DISSERTATIONES
SOCIOLOGICAE
UNIVERSITATIS
TARTUENSIS
27

HANS HÕRAK

Development of a computer vision-based privacy-preserving automatic observation method for measuring physical activity in school

DISSERTATIONES SOCIOLOGICAE UNIVERSITATIS TARTUENSIS
27

DISSERTATIONES SOCIOLOGICAE UNIVERSITATIS TARTUENSIS
27

HANS HÕRAK

Development of a computer vision-based privacy-preserving automatic observation method for measuring physical activity in school

UNIVERSITY OF TARTU
Press
1632

Institute of Social Studies, University of Tartu

Dissertation accepted in fulfilment of the requirements for the degree of Doctor of Philosophy (in Sociology) on May 10, 2023, by the Council of the Institute of Social Studies, University of Tartu.

Supervisors: Professor Triin Vihalemm
Institute of Social Studies
University of Tartu

Professor Gholamreza Anbarjafari
Institute of Technology, iCV lab
University of Tartu

Opponent: Professor Arūnas Emeljanovas
Lithuanian Sports University

Commencement: June 19, 2023, at the White Hall of the University of Tartu Museum

The publication of this dissertation is granted by the Institute of Social Studies, University of Tartu, the Doctoral School of Behavioural, Social and Health Sciences Created Under Auspices of the European Social Fund, by the Centre of Excellence in Estonian Studies (European Union, European Regional Development Fund). This research was also supported by a grant from the development fund of University of Tartu.

ISSN 1736-0307 (print)
ISBN 978-9916-27-229-9 (print)
ISSN 2806-2590 (pdf)
ISBN 978-9916-27-230-5 (pdf)

University of Tartu Press
www.tyk.ee

CONTENTS

LIST OF ORIGINAL PUBLICATIONS AND REPORTS

Study I. Hõrak, H. (2019). Computer vision-based unobtrusive physical activity monitoring in school by room-level physical activity estimation: A method proposition. *Information, 10*(9), 269.

Study II. Hõrak, H., Jermakovs, K., & Haamer, R. E. (2022). Modeling Physical Activity in Children by Combining Raw Hip-Worn Accelerometry, 2D Pose Estimation, and Direct Observation. *IEEE Access, 10*, 39986–40000.

Study III. Hõrak, H., Reis, M. S., Jermakovs, K., & Haamer, R.E. (2022). Physical activity observation sensor prototype: project report. *Zenodo*. https://doi.org/10.5281/zenodo.7725737

Demo video:
https://www.youtube.com/watch?v=RQHw2Z22pWc&t=1s&ab_channel=KEKA

AUTHOR'S CONTRIBUTION

All three studies were conceived and designed by the author. Data was collected by the author and in part annotated by the author. In **Studies II** and **III** the author conducted majority of the research, analyses, and writing while leading a team of engineers who had the major role in software implementation of deep learning and real-time optimization on the prototype.

ACKNOWLEDGMENTS

I would like to thank supervisor prof. Triin Vihalemm and the institute of social studies for taking the risk and allowing me into the sociology PhD programme with this kind of thesis topic. Thanks goes to cosupervisor prof. Gholamreza Anbarjafari (Shahab) for support and allowing access to the resources and expert knowledge of the iCV laboratory. I would like to thank the whole engineering team, especially Mateus Reis whose outstanding skills allowed the successful implementation of the sensor prototype. Big thanks to Helis Ojala for lots of tedious annotation and help with data collection. Special thanks to Sirje Ange from Põltsamaa Ühisgümnaasium for help with subject recruitment and data collection. Finally, a big thankyou to all participating children and their parents.

1 INTRODUCTION

In many contemporary societies, people find themselves in a prosperous environment with lots of sedentary jobs and leisure activities, and an online environment where different firms compete for the attention of people sitting on chairs and looking at screens (Bonnet & Cheval, 2022). On the caloric intake side, free market capitalism provides favorable circumstances for developing a food and beverage industry which maximize their profits by taking advantage of our preference to consume fast carbohydrates (Drewnowski et al., 2012; Lustig et al., 2012). This environment has appeared very quickly while our bodies are still largely adapted to constrained diets entailing much less sugar, and to a lifestyle of hunting, gathering, and perhaps some agricultural work (Lieberman, 2015; Lee et al., 2016; Li et al., 2018). These adaptations lead to energy frugality in a setting where energy is abundant – avoiding movement more than necessary, eating more, and storing more energy in body fat than we need. The mismatch of our adaptations to contemporary society reflects well in the trend of last-mile food delivery: the customer can obtain very tasty calorie-dense food by only standing up from their chair to pick up the food at the front door. In this case at least the delivery people using non-electric bicycles utilize their muscle tissue vigorously, but even these workers may eventually be replaced by drone operators sitting on chairs. Technological and social innovations of the past two centuries lead to unprecedented increases in wellbeing and comfort in our daily lives. Instead of foraging, chasing prey, working long hours in the field or factory, increasingly more people can earn a living while working shorter hours often completing their work tasks in the comfort of a chair. Indeed, owning a comfortable chair and being employed at a well-paying desk job can be seen as symptoms of human development, some of the distinguished accomplishments of our civilization. Significant parts of economic growth can likely be attributed to the will to do work easier (e.g., automating manual labor) and to live a comfortable life: the desire[1] to earn a high enough income that would allow to purchase products that increase comfort and to hire other people to do the uncomfortable tasks. But when such human development leads to epidemiological findings such as physical inactivity causing 7.2% of all-cause mortality globally (4.4% in low-income, 6.8% in middle-income, and 9.3% in high-income countries) or 3.6 million deaths in 2016 (Katzmarzyk et al., 2022), the value of all this comfort appears questionable. Modern prosperity and socio-technical environment (e.g., affordable computers and smartphones) combined with an energy conservation tendency (Booth et al., 2017) and a sweet tooth have led to a global public health crisis (French et al., 2001). Overweight and obesity are becoming increasingly more prevalent (Afshin et al., 2017) including in children (Sahoo et al., 2015). Physical inactivity, affecting cardio-vascular health directly and through contributing to overweight,

[1] This is not to downplay the goal of social status attainment in the desire to earn a high income, but comfort is relevant as well.

itself is seen as a global pandemic (Kohl et al., 2012; Guthold et al., 2018) and a major contributor to the burden of non-communicable disease (Ding et al., 2016; Katzmarzyk et al., 2022; World Health Organization, 2022). Increasing physical activity (PA) levels could reduce the risk of many types of cancers (Moore et al., 2016) and provide various mental health benefits (Biddle et al., 2019).

To mitigate the negative health effects of such an environment, societies would be wise to prepare children accordingly. Many years of healthy life could be added to the population if the education- and social protection systems were able to effectively cultivate healthy habits for coping with the energy conservation pressures in this energy-abundant environment.

There is some evidence that school-based PA interventions can be sustainable (Lai et al., 2014) and cost-effective (Abu-Omar et al., 2017) pathways for increasing PA among youth, but the evidence for long-term cost-effectiveness is limited (Batorova & Sørensen, 2019). Recent meta-analyses imply the effects of school-based intervention, even if relatively cost-effective compared to some other intervention strategies, have so far been minuscule to small (Love et al., 2019; Jones et al., 2020; Neil-Sztramko et al., 2021; van Sluijs et al., 2021). Despite limited success of school-based PA interventions, childhood and youth are considered high-priority intervention targets for life-course health behavior (GAPA, 2012; Sawyer et al., 2012), and schools as already existing standardized educational infrastructure should be an obvious intervention setting (Pate et al., 2006). Just as schools should teach reading skills and habits of critical thinking, they should as well teach physical literacy[2] (Whitehead et al., 2018) and promote healthy habits. Public health authorities are stressing the importance of developing a theoretically well-grounded (Gourlan et al., 2016) evidence base (Lewis et al., 2017) for PA interventions so that the best practices could be scaled up for maximum public health impact (Reis et al., 2016; Ding et al., 2020).

While it is important to measure theoretically grounded mediating variables (e.g., enjoyment/discomfort, attitudes, and motivation), the efficacy of PA intervention is ultimately judged by measuring the (sustained) change of actual movement behaviors. To generate the evidence base for school-based PA intervention, accelerometers, pedometers, and to a lesser degree, questionnaires can be used to estimate average PA intensity of students during the school day. However, such individual measures are intrusive and do not provide any information on where in this specific environment are children active or sedentary (**Study I**). Additionally, the datafication of society brings increased sensitivity to privacy issues and new strict data protection regulations (e.g., Regulation 2016/679/EC) which complicate individual-level privacy-intrusive research with human subjects. Intrusiveness of using wearable sensors burdens the subjects and requires researchers

[2] Physical literacy can be defined as the interactive and simultaneous consideration of competence in physical skills, confidence, motivation towards physical pursuits, and the valuing of physical movement and/or interacting with the physical world (Edwards et al., 2017). It is expected that high physical literacy means a person can maintain a state of mind and body which leads to being physically active to a healthy degree throughout their life course.

to inform and obtain consent from the parents of all the subjects. Then the researchers must rely on the children adhering to accelerometer wear protocols. The intrusiveness of wearable sensors also limits the possible duration of continuous measurement. This leads to research designs where long-term sustainment is assessed with a follow-up study (possibly requiring a second round of obtaining informed consent) instead of continuously monitoring PA to detect sustainment or relapse. School is a very specific environment (schoolhouse, sports infrastructure, and playground) with specific behavioral patterns (mandatory education often with standard curriculum delivered in a similar manner). When administering and researching PA intervention in schools, this specificity of the setting can be exploited in the design of interventions and intervention studies. Continuous long-term ambient measurement of PA in the school building could provide a new perspective on the implementation, adoption, and sustainment of interventions and their effects. Knowledge on the spatio-temporal distribution of PA in the school environment could carry important information: finding areas that facilitate or hinder PA and detecting changes in PA patterns associated with intervention efforts. Ambient measurement with wall-mountable sensors can circumvent the problem of accelerometer wear protocol adherence. If such ambient measurement was conducted in a data-secure and privacy-preserving manner, then this method could be more ethical, possibly removing the necessity for obtaining informed consent. Privacy-intrusive methods may require more time and effort to navigate research ethics and data protection requirements, but privacy-preserving methods could alleviate this bottleneck. Increasing methodical diversity in PA intervention research should generally support developing the evidence base necessary for improving public health. School-level ambient PA measurement could be especially useful for discovering simple and efficient interventions that affect the largest share of students throughout the school day – the critical interventions that would be scalable even in low-income regions. Acknowledging the potential utility of such long-term spatial/ambient measurement of PA and observing concurrent explosive development of computer vision and deep learning technologies (the "artificial intelligence/AI revolution"), **this thesis set out to develop a methodology and a real-time video analysis sensor for measuring PA of children in the field of view of a camera without violating their privacy (blind observation[3])**.

To achieve this goal, the following research questions were posed:

RQ1: How to model physical activity intensity of children in video data?

RQ2: Which video analysis approach can provide a stable physical activity signal?

RQ3: How does automatic blind observation compare to human observation methods on a fundamental level?

[3] The concept, introduced in this work, of collecting observational data through the visual modality without anyone having access to the visual information.

RQ4: How viable is the proposed method for research practice?

RQ5: How to use such sensors in physical activity intervention research?

As blind observation is a novel approach to collecting observational data on humans, another goal of this thesis is to reflect on the various implications of such methods (privacy, ethics, and trust).

The substance of this thesis is generally methodological entailing the use of several technologies and methods to model and measure movement behaviors. The goal is to leverage recent technological advances to develop a method with desirable novel properties: unintrusive, privacy-preserving, location-specific PA estimation with a theoretically unlimited measurement duration. Contributing novelty to its arsenal of methods could move forward the research field currently struggling to promote PA.

While this work develops a method for use in research tackling a specific public health problem[4], some properties of the method could have implications for the sociology of scientific knowledge as well. For knowledge produced by traditional observational research, scientists and publics have had to rely on trusting the senses and reasoning of human observers. Automation of observational methods by computer vision and machine learning (ML) entails different trust relations[5] and practices of scrutiny[6] in the production of scientific knowledge. Automatic observation, when conducted in a privacy-preserving manner (no video frames are exposed to human eyes), also has positive implications for research ethics. Privacy preservation is especially useful when studying children who may not be capable of truly informed consent to privacy-intrusive data collection methods. Instead of obtaining written informed consent from the parent and child, the proposed method should allow to circumvent consent altogether and make it easier to obtain human research ethics approval.

The research conducted for this thesis is inherently multidisciplinary. The nature of the problem regards health sociology, particularly the change of sociotechnical systems in modernization which lead to the proliferation of physically inactive lifestyles. The rationale to develop the method stems from epidemiology (the extent of inactivity and its health outcomes) and behavioral science (intervention into health behavior). At the core lies kinesiology – quantifying the kinetic energy produced in human muscle tissue. Computer science sub fields of

---

[4] While the proposed method is designed for school-based PA intervention research, it could also be used in health sociology: comparing long-term PA patterns of schools in rich and poor regions/neighbourhoods or comparing urban and rural schools or schools of different societies.

[5] As social creatures, scientists may be inclined to trust a poorly performing but well-mannered smiling human observer more than a better performing machine which does not express social cues.

[6] The skill of human observers can be evaluated against test data or a known skilled observer, but this does not guarantee they will perform as well in the field throughout the whole observation period. In case of a ML model, the training data can be thoroughly analysed, and the model can be tested under controlled conditions by independent research teams. Additionally, researchers can be certain that the skill of the system is consistent over time.

machine learning, pattern recognition, and computer vision provide the knowledge base and technologies to realize the proposed method in form of a privacy-preserving video analysis sensor prototype.

Research started out with a multidisciplinary scoping review article with an empirical section exploring correlations of hip-worn accelerometer signals with motion information in video (**Study I**). After the first batch of multimodal data had been collected and partially annotated, **Study II** was conducted to calibrate the measurement construct in the computer vision data set being developed. This was a knowledge transfer exercise where PA researchers' domain knowledge was used to define the moderate to vigorous PA (MVPA)[7] threshold in the eventual deep learning data set by using an online survey of video classification. **Study II** also explored hip and shoulder joint angle changes computed from 2D pose-estimated kinematic skeletons as a novel PA intensity indicator. Expert knowledge was extracted by asking researchers to classify PA intensity in 24 short videos synchronized with hip-worn accelerometers. Then the expert group's understanding of the MVPA threshold was extrapolated to the rest of the data set using the same accelerometers and pose-estimated hip angle features. **Study III** entails realization and demonstration of the method proposed in **Study I** in form of a real-time video analysis sensor prototype deploying a deep neural network trained on the collected data set. After demonstrating the possibility of privacy-preserving direct observation of PA, the discussion chapter of the cover article tackles privacy and trust issues concerning the development and deployment of such sensors in schools.

The cover article is structured as follows. The theory section first covers the main theoretical paradigms underlying PA intervention research and provides speculations on the potential uses of video analysis under these paradigms. The theory section continues with an overview of PA intensity as the measurement construct for the developed method. After this, classical observation methods are discussed and compared to a hypothetical privacy-preserving video analysis approach as an automatic equivalent to obtaining observational data. The theory section concludes with a brief overview of supervised machine learning and artificial neural networks as the basis for the video analysis approach. The methodology section covers majority of the research and development work conducted for this thesis: from developing the training data collection method up to the deployment of the trained model on the hardware of the prototype. The methodology section concludes with a reflection on the mistakes in the development process and offers considerations for future development. The results section provides assessment of the automatic observation processing pipeline and introduces indicators that can be used in eventual application of the method. The discussion section first comments on the viability of the developed method and proceeds to discuss issues of privacy, trust, and ethics related to development and potential application of blind observation methods. The cover article concludes with answers to the research questions.

---

[7] A widely adopted PA intensity construct discussed in Chapters 2.2 and 3.3

2 THEORETICAL FRAMEWORK

The following section covers some major theoretical paradigms in PA intervention and how an automatic video analysis approach could be used in research under these paradigms. This is followed by sections on the measurement construct (PA intensity), a theoretical comparison of human and automatic observation, and a brief overview of supervised machine learning and artificial neural networks as the basis of the method.

2.1 Theoretical bases for physical activity intervention

In the mid-70s, the discourse in public health started shifting from curing to preventing illness (Parish, 1995). The concept of health promotion stated to emerge from policy documents, but at first, the focus was on individual's responsibility to adopt healthy behaviors with the public health authorities assuming an educational role – informing the populace on positive and negative health effects of various behaviors. As a turning point, Parish (1995) notes the World Health Organization's (1985) *Health For All* programme which acknowledged health inequalities and introduced proactive concepts such as improving access to health, developing an environment conducive to health, and promoting positive health behavior and appropriate coping mechanisms. Baum and Fisher (2014) note that throughout the history of public health policy, there have been two conflicting views on health promotion: whether efforts should be focused on modifying unhealthy behaviors or whether state intervention should focus on the underlying social and economic factors as the primary determinants of health behavior. This chapter is divided by the same logic applied to the theoretical bases of PA intervention. First, I present theories that explain movement behaviors and sedentariness at the level of the individual. Then I cover theories where the individual and their movement behaviors are seen as situated in social and ecological context. To give an overview and comparison of the theories and their application in PA intervention, I have compiled Table 1.

