Information Polity 16 (2011) 339–353
DOI 10.3233/IP-2011-0252
IOS Press

The thinking eye is only half the story: High-level semantic video surveillance

Christoph Musik
Recipient of a DOC-team-fellowship of the Austrian Academy of Sciences at the Department of Social Studies of Science, University of Vienna, Universitätsstraße 7/Stg. II/6. Stock, A-1010 Vienna, Austria
Tel.: +43 1 4277 496 15; E-mail: christoph.musik@univie.ac.at

Abstract. An increase in video surveillance systems, paired with increased inquiry for efficiency, leads to the need of systems which are able to process and interpret video data automatically. These systems have been referred to as 'algorithmic video surveillance', 'smart CCTV', or 'second generation CCTV surveillance'. This paper differentiates and focuses on 'high-level semantic video surveillance' by referring to two case studies: Facial Expression Recognition and Automated multi-camera event recognition for the prevention of bank robberies. Once in operation these systems are obscure, therefore, the construction process of high-level semantic VS is scrutinized on the basis of a 'technology in the making' approach.

Keywords: Algorithmic video surveillance, smart CCTV, technology in the making, facial expression recognition, event recognition, computer vision

1. Introduction

> "Look, Dave, I can see you're really upset about this. I honestly think you ought to sit down calmly, take a stress pill, and think things over."

There is nothing extraordinary about this quotation. Somebody is telling a person called Dave that he looks really upset, and that he should sit down and take a stress pill. This type of situation is likely to be an everyday occurrence in western society, movie fans however will easily recognize the quotation as the computer HAL 9000 from Stanley Kubrick's and Arthur C. Clark's classic "2001: A Space Odyssey". In this science fiction film from 1968 the intelligent computer HAL 9000

> "displayed image understanding capabilities vastly beyond today's computer systems. HAL could not only instantly recognize who he was interacting with, but also he could lip read, judge aesthetics of visual sketches, recognize emotions subtly expressed by scientists on board the ship, and respond to these emotions in an adaptive personalized way." *Rosalind W. Picard 2001 [29]*

HAL 9000 had capabilities that strongly refer to Facial Recognition Technologies (FRT), Automatic Lip-reading, Motion Analysis, Behavioural Pattern Analysis (BPA), or Facial Expression Recognition. All these technologies can be ascribed to the field of Pattern Recognition and Computer Vision (a subfield of Computer Science) and many of these technologies have been mentioned and analyzed by scholars in the field of Surveillance Studies. In this context Norris and Armstrong coined the term 'algorithmic surveillance' [27], exemplifying it with 'intelligent scene monitoring', 'digital facial recognition systems', and 'license plate recognition'. The term has been adopted by Introna and Wood for the analysis of the Politics of Facial Recognition Systems [12]. They define algorithmic surveillance in a literal sense as surveillance that makes use of automatic step-by-step instructions, especially of computer systems, to provide more than the raw data observed. Another popular term for algorithmic surveillance is 'Smart CCTV', recently used by Gates when analyzing the failure of FRT in a CCTV system in Tampa, Florida [7] and also used by Introna and Wood [12]. Surette, using the metaphor of 'The thinking eye', introduced the term 'second generation CCTV surveillance systems' which are 'smart' and exploit digital technologies for artificial intelligence scene monitoring, e.g. the detection of people, unauthorized traffic, or unusual behaviour [37]. In contrast to this there are 'first generation CCTV surveillance systems' that are 'dumb', and based solely on human monitoring. The terms algorithmic surveillance, Smart CCTV, and second generation CCTV surveillance systems have been widely used synonymously. The covering term to use is 'algorithmic video surveillance', the term 'smart CCTV' should be avoided, because it is likely to promote CCTV as being brilliant, clever, effective, or knowing (to name only some of its synonyms). Also the term 'second generation CCTV surveillance system' fixes a limit where there is none and moreover is too imprecise: For example, the detection of unusual behaviour entails much more than simply detecting a person, but would also be considered second generation. Thus, 'algorithmic video surveillance' should be used as an umbrella term, but it does make sense to distinguish precisely within this term. There is a big qualitative difference between automatically detecting a person in a specific scene, recognizing who this person is, detecting in which direction this person is moving, or detecting that the person's behaviour does not fit defined norms. Computer scientists, such as Turaga et al. [39] note that human actions and activities can be recognized on four different levels:

1) Input video or sequence of images
2) Extraction of concise low-level features (e.g. tracking and object detection)
3) Mid-level action descriptions from low-level features (action recognition modules)
4) High-level semantic interpretations from primitive actions

Following this scheme, *low-level* 'algorithmic Video Surveillance' can detect that there is a person present and track this person in a specific area of surveillance. *Mid-level* surveillance systems can use the extracted low-level information and recognize that the person is walking or running. Finally, on the *high-level* this walking or running can be interpreted under certain circumstances as suspicious behaviour, to give one example.

Especially the development and deployment of such high-level semantic Video Surveillance (VS) changes the relationship between humans and machines, because interactivity between the two is itself changing. Machines are gradually becoming more able to act autonomously and gain a higher grade of agency [32]. Once those systems are in operation they are obscure [12]; the system's mode of decision-making is black-boxed, the consequences however can be extensive.

1.1. Aim and focus

This article argues that computer and machine vision technologies referred to as 'algorithmic video surveillance', 'smart CCTV' or 'second generation CCTV surveillance', especially high-level semantic video surveillance systems are not fully comparable with human vision abilities and therefore the metaphor of the thinking eye for the connection of a video camera to computer hard- and software is only half the story. Nevertheless, this comparison between machine vision and human vision is widely drawn. The article aims to show based on a 'technology in the making' approach, that state-of-the-art high-level semantic Video Surveillance (VS) is able to accomplish certain tasks, but does not fulfill expectations of simulating human vision abilities. Instead the article shows that it reduces and oversimplifies human vision abilities, because on the one hand it is not technically feasible yet and on the other hand lacks 'social knowledge'.

