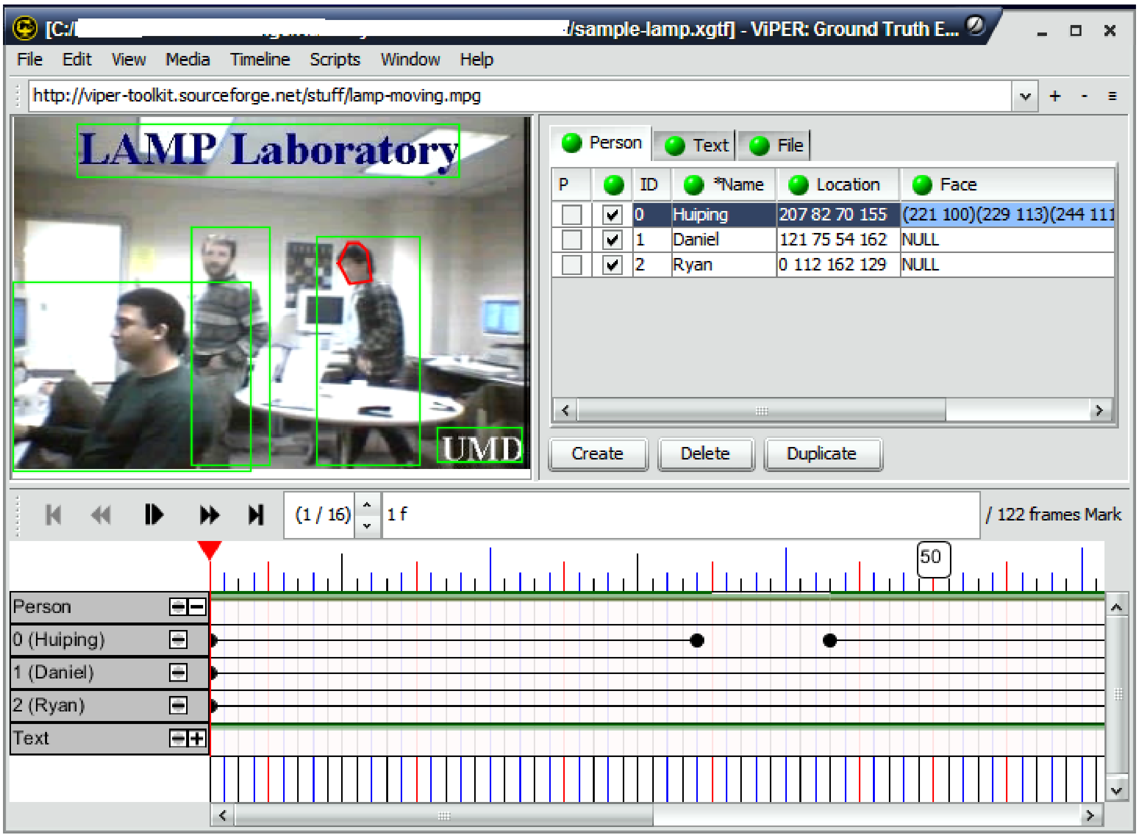
4.3.5 ViPER - GT

The Video Performance Evaluation Resource – Ground Truth Authoring Tool (ViPER-GT) is a Java-based video annotation tool originally developed to function as a ground truth authoring tool for testing automated analysis software performance. It allows annotations with metadata including descriptions of content of the document, such as date of filming and keywords about its content. Included are also concrete features such as scene breaks and bounding boxes around objects, such as people. The tool is designed for editing spatial annotations and implemented with a graphical user interface. The forms of spatial annotation regions include points, bounding boxes and oriented rectangles, ellipses, polygons and circles. There are also types without visual element, including text strings, numbers and boolean values. The tool supports operations for propagation and interpolation of spatial annotations, but the tracking has to be done manually.