2.1.1 Individual-level theories

Majority of theory-based PA interventions are individual-level approaches (the four first entries in Table 1) (Rhodes et al., 2019) focusing on how and why individuals move or avoid movement; how can they be influenced to move or to avoid longer periods of sedentariness? Rhodes and colleagues (2019) give an overview of the historical development of theoretical bases for PA interventions. They describe how the social cognitive framework developed in the mid-20th century and has since replaced **behaviorism** and become the dominant paradigm for explaining PA and informing interventions. While automatic responses to stimuli can explain some PA behaviors (e.g., rushing across the street when observing an approaching vehicle), physical labor and exercise require a more

cognitive explanation. When psychologists realized the inability of behaviorism to explain many complex mental states and behaviors, **cognitive theories** started to emerge. The theory of reasoned action (Fishbein & Ajzen, 1975) suggested that behavior is a consequence of behavioral intention which itself is determined by a combination of subjective norms and attitudes towards the behavior and/or its expected outcomes. Out of this grew the theory of planned behavior (Ajzen, 1991) which added the concept of perceived behavioral control as a third antecedent of behavioral intention – if one has a positive attitude towards jogging and deems it socially desirable, but believes that they cannot find the time to jog, they may not go jogging. Somewhat like the construct of perceived behavioral control, Bandura's (2004) social cognitive theory proposes the concept of self-efficacy as a major factor influencing health habits directly and through its impact on goals, outcome expectations and perception of sociostructural factors. Self-efficacy is an individual's belief in their capacity to perform a particular task (Bandura, 1997) and has been widely studied in context of PA promotion (Williams & French, 2011; Ramirez et al., 2012; Tang et al., 2019). Constructs of social cognitive theories have been estimated to explain one-third of PA in adolescents (Plotnikoff et al., 2013). Rhodes and colleagues (2019) suggest the lasting popularity of social cognitive theories in PA intervention is partly due to the methodological efficiency they allow: attitudes, subjective norms, perceived control, and intentions can be studied with questionnaires.

More recent theories try to consolidate behaviorism and cognitivism by acknowledging that some bodily movements (or their lacking) can result mainly from cognitive processing while others may have more non-conscious reasons (Rebar et al., 2016; Ekkekakis, 2017; Strobach et al., 2020). A survey respondent claiming to have motivation and intention to go to the gym does not guarantee that they will as demonstrated by the intention-behavior gap revealed in a meta-analysis by Rhodes and Dickau (2012). **Dual-process theories** could help to explain habitual PA (e.g., developing a habit for active transport when motorized options are available or vice versa) and sedentariness [e.g., neurochemical processes associated with various screen-viewing behaviors (Burhan & Moradzadeh, 2020; Lindström et al., 2021; Westbrook et al., 2021; Aru & Rozgonjuk, 2022) and the energy conservation tendency (Lee et al., 2016; Cheval et al., 2018)]. Cognitivist theories are more relevant to exercise behaviors which constitute just one very specific part of the overall PA dose. Re-introducing stimulus-reward learning allows to consider multi-target interventions where some components try to induce cognitions conducive to life-long participation in regular exercise while other components try to induce persistent PA responses to common cues (e.g., a habit of taking the stairs or taking regular breaks from prolonged sitting). The automatic pathway of behavior regulation could be exploited for associating movement behaviors generally with positive affective experiences [e.g., inducing reward when significant movement is performed or exercise discomfort is felt (Conroy & Berry, 2017; Maltagliati et al., 2022)].

Table 1. Major theoretical bases for PA intervention.

Theoretical paradigm	PA promotion pathway	Negative features	Positive features	Intervention effect estimation
Behaviorist	Associating PA with positive reward, sitting with negative reward	Inducing reward at exactly the right time is difficult; measuring affect is hard	Potential lasting behavior change (developing PA-seeking habits)	Observe behavioral reactions related to stimulus before and after intervention
Social cognitivist	Teaching to value PA and promoting self-efficacy	Improving attitudes do not guarantee improved behavior	Relatively easy to administer and measure	Knowledge and attitudes before and after intervention; PA measures
Dual-process theories	Targeting both cognitions and affects, reflective and automatic processing associated with movement behaviors	More complex intervention and study design; measuring affect in general and at the right time is hard	Considers both behavior regulation pathways	Test both affective and cognitive variables, PA measures
Humanistic/ organismic (mainly SDT*)	Alluding to universal psychological needs when promoting PA	Associations between intervention stimuli and psychological needs are difficult to operationalize	Universality: everyone desires autonomy, competence, and relatedness	Measure motivation-related variables with a questionnaire; PA measures
Social practice theories	Influencing practices and/or their context to increase PA	Hard to operationalize concrete constructs	Sustaining intervention delivery by shifting culture /habituating practices of PA intervention	Study qualitatively, quantitatively or observe changes in practices, signs of resistance and adoption; PA measures
Socio-ecological models	Facilitating PA through as many social-environmental factors/ levels as possible	Expensive; hard to operationalize concrete constructs; hard to disentangle intervention effects of different levels and their interactions	Potentially largest and widest impact on PA	Monitor population-level PA for overall intervention effects, study efficacy of specific components separately

Compiled by author based on literature cited in this chapter.
* Self-determination theory

Recently a **dual-process theory** has been proposed specifically for movement behaviors. The theory of effort minimization in PA (TEMPA) (Cheval & Boisgontier, 2021) combines the controlled and automatic evaluations of movement-related cues with effort minimization – a neuropsychological process that optimizes for cost-efficiency of behaviors. TEMPA suggests that movement-related cues are perceived as effortful which, in combination with the organismic tendency to conserve energy, provides a cost context for the automatic and cognitive evaluation systems to predict whether the behavior is worth the effort. When noticing a juicy apple on a branch, the decision to climb the tree could depend on several considerations. It can likely be influenced by affect towards apples of that color based on prior experience, and the blood glucose level at the time of observing the apple. On the controlled evaluation side, one may consider potential negative outcomes ("Can I climb the tree safely without injuring myself?"), alternatives ("Where is the nearest food place?"), or social acceptability ("Am I allowed to get that apple? Is it appropriate for my age and social status?"). By the effort minimization theory, these evaluation processes happen in context of perceived effort – how hard/strenuous the task seems. The person could decide that getting the apple is desirable, acceptable, and worth the effort so they start climbing. The experienced effort of the behavior and associated discomfort immediately feed back to the dual processes to constantly re-evaluate the cost-effectiveness of the endeavor. Indeed, one may decide, after a few attempts, that the apple is not worth the effort. A more cognition-heavy example behavior on a longer time scale could be choosing between job offers based on the description of work tasks and later considering the experienced effort in employment continuation decisions. TEMPA could help explain how school-based interventions have not been able to significantly reduce daily sedentary time or increase PA. It has been suggested (Ridgers et al., 2014; Jones et al., 2020) that increased PA during the school day may lead to compensatory behaviors after school. From the effort minimization perspective, it makes sense that engaging in high-effort movement behaviors during the day can lead to physiological states (fatigue) which in turn leads to perceiving increased effort in movement-related cues later in the day. Intervention components could possibly be designed to target effort perception of common movement-related cues (Maltagliati et al., 2022) appearing in school. If particular cues are fixed in space (e.g., an exergame[8] interface or a set of stairs next to an elevator) or space and time (e.g., teachers enforcing active breaks during class), then visual room-level estimation of PA can reveal the PA reactions to these cues and changes in these reactions over time.

The fourth individual-level paradigm approaches PA intervention through universal **humanistic/organismic** attributes (Rhodes et al., 2019). Maslow's hierarchy of needs (Maslow, 1943) was among the first theories describing universal psychological mechanisms of behavior motivation although its successors have seen wider application in PA intervention. Self-determination theory (Deci & Ryan, 1985; Ryan & Deci, 2000) suggests that people want to feel autonomy

[8] Exergames are videogames which require significant PA e.g., Staiano et al. (2013).

(e.g., engaging in PA on one's own volition rather than as a reaction to external power), competence (e.g., higher motivation to play basketball if one feels they are good at it), and relatedness (e.g., higher motivation to play basketball with friends than with strangers). In the PA domain, self-determination theory has been applied mainly to exercise behaviors and with consistent success (Teixeira et al., 2012). From the perspective of state intervention into public health, such universalist approaches appear desirable as they imply a "one size fits all" strategy which should be theoretically equitable and relatively easy to implement top-down.

2.1.2 Social- and environmental theories

The dominating individualist approaches to PA intervention have been criticized for being theoretically blind to the social context and -nature of sedentary and active behaviors (Spotswood, Wiltshire, et al., 2021). When focusing on the individual and their internal processes - trying to influence a child to move more and sit less – one may neglect important social-environmental determinants of PA. **Social practice theories** focus on organized/coordinated and routinized forms of activity (practices) where individual behavior is seen an expression of taking part in social practices, performing a (role in a) practice or carrying a practice (Welch, 2017). Welch (2017) argues that practice theories promise a reframing or even a resolution to the attitude-behavior gap which he sees as a clear sign of failure of the cognitivist paradigm to explain human activity – these theories ignore the possibility of various barriers unrelated to norms, attitudes, and values. From a practice theory viewpoint, the physical inactivity pandemic could be explained by global changes in existing practices and/or the appearance and proliferation of new practices with a smaller PA content. Applying such thinking to school-based PA intervention can shift the target from the behavior of individual students to the practices or the social environment and -structures that enable or encourage practices of various PA content. Instead of asking "Why is this child sedentary? or "Why is this child active?", it may be more useful to ask questions like: "In which contexts do children tend to move and in which do they tend to sit?"; "Which common practices entail more movement and which more sitting?"; "How and in which contexts do these practices emerge?"; "Which properties of those practice contexts are conducive to movement or sedentariness?"; "Can we manipulate these contextual properties to induce healthier behaviors?" and most importantly: "How can we scale up inducing contexts where healthier practices appear?". In school-based PA intervention, part of the social-environmental context at intervention time is fixed. Children of the same society are sent to the same institution for the same reasons at the same times, then attempting to get them to engage in the same practices in hopes of achieving the same results: educated, well-socialized, healthy, and happy citizens. This property should allow research designs where various contextual elements can be manipulated and changes (if any) in practices, their performance, and their PA content observed. Whole-of-school interventions (Colabianchi et al., 2015; Mooses et al., 2021; Pulling Kuhn

et al., 2021; McMullen et al., 2022; Webster, 2022) aim to increase PA of youth before, during, and after school by engaging school staff, students, parents, and the wider community. A practice theory lens can help understand the resistance to and adoption of new practices necessary for sustained implementation and eventual habituation of intervention components delivered by school staff (Spotswood, Vihalemm, et al., 2021). Combining practice theory analysis of intervention delivery with automatic PA observation could provide valuable information on the dynamics of adoption and sustainment of whole-school interventions and their effects on the average level and distribution of PA in the schoolhouse throughout and after the school day.

Finally, there are intervention strategies attempting to cover a whole range of social-environmental influences on movement behaviors. **Ecological models** (Sallis et al., 2006; Sallis & Owen, 2015) view behavior as resulting from various environmental influences from multiple levels acting on and interacting with the intrapersonal core (e.g., demographics, phenotype, health status, psychology of the individual). Social-environmental influences can range from the level of the immediate social and physical environment (e.g., home, parents, school route, school, classmates, teachers, access to a smartphone) to the organization (e.g., school PA policy and resources), community (e.g., availability of football fields and football players), state (e.g., youth sport participation subsidies), up to the global environment (e.g., climate, technological megatrends, internet, World Health Organization policy, global economy, and energy prices). In addition to covering the PA effects of the physical environment, this perspective unites the intervention efforts of actors at multiple levels into a single framework: public health policy at global, international, national, local, and organizational levels together with PA intervention efforts of parents, teachers, community members, and society in general (rituals, traditions, and cultural attitudes related to various sedentary or active behaviors like smartphone usage or different modes of transportation). While they may provide stronger and wider public health impacts, implementing such multi-level interventions and obtaining evidence on intervention effects is difficult and expensive especially when considering interactions of effects at multiple levels and differences at the intrapersonal level (Sallis & Owen, 2015). In a holistic multi-level PA intervention program encompassing state policy, media campaigns, spatial planning, community, school, and family levels it could be nearly impossible to disentangle the PA effects of stimuli at different intervention levels specific to age, gender, and socio-economic status. Paradoxically, the more thorough and successful an intervention program is (maximizing the number and quality of intervention stimuli at all levels), the harder it will be to isolate its unique effects: with increased integration and mainstreaming of health promotion comes decreased visibility of particular interventions (Dooris et al., 2007, p. 339). To enhance the effectiveness of multi-level ecological PA interventions, individual components could be studied separately using (quasi)experimental designs. Computer vision-based automatic observation could be useful for testing the efficacy of certain location-specific intervention components (school gym and playground) and comparing different con-

figurations of the physical environment (temperature, markings, stationary PA equipment, playground, and its parts).

As ecological models can theoretically cover all the environmental influences on our movement behaviors, what remains is our genes which have been found to explain from 20% to 90% of the variation of PA in adults (Lightfoot et al., 2018). A comprehensive PA promotion strategy would likely combine knowledge of genetic correlates (Z. Wang et al., 2022) of movement behaviors (personalized preventive medicine), known risk groups [e.g., the inactivity tendency of adolescent girls (Duffey et al., 2021)] while maintaining a unified strategy for the total population (e.g., universal PA-promoting environmental configurations, interventions targeting the universal preference for low effort behaviors and universal psychological needs as per the self-determination theory).

2.1.3 Synthesis

Privacy-preserving location-based automatic PA observation could be useful for PA intervention research in all the described theoretical paradigms given an appropriate setting and research design. In school-based intervention (and possibly other interventions where the same population appears consistently in the observed area), constructs of individual-level theories can be monitored with questionnaires (less so for affective variables) at multiple time points while continuously monitoring PA levels in the building using video analysis. Practice theory-based interventions could benefit similarly but perhaps using more qualitative methods to monitor practices. In case of ecological models, continuous privacy-preserving location-based monitoring of PA draws an obvious connection to research concerning the built environment and stationary equipment meant to induce PA. Surveilling a larger area of interest for a longer period using many video analysis sensors can provide insight into the spatio-temporal dynamics of movement behaviors in a specific ecological setting throughout the seasons. An understanding of possible seasonality of PA at different time scales (PA in the morning and after lunch, during colder and warmer months) could enhance the evidence base for PA intervention by contributing to theory and informing the timing of certain types of interventions.

To synthesize a PA intervention policy and research strategy, combining an ecological model of health behavior (Sallis & Owen, 2015) with the theory of effort minimization in PA (Cheval & Boisgontier, 2021) seems appealing. Ecological approaches have the potential for the widest public health impact, and are easily communicable to policymakers in charge of sustaining and changing the environment we live in. The ecological model also implies a role and responsibility of parents, teachers, and the wider community in the health behavior of children and youth. TEMPA is favorable due to its specificity to bodily movement when the other individual-level theories were developed for different purposes and fit better to specific meaningful actions. Cognitivist theories deal with attitudes, which, by definition, require a concrete object – attitude towards something specific such as the act of watching television or taking part in dance classes.

TEMPA on the other hand deals with PA and its avoidance in general – the tendency to prefer actions that require less energy expenditure. Planting TEMPA at the intrapersonal core of a holistic ecological model could already hint at intervention strategies: what can various actors at various places/settings do to reduce effort perception of various movement-related cues perceived by children and youth? Can we build automatic effort perception modification systems into the physical environment? Can we adopt a pedagogy and culture of convincing children how easy and fun it is to perform all kinds of physical activities? Can we consistently provide enjoyable experiences and/or outcomes whenever and wherever children engage in vigorous movement?

2.2 Physical activity and its intensity

When developing a method for measuring anything, one needs a thorough understanding of the measurement construct. Caspersen and colleagues (1985, p. 126) defined PA as "*any bodily movement produced by skeletal muscles that results in energy expenditure*". The dose of PA consists of the components of activity type/mode, duration, frequency, and intensity. As this thesis develops a method for location-based PA estimation, the frequency and duration components are not discussed further – these are not captured in a stationary video camera feed. **Study I** provides an overview of the various methods used for measuring PA, but here I focus on inferring PA intensity from wearable triaxial accelerometers (Figure 1) which are used for developing the proposed method. Intensity of PA refers to the quantity of energy expended on bodily motion often expressed relative to the resting energy expenditure[9] of the person – that is in metabolic equivalent of task units (METs). When an activity requires twice the calories burned when sitting quietly, the intensity of that activity is two MET. Strictly, the MET concept is individual as people have different bodies and metabolisms, however an oxygen utilization rate of 3.5 ml per kg of body mass per minute has been used as convention for one standard (adult) MET. Resting metabolism is higher in prepubescent children and generally higher in males (Harrell et al., 2005) so using the standard adult metabolism to define PA intensity is not very meaningful in the school-age population. For PA measurement in youth, the variability of anthropometrics, metabolisms, and the rate of their change (girls reaching puberty earlier) make it very difficult to translate accelerometer signals to a standard PA intensity construct (Figure 1). Some have suggested an age-based approach (Freedson et al., 2005) to account for differences in metabolic development and anthropometrics. In such cases however, the same accelerometer cut-off value is applied to people of same age but different body composition and fitness level while actually, a heavier person needs to expend more energy to achieve the same acceleration (Raiber et al., 2017; Raiber et al., 2019). Due to the complexities of

[9] Resting energy expenditure (REE) is the absolute energy expenditure in the resting position and resting metabolic rate (RMR) is REE per unit of body mass.

translating accelerometer data to standard PA metrics in youth (McMurray et al., 2015), a compendium-based approach (Butte et al., 2018) has been developed: anchoring PA measurement to a collection of energy costs of many common activities (e.g., reading, playing board games while standing, trampoline, etc., altogether 196 activities in 16 categories) measured in several age groups. Even this approach is not strictly correct because instead of measuring resting metabolic rate, it relies on defining the MET based on Schofield's regression equations (1985) for predicting basal[10] metabolic rate in youth from age, sex, and body weight. For developing the machine learning data set used for the video analysis method, this thesis opted to rely on expert knowledge in dealing with the uncertainty and complexity of translating accelerometers to PA intensity (**Study II** and Chapter 3.3).