The main concern of the article is to disclose what kinds of reductions of complexities of vision and perception are made and how this impacts our conception of how we see and perceive. This allows us to reflect on what an integration of such technologies into contemporary societies means and what kinds of new orderings will take place once these technologies are integrated into social life.

1.2. Structure

On the basis of a 'technology in the making' approach the article presents and discusses two high-level semantic Video Surveillance (VS) applications, which highlight different aspects of this kind of VS.

First, the 'technology in the making' approach is outlined, which tries to combine concepts from Science and Technology Studies (STS) with the Surveillance Studies' examination of the construction of code. Then two case studies are presented: 'Facial Expression Recognition' and 'Automated multi-camera event recognition for the prevention of bank robberies'. The first of these is located more in basic computer vision research, but also incorporates applied research elements. The second case is an applied research project designed for a very specific task (recognition of exploring bank robbers) and place (Austrian bank branches), but does also deal with basic computer vision research questions like multi-camera tracking. Therefore the first case has a more far-reaching universal claim, the second is limited to a concrete environment. Furthermore, 'Facial Expression Recognition' can draw on the tradition of facial expression research, which can be traced back to the 19th century, whereas the automated event recognition project had to generate new empirical data through interviews, observation and document analysis.

The two different types of high-level semantic VS have one crucial aspect in common: in both cases the so-called 'ground truth', the basis for teaching a machine to see had to be created. In the first case the 'ground truth' corresponds with the question 'what a specific facial expression, e.g. anger, looks like'; in the second case the 'ground truth' corresponds with the question 'what suspicious behaviour of an exploring bank robber looks like'. In the empirical section the article traces the construction of both cases 'ground truth' to show how categories of both facial expressions and "normal" or "suspicious" behaviour are created. In the concluding section of the paper the impact of complexity reductions of vision and perception and associated policy issues are discussed.

1.3. Methods

While both case studies aim at analyzing 'technology in the making' they nevertheless differ in the methods applied. For the first case five explorative in-depth interviews with computer scientists and behavioural scientists from Germany and Austria working in the field of Facial Expression Recognition were conducted in 2009 and basic papers in this area of research were surveyed. In addition, the history of facial expression research has been traced back to its beginnings, because the knowledge applied in Facial Expression Recognition systems is strongly grounded in its history.

The second case study 'Automated multi-camera event recognition for the prevention of bank robberies' presents empirical social scientific findings that have been produced by the author of this article in the framework of an interdisciplinary research project within the Austrian security research programme

KIRAS.[1] In this programme projects developing security technology are obliged to integrate a partner from the Social Sciences and Humanities in order to ensure socio-political compatibility. In this case the project consortium was managed by a Software Consulting company and was performed in cooperation with computer scientists, a commercial bank, social scientists and the Austrian Federal Criminal Police Office, Department of Crime Prevention and Victim Aid. For the social scientists the project was methodologically challenging as their role was far from being obvious at the beginning of the project. This role could be described as ranging from 'figleaf' or 'annex' to a fully integrated and critically reflecting partner of the technological development.

A key question for the social scientists emerging over the course of the project was whether it is possible to identify and define suspicious behaviour in the context of a bank (more precisely the behaviour of exploring bank robbers) and if so, how this could be translated into the programme of the technical system. This question was addressed by observing "normal" behaviour, describing activities of bank customers in detail. Observations in two different bank branches, as well as video analysis of seven project-cameras installed in one additional branch, were performed. The method of non-participant observation was used, combined with video-analysis in Social Research [17]. Within four observation sessions a sample consisting of 236 people was observed. To contrast the observations of "normal behaviour" of bank clients with the behaviour of exploring bank robbers records of interrogation footage of apprehended bank robbers were surveyed.

2. 'Technology in the making' and the co-production of technology and society

It is incontestable that nowadays technology plays a crucial role in surveillance contexts. In the past visual surveillance has been a matter of face-to-face communication, now it is also characterized by high-technology applications [24]. As can be observed for surveillance [27], technology is pervasive in all areas of everyday life, and is an ever-present part of social reality. In most cases of daily life technology works in the way we expect it to and therefore we are usually not interested in how exactly a specific technology functions. If a technological artifact is not performing adequately more often we are not able to fix it ourselves, because we do not have sufficient knowledge and the skills to do so. The technological artifact then appears as a 'black box' and as 'ready made technology', which seems to operate in a fixed and predictable manner [35].

This article's entry into science and technology will be through the back door of 'science and technology in the making' and not through the more grandiose entrance of 'ready made science and technology' [19]. The main purpose is the understanding of how technology, in this case high-level semantic Video Surveillance (VS) technology, is being constructed in computer scientists' laboratories. Of course this does not mean to forget the wider societal context in which these construction processes occur. The development and deployment of high-level semantic VS and the production of society are connected in a seamless web, so one can speak about a process of co-production. Generally technology "both embeds and is embedded... in all the building blocks of what we term the social" [14, p. 2]. This means that expectations and imaginations of high-level semantic VS are inseparable intertwined with the "ways in

[1] KIRAS (acronym from the Greek kirkos for circle and asphaleia for security) supports national research projects which aim to increase the security of Austria and its people. The protection of critical infrastructure was selected as the first thematic focus. The programme started in 2005 and is scheduled for a duration of 9 years. KIRAS is an initiatve of the Federal Ministry of Transport, Innovation and Technology (BMVIT) managed by the Austrian Research Promotion Agency FFG. For more information see http://www.ffg.at/en/kiras-security-research.

which we choose to live in" [14, p. 2]. The development of technology both shapes as well as being shaped by the specific societal context in which it is embedded. Analyzing high-level semantic VS 'in the making' allows grasping these co-production processes of technology and society in detail. In the following a theoretical discussion of high-level semantic VS technology in the context of co-production is presented.