Descriptors are used to describe elements of the video. A descriptor has an object that conforms to user defined schema, it is composed of several, typed attributes and has an unique ID. Each descriptor has one of three types: File, Content or Object. File refers to content-independent data, Content type instances has a time span and a set of attributes. Only one instance of the type Content can occur at a time, and instance can not change in time. Each descriptor has several attributes that can be used to build a simple hierarchy. Attribute values support various predefined data types mentioned above. Core elements of the ViPER; video annotation file format and data types have been defined with XML Schema in separate schema documents. Annotations are stored into a XML file. The graphical user interface is illustrated in Figure 16.

Expressiveness

- Temporal - Free overlapping intervals
- Spatial - Dynamic regions
- Semantic - Field-based descriptors (XML schema)



[C:/[redacted]/sample-lamp.xgtf] - ViPER: Ground Truth E...

http://viper-toolkit.sourceforge.net/stuff/lamp-moving.mpg

P	●	ID	*Name	Location	Face
□	☑	0	Huiping	207 82 70 155	(221 100)(229 113)(244 111
□	☑	1	Daniel	121 75 54 162	NULL
□	☑	2	Ryan	0 112 162 129	NULL

Figure 16. *The user Interface of Viper-GT*

5 Annotation of the stimulus movie

5.1 Introduction

This chapter describes the the practical part of this thesis, which was to annotate visual features of the stimulus movie used in the neuroscientific studies. The purpose of the annotation is to acquire simplified model of the complex visual stimulus. The visual features are annotated on the basis of the brain activation data.

5.2 Methods

The movie

The annotated movie was a re-edited version of The Match Factory Girl (Aki Kaurismäki, 1990), which was used as a stimulus movie in the studies of Lahnakoski (2010) and Nieminen (2010). The movie has the length of 31 min 10 s with a frame rate of 25 frames per second.

The visual content of the movie is minimalistic; camera movements are low, which emphasizes the meaning of also minimal movements of the objects. The storytelling is straightforward and clear - the meaningful events in principle do not occur simultaneously, and the viewers attention is directed with intuitive methods.

Annotated features

The annotated features were selected in a data driven manner by examining the acquired IC components. Selected features to annotate were main characters; the spatial locations of the following body parts:

- Body
- Head
- Eyes
- Hands

Locations of objects hold great deal of information about the stimulus movie. This information includes the relative sizes of the objects. If the image has more than one object the spatial relationships between the objects and distances can be inferred from these features.

While form and location of objects are low-level features, these features can be used to infer high-level features. For example studying social interaction can be inferred from first of all from the number of human characters in the movie image. Secondly the relative size and the distance between the objects can reveal details about the nature of the interaction.

Annotation tool

BECS Low-Level feature toolbox (BECS- LLFT) was used for the annotations. It is an video annotation tool with graphical user interface. The toolbox has been developed at BECS by Jussi Rytkönen and Juho Kettunen. It is written with Matlab. The toolbox supports dynamic spatiotemporal annotations with free overlapping intervals. The user interface consists of three parts: timeline, video player, and annotator. In Figure 17 is the user interface of the tool. The tool extracts automatically the structural features of the movie; shot boundaries (cuts) and scenes. Is is also capable of pattern recognition, which makes it capable of semi- automatic movement annotations of objects.

The tool supports spatiotemporal annotations and because of its pattern recognition abilities it is capable of tracking movement of objects quite reliably. The object tracking is not perfect, thus the annotations can be done semi-automatically. Each one of the descriptors forms a stratum of its own, but the timeline does not support visualization of all the strata simultaneously.

This denotes in practice very efficient spatial and temporal annotations, but high-level feature annotations and inference require more semantic expressiveness.