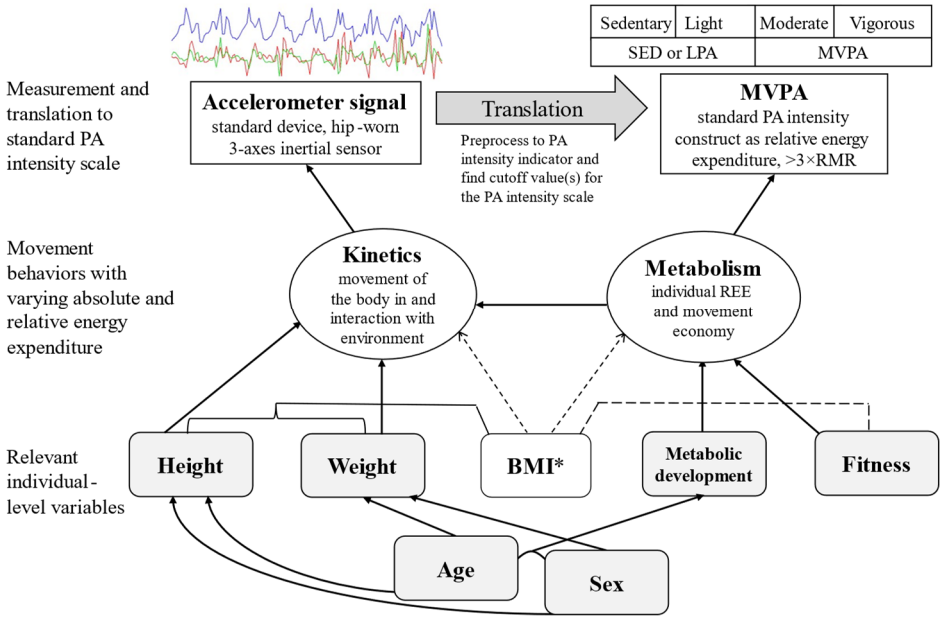

Figure 1. Some factors influencing the accelerometer signal and its translation to a standard PA intensity scale based on (individual) MET. As this diagram is only meant to give a brief overview of the complexity of measuring PA in the school-age population, not all the relations and interactions between all these elements are brought out or elaborated on.

* BMI – the body mass index (weight/height^2 in kg/m^2) as an indicator body fat.

[10] Since basal metabolic rate refers to the metabolism in a state of total relaxation (the number of calories per kg body weight necessary for the most basic life-sustaining functions), it is lower than resting metabolic rate measured from sedentary position which is used for the classic definition of MET. Hence it is called the youth MET – MET$_y$. Also, the regression equations are not guaranteed to predict the true basal metabolic rate (Bottà et al., 2020).

Associating PA intensity levels, their prevalence, and distribution in daily life to health outcomes (dose-response relations) is a complex research topic (Shephard, 2001; Warburton et al., 2006; Arem et al., 2015; Oja et al., 2017; Ekelund et al., 2019). Performing a few short bouts of intensive PA may have different health effects than doing a longer bout of less intense PA. These effects could be different between sexes, age groups, body types, and fitness levels. How does one even draw the line between not so healthy lower PA intensities and the healthier higher intensities? These questions have been tackled by various public health authorities and their advisory committees based on reviewing state of the art research discussed below.

While prior public health guidelines were focused on promoting vigorous exercise, by the mid-90s evidence had accumulated that convinced leading experts to start stressing the importance of moderate intensity PA (MPA) as activities with an intensity of 3–6 MET (Pate et al., 1995). The Physical Activity Guidelines Advisory Committee (2008) determined that roughly 500 to 1000 MET-minutes of moderate to vigorous PA (MVPA) per week should bring sufficient health benefits for the majority of people (walking for 200 minutes at 3 MET intensity would be 600 MET-minutes). Instead of using the concepts of METs and MET-minutes, the guidelines developed from this research (U.S. Department of Health and Human Services, 2008) encourage to communicate MPA as the equivalent of brisk walking. For children and youth, at least 60 minutes of daily MPVA has been recommended (Janssen & LeBlanc, 2010), however the authors admit that there is a lack of evidence on the health effects of lower intensity activities and that the MPA threshold is inconsistently set at either 3 or 4 MET[11] in youth. As such, the MVPA intensity threshold and prescribed dose used in public health recommendations are somewhat arbitrary. Oversimplifications made in the interest of communication may introduce a problem. Blair et al. (1992) noted an apparent incorrect dichotomous view of PA intensities among professionals and the general public – as if PA has health benefits only when exceeding the daily recommended dose. In actuality the dose-response relationship is gradual: even some PA is better than just sitting all day (Lee, 2007). The WHO 2020 guidelines for PA (Bull et al., 2020) acknowledge these issues, putting less emphasis on minutes of MVPA while stressing that some PA is better than none and more PA is better for optimal health outcomes. These latest recommendations also particularly prescribe muscle-strengthening activities and focus more on reducing sedentary behaviors (sitting, reclining, or lying) as a separate target not just considering sedentariness as low intensity PA to be treated similarly to standing.

[11] Assumptions on both the numerator (PA intensity threshold for MVPA - how brisk a walk should be to constitute MVPA) and the denominator (resting energy expenditure) in MET definition can lead to these discrepancies. When assuming a low resting energy expenditure, brisk walking may require 4 MET. When measuring the actual resting energy expenditure in prepubescent children, brisk walking could possibly be achieved even below 3 MET.

2.3 Direct observation and its automation

Humans are highly visual creatures (Kaas & Balaram, 2014) so one might assume that observation as a research method should come naturally to us. Indeed, we have evolved brain structures for visual processing unprecedented in mammals (Kaas & Balaram, 2014). However, we are also social creatures with adaptations for processing visual information specifically from other people and their behaviors (Pitcher & Ungerleider, 2021). Taken together we have powerful means for processing visual information in general, but for methodically observing our own species, there may be blind spots to some visual cues (e.g., those which have not provided a fitness advantage if noticed quickly) and over-attention to others (e.g., noticing signs of aggression). It has been shown that during action observation, the mirror neuron system generates motor simulations of observed actions and this process is modulated by various factors including the actor, the observer, their relationship (even race, ethnicity, in-group/out-group membership) and the context (Kemmerer, 2021). It is not clear how the involvement of the mirror neuron system and its motor simulations impact the quality of observational data. But intuitively, observation of other animals which do not activate the mirror neuron system at all should be less biased.

Observation methods can vary greatly in data collection procedures, observed behavioral categories, and the type of inference they allow. On the one end, there is strict ethological observation of natural behavior which aims to fully describe behavioral sequences constrained in time and space based on a predefined ethogram (e.g., Jones et al., 2016) of mutually exclusive behavioral categories. On the other end there is participant observation in sociology where the observer, instead of taking a cold external view, takes part in the activities of the group they study and where behavioral categories[12] are not determined *a priori*, but emerge gradually throughout fieldwork and are often combined with interviews (Platt, 1983).

Already the pioneers of ethnography and cultural anthropology noted the conflicting nature of methodical observation of human behavior. Bronislaw Malinowski stated in "Argonauts of the Western Pacific":

> *As to the actual method of observing and recording in field-work these imponderabilia of actual life ["…a series of phenomena of great importance which cannot possibly be recorded by questioning or computing documents, but have to be observed in their full actuality"] and of typical behaviour, there is no doubt that the personal equation of the observer comes in here more prominently, than in the collection of crystallised, ethnographic data.* (Malinowski, 1922/2017, p. 21).

While naturalistic observation promises objective knowledge in the sense of recording behaviors as they occur in their natural setting (except for the presence of the observer), the "personal equation" of the observer still influences attention

[12] Quantifying the prevalence of specific behaviors is not the goal. In participant observation, visual information on behaviors and interactions are combined qualitatively with other sources of data to reveal (social) meanings.

and thereby also the data. If different observers are predisposed to notice different cues and aspects of behavior, then they are also likely to produce different recordings. While such individual differences may not be a major source of error in ethological studies of humans observing other animals, the life experience, cultural background, and social position of the observer can affect results much more when studying humans. For this, social scientists are taught to "switch off" their cultural background as much as possible to reduce bias in observation and interpretation of the phenomena being studied. Human observation of humans is also complicated by our own familiarity with human behavior such that research-wise relevant and visible behavioral information may be missed or left un-recorded since it seems so natural and trivial (Richer, 2017). Striving to minimize subjectivity by viewing the observed person as a biological machine, as opposed to a moral agent, can feel uncomfortable and amoral as we are not used to such cold objectification of people (Richer, 2017).

Aside from the methodological problem of gaining objective knowledge from subjective experience, observational research also entails a difficult ethical problem. While it was not considered a major issue during the early days of cultural anthropology, privacy is now considered a fundamental right. As such, the very act of observing a person could be considered a rights' violation. Researchers can get past this by obtaining consent from the research subjects [this introduces potential subject reactivity – the Hawthorne effect (McCarney et al., 2007)] or by doing the study in a public space where people observing each other is unavoidable.

Direct observation has been used to assess children's PA often with methods designed for specific contexts (e.g., recess, physical education class, playground, or at home) and focusing on context-specific research questions (McKenzie, 2002). Since observation allows to capture rich contextual information associated with the observed behaviors, it is especially useful for studying cognitive-behavioral and ecological antecedents of PA (Sallis & Owen, 2015). As such, observation is a natural fit to school-based PA intervention research that aims to understand which properties of the school environment and which strategies of intervention facilitate or hinder PA.

There are several possible sampling protocols for PA observation: one could mark down whether a behavior occurs during a time interval [partial time recording or interval recording (McKenzie & van der Mars, 2015)] or one could try to estimate the average PA intensity level for the time interval similar to **Study II**. But most commonly momentary time sampling is used: collecting the observation sample at the end of the time interval (PA intensity of the observed subject at the moment of the audio cue), allowing more attention for observing contextual factors (McKenzie, 2002). In several established instruments, the intensity of PA is defined by a combination of semantic categories of body position, locomotion (foot-to-foot movement), and a more obscure notion of energy expenditure for the higher intensity class. In the System for Observing Fitness Instruction Time (SOFIT) (McKenzie, 2015), the activity levels are defined as (1) lying down, (2) sitting, (3) standing, (4) walking (considered as MPA), and (5) vigorous which

corresponds to energy expenditure beyond what is needed for ordinary[13] walking. McKenzie (2015) specifies: "*When the student is in transition from one category to another, enter the code for the higher category*". From the sampling method and observed categories, time spent in MVPA is estimated. SOFIT was shown to be valid for measuring children's PA by using heart rate as criterion measure (Rowe et al., 1997). It has been suggested that combining SOFIT observations with subjects' body weight can even provide a relatively robust measure of PA energy expenditure (Honas et al., 2008). SOFIT has been widely used in physical education research (Smith et al., 2018).

With the rapid development of deep learning and computer vision, the question arises whether direct observation methods could be automated. Human observers require food, bathroom breaks, sleep, and salary. As with the automation of manual labor, machines promise to overcome these basic human limitations. To the author's knowledge, the first investigation into the potential of video-based PA measurement was conducted by Silva et al. (2015). They used a top-down view camera in a gymnasium combined with a tracking algorithm. Anchoring the pixels of the scene to real-world coordinates on the basketball court allowed them to obtain a measure of PA by computing the velocities of the tracked subjects. This system was designed as a semi-automated observation tool where video was recorded for later analysis by the researcher using the tool. Hence, full automation of PA measurement with privacy preservation was not the goal of this system. Carlson and colleagues (2017; 2020) were the first to develop a PA observation video analysis approach using deep neural networks (discussed in **Study II**). Their system outperformed human observers both in counting people and assessing the share of people active above the MVPA threshold determined by accelerometers (Carlson et al., 2020). **Study I**, conducted at a time when the new General Data Protection Regulation (Regulation 2016/679/EC) was causing concerns among European researchers working with human subjects, argues that automating observation can also provide major ethical and data security benefits if the video analysis was done at real-time speed without recording any of the pixels carrying personal information of the subjects.

Inherent advantages of human direct observation methods for PA measurement include flexibility, high internal validity, low inference, and low participant burden (McKenzie & van der Mars, 2015). On the negative side, McKenzie and van der Mars (2015) note the necessity for careful observer training and recalibration, inaccessibility to certain environments, and potential subject reactivity. Table 2 considers these advantages and disadvantages from the perspective of automating the observation method using a stationary video camera and a privacy-preserving video analysis system. Which of these advantages can be maintained and which would be diminished? Which disadvantages can be improved on, and which will remain when replacing the human observer with a real-time video analysis machine?

[13] This method classifies "ordinary walking" as MVPA while public health authorities recommend drawing the line at "brisk walking". An example of the measurement construct definition problem addressed in **Study II**.

Table 2. Advantages and disadvantages of measuring PA by conventional trained human observers and privacy-preserving video analysis sensors.

Method property	Comparison of direct observation of PA by humans and privacy-preserving automatic video analysis	+/– *
	Advantages of human direct observation of PA	
Flexibility	While human observers can be trained to observe other relevant behaviors or contextual elements (including antecedent-behavior-consequence observations), a simple video analysis sensor is only recording PA intensity. Contextual/environmental factors of interest would have to be controlled by experimental design or other means.	–
High internal validity	Instead of trusting the observations of trained human observers, in the video analysis case, researchers need to trust the camera placement and the capacity of the ML model to perform in this setting as well as it did on the validation set (equivalent of trusting the human observer to perform as well as they did during training/recalibration). A ML model can fail in some cases which would be trivial for humans (occlusion and unusual appearance). However, given a sufficiently large and varied training data set, a video analysis sensor could likely be more accurate on borderline cases.	–
Low inference	When both human- and automatic observation methods attempt to capture the same PA intensity construct from visual information, both can be considered low inference direct measures. However, an abstract probabilistic MVPA threshold learned by a ML model could fit theory (MVPA defined as >3 MET) better than an arguably lower inference semantic category ("walking" or "brisk walking") learned by a human observer. Still, the researcher would have to trust a (thoroughly testable) black box model.	–
Low participant urden	Human observation of PA is considered to cause low participant burden when compared to attaching accelerometers to the subjects or asking them to fill out surveys. A privacy preserving real-time video analysis approach causes even less participant burden as it removes the human observer and their intrusive gaze.	+
	Disadvantages of human direct observation of PA	
Observer training and recalibration	Compared to training and recalibrating human observers, a ML model can be trained once, and the biases of the observation system are fixed in the model. This allows different researchers to use the same observer comparably and to trust each of the observation sensors having stable bias throughout the study. Creating a trustworthy ML data set as well as developing large 3D convolutional neural networks are still costly.	+
Inaccessibility to certain spaces	If privacy preservation is certified and trusted, the sensors could be used in various spaces not accessible or viable for human observation. However, camera placement can restrict observation under some conditions where humans could perform easily (real-time processing restricts us to lower resolution video data which directly limits the maximum distance from the camera where person detection is viable).	+
Subject reactivity	Compared to the human gaze (Ricciardelli et al., 2000), a video analysis sensor is harder to notice so the behavioral reactions to observation should be much less. However, the presence of a camera, even when not recording any video, may still cause some behavioral reactions (van Rompay et al., 2009; Pfattheicher & Keller, 2015).	+

Table 2. (Cont.) Advantages and disadvantages of measuring PA by conventional trained human observers and privacy-preserving video analysis sensors.

Unique advantages of automatic observation of PA		
Stable subjectivity/ reliability[a]	Automating observation by machine learning solves a core reliability problem of human observation: dynamic subjectivity of the observer/ observer drift (observation performance can change over time). The subjectivity of a ML-based video analysis sensor at the deployment site is stable throughout the observation period (the weights and biases of the model are fixed).	+
Research ethics[a]	Real-time automatic video analysis removes objective privacy intrusion. Subjective privacy harm (Calo, 2010) may still be an issue – e.g. if one does not believe that nobody will see the video or if facial recognition and identification are suspected.	+
Sampling rate/data consistency[a]	Human observation instruments only sample PA momentarily in relatively long epochs (10–15 seconds observation and the same duration to take notes) or record whether/how often a behavior occurs within the long epoch (commonly a minute) (McKenzie, 2002). A video analysis approach would observe PA continuously, striving to classify all visible behavior in terms of PA intensity.	+
Length of continuous observation[a]	Basic human needs severely limit the possible duration of continuous observation while a video analysis device can generate observational data indefinitely provided stable power supply, cooling, and integrity of its components.	+

[*] "+" indicates that automatic observation has an advantage compared to human observation, "–" a disadvantage.
[a] indicates method properties added by the author to ones listed by McKenzie and van der Mars (2015).

In summary, these comparisons (Table 2) indicate that automating PA observation diminishes the flexibility allowed by human observation: ability to record subjects' interactions and phenomena directly relevant to intervention research (how the observed behavior relates to intervention stimuli). Automation also ameliorates or removes completely some of the main weaknesses of human observation by enabling lower subject reactivity and stable subjectivity. As an additional benefit for interpreting research results, this subjectivity can be studied by analyzing the potential biases in the training data set and model performance under various conditions of interest. If such a thoroughly analyzed ML model would be available for researchers globally (and assuming the training data set is large and varied enough to be considered globally representative), then this research should be more-or-less comparable. Of course, different researchers could set up the sensors differently and the lighting conditions and their temporal patterns could be different around the world, but **the MVPA threshold would be the same**. Hypothetically, with enough effort, a ML data set could even be developed which reflects the correct MVPA threshold for different body types across the school-age population. On the negative side, a video analysis sensor would be susceptible to adversarial attack – a mannequin or a large poster of a person could register as a person thereby corrupting the data.

2.4 Supervised machine learning

Automating the task of observation can be seen as building a machine that can fulfill a similar function as a human observer – given the same input, to generate the same output as a human would. In other words, the method developed here requires approximating a function that could be considered intelligent[14]. As such, supervised machine learning and artificial intelligence form the basis of the proposed automatic observation method.

Supervised machine learning entails learning a function that maps inputs to outputs based on observing valid input-output pairs (Russell et al., 2010). For example, one may be interested in learning a function that maps an input vector of health indicators (data/input) to a diagnostic output of whether the patient has a particular disease or not (label/output). To learn this function in a supervised manner, one needs examples of the health indicators for people who have the disease and for those who do not. These data-label pairs of known positive and negative cases form the machine learning data set. The data set is split into training and test sets and combined with a learning algorithm to generate the input-output mapping – this is called training or learning the model. In this example, the mapping (the function approximated by the supervised machine learning model) is a diagnostic classifier that one hopes is generalizable to new patients. For estimating generalizability, the model learned from training data is fit to the test set, to evaluate its performance on unseen cases. As this example of supervised machine learning approximates a function mapping a vector to a binary output variable (the diagnosis), a confusion matrix of true positives, true negatives, false negatives, and false positives can be used to evaluate model performance. The same general principles of function approximation apply to mapping inputs to multinomial and continuous (regression) outputs, but the latter case is evaluated with error metrics.