2.1. On the co-production of technology, knowledge, code, and society

Studying high-level semantic VS technology initializes the question of what kind of knowledge and computer codes are applied, transformed, and co-produced in the same way. Both can be regarded as social, active, not natural, and both are constructed, produced, manufactured. Here one can draw on laboratory studies [18,22]. Laboratory studies analyze the manufacture of techno-scientific facts and knowledge in situ in scientists' laboratories. "Facts are not something we can take for granted or think of as the solid rock upon which knowledge is built" [18, p. 1]. Knorr Cetina gives meaning to the "decision-ladenness" and selectivity of fact-fabrication. Thus, it is important "to study the process by which the respective selections are made" [18, p. 7].

In the context of high-level semantic VS we have to bring to mind the specificity of knowledge. Basically it concerns the pressure to translate implicit into explicit knowledge. This pressure, which can be found in more and more areas of life, is generated through the increasing application of Information Technologies (IT) on the basis of computers. So far, most decisions and activities have been based on implicit or tacit knowledge of the people involved. By tacit knowledge Polanyi [30] means that 'we can know more than we can tell'. Tacit knowledge is not captured by language or mathematics, but has to be performed.

Nowadays these activities and decisions based on implicit or tacit knowledge are increasingly delegated to IT systems. In this process, the implicit or tacit knowledge has to be made explicit. Thus, rules of activities and decisions have to be identified and specified in a way which answers to specificities of computer programmes. In further consequence, they have to be formalized and codified [32].

The process of making tacit knowledge explicit has been described as an issue of reduction; this issue especially refers to FRT [12,15], the process of reducing complexity has consequences. In the case of FRT, for example, minorities are easier to recognize [12]. The problem with algorithmic surveillance systems is, that the issue of reducing information is a requirement, because systems only operate with the binary codes of 1s and 0s [8]. Getting inside the production of these computer codes – that distinguish between one group and another – is becoming more and more important [24]. Moreover, these processes are now the only parts "completely open to human discretion and shaping" [8]. This is especially important when that coding is done from afar, removed from the point of application [23] and therefore ignores the specialities and peculiarities of this point of application. The crux of the matter is that coding, especially relating to classification such as social sorting [24] never occurs in an objective or neutral way, but is embedded in specific social practices. Bowker and Star see software in many ways as "frozen organizational and policy discourse" [1], in which policy is coded into software. In this view software, like technology, is 'society made durable' [20]. This means that specific social practices, normative notions of good behaviour, political assumptions, and cultural values are either consciously or tacitly inscribed in the software [8]. Moreover, "algorithmic systems thus have a strong potential to fix identities as deviant and criminal" – what Norris calls the technological mediation of suspicion [25]. However it is not only the individual person that is singled out for attention, in some circumstances coding and classification processes may have profound effects on the shaping and ordering of human life in general, creating new social classes [24].

3. High-level semantic video surveillance: Case studies

In this section the two case studies of high-level semantic VS are presented by describing the extent of research, possible application areas, and especially by understanding the construction of both cases 'ground truth'.

3.1. Facial expression recognition

A relatively new field of research in computer vision are technologies of Facial Expression Recognition. These technologies aim at determining the mood and emotions of a person automatically and in real time. A specific facial expression is related to a specific basic emotion, like happiness or anger. First approaches for facial expression recognition emerged in the 1990s, today we can find research in this area in at least 70 research institutions around the world. The contextualization of facial expression recognition exists especially in two areas: Human-Machine Interaction (HMI) and Video Surveillance (VS). In the case of HMI machines (robots, computers etc.) are required to be able to detect and interpret human facial expressions automatically. The aim is to improve interaction between humans and machines in general [13], because it is argued that humans expect machines to behave like humans [40]. Facial Expression Recognition technologies could, for example, be integrated in ticket machines or personal computers, to recognize when the user becomes frustrated and then to provide help as a result of the recognition.

The second area of application is Video Surveillance. Facial Expression Recognition is intended to become part of workplace monitoring systems, research on the impact of advertisements on consumers in public as well as in private space, consumer research (one example is the commercial software FaceReader$^{\mathrm{TM}}$[2]) and in the detection of terrorists, e.g. under the US security program SPOT (Screening Passengers by Observational Techniques), which was introduced in 14 US airports in 2006.

3.1.1. Historical embedding of facial expression recognition

For a long time science has tried to make human beings, and especially the human body, readable. The human face was, and still is, of special interest. It has been measured not only for identification claims, but also in the hope of gaining access to the 'inside' of human beings. One can look back upon the ideas of the ancient worlds from Aristoteles' *Historia Animalium* to pre-Confucian China, with its face readers, to meet with physiognomy, "the study of the systematic correspondence of psychological characteristics to facial features or body structure."[3] In the past physiognomy has been situated between the poles of the sciences and the arts, and is today said to be non-scientific. On the other hand it is firmly grounded in daily life. We are not able to go through life without being confronted with physiognomy [34].