Expressiveness:

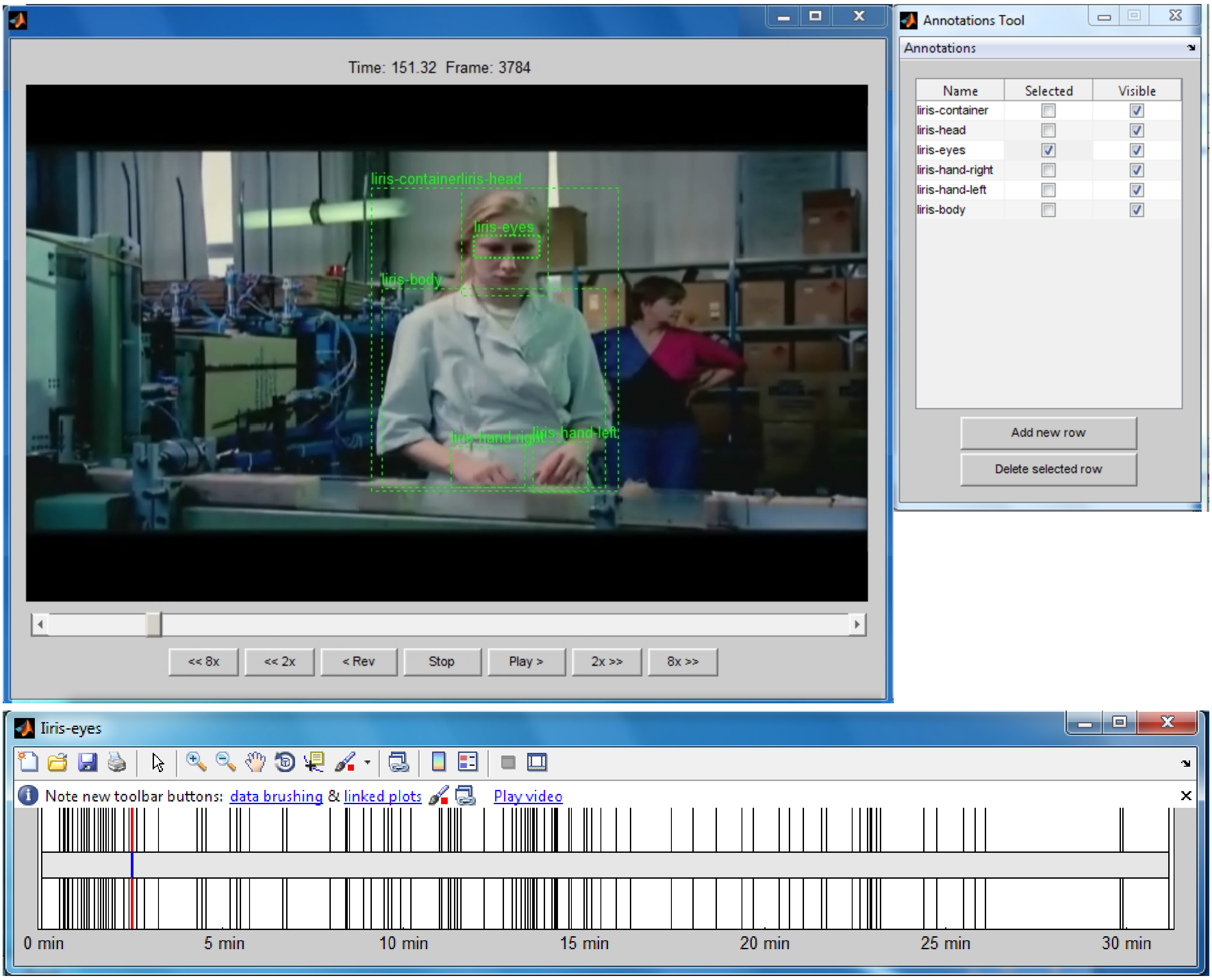
- Temporal - Free overlapping intervals
- Spatial - Dynamic regions
- Semantic – Single, Free-text descriptor

5.3 Procedure

Annotations

A lexicon for the annotations was determined. The annotations were organized by the tool were in stratified structure with no hierarchy. For this reason hierarchy was artificially added to the structure by naming the strata so that the in the case of hierarchically linked strata, the name of a parent stratum was added to the beginning of the name of the child stratum.