Well-known algorithms like linear- and logistic regression can be used in a supervised learning framework, but these simple algorithms are not able to map an input of image or video data to an output of interest. To understand how to approximate functions given such inputs, some critical developments in the history of machine learning are noted. In 1943 the first mathematical model of an artificial neuron was published (McCulloch & Pitts, 1943) laying the foundations for modern artificial neural networks (ANNs). This simple mathematical model depicted the neuron as weighing each input, summing them, and comparing the result with a threshold value to determine whether the neuron "fires" (propagates the signal to the following neurons) or not. In 1957 Frank Rosenblatt built the first ANN which was an image recognition apparatus with 400 photocells connected to physical artificial neurons with weights encoded in potentiometers (Rosenblatt, 1958). As a single-layer network, it was able to learn only linearly separable patterns by tuning the weights of just one layer. Later it was shown that adding more

[14] Here I mean intelligent in the sense that one could teach a human to conduct PA observation, but not a dog.

(hidden) layers increases the range of functions such networks can approximate (Ivakhnenko, 1971). Eventually it was demonstrated that stochastic gradient decent backpropagation (Linnainmaa, 1976; Rumelhart et al., 1986) is a viable learning algorithm for networks with multiple layers. As one of the key breakthroughs enabling the current deep learning revolution, Krizhevsky and colleagues (2012) demonstrated the power of including additional hidden layers and the viability of training these deep networks efficiently on graphics processors (GPUs).

Convolutional neural networks (CNNs) are a class of ANNs extensively used in computer vision. These networks use the convolution operation to produce various representations (feature maps) of digital image input. The convolution kernels which produce the feature maps are learned using gradient descent back-propagation (Gu et al., 2018). A convolutional layer is usually followed by a pooling layer which reduces the resolution of the feature maps before the following convolutional or fully connected layer. LeCun and colleagues (1989) developed the first CNN and used it for handwritten digit recognition. This basic concept of convolution and pooling has been at the core of the computer vision revolution until transformer models (Vaswani et al., 2017) started to take over (Khan et al., 2022). Convolution and pooling can also be implemented in three dimensions which is especially useful for video analysis. While 2D CNNs aggregate the RGB image data spatially to form the two-dimensional feature maps used in object detection, 3D CNNs (Ji et al., 2013; Tran et al., 2015) can also convolve over the time dimension in video allowing to form 3D spatio-temporal feature maps useful for recognizing actions, processes, and their properties such as intensity of movement.

It has been formally proven that multilayer feedforward networks can approximate any measurable function from one finite-dimensional space to another given a sufficient number of hidden units even with just one hidden layer (Hornik et al., 1989). One can oversimplify this proof such that given a sufficiently deep and/or wide neural network, any measurable function between two finite-dimensional spaces can be approximated. Of course, this says nothing about how to do the mapping or whether it is technically viable. But it does indicate that a mapping of a video format to PA intensity levels of people visible in the video should be possible as PA is a visually measurable phenomenon.

3 METHODOLOGY

The overall research and development process of the blind observation method follows roughly the design science research methodology for information systems research (Peffers et al., 2007). Design science is concerned with designing artifacts with certain desired properties in order to address a problem (Simon, 1988). In the present case, the target artifact is a video analysis sensor for privacy-preserving PA intensity estimation in the field of view of a camera to advance the field of PA intervention research. As the function of the developed device is to measure human behavior for scientific research via the visual modality, ethics and data security are prioritized from the beginning: a *privacy by design* approach (Cavoukian, 2009; Schaar, 2010). The six steps of the adapted methodology (Peffers et al., 2007) are presented in Table 3 with respect to the following three main criteria: a) a set of sensors should be able to reveal the spatio-temporal distribution of PA in school; b) the sensors must preserve the privacy of the people being sensed; and c) striving for general data security.

Table 3. Design science research methodology for PA sensor development

0	**Initiation**	Observing the need for measuring PA in school-based PA interventions, preferably with low participant burden.
1	**Define problem**	Need to measure the (a) spatio-temporal distribution of PA in schools in a (b) privacy-preserving and (c) data-secure manner.
2	**Define objectives**	Build (a) a video action detection system able to classify MVPA (b) at real-time speed (c) on a low-power single-board computer.
3	**Design and develop**	(a) Develop a data set reflecting MVPA classification using accelerometers, 2D pose estimation, and expert survey of video classification; (b) Identify neural network architectures with good speed-accuracy trade-offs and available code; (b) design and optimize PA detection pipeline for real-time processing given the (c) hardware constraints.
4	**Demonstrate**	Video proof of real-time multiple concurrent action detection.
5	**Evaluate**	Test set performance for the PA intensity classifier; ball-park estimate of the whole action detection pipeline; qualitative examination and worst-case analyses of person detection and assessing computational cost in crowded scenes. Loop back to step 3 in future research and development.
6	**Communication**	Published paper on designing the PA intensity data set; presented results and video proof at one international and one local conference. Publish thesis.

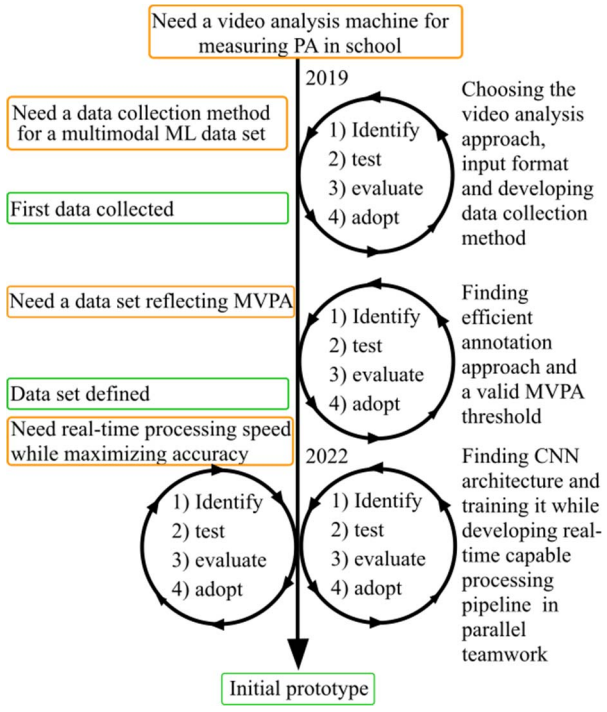

Figure 2. Research and development process.

The research and development process entailed several crucial design decisions requiring iterative search and testing to come to the eventual solution (Figure 2).

For the overall video analysis approach a two-step spatio-temporal action detection method was adopted (for an overview of action detection approaches see Vahdani & Tian, 2022). A two-step approach first detects and tracks all people in a video and then classifies these tracks into behavioral categories as opposed to a single-step system which detects actions directly from video without prior person detection (e.g., detecting falling in a retirement home instead of constantly classifying all visible people in terms of falling). When the goal is to measure all visible PA via a stable signal (a fixed video analysis epoch/ sampling frequency), a two-step action detection approach has several benefits:

1) As opposed to the scene-level approach of Carlson et al (2017, 2020), individual-level PA modeling should intuitively be more valid since the PA intensity concept is individual.

2) It is a natural fit for modeling PA intensity with the requirement of a fixed epoch length: whatever the goal or meaning of the behavior, we want to know the PA intensity of that person during that epoch. While there is a person, there is a PA intensity level.

3) As opposed to a single-step approach, it allows to develop the data set as a set of action tubes instead of a set of fully annotated video scenes. This means

that annotation can be selective[15] and it enables the exclusion of borderline cases (Figure 3) to learn a more robust classification model while providing more reliability for interpreting test set performance.

4) It allows to optimize person detection/tracking and PA classification separately thus increasing overall computational efficiency for achieving better performance at real-time processing speed on the constrained hardware of the prototype.

The development process and the rationale for various design decisions are described in the following subsections.

3.1 Data collection

Considering that wearable accelerometers are the go-to approach for objectively measuring PA in public health research, it was an obvious choice for assigning PA intensity labels for the video data set. I explored literature (**Study I**) to determine the most widely used wearable accelerometers and sampling frequencies to find a device that would allow good comparisons with existing and future PA research. While wrist-worn activity monitors could be preferable in free living studies of children and adolescents due to better adherence to device wear protocols (Fairclough et al., 2016), hip placement was chosen as it has been found to perform better in PA intensity classification and energy expenditure estimation (Migueles et al., 2017). ActiGraph devices (in this case the wGT3X-BT model) (ActiGraph LLC, Pensacola, FL) with a 30 Hz frequency appear widely used and validated while conveniently fitting the 30 (29.97) FPS video frame rate standard. Despite this match, Gholamreza Anbarjafari suggested that PA intensity classification should be viable at lower frame rates as well – slow walking and brisk walking look different enough (e.g., motion blur and background change resulting from displacement) to reduce the frame rate to 10 FPS. The decision to reduce the frame rate by 2/3 simplified work by reducing the size of the data set and greatly increased the chance of achieving real-time processing capacity necessary for privacy preservation. In the interest of real-time processing while also considering the intended use in mainly indoor settings, I selected a 1280×720 spatial resolution. In hindsight, a higher spatial resolution would have been a better choice as hardware and algorithms keep improving. The higher the spatial resolution, the larger the perceptive field of the sensor can be. Initially I collected data using a Logitec c922 webcam with autofocus disabled, and a custom script attempting to achieve a stable frame rate of 10 FPS while avoiding lossy compression. For the last two data collection sessions (REC6 and REC7), Rain Eric Haamer developed a second custom camera using the eventual sensor prototype

[15] This allows to use sections of video where a person without an accelerometer is in view: one can just annotate and cut out action tubes for the subjects wearing accelerometers. This property should also allow to repurpose existing video data for creating PA data sets via 3D monocular pose estimation techniques as argued in **Study II**.

hardware. The second camera had some barrel distortion and a different recording method which achieved a more precise and stable frame rate but introduced some lossy compression.

Having chosen the data format, I developed the data collection procedures. These choices try to obtain high validity and reliability for the sensor. Besides the size and quality of the data set, validity for a sensor using a supervised ML model can be gained from the general similarity of training data to the eventual application setting. The method was developed with the Estonian school system in mind where grades one to six are often together in one schoolhouse. For this, I chose the age group of 7–14 years. In the beginning, I recruited subjects by convenience sampling as it is much easier to approach friends and acquaintances to allow video recording of their children than it would be strangers. I conducted one major data collection session (REC6) in a school gymnasium with a school contact (Sirje Ange) helping to organize the informing and recruitment of subjects. As a lucky coincidence in convenience sampling, the last recording session entails twin brothers with substantially different body mass. The weight difference of these identical twins provides unique value to the data set by allowing to interpret the effects of anthropometrics on the hip-worn accelerometer signals, but this is outside of the scope of this thesis.

In such multimodal data sets, the data quality is highly dependent on the synchrony between the video feed and the accelerometers worn by subjects. The video capture system was implemented to record the 10 FPS video frames as .png images with the time stamp in the image name. I plotted the acceleration signals and explored time stamps to assess how to achieve best synchronization. It appeared that even for just a half-hour recording some minor accelerometer time drift appeared (largest discrepancy between two accelerometers was observed 567 ms over a 32 min period) and occasionally the webcam skipped a frame or several. To synchronize the accelerometers with each other and with the video feed, I adopted a strategy where I attached all accelerometers to my waist and jumped up and down repeatedly at the beginning and the end of each data collection session. The brief period of weightlessness before falling back toward the ground is clearly visible in acceleration signals and can be seen in video as well (Figure 2 in **Study III**). Plotting the accelerometers and visually assessing the time-stamped video frames allows to achieve near-perfect synchronization at the beginning and end of each session. I performed synchronization corrections by first imputing missing frames by copying the previous frame when the time elapsed between two consecutive frames exceeded 166 ms (two frames inserted if lag is above 300 ms, three inserted if lag is above 400 ms etc.). This introduces some noise to the data set since the eventual application camera does not repeat frames. Then I aligned the accelerometers at the synchronization moments in the beginning and end of the session revealing the extent of relative time drift. I imputed acceleration values as the median of each axis or removed with a strategy such that corrections are distributed equally in the period: if one value needs to be removed or imputed, it is done at the mid-point between synchronization moments, if two corrections need to be made, then the first around the 33% mark and second at 66% etc. Before committing to machine learning, I reassessed the synchrony of all data by

visualizing the acceleration signal in video as a circle on top of the bounding box with the radius defined by the acceleration signal (custom code written by Rain Eric Haamer).

Variability in the ML data set plays a large role in the generalizability of the ML model and thereby the reliability of the sensor being developed. To enhance variability, I asked the children to bring a coat, school backpack, and an extra pair of indoor shoes. During data collection I asked the subjects to change their appearance while attempting to cover the whole PA intensity scale (sedentary… vigorous) for each change of appearance. Having material for the same subject wearing socks, soft-sole shoes, and heavier boots can provide visual variability in gait while also possibly providing variability in accelerometer signals for the same PA intensity category (hypothetical effect of footwear on the hip-worn accelerometer signals resulting from feet impacting the floor). Additionally, I slightly repositioned the camera(s) at one or two time points during recording. Depending on the recording location, the lighting of the area was also modified throughout the session where possible (turning lights on/off, closing/opening curtains). During the sessions, I asked the children occasionally to walk or run at a certain pace in a certain direction and to sit in certain positions towards the camera in hopes of recording as many different expressions of the PA intensity categories for each subject as possible. Conflicting with the guided part of data collection, another goal was to also capture natural movement behaviors in free play by allowing longer uninterrupted periods[16].

I collected the data in seven recording sessions but with greatly varying amounts of data per session (e.g., REC2 and REC3 only entail one child filmed with one camera for 15–20 minutes while REC6 entailed three groups of children in one location filmed with two cameras). At most there are five children wearing accelerometers per scene. Altogether, 24 children (15 girls and 9 boys) participated, however, two boys take part in two different sessions (one year apart, different clothing and hair) so the two instances are considered separate subjects. Anthropometrics of the sample are reported in Table 4 and an overview of the scenes is provided in Appendix I.

Table 4. Age, weight, and height of subjects in train and test sets. Mean (SD).

	Train (n= 19 subjects)	**Test (n=7 subjects)**	**Weighted full train set (n=14228 samples)**	**Weighted full test set (n=7051 samples)**	**Weighted total (21279* 2-s samples)**
Age, y	10.3 (1.9)	10.9 (1.2)	10	10.9	10.3
Weight, kg	40.6 (11.5)	41.4 (9.4)	40	42.8	40.9
Height, cm	147 (12.5)	150 (9.3)	146.2	148.5	147

*This reflects the amount of unique material in the data set excluding partially occluded annotations and before various augmentation steps described in Chapter 3.4.

[16] Capturing natural behavior and free play was easiest for the youngest group (REC6 a) while the teenage groups (REC6 b and c) appeared more uncomfortable and rigid.

3.2 Data annotation

Compared to the scene-level approach of Carlson and colleagues (2020), a spatio-temporal action detection approach is much more work-intensive when it comes to data annotation. Instead of just assigning a people count and count of people active above the MVPA threshold to a video sequence, spatio-temporal action localization requires bounding box annotations to create the training and test samples in form of action tubes (e.g., Figure 1 in **Study I**). Computer Vision Annotation Tool (CVAT) (Sekachev et al., 2020) was used for action tube annotation but the exact technique of annotation was modified during the project to reduce manual work. For the first session (REC1 – the data used in **Study II**), I used CVAT automatic annotation functionality[17] to generate person bounding boxes. Then I annotated the people missed by the object detection model and adjusted some of the automatically generated boxes to be tighter around the persons and smoother/ more consistent across frames. REC 2-4 were annotated manually by Helis Ojala in video track mode providing frame-wise tighter and across-frames smoother action tubes. For the last recordings (REC 5-7), Klavs Jermakovs applied the FairMOT multiple person tracking model (Zhang et al., 2021) for automatic annotation with manual annotation of gaps and assignment of subject ID later on by Helis Ojala. Since the annotation approach and the annotator changed throughout the project, the overall consistency of annotations in the data set is relatively poor. Annotation instructions included annotating partially occluded persons when roughly 1/3 to 2/3 of the body area was occluded or out of frame. When more than 2/3 of the body was occluded, no box was annotated. At first, the partially occluded bounding box annotations were not used in training the model with the rationale to ease learning. Indeed, the model learned very quickly (Chapter 3.4, Figure 4 B), but the choice to discard the partially occluded annotations likely ended up seriously affecting PA classification accuracy when deployed in a setting where people do occlude each other (Chapter 4). Further models were trained including the partially occluded bounding boxes.

3.3 Defining the MVPA threshold

To achieve MVPA recognition in video analysis, a machine learning data set of bodily movement sequences paired with PA intensity labels is required. Given an arbitrary sequence of bodily movement, one must reliably assign a valid label corresponding to the PA intensity associated with that movement. For prescription purposes in medicine, one can describe the MVPA threshold as being equivalent to brisk walking. One way of defining the ground truth would be to have an expert look at the video samples and classify whether the bodily movement entails more or less energy expenditure than that of their understanding of brisk walking or 3 METs. When considering the volume of data involved in deep

[17] Faster RCNN Inception Resnet v2 Atrous Model (Ren et al., 2015) trained on the COCO object detection data set (Lin et al., 2014).

learning, in this case almost 12 hours of PA sequences, then this approach does not seem viable.

Another approach would be to have the subjects in the video wear an activity monitor/accelerometer which has been previously calibrated in terms of MVPA and validated in a similar population (metabolic development, height, and mass of the accelerating body in a particular age group), using the same attachment site (hip accelerations), and acceleration signal processing method. Due to the very specific and unconventional requirements in our use of the ActiGraph accelerometers, we were restricted to using raw acceleration data[18]. Triaxial accelerometer data can be reduced to acceleration vector magnitude by taking the Euclidean norm of its axes: $\sqrt{x^2 + y^2 + z^2}$. When standing perfectly still, the acceleration vector magnitude should show one gravitational unit (1 g). To better reflect the forces resulting from PA (van Hees et al., 2013), the gravitational component can be subtracted and resulting negative values rounded up to zero (Euclidean Norm Minus One g – ENMO). Hildebrand and colleagues (2014) developed MVPA cut points for ENMO using an earlier model of the ActiGraph device with hip placement in a group of children (30 7–11-year-olds) using indirect calorimetry. This study provided the initial anchor to other research, allowing to start interpreting the acceleration signals in terms of MVPA classification. The simplest way to assign PA intensity labels to the data set would be to just apply the 3-MET ENMO cut-point of 142 mg (Hildebrand et al., 2014). However, they used an accelerometer of a previous generation with a smaller dynamic range, their models are in part based on treadmill walking/running which may produce different signals than walking on a hard floor, their sample is younger, and they found a resting metabolic rate of 1 MET = 6.0 mL $O_2 \cdot kg^{-1} \cdot min^{-1}$ which is on the higher end compared to what others have found (Saint-Maurice et al., 2016).