In late 18th century the founder of scientific physiognomy, Swiss Johann Caspar Lavater, wanted to be able to recognize human character in the outlines of the human face on a scientific basis. Lavater worked with graphics and illustrations that were produced by different artists. These artists also had the task of standardizing and homogenizing the heterogeneous material for further usage (the German term 'Umzeichnen' in the words of Swoboda [38]). The artistic image had to be transformed into a scientific image for further analysis. Lavater also produced graphics on his own; most important were pictures

[2] According to the producer... "the world's first tool that is capable of automatically analyzing facial expressions" Noldus Information Technology, http://www.noldus.com/human-behavior-research/products/facereader

[3] Encyclopaedia Britannica, http://www.britannica.com/EBchecked/topic/458823/physiognomy.

of the silhouette, which he produced with the help of a special machine, objectivity was reached by mechanical picture-making [3]. The next step was to produce lines and angles that allowed mathematical calculations, classifications, and a specific order [38].

The era following Lavater can be characterized as the pathway from physiognomy to mimic and facial expressions, especially Charles Darwin's studies of facial expressions. Darwin's book *The expression of the Emotions in Man and Animals* from 1872 has to be read in physiognomical tradition, even though there is a radical change [28] away from the steady parts of the body and physiognomy (bodily frame and the bones), to the flexible parts of the body and the face, pathognomy and mimic [2]. On a more direct route classical physiognomy was continued, particularly in the phrenology of Franz Joseph Gall [2,34].

Darwin's book *The expression of the Emotions in Man and Animals* – which was published only four months after *The Descent of Man,* and which was actually planned to be published only as a chapter of the latter – was revisited by American psychologist Paul Ekman less than 100 years later in the mid 1960s when he started his research on facial expressions and emotions. Ekman and his colleagues created the Facial Action Coding System (FACS) on which virtually all efforts to recognize facial expressions are based. At the beginning of Ekman's research the fundamental question was if facial expressions are universal or specific to each culture. The result was that specific facial expressions are recognized as a specific emotion in every examined culture [4]. According to Ekman there are six basic emotions that are expressed in the same way in every culture worldwide: Anger, Disgust, Fear, Happiness, Sadness, and Surprise. However emotions are not only determined biologically, they are also culturally influenced and there are different display rules in every culture. Display rules are informal norms about when, where, how, and to whom one should express emotions [4]. Subsequently, Ekman focused on physiology and especially on facial muscles. In 1978 Ekman, together with Wally Friesen, developed a tool for measuring the face – the Facial Action Coding System (FACS) – which was revised in 2002 by Ekman, Friesen and Hager [6]. The FACS is a mode of coding the over 10.000 possible facial expressions of human beings and is based on the human anatomy of facial musculature. According to Ekman, FACS today is used "by hundreds of scientists around the world to measure facial movements" [5]. In addition to it "computer scientists are working hard on how to make this measurement automatic and speedy" (ib.). The aim of FACS was to create a comprehensive system of categories that can be used for defining all muscle movements of the face that are distinguishable with the human eye [4]. The movements of the face have been summarized into 44 Action Units. With the help of FACS, experts can describe facial movements; these have to be measured, classified, and then a specific emotion can be interpreted.

3.1.2. The 'ground truth' of Facial Expression Recognition – A matter of selection

The basis for teaching a machine to recognize facial expressions is the engineering of a so-called 'ground truth' or 'ground reality' of 'what a specific facial expression, e.g. anger, looks like'. But what does ground truth mean? In the following interview passage a computer scientist explains:

I3a:[4] "... And maybe back to the ground truth question. That is, I said all our algorithms are based on machine learning and for machine learning you supervise machine learning that means that you give the machine example data to train from. So for instance if you want a machine to recognise a specific person then you show the machine images of this person and you tell the machine that this image shows that person. You give the correct answer already in the training phase. If you want to

[4] Quotes from the interviews mentioned in section Ç1.3 Methods' are coded with I (for Interview) and the number of the Interview (1-5) at the beginning of the interview passage in bold letters. Interview 3 (I3) consists of two persons, which are marked with I3a and I3b.

recognize laughing or fear or whatever you show the machine images of laughing or afraid persons and you tell the machine these images show laughing or afraid persons. And so the machine can recognize it later. But in the training phase this information has to be given and this is called ground truth."

The ground truth does not exist from the beginning, but has to be generated. The machine has to learn from the computer scientist first. The computer scientist teaches the machine what the ground truth looks like, for example what the facial expression of fear looks like. In another passage, the comparison of machine learning and human learning is quoted as an example:

I3a: "... but it's pretty much to human learning. If you learn vocabulary then you have been given vocabulary. You have to match the German word to the English word. If you don't know the vocabulary and you hear the English word, you know the German word, you don't have to see it anymore. But during learning, of course, you have to match it. That's what the machine does."

Two things that mean the same have to be matched. Just like a German word has an equivalent in English and vice versa, an emotion, e.g. fear, has an equivalent in a facial expression, displayed on a digital image. But what does this equivalent look like? Who tells the machine which specific facial expression corresponds to which emotion? In the interview data can be found two different approaches: One is the 'FACS expert approach' and the other is the 'computer scientist layperson approach'.

a) Facial Action Coding System (FACS) expert approach:

I3a: Cohn-Kanade. That's a really standard database. Many, many people are working with that.

I3b: These databases are generated on the basis of ground realities. Cohn-Kanade facial expressions database is connected with the Facial Action Coding System.

INTERVIEWER: What does it mean ground reality?

I3b: For facial expressions there is a full coding of a face. That if you move your one eye up and if you are smiling so your lips are going up. Databases are generated by persons sitting in front of a camera.

INTERVIEWER: But people are said to do facial expressions?

I3a: In Cohn-Kanade they are advised to give an extreme facial expression. So if they have to smile, in the first streams they are neutral and in the ending they are really smiling. So they generate smile, it is not natural as when I am talking with you, but I am really forced to laugh.