Also the spatial annotations were constructed so that for each character an extra annotation, called container was assigned to contain all the body parts. This was done to help to organize the annotation. Next low level features (content-dependent metadata) annotation was performed using the annotation tool. Segmentation based on shot boundaries extraction was executed automatically and the partially manual annotations included movement annotations of main characters of the movie. An example annotation with the user interface of the BECS – LLFT is illustrated in Figure 17. High level features (content-descriptive metadata) were annotated in the form of describing the main characters and their body parts.



Time: 151.32 Frame: 3784

Name	Selected	Visible
liris-container	☐	☑
liris-head	☐	☑
liris-eyes	☑	☑
liris-hand-right	☐	☑
liris-hand-left	☐	☑
liris-body	☐	☑

Note new toolbar buttons: data brushing & linked plots Play video

Figure 17. *The user Interface of the BECS- LLFT with an example annotation of the selected features.*

6 Discussion & Conclusions

Studying the brain with movie as a stimulus is possible, thanks to fMRI and analysis methods like ICA and ICS. Using a movie as a naturalistic stimulus is well justified, since in many cases movies mimic reality. However, understanding the multiple levels of this complex stimulus is problematic. A film historian Christian Metz (1931-1993) has stated “Film is hard to explain, because it is easy to understand”. This problem can be alleviated by understanding the cinematic code - the means of communication of the film medium.

In neurocinematics annotations are used to model the stimulus movie. It is challenging to produce reliable and extensive annotation with feasible costs. Low-level features can be reliably annotated by using automatic annotations, but since the domain of movies is broad manual annotations are needed for high level features. However, this semantic gap is bridging slowly due to the ongoing development of video analysis methods.

The process of annotation was clarified and the studies revealed that organizing annotations is often an overlooked problem. The level of organization is decided with the annotation approach as explained in section 3.4. The selected approach has a straight effect on the expressiveness of the acquired model. The levels of structure in annotations range from free text to highly structured ontology-based annotations. Organizing the annotations can have various advantages as explained in section 3.4.4. In neurocinematics an ontology-based annotation approach could be used to build a highly expressive model of the relations between stimulus and the brain activation data. The model could also be used to computationally infer about correlations between the activation and the movie features. On the other hand, annotations could also become too complex; the quality of annotations depends on the context of use. Thus the suitability of ontology-based annotations for neurocinematics should be studied further. Some of relevant technologies for building an ontology-based annotation framework were reviewed in the section 3.4.5.

Annotations are found everywhere and reasons for using them are numerous. New annotation tools and technologies are being developed all the time. Much of this is due to the ever increasing amount of data in the WWW. At the moment of the executed survey on video annotation tools there was not yet perfect tool available to be used for our purposes.

Bibliography

Agosti, M. et al. (2004) Annotations in Digital Libraries and Collaboratories – Facets, Models and Usage. In Research and Advanced Technology for Digital Libraries ECDL 2004. Springer Berlin / Heidelberg, pp. 244-255.

Allemang, D. & Hendler, J.A. (2008) *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL.* Burlington(MA):Morgan Kaufmann. 330p.

Antoniou, G. & van Harmelen, F. (2008) *A Semantic Web Primer* 2nd., Cambridge(MA): The MIT Press. 238p.

Bartels, A. & Zeki, S. (2005) Brain dynamics during natural viewing conditions- a new guide for mapping connectivity in vivo. NeuroImage, 24, pp. 339-349.

Bartels, A. & Zeki, S. (2004) Functional brain mapping during free viewing of natural scenes. *Human Brain Mapping*, 21, pp. 75-85.

Bartels, A. & Zeki, S. (2004) The chronoarchitecture of the human brain-natural viewing conditions reveal a time-based anatomy of the brain. NeuroImage, 22(1), pp. 419-433.