Considering these uncertainties and the visual nature of the data, I decided that rather than blindly trusting the accelerometer cut points, reaching out for expert opinion could help produce a data set that better reflects the MVPA construct. Experts in the field of PA research would presumably have a relatively deep understanding of the PA intensity level which has been associated with health benefits. The approach to defining MVPA in the data set combines elements of expert observation and accelerometer thresholding while introducing a novel PA intensity indicator: 2D pose-estimated hip angle changes between video frames (**Study II**). The hip angle change indicator [computed using HRNet (Sun et al., 2019) from the AlphaPose repository (Fang et al., 2021)] showed almost as good discriminative power as ENMO in the small sample of **Study II**, but it is inherently flawed because it represents only the 2D projection[19] of the body pose. Still, the hip angle features should increase the quality of the ML data set because

[18] The widely used ActiGraph counts algorithm (Neishabouri et al., 2022) was held secret at the time and ActiLife software did not allow to import the precisely synchronized raw data files to compute the counts.

[19] As argued in **Study II**, 3D monocular pose estimation (Liu et al., 2022) could get past this allowing to compute joint angle changes from three dimensional poses. This approach could allow to translate existing pose estimation data sets and other existing video to PA intensity recognition data sets without the use of accelerometers.

it represents a different, likely complementary aspect of PA compared to accelerometers. Complementing the manually synchronized acceleration signals with an indicator computable directly from video could as well help smooth out asynchrony[20]. In **Study II**, the MVPA threshold was estimated in the ENMO and hip angle change space based on expert classification of 24 short PA sequences. Then this MVPA threshold was extrapolated to the rest of the data set while also removing borderline cases in **Study III** (Figure 3). Knowledge engineering is classically seen as a process of finding rules and pattern features to imitate the deductive reasoning of the expert (Brey & Søraker, 2009). The method applied here is somewhat similar but perhaps better described as mental model engineering/ estimation: accelerometers (validated in similar population) and pose-estimated hip angle changes (existing ML model of the human body validated on large benchmark data sets) in video are considered as stable PA intensity signals, expert consensus is extracted via a survey of video classification, and applied as cut-off values on each of the PA indicators and as a plane through the combined label space (Figure 3).

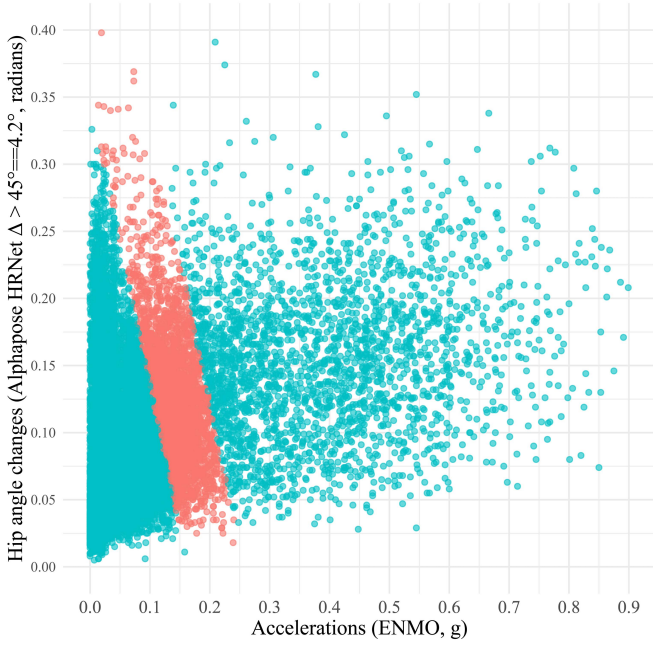

Figure 3. Removing borderline cases (red). Each dot represents a two-second action tube. Results of **Study II** and testing of the eventual prototype (**Study III**) imply that using the ENMO cut point[21] by Hildebrand et al. (2014) could have led to a sensor that classifies the majority of walking as MVPA. The users of such sensors are likely more interested in PA intensities higher than that of comfortable walking.

[20] The hip angle change indicator also proved useful for detecting synchronization problems and mix-up of subject IDs during the development of the data set.

[21] Modeling expert opinion produced a cut point of 0.17 g ENMO, substantially higher than 0.14 g found by Hildebrand et al. (2014).

3.4 Video analysis processing pipeline development

For automatic observational data to be meaningful, a fixed time sampling method is required. Carlson and colleagues (2017; 2020) chose a one-second epoch length and even though their data set was large, the PA intensity estimation capacity of their automatic observation system was moderate [0.55 concordance correlation with the ground truth (Carlson et al., 2020)]. In **Study II** we speculate that increasing the epoch length could increase the PA intensity classification performance by reducing synchronization errors resulting from accelerometer time drift and by generally providing more temporal information on the observed behaviors. As a trade-off, a longer epoch length brings higher computational cost and increases chances of people stepping into or out of the scene in the middle of an epoch. The latter is more of an issue in scene-level systems where each epoch requires a people count label: how many people are there really if two step in and one steps out during an epoch? For a spatio-temporal action detection approach with a fixed epoch, stepping into/out of the scene in the middle of an epoch can be handled by setting a minimal acceptable length for a valid action tube (**Study I**, Figure 1) and applying temporal padding on each end of the short tube (inserting blank frames to match the input shape of the action tube classifier). In **Study III**, I chose a two-second epoch length resulting in the video analysis unit of 1280×720×20 every two seconds.

Given the final data format, the two-step spatio-temporal action localization approach can be specified. The first stage requires detecting and tracking people in these two-second video samples. Then these tracks need to be formatted to a data structure which is the input of the PA intensity classifier (Figure 5 D). Person detection and tracking models are widely available because multiple object tracking (MOT) research is largely based on person detection and tracking benchmark data sets (Luo et al., 2021). Mateus Reis tested several freely available multiple-person tracking models on some of our videos (Bergmann et al., 2019; Wojke et al., 2017; Yang, 2020; Y. Zhang et al., 2021) and we chose one with favorable speed-performace trade-off [ByteTrack by Y. Zhang et al. (2022)] without fine-tuning the person detector on our data set. ByteTrack (Y. Zhang, 2021/2022) allows using different versions and configurations of the YOLO family of object detection models[22] which makes it easy to find a model that uses optimally any computational overhead left over by the rest of the processing pipeline on particular hardware (Chapter 3.5, Figure 5). The sensor technology could likely benefit from fine-tuning the person detector on data form children as play behavior entails more unusual body positions than there are in MOT data sets of public spaces like malls and train stations. As the goal of **Study III** was to develop the first prototype, not a final product, work focused more on the second step of classifying the action tubes into PA intensity categories.

The choice of neural network architecture for action tube classification was somewhat theoretical but mostly pragmatic. I explored video action recognition

[22] The AlexeyAB/darknet GitHub repository (Bochkovskiy, 2016/2023) covers most of the YOLO family of models initially developed by Redmon and colleagues (2016).

literature and GitHub repositories to find a model with a good speed-accuracy trade-off and freely available code with favorable license terms. As a lucky coincidence, the code for the highly efficient X3D architecture (Feichtenhofer, 2020) was published as part of the PyTorchVideo repository (Fan et al., 2021) at a critical time during **Study III**. This model was advertised to perform at real-time speed on a smartphone, and the code appeared to be of high quality, so the choice was made. A TensorFlow implementation of the model (Ogidi, 2022) was adopted however as it was more familiar to Klavs Jermakovs. In the original work, Feichtenhofer (2020) expands the X3D architecture gradually across various axes (hyperparameters) to find a good accuracy-efficiency trade-off. These expandable hyperparameters include the temporal length and frame rate which make this architecture easy to adapt to our custom data format. In the interest of processing speed, the smallest version, X3D-XS, was chosen as the base model and Klavs Jermakovs adjusted the temporal parameters to fit our custom format.

Transfer learning is the concept of using knowledge learned from one data set (pre-training) to increase the performance of a target model ultimately trained on a different data set (fine-tuning) (Zhuang et al., 2021). Instead of training a randomly initiated model directly on our rather limited data, the X3D network was first pre-trained on Kinetics 400 (Kay et al., 2017). This data set of human actions entails several categories which should intuitively be useful for transfer learning PA intensity classification. If the neural network learns to differentiate human actions, some of which also differ by their PA intensity (e.g., "presenting weather forecast" compared to "playing tennis"), then the network should already have learned some kind of representation of the human body and its movement intensity before starting to learn on our limited data set. Klavs Jermakovs down-sampled the Kinetics 400 data to match our frame rate and trained the model for 18 epochs until it surpassed 30% accuracy (Figure 4 A). Pre-training was spread over four Tesla-V100 GPUs and fine-tuning used one.

Before fine-tuning this model for MVPA classification, several pre-processing steps were performed on our raw data set. In iterative development, these pre-processing steps were modified producing two versions of the data set. The pre-trained X3D backbone was used to train altogether three models (Figure 4 B, C, and D) each time fixing some shortcomings of the previous attempt. Klavs Jermakovs fine-tuned the initial model (Figure 4 B) on a version of the data set which excluded the partially occluded bounding box annotations. Testing this model deployed on the prototype showed increased predicted probability of MVPA in cases of people walking past each other which I suspected was caused by excluding the partially occluded bounding boxes and not having any samples shorter than the full 20 frames (both conditions which can be expected in the application setting). To address this issue, Mateus Reis sampled a second version of the data set as described below.

The continuous nature of PA allows to use the same video sequence to make several samples. A three-second action tube (30 frames) can be turned into two "unique" two-second action tubes by treating the last half of the first tube as the first half of the next tube. This shifted resampling strategy was used to roughly double the size of the data set to ~42700 action tubes (48003 including partially

occluded bounding boxes in the second version). After removing borderline cases (Figure 3), the PA intensity label (SED or LPA/ MVPA) was assigned to each of the remaining action tubes (~35100 in the initial version and 39600 in the second version with partially occluded samples). Then the data set was split into training and test sets such that the most thoroughly annotated and analyzed scene (REC1), and the last recording with twins (REC7) were assigned to the test set. Due to imbalance in our data set (~14% MVPA), many training samples had to be discarded to avoid overfitting to the lower PA intensity category. The train set also required subject-wise balancing to avoid overfitting to specific individuals (e.g., the model learning to associate MVPA with red color if a subject with a red shirt was overrepresented in the MVPA class). These steps resulted in a training set of 3253 (4330 in the second version) samples of sedentary behavior or light PA and an equal amount of MVPA samples such that each subject has an equal number of samples from each class. The test set remained unbalanced entailing 12326 samples with 14.0% MVPA in the original and 13619 samples with 15.1% MVPA in the second version. The training set was doubled to 13012 samples (17320 in the second version) by data augmentation. Each action tube was either brightened by 30% or dimmed by 30% and each sample had 50% chance to be flipped horizontally. For training the initial model, each tube was also rotated randomly from one to five degrees left or right. For the latter two models, the rotation augmentation is replaced by pseudo shortening and temporal padding of the action tubes: each sample had an equal chance of replacing one to ten frames from the end(s) of the action tube with blank frames to simulate shorter action tubes likely encountered in the wild. In case of an odd number of frame replacements, the extra replacement was performed at the tail end of the action tube.

Due to the complexities of developing custom ML training pipelines, Klavs Jermakovs fine-tuned the initial model using the 400-class output layer from pretraining. The first two nodes of the output were assigned to the classes relevant to MVPA classification, but loss and accuracy were measured from the whole 400-class head. This pseudo binary classification set up can cause below 50% accuracy in the first epochs of fine tuning (Figure 4 C) – at first the model is outputting higher probabilities for any of the 398 irrelevant output nodes and only later learns to reduce loss by focusing on the two relevant nodes. The first fine-tuned model (Figure 4 B) learned from a data set of near perfect action tubes which is also reflected in the learning curve. As can be seen, test accuracy surpassed 80% already by the end of the first epoch while training accuracy was barely above 50%. This could be explained by the test set not being balanced and including many samples of sedentary behavior which are much easier to differentiate from MVPA than LPA is (i.e., first learning to associate sedentary position and/or presence of a chair with the lower PA intensity category and later learning to differentiate slow walking from brisk walking which look more similar than sitting looks compared to MVPA). To test whether including the partially occluded bounding boxes and simulating shorter tubes can fix the issues of the first model, Mateus Reis trained the next model with the second version of the data set and using the same 400-class head (Figure 4 C). This time it took several training epochs before the model even started associating the relevant

output nodes with the task at hand. After the fourth epoch, the model started learning with the test accuracy again increasing faster, but overall learning slower than the initial model trained on a smaller but simpler version of the data set. For the final model (Figure 4 D), Mateus Reis added a binary output layer after the 400-class layer to achieve a proper binary classification model. The same tendency appears where test set performance increases faster. The final model also reveals a potential pattern of overfitting to training data after the test set accuracy levels off around 87% accuracy. Further metrics on test set performance of the action tube classifiers are provided in the results section along with preliminary performance assessment of the whole automatic PA observation pipeline.

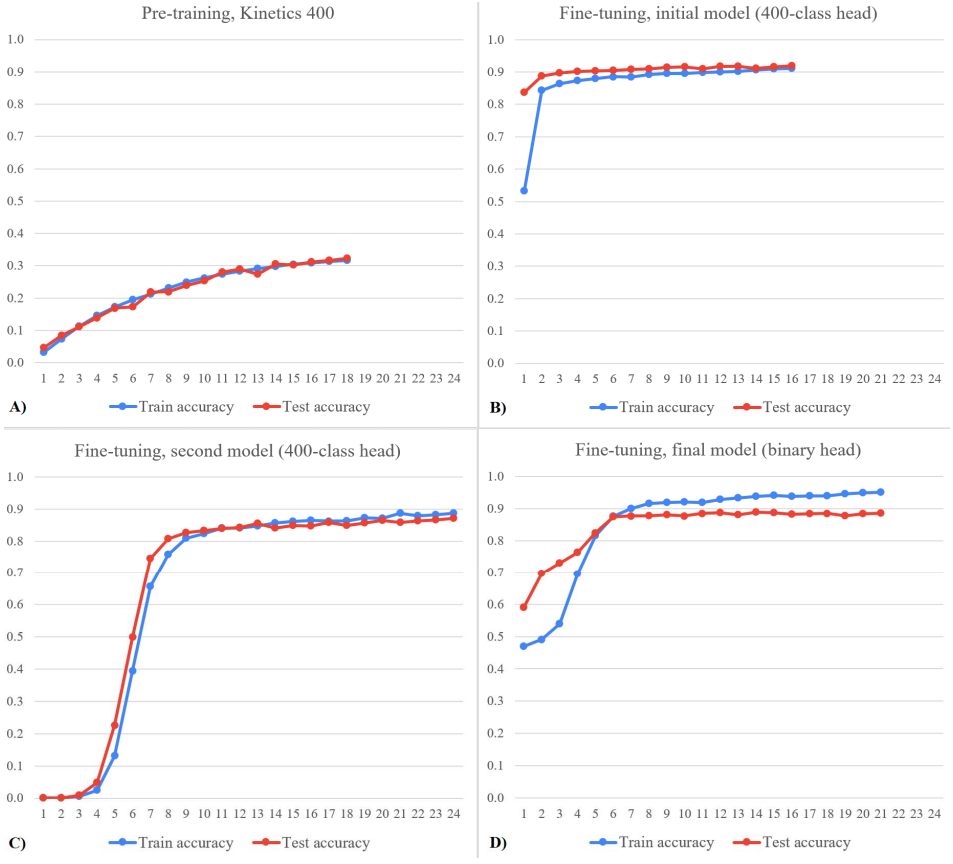

Figure 4. Pre-training custom X3D on Kinetics 400 (A); fine-tuning on our initial data set without partially occluded bounding boxes or temporal padding augmentations and using the 400-class output layer (B); fine-tuning on the second version of the data set including partially occluded samples, temporal padding augmentations, and a 400-class output layer (C); fine-tuning the final correct model including partially occluded samples, temporal padding augmentations, and using a binary output layer (D). Top-1 accuracy of the X3D model on 400-class action recognition task (A), pseudo binary MVPA classification task (B and C), and binary MVPA classification task (D).

3.5 Implementation of the sensor prototype

Mateus Reis deployed the video analysis processing pipeline on a Nvidia Jetson Xavier NX development board[23] (**Study III**). It has a Nvidia Volta GPU, a 6-core ARM CPU, two NVDLA deep learning accelerators, and 8GB of memory. The computational cost of analyzing one 2-second sample of video depends on the number of people visible. The more people are detected and tracked concurrently (K in Figure 5) the more action tubes need to be classified. To maximize the number of possible concurrent detections while maintaining real-time processing speed, the neural networks (the person detector and the MVPA classifier) were optimized using TensorRT[24].

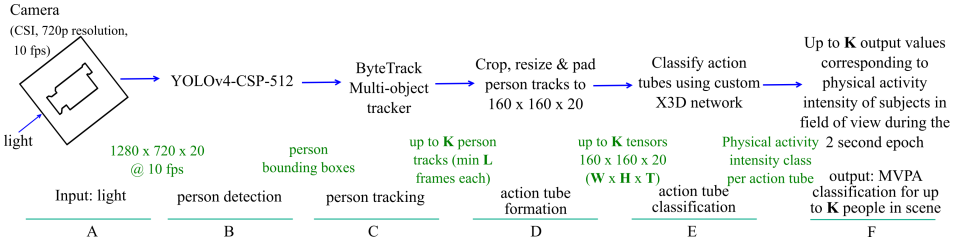

Figure 5. Video analysis sensor processing pipeline.