INTERVIEWER: Is any expert controlling these expressions? Like in an experiment if anybody tells me to smile and is there anybody who says yes that's a correct smile or too much fake?

I3a: It depends on the database. In the Cohn-Kanade database it is like that. There is an annotation which tells you in which ways the person is smiling and it has been annotated by an expert to give evidence.

INTERVIEWER: And do you know more about these experts, from which profession do they come from? Are they psychologists?

I3a: Usually they are FACS experts, usually they annotate such data.

The ground truth is produced by experimental artificial data that has to be annotated by experts. People are asked to give an extreme facial expression in a laboratory. Even if this is not a "natural" expression experts annotate facial expressions that are said to be naturally and biologically caused [5]. The ground truth is co-constructed by "laypersons" and FACS experts in a laboratory. What counts as an emotion, e.g. fear, is thus a co-product of "artificial" facial expressions and expertise based on biological and natural facial expressions.

b) Computer scientist layperson approach

In the computer scientist layperson approach, the pictures used are not from a FACS database, but rather from several other picture collections:[5]

> **I4**:[6] It is a mixed collection of images from very different sources. That starts with any databases, progresses with self-photographed pictures, also pictures collected from the internet and ends with pictures that we win from the TV...
>
> **I4**:[7] . . . Our procedure of training the library is a fully layperson approach. This means that people are not trained, they are just like you and me. They are doing the annotation on the basis of their practical knowledge.

This is a layperson approach, operating with practical and tacit knowledge. The individuals annotating the data, for example stating that they recognize fear in a specific image, are computer scientists and students. They have no special training on facial expression recognition. The ground truth is based on a library of pictures from very different sources. The aim of the picture library is to have a variety of pictures from many different sources and not to have pictures that have been produced in laboratories under specific conditions and have been annotated by FACS experts. The computer scientist explains the rough estimation of facial expressions data with the absence of FACS experts in the annotation process. On the other hand real-time ability is more important than exact results and therefore the system has to work with simple features. The essentiality for the system's real-time processing is a demonstration of the "naturalness" of the system. The system needs to be as fast as humans are, exact expert knowledge could block real-time processing.

What we see in the two different approaches ('FACS expert approach' and 'computer scientist layperson approach') is that the construction of the Facial Expression Recognition 'ground truth' is a matter of selection. There is no truth from which to start with, it has to be constructed. The two different approaches to constructing a ground truth show that it is necessary to choose which knowledge is workable and in what way it should be processed and codified. The word ground truth itself refers to the longing for truth that can be found in the face and in the biological body. But who is right? Is it expert' knowledge of knowing exactly where to look in order to be able recognizing a real emotion? Or is it the practical knowledge of laypersons that is used to recognize emotions in facial expressions in everyday life? Or is it the system itself that constructs a definition of a specific emotion on basis of statistical models?

3.2. Automated multi-camera event recognition for the prevention of bank robberies

The second case study of dealing with 'Automated multi-camera event recognition for the prevention of bank robberies' could not draw on existing knowledge, but had to instead generate new empirical data

[5] The following interview passage is originally in German (see below), English translation by author.

[6] Original text in German: Das ist eine Mischsammlung aus allen möglichen Quellen, die man so auftut. Und das fängt an bei irgendwelchen Datenbanken, geht weiter bei Bildern, die wir selbst fotografiert haben, das sind Bilder, die wir aus dem Internet sammeln und endet bei Bildern, die wir aus dem Fernsehen gewinnen.

[7] Original text in German: Bei uns und bei dem Verfahren, wie wir diese Bibliothek trainiert haben ist ein völlig laienhafter Zugang, d.h. die Leute sind nicht trainiert, das sind einfach nur Menschen wie du und ich, die einfach aus ihrer Erfahrung heraus die Annotation durchführen.

through interviews, observation, and document analysis. These methods have been mentioned before (see 1.3 Methods). In the following section, the negotiation process of the 'ground truth' of suspicious behaviour of exploring bank robbers in Austrian bank branches is described.

3.2.1. The 'ground truth' of Suspicious Behaviour – A matter of context

Norris and Armstrong demonstrated how categories of suspicion are constructed by CCTV control room operators [27]. With high-level semantic VS systems this process is brought forward to programming, and therefore computer scientists seem to act in place of CCTV control room operators in constructing categories of suspicion. In this specific case it is not only the computer scientist constructing suspicion (and in the same way normality), but also security experts of a commercial bank, Police experts, as well as social scientists.

The contribution of the social scientists tended to be twofold: on the one hand the collection of data for generating a ground truth was what the computer scientists asked for, on the other hand the process of generating ground truth was critically reflected.

At the beginning of the project the bank's security experts and the Police security experts strongly put forward the assumption that it is possible to uncover potential bank robbers in their process of exploring bank branches when deciding which branch they will rob. Furthermore they suggested that potential robbers will desist from robbing one particular branch if they are addressed in person in or in front of the respective branch. In their view automatic detection of such critical behaviour could assist in averting anticipated criminal action. Security experts have been interviewed to make use of their "gut feeling" when designing an applicable system that automatically detects suspicious behaviour.

The outcome was a predefinition of a set of suspicious behaviour terms in collaboration with all project partners. This set includes lingering time in the bank foyer without using a machine (e.g. the ATM), or interacting with a member of staff over an extended period of time, or also staying at a machine for an unusually long period of time. However, it was not possible to gather accurate knowledge about the actual behaviour of bank robbers exploring objects. Thus the determined criteria remained questionable regarding their relevance for implementing an effective and applicable system. Selections were made, on the one hand because attendance time and interaction/interactivity had been mentioned as potential indications for suspicious behaviour, on the other hand because it was technologically feasible.