Beckmann, C.F. & Smith, S.M. (2004) Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging. FMRIB Technical Report TR02CB1, pp. 1-27.

Blanken, H. et al. (2007) Introduction. In H. M. Blanken et al. *Multimedia Retrieval*. New York:Springer, pp.1-22.

Chaudhary, A. (2008). *Video Annotation tools*. M.Sc. thesis. Texas A&M University. 45 p.

Chua, T., Chen, L. & Wang, J. (2002) Stratification Approach to Modeling Video. Multimedia Tools and Applications, 16, pp.79-97.

Daconta, M.C., Orbst, L.J. & Smith, K.T. (2003) *The Semantic Web: A Guide to the Future of XML, Web Services and Knowledge Management*, Wiley Publishing, Inc. 281 p.

Del Bimbo, A. (1999) *Visual Infromation Retrieval*, San Fransisco(CA):Morgan Kaufmann. 270 p.

Desmukh, A.V. & Gadre, V.M. (2008) *Functional Magnetic Resonance Imaging - Novel Transformation Methods.* Delhi:Narosa Publishing House. 106 p.

Dong, A. & Li, H. (2008) Ontology-driven annotation and access of presentation video data. Journal of Theoretical and Applied Information Technology.

Eastaway, M. et al. (2007) Blind Source Separation Via ICA: Implementation. Available at: http://cnx.org/content/m15639/1.2/.

Edyburn, D.L. (2008) collecting and coding observational data. Journal of Special Education Technology, 23(2).

Geurts, J., Ossenbruggen, J.V. & Hardman, L. (2005) Requirements for practical multimedia annotation. Workshop on Multimedia and the Semantic Web, pp.4-11.

Hasson, U. & Malach, R. (2005) Human Brain Activation During Viewing of Dynamic Natural Scenes. Novartis Foundation Symposium NF, 270, pp.203-212.

Hasson, U. et al. (2004) Intersubject synchronization of cortical activity during natural vision. Science, 303(5664), pp.1634-1640.

Hasson, U. et al. (2008) Neurocinematics: The Neuroscience of Film. *Projections*, 2(1), pp.1-26.

Hasson, U., Malach, R. & Heeger, D.J. (2009) Reliability of cortical activity during natural stimulation. Trends in Cognitive Sciences. 2009, Vol.14(1), pp.40-48.

Heggland, J. (2005) OntoLog - Ontology-Based Annotation of Video and Audio.[online]. Available at: http://www.idi.ntnu.no/~heggland/ontolog/. [Accessed 21 April 2010]

Heggland, J. (2005) OntoLog: Flexible Management of Semantic Video Content Annotations. Doctoral thesis. Norwegian University of Science and Technology, Department of Computer and Information Science. Trondheim. 254 p.

Hillmann, D. (2007) Using Dublin Core. [online] Available at: http://dublincore.org/documents/2005/11/07/usageguide/. [Accessed 21 April 2010]

Himberg, J., Hyvärinen, A. & Esposito, F. (2004) Validating the independent components of neuroimaging time-series via clustering and visualization. NeuroImage, 223 pp. 1214-1222.

Huettel, S.A., Song, A.W. & McCarthy, G. (2004) *Functional Magnetic Resonance Imaging* 1st., MA:Sinauer Asociates, Inc.

Hyvönen, E., Styrman, A. & Saarela, S. (2002) Ontology-Based Image Retrieval. Helsinki Institute of Information Technology (HIIT), Finland. Number 2002-03 in HIIT Publications. pp.15-27

Iacoboni, M. (2008) *Mirroring People, The New Science of How we Connect With Others.* Helsinki:Terra Cognita. 230p.

Jääskeläinen, I.P. et al. (2008) Inter-Subject Synchronization of Prefrontal Cortex Hemodynamic Activity During Natural Viewing. The Open NeuroImaging Journal, 2 pp.14-19.

Kipp, M. (2008) Spatiotemporal Coding in ANVIL. Proceedings of the 6th international conference on Language Resources and Evaluation (LREC-08), 2042-2045.