We assessed the sensor's performance while setting the minimum length of a viable action tube (L in Figure 5) to 10 and 15 frames. The former managed to capture more of the visible behavior and the latter caused a drop in overall detections.

Due to earlier frugal decisions (selecting a low resolution and frame rate for video and the smallest version of X3D), Mateus Reis achieved real-time processing capacity without excessive parallelization of the various processing steps while using just python code. This implies that cheaper hardware could suffice for the current real-time automatic PA observation pipeline if a faster language was used, and the computational resources were exploited optimally.

Figure 6. Prototype on a tripod

[23] https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-xavier-nx/

[24] Nvidia's framework for optimizing ML models for fast processing on Nvidia hardware.

3.6 Limitations, ethics, and reflections

A *privacy by design* approach in this case led to the mistake of early optimization due to overestimating the computational cost of the video analysis task and underestimating the power of available hardware. This leads to an educational moment of really appreciating Donald Knuth's famous quote: *"Premature optimization is the root of all evil"* (Knuth, 1974, p. 268). Even though the original quote refers to wasting labor in writing programs, the general sentiment seems to apply. The fear of not achieving real-time processing necessary for privacy preservation led to some poor decisions. A higher video resolution and frame rate could have increased the long-term value of the data set since the processing power of GPUs keeps growing. Super-resolution techniques (Anwar et al., 2020) could be used to artificially increase the resolution, but this could introduce unpredictable bias to the data when combined with real high-resolution recordings. Also, a larger, more capable X3D model could have been used if the work started out by getting a good understanding of the available computational resources and the computational cost of different sized networks when deployed on this hardware.

The semi-automatic annotation method should have been clearly defined and chosen at the beginning of the project and the same person should have done all the manual parts of annotation. In the current form, the annotations are not very consistent across the recordings. The objective of capturing natural behavior and free play lead to large imbalance in PA intensity categories (only ~14% MVPA) which causes waste of laboriously annotated data since the training set must be class-wise balanced. This should have been a predictable issue and measures should have been taken to avoid waste: leaving some sections of low activity without annotation and/or artificially inducing more intense movements during data collection. Another issue is low variation in the skin tone of the participants – one should not assume that a ML model trained on data from a majority white country could generalize very well to more varied conditions. Intuitively, PA intensity classification performance with 3D CNNs should be relatively robust to skin tone, but this is a risky assumption on the inner workings of a black box model. A globally representative video data set should cover more skin tones, school uniforms, and traditional dress.

The processing pipeline in its current form is not end-to-end[25] trainable. This makes it hard to evaluate its performance – we cannot compute one clear indicator to reflect the performance of the whole automatic PA observation system. Even if the test set was fully annotated[26] to allow validating an end-to-end PA detection method, the current approach is incompatible as it relies on removing borderline cases. A 100% annotated end-to-end approach excludes the possibility of

[25] An end-to-end system would mean a ML task where the loss function covers the whole PA measurement pipeline from the 1280×720×20 video sample as the input layer to the action tube coordinates and PA intensity class as output. This would require a loss function with at least two parts: one measuring distance from the action tube coordinates of the sample and the other for the PA intensity class.

[26] This would require all people in all scenes to wear accelerometers and have valid PA intensity labels.

increasing the model's robustness by removing the low confidence cases. For further development of such methods, a regression approach might be preferable since capturing the correct MVPA threshold via binary classification does not provide much value if the goal is to detect the increase and drop of school-level PA intensity over long periods. Treating PA as a continuous phenomenon (very low PA…very intense PA) might not allow to interpret sensor output in terms of MVPA but it would also eliminate the problems arising from setting such a threshold – the category between slow and brisk walking is a real PA intensity level and can be treated as such in regression.

Concerning research ethics, this work likely suffers from similar issues as any research involving underage children. Even informing the parents of the data subjects required certain effort – explaining the difference between research data for studying the data subjects and ML training data for developing video analysis applications. It would be too much to expect from an 8-year-old to comprehend the purpose of collecting this video to really be able to informedly consent. Even though the consent form was worded in a way that expressed intent of potentially publishing the data (Appendix II), this could be problematic for reasons covered in Chapter 5.2. Ethical issues concerning privacy in the potential application of blind observation methods are discussed in detail in Chapter 5.1.

4 RESULTS

Testing the prototype while using the initial fine-tuned X3D model and different-sized and differently optimized YOLO person detectors implies that there is significant computational overhead. With a relatively large person detection model [YOLOv4-CSP-512 (C.-Y. Wang et al., 2021)] at FP16 precision, a maximum of 15 people were detected, tracked, and classified within one 2-second epoch[27]. This performance was achieved without parallelization of the various processes and without excessive code optimization. This implies that using a newer generation device and optimizing code thoroughly, a larger X3D model should be viable and at a higher video resolution possibly even on a cheaper device. This in turn shows the maturity of hardware and algorithms necessary for real-time video analysis on edge devices – visual information on PA can be analyzed efficiently without recording the video and without humans having access to the video.

Testing the initial model also revealed a potentially critical concern: when people walked past each other, ByteTrack was surprisingly good at tracking the correct person, but at the moment of passing each other, the predicted probability of MVPA increased for both (demo video 0:16-0:38). This was likely caused by excluding partially occluded bounding box annotations from the data set and possibly by not simulating short action tubes with temporal padding during training. If the model is trained and tested on "perfect" action tubes without people (partially) occluding each other, the momentary occlusion could increase variance along the action tube during inference which the 3D-convolutional model could associate with the higher PA intensity class (the more the body is moving, the more the background is also relatively moving thereby associating variance along the time dimension with higher PA intensity). The effect of not learning on temporally padded tubes but running inference on temporally padded tubes is harder to predict – in the padded case, there is no variance in the beginning and end of the action tube, but a sharp crossover from an image to blank frame(s). In hopes of fixing this, the second model was trained on a version of the data set including the partially occluded bounding boxes and simulated shorter, temporally padded action tubes. To test whether this bias could be fixed by data set design, the bias was quantified by repeatedly walking and running past a cardboard cutout and seeing how much its predicted probability of MVPA changed compared to the cutout standing alone. Figure 7 shows the comparison of the three models in this test. Using the original model, the cardboard cutout's predicted probability of

[27] Due to the way the sensor operation was visualized during filming of the demo video, computational resources of the prototype were constrained. In the video, the sensor uses a smaller person detection model quantized to INT8 precision to maintain real-time processing, hence the poor person detection. Later testing by connecting a monitor directly to the prototype during a crowded event, major computational overhead appeared showing great potential for the processing pipeline. Demo: https://www.youtube.com/watch?v=RQHw2Z22pWc&ab_channel=KEKA

MVPA increased on average by 0.21 when walking or running past it while using the directly comparable second model, it increased by 0.11 – the bias was reduced by a half. The final binary model reduced this effect further (average difference 0.05 probability points). This means that proper data set design and augmentation techniques allow reducing occlusion-induced PA overestimation biases. The final model however consistently predicted a 0.5 chance of MVPA for the cutout alone – standing up perfectly still appears to be the MVPA threshold while it should be slightly below brisk walking. This could be specific to the body pose or appearance of this cardboard cutout since standing still in different upright positions showed lower probabilities as well. The probability of MVPA drops significantly when sitting or squatting and increases starkly in brisk walking, so the model still appears to measure PA intensity. When running, the predicted probability of MVPA is close to 1.

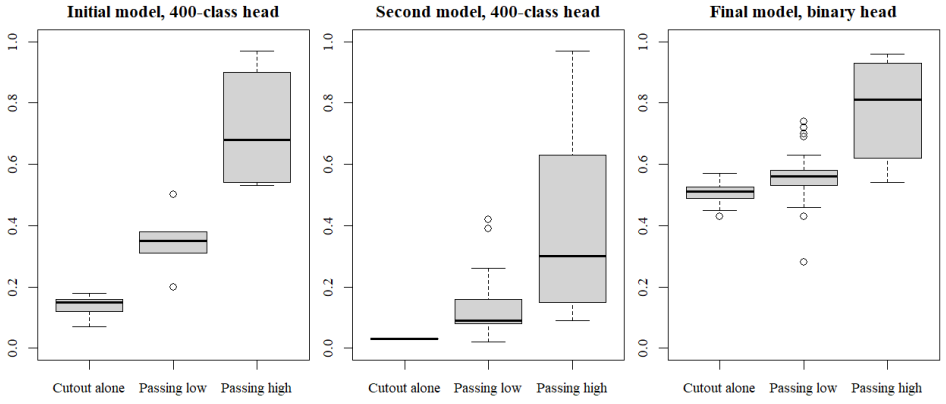

Figure 7. PA intensity overestimation bias in situations where people walk past each other. "Cutout alone" represents the predicted probability of MVPA of a cardboard cutout alone in view. "Passing low" represents the predicted probability of MVPA for the cutout in situations where I walk or run past (both front and behind) the cutout in the middle of an epoch. "Passing high" represents the predicted probability of MVPA for me in these passing epochs. A perfectly unbiased model would show equal means for the cutout alone and during passing.

Different versions of the X3D classifier (Figure 5 E) achieved good performance on the test set overall but had more confusion with MVPA (Table 5). Large share of the test set entails sedentary behavior which is under the same category as standing and light activity in this data set. The high accuracy on sedentary and light intensity cases contrasting with mediocre performance on MVPA could point to the model being near perfectly accurate at differentiating sedentary behaviors from any standing and walking but properly discriminating light activity from moderate requires more/better data. Applying different accelerometer thresholds for smaller and larger subjects could possibly improve performance, but research-grade validity and reliability likely require a larger and more varied data set. The models trained on the second, more difficult version of the data set reach

performance on par with the initial model while reducing the occlusion-induced bias towards MVPA. This reflects the general power of 3D convolutional neural networks and highlights the importance of properly annotating partially occluded samples[28].

Table 5. X3D action tube MVPA classification performance on the test set.

	Precision	Recall	F1-score	n samples
Initial model: no occluded samples, 400-class head				
SED or LPA	0.95	0.96	0.95	10594
MVPA	0.74	0.68	0.71	1732
Macro average	0.84	0.82	0.83	12326
Weighted average	0.92	0.92	0.92	12326
Second model: occluded samples, 400-class head				
SED or LPA	0.93	0.96	0.95	11565
MVPA	0.75	0.59	0.66	2054
Macro average	0.84	0.78	0.80	13619
Weighted average	0.90	0.91	0.90	13619
Final model: occluded samples, binary head				
SED or LPA	0.94	0.96	0.95	11565
MVPA	0.76	0.67	0.71	2054
Macro average	0.85	0.82	0.83	13619
Weighted average	0.92	0.92	0.92	13619

Because the video analysis processing pipeline is not end-to-end trainable, and our test set is only partially annotated, the PA observation performance has not been rigorously validated. To proximally estimate the performance of the whole system, we refer to the multi-object tracking accuracy reported for ByteTrack (Y. Zhang, 2021/2022; Y. Zhang et al., 2022) on person tracking data sets (0.80 and 0.78 on MOT17 and MOT20 data sets respectively). This performance metric is measured from detecting and tracking people (Figure 5 B and C) in rather crowded data sets and may not perfectly reflect the application setting of the sensors. Still, our best estimate for the performance of the whole pipeline is to multiply the macro F1-score of our PA intensity classifier (assuming that both PA intensity classes are equally important) with the multi-object tracking accuracy of ByteTrack arriving at a ball-park performance estimate of ~0.66 for the prototype. This estimate assumes that the PA intensity labels assigned based on accelerometers and 2D pose-estimated hip angle changes correspond to the true

[28] Ideally the same person detection model and tracking algorithm should be used for semi-automatic annotation of the data set as is used in the eventual automatic observation pipeline. The manually annotated bounding boxes for partially occluded cases should imitate the person detector and tracker – if only the upper part of the body is visible, one should still annotate the bounding box for the whole body area if that is how the detector and tracker behave.

MVPA construct. We can assume that some activities with high energy expenditure but low kinetic movement (e.g., slow push-ups or squats) are considered as low intensity because accelerometers cannot capture the energy expenditure of such activities either. The method should capture aerobic PA well, and relatively low prevalence of strength exercise behaviors can be assumed in primary school hallways.

Five days of the first blind observation data are presented on Figure 8.

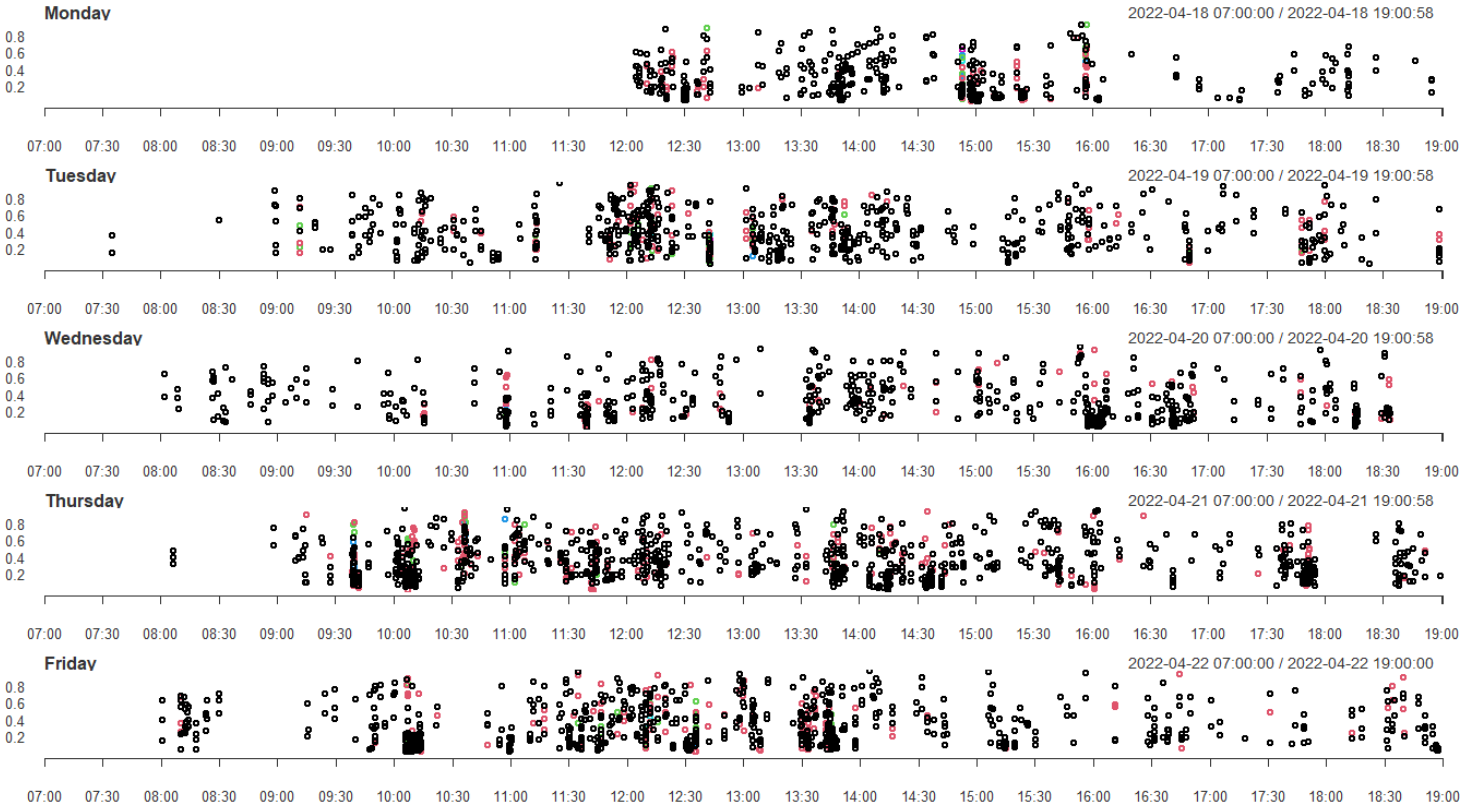

Figure 8. Five days of observations from the prototype at the University of Tartu Delta center 2nd floor atrium in front of the iCV lab. Mon-Fri 7:00–19:00 (missing data Monday 07:00–~12:00). The *y*-axis shows the predicted probability of MVPA for the detected person (the initial X3D model trained without partially occluded samples). Different colors for concurrent detections (black = first detected person).

4.1 Computing indicators from sensor output

This section introduces some indicators and approaches to analyze automatic observational data.

A set of sensors $S = \{s_1, s_2, \dots, s_k\}$ start recording at time t_1. Let $d_{ij} \in D_j = \{d_{1j}, d_{2j}, \dots, d_{nj}\}$ be the PA intensity classification (the predicted probability of MVPA) for the i'th detected person during the j'th 2-second observation epoch where n_j denotes the number of concurrent detections. Then $D_j \in T_{sm} = \{D_1, D_2, \dots, D_m\}$ is the sensor output at epoch j within the data set T_{sm} representing the observation period $\Delta t_m = [t_1, t_m]$ of duration $m \cdot 2$ seconds at sensor $s \in S$. Total PA intensity observed at epoch j, x_j, is the sum of the elements of D_j when at least one person is detected. When no people are detected, total PA intensity is set to zero.

$$x_j = \begin{cases} \sum_{i=1}^{n_j} d_{ij}, & \text{if } n_j > 0 \\ 0, & \text{if } n_j = 0 \end{cases} \tag{1}$$

$$x_j \in U_m \quad \text{where} \quad |U_m| = m \tag{2}$$

Combined with the number of concurrent detections n_j, x_j can be used for spatial data visualization on the floor plan of the observation area where the size of the circle reflects the number of people present and a color scale defined by the range of x_j represents group-level PA intensity. Creating a heatmap or animating these visualizations across all sensors over a period (whether it is a school day, semester, or year) can reveal spatio-temporal clusters of activity and sedentariness. Average PA intensity observed at epoch j, $\bar{x}_j$, is computed only for epochs where at least one person is detected. Epochs without people are removed, leaving a set of observations V_m of size v_m.

$$\overline{x}_j = \begin{cases} \frac{x_j}{n_j}, & \text{if } n_j > 0 \\ none, & \text{if } n_j = 0 \end{cases} \tag{3}$$

$$\overline{x}_j \in V_m \quad \text{where} \quad |V_m| = |\{n_j > 0\}| = v_m \tag{4}$$

Two main indicators can be computed for periods of interest. Average total activity intensity, $\overline{U}_m$, can be a very small value when analyzing a period with very few people present (e.g., over night) and requires contextual information for meaningful interpretation: comparison with other sensors in the surveilled area while considering the spatial distribution of sensors, the size of their perceptive fields and the temporal dynamics of the number of people present inside the building – controlling for student density estimates as proposed in **Study I**.

$$\overline{U}_m = \frac{\sum_{i=1}^{m} x_j}{m} \tag{5}$$

Average activity intensity when people are present, $\overline{V}_m$, is easier to interpret without much knowledge of the temporal dynamics of crowdedness and can be used in experimental studies with only a few sensors.

$$\overline{V}_m = \frac{\sum_{i=1}^{v_m} \overline{x}_j}{v_m} \tag{6}$$

All sensors could be plotted in a two-dimensional space of these indicators (Figure 8) and the movement of these data points throughout a longer observation period (e.g., before, during, and after intervention) can reveal changes in PA intensity and crowding at the sensor locations.