In a further step the behaviour of people staying in a bank branch was analysed to learn about "usual" or "normal" behaviour of bank customers. The average time people stayed in the bank foyer (where machines like ATM, bank statement printer and bank counters can be found) is 03min 08s (median 01min 58s). 38 out of 236 people (16%) stayed longer than 5 minutes and 10 out of 236 (4%) stayed longer than 10 minutes. The outlier percentage concerning attendance time of bank clients is rather high, so a simple detection of time is not practical and there is no evidence indicating longer attendance time to be unusual or even suspicious behaviour. In fact, all longer attendance time periods can be explained and appear as usual behaviour; many bank clients had to wait in the line in front of the ATM machine or bank counter, others had problems with a machine, requiring the assistance of a bank officer and some just took their time filling in a transfer form.

Facing people staying at the bank foyer without using a machine or without interacting with a member of staff over an extended period of time could be another interesting point to examine 17 out of 236 (7%) people observed did not exhibit any usual activity; almost 50% of these accompanied somebody performing usual activity. The other half was inside the bank foyer for a very short period of time. Most of these clients just glimpsed inside, observed the line and left.

One main conclusion of the observation is that usual behaviour in bank foyers is very diverse, although the specific context of a bank determines the expected human behaviour to a considerable extent. There is a great range of different behaviour patterns, making the detection of unusual or suspicious behaviour difficult. One way could be to detect those with specific deviation from the average attendance time, but this is questionable, because there is no evidence that those differing from the average attendance time are suspicious. Additionally, there is information that many bank robbers exploring a bank branch behave like ordinary customers or are in fact bank customers. Then, in the case of detecting those with specific deviation from the average, the usual would become the normative. This may have serious consequences for the watched. The pressure to adapt may increase for those entering a bank foyer and if they do not behave flawlessly they might provoke adverse consequences. Those simply diverging from the norm would attract attention instead of real bank robbers. One must consider that the detection of suspicious behaviour of bank robbers is like finding a needle in a haystack. In Vienna's 512 bank branches there are estimates of 70 million people entering and leaving a bank branch in the course of one year. By contrast there were 63 bank robberies in Vienna in 2008.

4. Concluding discussion

The main concern of this article was to disclose what kinds of reductions of complexities of vision and perception are made when developing high-level semantic Video Surveillance technology. This was obtained by exploring this kind of technology 'in the making' in two case studies to be able to understand how it is embedded in social practices and how it impacts our conception of how we see and perceive.

The first case study 'Facial Expression Recognition' showed that there are different ways to construct and reach the 'ground truth' of specific facial expressions. Thus the 'ground truth' of facial expressions is a matter of selection and it is not clear what the "right" approach is. Moreover, both approaches have different ambitions: the first one pursues precision, the second one promptness and real-time ability.

The second case study 'Automated multi-camera event recognition for the prevention of bank robberies' expressly underlined the importance of context information. It was not possible to define clear categories that represent "suspicious" or "normal" behaviour of exploring bank robbers. This may stem from missing knowledge about the exploring behaviour of bank robbers and the fact that "normal" behaviour of bank clients is too diverse. Though it is technically feasible to automatically measure the attendance time of bank clients this information does not provide substantial information about "suspicious" or "normal" behaviour, because a deviation from the average attendance time can bear different meanings.

4.1. Reduction of complexity

Both high-level semantic Video Surveillance cases presented in this paper contain different steps of reductions of complexity:

1. On the level of conception it is the fragmentation of the body and missing integration of contextual information. For example, in the case of Facial Expression Recognition only the face is incorporated. All other body parts and the situation in which the Facial Expression Recognition occurs are neglected.
2. On the level of engineering reduction has especially to do with the displacement from one frame of reference to the next. This means that in comparison to face-to-face interaction the frame of reference moves away from the complex situation in reality, across the way of the digital picture to representation in numbers. For example, in the second case study the event recognition of "suspicious" behaviour was attempted by measuring attendance time of persons, as this task was technically feasible.

3. The third level of reduction is the immediate processing of data. During this process the observed data has to be filtered and smoothened to obtain unique findings, which in turn means specific thresholds have to be predefined by the computer scientists, for example when precisely a facial expression represents the emotion anger.

4.2. Context is crucial

In summary it can be stated that in both cases the technical and rule-governed component of vision and cognition dominates and therefore has largely ignored the social and interpretative component. The whole complexity of human vision and cognition is simulated in its structure and framework. Widely denied is the involvement of complex information about the contextual face-to-face situation. This is of importance, because information as well as human action and activity are not self-explanatory, but rather are negotiated out of social context [33]. This also concerns face-to-face control, which is negotiated, not absolute. Furthermore it is "based on a complex moral assessment of character which assesses demeanor, identity, appearance and behavior through the lens of context-specific relevancies." [26, p. 276].

Technology that only focuses on visually observable objects and body movements hides the processes of negotiating the meaning of these visually observable objects and body movements in face-to-face interactions. Human vision is not the sum of isolated observed components. Instead we can start from the premise that vision is subject to change, culturally and historically [15]. Charles Horton Cooley, a precursor of *symbolic interactionism*, distinguished between spatial and social knowledge. The former, based on sense perceptions, gives rise to exact or quantifiable natural science. The latter only emerges in the negotiation and communication of other people's way of thinking [9]. Current high-level semantic VS systems do not integrate this kind of social knowledge, which is of outstanding importance to understand human behaviour or facial expressions in a specific situation. This is why the metaphor of the thinking eye for the connection of a video camera to computer hard- and software is only half the story. These systems are not comparable with human vision abilities and it can be still stated that the film computer HAL 9000 displays "image understanding capabilities vastly beyond today's computer systems." [29]

Under these circumstances of systems reducing complexity significantly and focusing on quantifiable elements we can reflect on what an integration of such systems into contemporary societies means and what kinds of new orderings will take place once these systems are integrated into social life. For example, automatically measuring attendance time or facial expressions can provide an indication for "suspicious" behaviour or for the emotional constitution of a person. The crucial point is that it is not the synonym. "Suspicious" behaviour or the emotional constitution of a person are very complex entities constituted not merely of attendance time or facial expressions, but rather of many different elements. We have to make that difference clear. Furthermore, attendance time or facial expressions are ambiguous; they are context and situation dependent.