Koskentalo, K. (2007) *Brain Activation Networks During Natural Viewing Conditions*. Diploma thesis. Helsinki University of Technology, Department of Electrical and Communications Engineering. 69p.

Lahnakoski, J. (2010) *Functional Magnetic Resonance Imaging of Human Brain during Rest and Viewing Movies*. M.Sc Thesis. Aalto University School of Science and Technology, Faculty of Electronics, Communications and Automation. 59 p.

Lassila, O. & Mcguinness, D. (2001) The Role of Frame-Based Representation on the Semantic Web. Electronic Transactions on Artificial Intelligence, 6(005).

Martínez, J.M. (2004) MPEG-7 Overview. [online] Available at: http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm. [Accessed 21 April 2010]

McBride, D. (2009) An Introduction to RDF and the Jena RDF API. [online] Available at: http://jena.sourceforge.net/tutorial/RDF_API/index.htm.[Accessed 21 April 2010]

Mechouche, A. et al. (2009) A Hybrid System Using Symbolic and Numeric Knowledge for the Semantic Annotation of Sulco-Gyral Anatomy in Brain MRI Images. , 28(8), 1165-1178.

Mechouche, A. et al. (2008) Ontology-based annotation of brain MRI images. *AMIA* Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, 460-464. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18998967.

Musen, M. (2010) what is protégé? [online]. Available at: http://protege.stanford.edu. [Accessed 21 April 2010]

Nagao, K. (2003) *Digital Content Annotation and Transcoding.* Norwood (MA):Artech House.

Nieminen, J. (2010) *Brain activation in a complex stimulus.* M.Sc Thesis. Aalto University School of Science and Technology, Faculty of Electronics, Communications and Automation.58 p .

Nordberg, A. (1986) Elokuvan Ilmaisukeinot. In A. NordBerg *Eläviä Kuvia Videolla*. Oulun läänin taidetoimikunta, Keski-Suomen läänin taidetoimikunta, pp. 17-36.

Petridis, K. et al. (2006) M-OntoMat-Annotizer: Image Annotation Linking Ontologies and Multimedia Low-Level Features. 3rd European Semantic Web Conference.

Pirilä, K. & Kivi, E. (2008) *Leikkaus: Elävä Kuva - Elävä Ääni Toinen osa.* Helsinki: LIKE.

Pirilä, K. & Kivi, E. (2005) *Otos: Elävä Kuva - Elävä Ääni Ensimmäinen osa.* Helsinki:LIKE.

Prunes, M., Raine, M. & Litch, M. (2002) Yale Film Studies: Film Analysis Web Site 2.0. [online] Yale University, New Haven, CT 06520. Available at: http://classes.yale.edu/film-analysis. [Accessed 21 April 2010]

Ranguelova, E.& Huiskes, M. (2007) Pattern Recognition for Multimedia Content Analysis. In H. Blanken et al. Multimedia Retrieval. Springer, pp. 53-95.

Rossier, J. (2009) Wiring and plumbing in the brain. Frontiers in Human Neuroscience. Available at: frontiersin.org/neuroscience/humanneuroscience/paper/10.3389/neuro.09/002.2009/.

Styrman, A. (2008) *Ontology-Based Image Annotation and Retrieval.* M.Sc. Thesis. University of Helsinki, Dept. of Computer Science. 95p.

Tikka, P. (2008) *Enactive Cinema: Simulatorium Eisensteinense.* Publication Series of the University of Art and Design Helsinki A 89. 338p.

Tohka, J. (2008) Automatic independent component labeling for artifact removal in fMRI. *NeuroImage*, 39, 1227-1245.

Wollen, P. (1972) *Signs and Meaning in the Cinema.* London:Secker & Warburg. 176p.

Xu, C.Z. et al. (2009) A Hierarchical Video Annotation System. *2009 IEEE* International Conference on Semantic Computing, 670-675. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5298722.