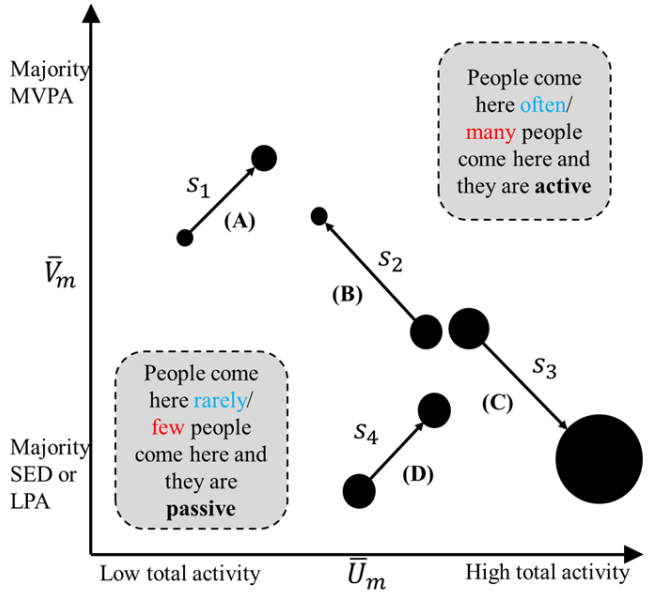

Figure 9. Hypothetical comparison of sensor locations in a schoolhouse pre- and post-intervention (beginning and end of the arrows) based on average total PA intensity ($\overline{U}_m$) and average PA intensity when people are present ($\overline{V}_m$). Size of the spheres indicates the average number of concurrent detections n_j for the period as an indicator of crowding.

Figure 9 A could represent a location with low crowding and above-average PA intensity. Compared to the first period (beginning of the arrow), average PA intensity at sensor s_1 increased and more people started coming to this area, potentially indicating successful intervention. Figure 9 B could represent a situation where average PA intensity at the location increased but people started going there less. Figure 9 C could depict a location where average PA levels dropped due to stark

increase in crowding (total activity increased from many additional detections with low predicted probability of MVPA). In contrast, Figure 9 D depicts a location where both indicators increased but without change in crowding/ visiting frequency – people just started moving more intensely at this location.

Hypotheses could be posed around these indicators and intervention. Covering two schools of similar size/architecture and population, one could be assigned as control and the other as test. Then hypotheses could be posed that e.g., banning smartphones or playing dance music in the test school will lead to increases in these indicators in the test school while they would not change in the control school. The sensors could also be used in smaller research projects where only one or a few locations are covered with sensors to test various location-specific intervention stimuli (e.g., removing benches or introducing stationary PA equipment).

5 DISCUSSION

This thesis demonstrates the viability of automatically observing PA intensity in indoor settings at real-time processing speeds on a 15-watt device. Technical solutions can be built on top to ensure data security – protecting the device from hacking to ensure near zero probability of any human seeing any of the video frames produced by the camera of the device. Based on the European Union definition, the technological readiness level (TRL) of the sensor technology is at a minimum of 3 – experimental proof of concept. Formally claiming TRL 4 would likely require either validating the method against well-trained human observers and/or measuring a clear error metric on a fully annotated validation data set. Even if humans could be more accurate at classifying a given 2-second sequence of bodily movement, superior performance of the sensor can be assumed confidently when considering long observation periods (basic human limitations and unstable subjectivity). Given a large, varied, and well annotated data set, ML models should outperform humans in validity of MVPA classification. Humans cannot compete at all when it comes to the possible number of concurrent observations nor the frequency of these observations. This implies the technical viability of applying blind observation to research with human subjects. However, while the enabling technologies are mature, there could be significant obstacles to the adoption of such methods for developing the evidence base for school-based PA intervention. The following discussion sections tackle ethics and trust of the general public and the research community in developing and adopting blind observation.

5.1 Privacy ethics in blind observation

Even though video analysis can be conducted without people having access to the video, assessing the (research) ethics of such data collection is not trivial. A naïve person noticing a device with a camera lens in a school hallway could assume a security guard has access to the video. If a sign was set up claiming "Nobody can see the video, no video is recorded, and no people are identified", the person would have to trust the truthfulness of the sign. In this section, I try to predict public acceptance of the blind observation method and evaluate its privacy ethics.

A popular approach to assessing the privacy ethics of data processing is Helen Nissenbaum's framework of privacy as *contextual integrity* (Nissenbaum, 2004; Nissenbaum, 2009). By this approach, the justice and ethics of data collection, processing, and transmission can be judged based on adherence to informational norms applicable to the context. Nissenbaum (2004) postulates that all information flows in all contexts are governed by norms of appropriateness (what kind of information about persons is appropriate to reveal in a specific context) and norms of distribution (to which parties is it appropriate to transfer the information). If the collection and communication of a datum adheres to relevant contextual norms, then contextual integrity is preserved, and the data flow can be deemed

privacy-wise ethical. As this framework implies a seemingly straightforward approach to assessing the ethics of all information flows, it has become popular in privacy research and the development of information systems (Badillo-Urquiola et al., 2018). However, when assessing a novel data generation technology, pinpointing the relevant contextual norms may not be easy. When dealing with truly novel information flows, the existence of relevant contextual norms appears questionable since the wider society has not been exposed to the phenomenon long enough to develop a new norm or to collectively agree on the applicability of existing norms. Which norms should govern the appropriateness of light reflecting from a student in a school hallway into the camera of a real-time video analysis sensor? Are there any norms relevant to automatic classification of pixels into PA intensity categories? Which norms should govern the flow of the ML model's output to researchers and different actors in school? If human observation requires informed consent to be deemed ethical, could the consent requirement be circumvented by removing the human?

Some have opted to discover contextual norms by asking people about their attitudes toward the collection and communication of various types of information in various imagined contexts (Apthorpe et al., 2018). While it would be convenient to develop research ethics guidelines based on such opinion surveys, this approach could lead to formal ethics based on the majority opinion in the survey sample fixed in time with the risky assumption that imagined context is a sufficient proxy for real context (tyranny of a momentary majority opinion pertaining to imaginary context at some point in the past). Rule (2019) argues that many norms are constantly contested, and when considering privacy-related norms in a rapidly evolving information society, they may be especially volatile. If the blind observation method was proposed before the use of excessive facial identification in China (Leibold, 2020), public attitudes towards it could have been different than now when people have acknowledged the real dystopian potential of computer vision technologies. If a total facial recognition ban was adopted by the European Union, Europeans' attitudes towards automatic observation methods might change in the opposite direction, reassured that identification of the students by the sensors is highly unlikely as it is illegal.

A more nuanced approach to probing contextual norms would be to inquire about attitudes towards collecting and communicating specific kinds of data while in the context of interest. Shikun Zhang and colleagues (2021) studied privacy preferences concerning various video analytics applications in real world settings via a mobile phone-based research design. When prompted, respondents preferred to deny data collection for generic video surveillance 33% and for anonymous people counting video analytics 49% of the times. While these differences were not statistically significant in their regression model, such a discrepancy could indicate certain reluctance to novel uses of video analysis in principle since there do not seem to be any legitimate privacy reasons to deny anonymous people counting more than standard surveillance recording pixels of faces. Across all scenarios, they did not find a statistically significant difference in allowing/denying data collection whether the video was kept for 30 days or deleted immediately. This could

also be an issue of statistical power as most scenarios stated 30-day retention of raw video footage (personal communication with first author). If, however this was a real phenomenon (people being relatively indifferent to the retention of raw video), then privacy preservation via real-time processing might not be as crucial to public acceptance of the method as assumed. Their study did indicate that generally the purpose of data collection was the most important factor in respondents' decisions to allow or deny data collection. Assuming most people value health promotion among the youth, this result could speak in favor of public acceptance of blind observation of PA in schools.

To further gauge potential public reactions to privacy-preserving video analysis, one might look at analogies or precedents. While not a perfect comparison, the transition from physical- to e-mail could be viewed as an analogy of automating observational methods in a privacy-preserving manner. In traditional mailing services, people have had to trust a whole chain of postal workers to not read their private letters. If a seal is broken when receiving the letter, relevant parties could suspect privacy violation by a postal worker and may take it up with the postal service. Instead of passing private letters through the hands of postal workers, sending an e-mail is automatic. The sensitive content of e-mails from criminals, spies, and politicians are automatically analyzed to detect spam. As a digital process, the recipient might not be able to even detect whether a human has accessed the content of the e-mail (i.e., a broken seal), yet there does not seem to be much panic around this issue. Tokson argues that processing of e-mails by spam classifiers is not deemed privacy-wise problematic because "*…our conception of a loss of privacy is bound up with the presence of a human observer"* (Tokson, 2010, p. 611). His conclusion – that exposing personal information to automated processing does not by itself constitute a privacy intrusion – is itself a claim on the existence of a social norm deduced in part from attitude surveys. Judging the acceptability of the automatic observation method based on the apparent acceptance of e-mail spam filtering could be problematic for several reasons. In the same work, Tokson brings the example of security cameras influencing people's behavior precisely because of an implied likelihood of human observation (Tokson, 2010, p. 614). The whole purpose of imaging technologies has been to create images for people to see, so even the claims of automatic processing and privacy preservation may be hard to accept. Furthermore, when considering the specific context – children in school – the presence of a camera could evoke associations to pedophilia which is a known subject of moral panic (Critcher, 2017). Even if most people would agree that automatic processing in general does not constitute a privacy violation, the type of information being processed in a context, and the history of similar technologies could induce suspicions. These problems could possibly be addressed by certification by a trustworthy institution such as a research ethics committee and/or a data protection agency. If a data security audit would result in official privacy preservation certification of the sensor and the deployment team, part of public concerns could be relieved. Even then it is not guaranteed that everyone will trust the privacy preservation capacity of the sensor or the purported goals of data collection.

Calo (2010) describes two types of privacy harms. Objective privacy harms are external to the victim and involve the forced or unanticipated use of personal information. Importantly, Calo (2010, p. 18) specifies personal information as information about a person (e.g., John Smith was engaged in MVPA at 08:45:32 at sensor #1) not as personally identifiable data (e.g., the video pixels where John's face is visible). Light falling on the image sensor can surely be unanticipated, but this thesis demonstrates that a device can be built that generally[29] does not cause objective privacy harm because the automatic observations cannot be associated with persons/identities. Subjective privacy harms however are outside of the researcher's control. These are internal to the victim and result from the unwanted perception of observation. If one correctly perceives a camera, they might also incorrectly perceive observation which could be unwanted and uncomfortable whether or not any actual (human) observation occurs. Combining certification with reassuring notice signs could reduce the risk of subjective privacy harms, but this may not suffice for particularly paranoid individuals (e.g., someone who might believe in the existence of mass surveillance microchips in vaccines). Such people could generally have a hard time coping with the abundance of sensors and data processing in the age of pervasive computing. One could hope that in the primary school population, such personality traits have yet to develop/mature and subjective privacy harm should be relatively rare and mild.

To conclude, provided that the sensors are certified by relevant authorities and reasonable measures are taken to minimize risks of subjective privacy harm, informed consent should not be necessary for the application of the proposed method. Given it is a location-based method, and no individual measures are required[30], obtaining informed consent itself presents an unnecessary privacy risk. The researchers doing the informing would have met the otherwise anonymous data subjects and there would be a stack of signed consent forms with the names of data subjects and their parent(s). If others should arrive at a different conclusion – that informed consent is necessary, even for privacy-preserving video analysis – then the method would be impractical since sensor deployment would require a 100% consent rate (including when a new student or employee joins the school during the research project).

[29] Objective privacy harm could occur when based on other information, the user knows that only one specific person was in the observation area at that time. Therefore, the school should only have access to aggregated data for relevant periods to rule out spying after employees.

[30] This is assuming a research design where researchers do not need to meet the students at all (e.g., testing the PA reactions to a school-wide smartphone ban). If researchers want to measure anthropometrics and conduct surveys before and during/after intervention, informed consent is required anyway.

5.2 Researchers' trust in method

I have argued that computer vision-based blind observation can provide unique advantages as a method for measuring PA (Table 2). Clearly human observers cannot compete with electronic devices when it comes to the possible duration of continuous observation. But humans could have an advantage when it comes to the primary researcher trusting the observational data and the research community trusting the results deduced from such data. The mind of a human observer and a deep convolutional neural network could both be seen as black boxes which map visual information to PA intensity categories. The processes of peeking inside these black boxes and deducing whether the output can be trusted could be markedly different. The formal practices are similar: testing observer(s) performance on example data and computing statistics to reflect reliability and validity[31]. But there could be a subjective difference.

Hidalgo and colleagues (2021) compared how people judge machines and humans in equivalent simulated scenarios. Their experimentation led to the conclusion that we judge humans more by their intentions and machines by their outcomes. It is unlikely that researchers would recruit observers whom they suspect of ill intentions. They would prefer to hire someone who leaves an impression that they really do intend to maximize data quality. Any mistakes or imprecision of such human observers could be more readily forgiven than for a video analysis sensor which is not ascribed intention, and which is judged purely by performance. Judging from the validation study of SOFIT (Rowe et al., 1997), even the formal approach to assessment of human observation might not be as rigorous as one would expect from ML experimentation. Validation of SOFIT (Rowe et al., 1997) excludes slow walking (technically in the same category as walking with moderate speed in this observation instrument) from analysis, yet the paper exhibits confidence in its conclusion that SOFIT categories are valid. The same blind spot to slow walking is present in a further study validating SOFIT for high-school students (Rowe et al., 2004). Pope and colleagues (2002) validated a modified version of SOFIT including slow walking/light PA as a separate category and realized that it fits the theoretical PA scale much better[32], but judging

[31] An important difference is that you can use the same test data only once for assessing the performance of a human observer since they could learn the data. Assessment of human observers relies more heavily on reliability analyses in form of inter-rater agreement when several observers are observing the same scene. Relatively high validity is likely assumed based on the original validation study of the observation method/PA categories and by trusting the vision and ability of the observer [e.g., the "What you see is what you get" assumption for assuming high internal validity for direct observation methods in McKenzie & van der Mars (2015)].

[32] Referring to the results of a three-year national trial (Luepker et al., 1996; McKenzie et al., 1996) where MVPA during physical education classes assessed by SOFIT increased significantly but without significant change in physiological variables. If majority of the increase in "MVPA" comes from light intensity movement, then one should not expect great reduction in BMI or increase in aerobic fitness.

from inter-rater agreement, differentiating between light and moderate intensity activities proved rather difficult for the observers. Provided high-quality training data, a ML model could likely be more accurate at differentiating slow walking from brisk walking. We can also run more thorough analysis on a ML model, but we should not expect researchers to take its imprecisions as lightly as for human observers performing the equivalent task.

Another aspect of trust in blind observation concerns the availability of the training- and test data. It would be harder to trust the black box model used in automatic observation if the research community would not have access to the training data. If the method was further developed by a company, then releasing the training data set would immediately devalue the company as everyone could train their own model on the data set and deploy it on their own hardware. On the other hand, when the development of the data set and the automatic observation method was conducted as open science, researchers would more readily trust the method, but the data set would have to be collected under conditions that allow open access publishing. Developing an open access data set of video depicting underage children could be difficult due to legal- and research ethics considerations. While parents informed consent should legally suffice to collect research data on children, publishing video open access without blurring faces might be deemed unethical: the data would already be freely circulating by the time the child obtains active legal capacity and cognitive ability to really understand the purposes of data collection and use. The European General Data Protection Regulation (Regulation 2016/679/EC) also gives data subjects the right to be forgotten – if a subject should at some point decide they do not wish to be in the data set, removing them from a freely circulating benchmark data set cannot be guaranteed. As such, collecting such ML benchmark video data sets in Europe seems inherently risky. Perhaps arguments could be made to not consider such video data as sensitive research data *per se* so that the full extent of data protection would not apply. Indeed, the purpose of such data is not to study the people in the data set, but to train and evaluate ML models. The data would represent PA intensities for movement sequences. Provided that age, height, and weight would already be accounted for during label assignment, these personal data would not have to be released together with the video data. However, the skin tone and general appearance of the data subjects would be clearly visible, which might be considered a sensitive personal data category of race. It appears that developing blind observation methods for measuring the behavior of underage children is currently viable only in secrecy while the research field and public health could benefit more from an open science approach. When it comes to visual information, the European Union's goals of open science and personal data protection are bound to conflict. A research ethics committee might however weight public (health) interest against the rigidity of data protection regulations and ethics of underage consent. If researchers really saw potential in developing a large and varied video dataset of PA, perhaps collecting and publishing such a data set could be excused. When considering the difficulty of measuring PA comparably in the school-age population (Figure 1) and the confusion in interpreting accelero-

meters (Migueles et al., 2019), perhaps a large visual data set could bring some much-needed clarity to the research field – everyone could look at the acceleration values and the visual appearance of some movement sequences and collectively decide how brisk a walk should really be to constitute MVPA. After all, society deserves to benefit from new technologies and data processing capacities. Rather than categorically blocking access to visual information, perhaps data governance should focus more on disincentivizing and punishing harmful uses of visual data.