4.3. Socio-technical implementation

For all practical purposes this asks for careful distribution of power and agency of high-level VS systems to machines and human users in consideration of managing possible risks and adverse effects. If such a system is proportional it should be implemented in a concrete context of application. Workplace studies have particularly demonstrated the importance of how technologies are being applied in situated actions, and how they can fail if they do not meet users' needs [16]. A consequence is that a high-level semantic VS system must support the complex sociality of the specific work setting, in which it is going to be implemented in [10]. It has to be purpose-built and one-off [11]. This also means reflecting about implications if employed in another place or at another point in time. Both sophistication and generalisability of high-level semantic VS have to be challenged.

4.4. Transparency and reflexivity

A high degree of transparency and full disclosure of reduction processes is a good basis for reflection. We have to make clear that current high-level semantic VS systems cannot act autonomously, but must be integrated into social settings with professional staff who understand how the algorithms applied work. The more they know about the used 'ground truth', tolerances, thresholds, and the reduction of complexity, the better they can handle this technology and minimize possible risks such as false positive findings.

Against this background high-level semantic VS systems can be regarded as 'upskilling' rather than 'deskilling': it can be assumed that along with the implementation of such systems, operators have to be trained in order to manage and work with these systems. A reduction in operators is unlikely, because human analytical skill is still required and inevitable.

Transparency is not only a matter of importance for operating staff, it is also a matter for the people being observed; at the very least they need to be made aware that their behaviour is being observed and analyzed by computer algorithms. It should be asked if and to what extent this has to be explicitly stated on the site of operation and in what way this information is established by law (e.g. data protection law). In addition it should also be discussed, if and how people have to opt in and consent to be observed and analyzed. This seems to be especially difficult in indefinite public places with a high volume of people and movements.

This paper's 'technology in the making' approach showed that transparency is the starting point for the reflection on high-level semantic VS, but it has to be noted that transparency and disclosing how it works is not enough. Computer scientists and engineers should be made aware of their work being part of social practices and shaping society and social life in a significant way. Of course, not all implications of high-level semantic VS can be predicted or affected by computer scientists, but their work is inseparable intertwined with the "ways in which we choose to live in" [14, p. 2].

4.5. Policy related issues

In conclusion two policy-related issues arising out of the 'technology in the making' approach of the paper are resumed.

Institutions (e.g. public services, private companies) planning to invest in or implement high-level semantic video surveillance are advised to consider the 'upskilling' nature of these. With these technologies we are confronted with the need for new forms of qualifications. Thus, they do not necessarily replace human labour, but have to be understood as co-produced with the 'upskilling' of human labour.

In this respect the integration of the social sciences and humanities in the technological development is vitally important. This is due to the complexity of the social setting of potential sites of operation. This complexity needs to be considered all along the process of technological development.

This insight invites us to balance the relationship between technological and social developments and thus, to really perform sociotechnology. For public research funding this implies to promote problem-centred instead of technology-centred research projects. Technology-centred projects make use of resources for the sake of developing one specific technology that is promoted as ready-made solution for a pre-defined problem from first to last. In contrast problem-centred projects would involve an open interdisciplinary engagement already at the problem definition level, which will potentially lead to different comprehensive sociotechnological solutions.

Acknowledgements

I would like to thank Ulrike Felt, Stephan Gruber and Daniel Messner as well as the two anonymous referees who have all helped with developing this paper further. I also benefited from presenting parts of the paper at various conferences, workshops and summer schools. Above all I am indebted to Sheila Jasanoff and Stefan Vogtenhuber. I am also grateful to my informants from the field of computer science and would like to thank Fiona Zisch for very helpful comments on the manuscript.

References

[1] G.C. Bowker and S. Leigh Star, Sorting things out. Classification and its consequences, Cambridge, London, The MIT Press, 2000.

[2] W. Brednow, Von Lavater zu Darwin, Berlin, Akademie Verlag, 1969.

[3] L. Daston and P. Gallison, Objectivity, Zone Books, 2007.

[4] P. Ekman, Gesichtsausdruck und Gefühl. 20 Jahre Forshcung von Paul Ekman, Paderborn, Junfermann, 1988.

[5] P. Ekman, Emotions Revealed. Recognizing Faces and Feelings to Improve Communication and Emotional Life, New York, Owl, 2007.

[6] P. Ekman, W.V. Friesen, Wallace and J.C. Hager, Joseph C., The Facial Action Coding System. Second Edition, London, Weidenfeld and Nicolson (world), 2002.

[7] K. Gates, The Tampa "Smart CCTV" Experiment, *Culture Unbound* **2** (2010), 67–89.

[8] S. Graham and D.Wood, Digitizing Surveillance: Categorization, Space, Inequality, *Critical Social Policy* **23** (2003), 227–248.

[9] H.J. Helle, Verstehende Soziologie und Theorie der Symbolischen Interaktion, Stuttgart, Teubner, 1977.

[10] J. Hughes, J. O'Brien, T. Rodden and M. Rouncefield, Ethnography, communication and support for design. in: *Workplace Studies. Recovering Work Practice and Informing System Design*, P. Luff, J. Hindmarsh and C. Heath, eds, Cambridge University Press, 2000, pp. 187–215.

[11] L.D. Introna and H. Nissenbaum, Facial Recognition Technology. A Survey of Policy and Implementation Issues, Report of the Center for Catastrophe Preparedness and Response, NYU, 2009.

[12] L.D. Introna and D. Wood, Picturing Algorithmic Surveillance: The Politics of Facial Recognition Systems, *Surveillance and Society* **2**(2/3) (2004), 177–198.

[13] S. Ioannou, G. Caridakis, K. Karpouzis and S. Kollias, Robust Feature Detection for Facial Expression Recognition, in: *EURASIP Journal on Image and Video Processing*, Article ID **29081** (2007), 22.

[14] S. Jasanoff, Sheila, States of knowledge. The co-production of science and social order, London and New York, Routledge, 2004.

[15] D. Kammerer, Bilder der Überwachung, Frankfurt am Main, Suhrkamp, 2008.

[16] H. Knoblauch and C. Heath, Technologie, Interaktion und Organisation: Die Workplace Studies, *Swiss Journal of Sociology* **25**(2) (1999), 163–181.

[17] H. Knoblauch, B. Schnettler, J. Raab and H.-G. Soeffner, Video Analysis: Methodology and Methods. Qualitative Audiovisual Data Analysis in Sociology, Frankfurt, Berlin, Bern, Bruxelles, N.Y., Oxford, Wien, Peter Lang, 2006.

[18] K.D. Knorr-Cetina, The Manufacture of Knowledge. An Essay on the Constructivist and Contextual Nature of Science, Oxford, N.Y., Toronto, Sydney, Paris, Frankfurt, Pergamon, 1981.

[19] B. Latour, Science in Action. how to follow scientists and engineers through society, Harvard University Press, 1987.

[20] B. Latour, Technology is society made durable, in: *A Sociology of Monsters: Essays on Power, Technology and Domination*, J. Law, ed., London, Routledge, 1991, pp. 103–131.

[21] B. Latour, Reassembling the social: an introduction to Actor-network theory, New York, Oxford, University Press, 2005.

[22] B. Latour and S. Woolgar, Laboratory Life: the Social Construction of Scientific Facts, Sage, Los Angeles, 1979.

[23] L. Lessig, Code – and Other Laws of Cyberspace. New York, Basic Books, 1999.

[24] D. Lyon, Surveillance Studies: An Overview, Cambridge and Malden, Polity Press, 2007.

[25] C. Norris, From Personal to Digital: CCTV, the Panopticon and the Technological Mediation of Suspicion and Social Control, in: *Surveillance as Social Sorting*, D. Lyon, ed., London, Routledge, 2002, pp. 249–281.

[26] C. Norris, From personal to digital CCTV, the panopticon, and the technological mediation of suspicion and social control, in: *Surveillance as Social Sorting. Privacy, Risk and Digital Discrimination*, D. Lyon, ed., London and New York, Routledge, 2003.

[27] C. Norris and G. Armstrong, The Maximum Surveillance Society: The Rise Of CCTV, Oxford, 1999.

[28] J. Person, Der pathographische Blick: Physiognomik, Atavismustheorien und Kulturkritik 1870-1930, Königshausen and Neumann, 2005.

[29] R.W. Picard, Building HAL: Computers that sense, recognize, and respond to human emotion, in: *Human Vision and Electronic Imaging VI*, part of IS&T/SPIE9s Photonics West, Society of Photo-Optical Instrumentation Engineers, ed., 2001.

[30] M. Polanyi, The tacit dimension, Garden City, N.Y., Doubleday, 1966.

[31] W. Rammert, Gestörter Blickwechsel durch Videoüberwachung? Ambivalenzen und Asymmetrien soziotechnischer Beobachtungsordnungen, *TU Berlin Technology Studies Working Papers*,TUTS-WP-9-2002.

[32] W. Rammert, Technik – Handeln – Wissen. Zu einer pragmatistischen Technik- und Sozialtheorie, Wiesbaden, VS Verlag, 2007.

[33] R. Richter, Verstehende Soziologie, Wien, Facultas, 2002.

[34] C. Schmölders, Das Vorurteil im Leibe: eine Einführung in die Physiognomik, Berlin, Akademie Verlag, 1997.

[35] I. Schulz-Schaeffer, Sozialtheorie der Technik, Frankfurt/Main, Campus, 2000.

[36] L. Suchman, Feminist STS and the Sciences of the Artifcial. in: *The Handbook of Science and Technology Studies*, E. Hackett, O. Amsterdamska, M. Lynch and J. Wajcman, eds, Third Edition, Cambridge and London, MIT, 2008, pp. 139–164.

[37] R. Surette, The Thinking Eye. Pros and cons of second generation CCTV surveillance systems, *Policing: An International Journal of Police Strategies and Management* **28/1** (2005), 152–173.

[38] G. Swoboda, Lavaters Linienspiele. Techniken der Illustration und Verfahren graphischer Bildbearbeitung in einer physiognomischen Studiensammlung des 18. Jahrhunderts, Dissertation, Universität Wien, 2002.

[39] P. Turaga, R. Chellappa, V.S. Subrahmanian and O. Udrea, Machine Recognition of Human Activities: A Survey, *CSVT* **18**(11) (2008), 1473–1488.

[40] M. Wimmer, C. Mayer, S. Pietzsch and B. Radig, Tailoring Model-based Techniques to Facial Expression Interpretation, in: *achi, First International Conference on Advances in Computer-Human Interaction*, 2008, pp. 303–308.