6 CONCLUSIONS

Computer vision algorithms and available low-power hardware allow estimating PA in the field of view of a camera via privacy-preserving real-time video analysis. The approach developed in this thesis can be susceptible to a bias of overestimating PA intensity in situations where subjects walk past one-another. This bias, however, can be reduced or possibly eliminated via careful data set design and augmentation during training. A fully annotated data set and an end-to-end trainable processing pipeline could be preferable for future development as it would allow to evaluate the method more clearly. In such a case, a continuous PA intensity scale and a regression approach could be preferable since removing borderline cases in a classification approach is not viable when using a fully annotated data set and an end-to-end trainable PA observation pipeline. Depending on the intended use of the sensors, an MVPA classification approach could be preferable (when the user is interested in the prevalence of MVPA at a location) but for measuring the effect of school-based PA intervention inside the school building, a regression approach should suffice. The PA research field could benefit most from developing such a data set and method if it was done as open science in international collaboration. This assumes some prior clarifications concerning human research ethics and data protection law.

Answers to the research question are provided below:

RQ1: How to model physical activity intensity of children in video data?

To allow comparability with other PA research, the most widely used accelerometers are recommended for assigning PA intensity labels for children in video. Even though wrist placement is preferable in typical research due to better compliance, a hip placement should be preferred for creating video data sets as it is closer to the body's center of mass. However, using both hip and wrist placement when creating such data sets could provide additional value if not so much for the ML applications, then more for interpreting wrist acceleration data in general. It is worth investing effort into precision synchronization of accelerometers and video to enhance data quality (some accelerometer time drift should be expected and not underestimated especially in longer data collection sessions). Data from twins with different body mass could possibly be used to design a data set that represents the correct MVPA threshold for children with different body types: shifting the accelerometer MVPA threshold for heavier and lighter children in the data set and observing model performance on the twins with different body mass until they are classified equally well. **Study II** showed some promise for using 2D pose estimation techniques for computing PA intensity indicators directly from existing video data. The author urges the research community to try out 3D monocular pose estimation methods for modeling PA intensity. If viable, these techniques could allow to translate existing video data to PA intensity detection training data.

RQ2: Which video analysis approach can provide a stable physical activity signal?

A fixed video analysis epoch sets the constraints for obtaining a stable PA signal: the sensor must provide automatic observations at a fixed frequency (two seconds appears more reliable than one based on **Study II**). Subjects may step into and out of the scene at any point during each epoch and this requires setting a minimum acceptable length for continuous detections (Figure 1 in **Study I**) and applying temporal padding for action tubes shorter than the epoch length. As the third main consideration for selecting the video analysis approach, the method should take advantage of the fact that a fixed amount of kinetic energy is produced in the muscle tissue of each person during each time interval: viewing PA as a property of any behavior as opposed to treating PA as a behavior. This allows using a person detection and tracking-based approach instead of directly detecting lower or higher PA intensity class from video (finding all people in an epoch and classifying their PA intensity as opposed to finding people engaged in SED or LPA and MVPA). This property also allows to use shifted resampling techniques to multiply the size of the data set. Based on these considerations it is demonstrated that a stable PA signal can be obtained using a spatio-temporal action localization approach based on multiple person tracking with a fixed epoch length (untrimmed but equal sized video spatiotemporal action localization task) combined with temporal padding of the action tubes feeding into a 3D convolutional neural network as the PA intensity classifier.

RQ3: How does automatic blind observation compare to human observation methods on a fundamental level?

Compared to human observation, the proposed method allows stable subjectivity, objective privacy preservation, continuous stable observation at high sampling frequencies, and enhanced comparability of observational data between studies (different researchers can use the same observer with the same stable biases). Compared to human observation, a computer vision pattern recognition system may not be able to differentiate humans from human-looking objects (e.g., a poster depicting a human). As opposed to human observation, in its current form, the method cannot provide intelligent inferences on interactions of subjects or their reactions to specific intervention stimuli. ML-based automatic observation systems can be assessed more rigorously than human observers, but researchers might be more critical towards them than the human equivalent.

RQ4: How viable is the proposed method for research practice?

The algorithms and the hardware for privacy-preserving real-time video analysis applications have achieved maturity and will likely become cheaper in the future. The current state of the data set does not allow clear measurement of the performance of the whole video analysis processing pipeline. But since the occlusion-induced bias towards MVPA can be demonstrably fixed by training on partially

occluded action tubes, there is no reason to assume less than human-level performance. However, a larger and more varied training data set is required to achieve a PA detection model which could be trusted to be generalizable. There could be significant research ethics and data protection hurdles related to open access publishing of the training data set and application of the privacy-preserving automatic observation sensors. If the issues described in Chapters 5.1 and 5.2 can be overcome, the blind observation approach could potentially be applied to measuring other types of behavior as well.

RQ5: How to use such sensors in physical activity intervention research?

Ideally one should find at least two schools of similar architecture and population where the sensors should be deployed with a similar distribution (similar size of the average field of view and a similar distance between sensors). Then one school can be treated as control and the other as test where an intervention of interest is applied. Scientific knowledge on the efficacy and sustainment of intervention can be gained by comparing the automatic PA observation indicators (Chapter 4.1) for the period leading up to intervention, the period during active intervention, and period(s) after the active part of intervention. Smaller experimental studies could be designed using only a few sensors to evaluate specific intervention stimuli.

REFERENCES

Abu-Omar, K., Rütten, A., Burlacu, I., Schätzlein, V., Messing, S., & Suhrcke, M. (2017). The cost-effectiveness of physical activity interventions: A systematic review of reviews. *Preventive Medicine Reports*, *8*, 72–78. https://doi.org/10.1016/j.pmedr.2017.08.006

Afshin A, Forouzanfar M. H., Reitsma, M.B., et al, and the GBD 2015 Obesity Collaborators. (2017) Health effects of overweight and obesity in 195 countries over 25 years. *New England journal of medicine, 377*(1), 13–27. https://doi.org/10.1056/NEJMoa1614362

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Anwar, S., Khan, S., & Barnes, N. (2020). A deep journey into super-resolution: A survey. *ACM Computing Surveys*, *53*(3), 60:1–60:34. https://doi.org/10.1145/3390462

Apthorpe, N., Shvartzshnaider, Y., Mathur, A., Reisman, D., & Feamster, N. (2018). Discovering smart home internet of things privacy norms using contextual integrity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(2), 59:1–59:23. https://doi.org/10.1145/3214262

Arem, H., Moore, S. C., Patel, A., Hartge, P., Berrington de Gonzalez, A., Visvanathan, K., Campbell, P. T., Freedman, M., Weiderpass, E., Adami, H. O., Linet, M. S., Lee, I.-M., & Matthews, C. E. (2015). Leisure time physical activity and mortality: A detailed pooled analysis of the dose-response relationship. *JAMA Internal Medicine*, *175*(6), 959–967. https://doi.org/10.1001/jamainternmed.2015.0533

Aru, J., & Rozgonjuk, D. (2022). The effect of smartphone use on mental effort, learning, and creativity. *Trends in Cognitive Sciences*, *26*(10), 821–823. https://doi.org/10.1016/j.tics.2022.07.002

Badillo-Urquiola, K., Page, X., & Wisniewski, P. (2018). Literature review: Examining contextual integrity within human-computer interaction. *Available at SSRN 3309331*. https://doi.org/10.2139/ssrn.3309331

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman/Times Books/Henry Holt & Co.

Bandura, A. (2004). Health promotion by social cognitive means. *Health Education & Behavior*, *31*(2), 143–164. https://doi.org/10.1177/1090198104263660

Batorova, D., & Sørensen, J. (2019). Methodological review of model-based cost-effectiveness analyses of school-based interventions to increase pupils' level of physical activity. *Journal of Physical Education*, *30*. https://doi.org/10.4025/jphyseduc.v30i1.3013

Baum, F., & Fisher, M. (2014). Why behavioural health promotion endures despite its failure to reduce health inequities. *Sociology of Health & Illness*, *36*(2), 213–225. https://doi.org/10.1111/1467-9566.12112

Bergmann, P., Meinhardt, T., & Leal-Taixe, L. (2019). Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 941–951).

Biddle, S. J. H., Ciaccioni, S., Thomas, G., & Vergeer, I. (2019). Physical activity and mental health in children and adolescents: An updated review of reviews and an analysis of causality. *Psychology of Sport and Exercise*, *42*, 146–155. https://doi.org/10.1016/j.psychsport.2018.08.011

Blair, S. N., Kohl, H. W., Gordon, N. F., & Paffenbarger, R. J. (1992). How much physical activity is good for health. *Annual Review of Public Health*, *13*, 99–126. https://doi.org/10.1146/annurev.pu.13.050192.000531

Bochkovskiy, A. (2023). *Yolo v4, v3 and v2 for Windows and Linux* [C]. [Accessed January 2022] https://github.com/AlexeyAB/darknet (Original work published 2016)

Bonnet, C. T., & Cheval, B. (2022). Sitting vs. standing: An urgent need to rebalance our world. *Health Psychology Review*, 1–48. https://doi.org/10.1080/17437199.2022.2150673

Booth, F. W., Roberts, C. K., Thyfault, J. P., Ruegsegger, G. N., & Toedebusch, R. G. (2017). Role of inactivity in chronic diseases: Evolutionary insight and pathophysiological mechanisms. *Physiological Reviews*, *97*(4), 1351–1402. https://doi.org/10.1152/physrev.00019.2016

Bottà, G., Binelli, G., Agostoni, C., Aliverti, A., Scarì, G., Manenti, R., & La Vecchia, C. (2020). Evaluating human basal metabolism: The erroneous and misleading use of so-called "prediction equations." *International Journal of Food Sciences and Nutrition*, *71*(2), 249–255. https://doi.org/10.1080/09637486.2019.1641472

Brey, P., & Søraker, J. H. (2009). Philosophy of computing and information technology. In A. Meijers (Ed.), *Philosophy of Technology and Engineering Sciences* (pp. 1341–1407). North-Holland. https://doi.org/10.1016/B978-0-444-51667-1.50051-3

Bull, F. C., Al-Ansari, S. S., Biddle, S., Borodulin, K., Buman, M. P., Cardon, G., Carty, C., Chaput, J.-P., Chastin, S., Chou, R., Dempsey, P. C., DiPietro, L., Ekelund, U., Firth, J., Friedenreich, C. M., Garcia, L., Gichu, M., Jago, R., Katzmarzyk, P. T., … Willumsen, J. F. (2020). World Health Organization 2020 guidelines on physical activity and sedentary behaviour. *British Journal of Sports Medicine*, *54*(24), 1451–1462. https://doi.org/10.1136/bjsports-2020-102955

Burhan, R., & Moradzadeh, J. (2020). Neurotransmitter dopamine (DA) and its role in the development of social media addiction. *Journal of Neurology & Neurophysiology, 11*(7), 1–2.

Butte, N. F., Watson, K. B., Ridley, K., Zakeri, I. F., McMurray, R. G., Pfeiffer, K. A., Crouter, S. E., Herrmann, S. D., Bassett, D. R., Long, A., Berhane, Z., Trost, S. G., Ainsworth, B. E., Berrigan, D., & Fulton, J. E. (2018). A youth compendium of physical activities: Activity codes and metabolic intensities. *Medicine and Science in Sports and Exercise*, *50*(2), 246–256. https://doi.org/10.1249/MSS.0000000000001430

Calo, R. (2010). The boundaries of privacy harm. *Available at SSRN 1641487*. https://papers.ssrn.com/abstract=1641487

Carlson, J. A., Liu, B., Sallis, J. F., Hipp, J. A., Staggs, V. S., Kerr, J., Papa, A., Dean, K., & Vasconcelos, N. M. (2020). Automated high-frequency observations of physical activity using computer vision. *Medicine and Science in Sports and Exercise,* 52(9), 2029. https://doi.org/10.1249/mss.0000000000002341

Carlson, J. A., Liu, B., Sallis, J. F., Kerr, J., Hipp, J. A., Staggs, V. S., Papa, A., Dean, K., & Vasconcelos, N. M. (2017). Automated ecological assessment of physical activity: Advancing direct observation. *International Journal of Environmental Research and Public Health*, *14*(12), 1487. https://doi.org/10.3390/ijerph14121487

Caspersen, C. J., Powell, K. E., & Christenson, G. M. (1985). Physical activity, exercise, and physical fitness: Definitions and distinctions for health-related research. *Public Health Reports*, *100*(2), 126–131.

Cavoukian, A. (2009). Privacy by design: The 7 foundational principles. *Information and privacy commissioner of Ontario, Canada,* 5, 2009.

Cheval, B., & Boisgontier, M. P. (2021). The theory of effort minimization in physical activity. *Exercise and Sport Sciences Reviews*, *49*(3), 168–178. https://doi.org/10.1249/JES.0000000000000252

Cheval, B., Tipura, E., Burra, N., Frossard, J., Chanal, J., Orsholits, D., Radel, R., & Boisgontier, M. P. (2018). Avoiding sedentary behaviors requires more cortical resources than avoiding physical activity: An EEG study. *Neuropsychologia*, *119*, 68–80. https://doi.org/10.1016/j.neuropsychologia.2018.07.029

Colabianchi, N., Griffin, J. L., Slater, S. J., O'Malley, P. M., & Johnston, L. D. (2015). The whole-of-school approach to physical activity: Findings from a national sample of U.S. secondary students. *American Journal of Preventive Medicine*, *49*(3), 387–394. https://doi.org/10.1016/j.amepre.2015.02.012

Conroy, D. E., & Berry, T. R. (2017). Automatic affective evaluations of physical activity. *Exercise and Sport Sciences Reviews*, *45*(4), 230–237. https://doi.org/10.1249/JES.0000000000000120

Critcher, C. (2017, March 29). Moral Panics. In *Oxford Research Encyclopedia of Criminology and Criminal Justice*. https://doi.org/10.1093/acrefore/9780190264079.013.155

Deci, E. L., & Ryan, R. M. (1985). Conceptualizations of intrinsic motivation and self-determination. In *Intrinsic Motivation and Self-Determination in Human Behavior* (pp. 11–40). *Perspectives in Social Psychology*. Springer. https://doi.org/10.1007/978-1-4899-2271-7_2

Ding, D., Lawson, K. D., Kolbe-Alexander, T. L., Finkelstein, E. A., Katzmarzyk, P. T., van Mechelen, W., Pratt, M., & Lancet Physical Activity Series 2 Executive Committee. (2016). The economic burden of physical inactivity: A global analysis of major non-communicable diseases. *The Lancet*, *388*(10051), 1311–1324. https://doi.org/10.1016/S0140-6736(16)30383-X

Ding, D., Varela, A. R., Bauman, A. E., Ekelund, U., Lee, I.-M., Heath, G., Katzmarzyk, P. T., Reis, R., & Pratt, M. (2020). Towards better evidence-informed global action: Lessons learnt from the Lancet series and recent developments in physical activity and public health. *British Journal of Sports Medicine*, *54*(8), 462–468. https://doi.org/10.1136/bjsports-2019-101001

Dooris, M., Poland, B., Kolbe, L., de Leeuw, E., McCall, D. S., & Wharf-Higgins, J. (2007). Healthy settings. In D. V. McQueen & C. M. Jones (Eds.), *Global Perspectives on Health Promotion Effectiveness* (pp. 327–352). Springer. https://doi.org/10.1007/978-0-387-70974-1_19

Drewnowski, A., Mennella, J. A., Johnson, S. L., & Bellisle, F. (2012). Sweetness and food preference. *The Journal of Nutrition*, *142*(6), 1142S–1148S. https://doi.org/10.3945/jn.111.149575

Duffey, K., Barbosa, A., Whiting, S., Mendes, R., Yordi Aguirre, I., Tcymbal, A., Abu-Omar, K., Gelius, P., & Breda, J. (2021). Barriers and facilitators of physical activity participation in adolescent girls: A systematic review of systematic reviews. *Frontiers in Public Health*, *9*, 743935. https://doi.org/10.3389/fpubh.2021.743935

Edwards, L. C., Bryant, A. S., Keegan, R. J., Morgan, K., & Jones, A. M. (2017). Definitions, foundations and associations of physical literacy: A systematic review. *Sports Medicine*, *47*(1), 113–126. https://doi.org/10.1007/s40279-016-0560-7

Ekelund, U., Tarp, J., Steene-Johannessen, J., Hansen, B. H., Jefferis, B., Fagerland, M. W., Whincup, P., Diaz, K. M., Hooker, S. P., Chernofsky, A., Larson, M. G., Spartano, N., Vasan, R. S., Dohrn, I.-M., Hagströmer, M., Edwardson, C., Yates, T., Shiroma, E., Anderssen, S. A., & Lee, I.-M. (2019). Dose-response associations between accelerometry measured physical activity and sedentary time and all cause mortality: Systematic review and harmonised meta-analysis. *BMJ*, *366*, l4570. https://doi.org/10.1136/bmj.l4570

Ekkekakis, P. (2017). People have feelings! Exercise psychology in paradigmatic transition. *Current Opinion in Psychology*, *16*, 84–88. https://doi.org/10.1016/j.copsyc.2017.03.018

Fairclough, S. J., Noonan, R., Rowlands, A. V., Van, V. H., Knowles, Z., & Boddy, L. M. (2016). Wear compliance and activity in children wearing wrist- and hip-mounted accelerometers. *Medicine and Science in Sports and Exercise*, *48*(2), Article 2. https://doi.org/10.1249/MSS.0000000000000771

Fan, H., Murrell, T., Wang, H., Alwala, K. V., Li, Y., Li, Y., Xiong, B., Ravi, N., Li, M., Yang, H., Malik, J., Girshick, R., Feiszli, M., Adcock, A., Lo, W.-Y., & Feichtenhofer, C. (2021). PyTorchVideo: A deep learning library for video understanding. *Proceedings of the 29th ACM International Conference on Multimedia*, 3783–3786. https://doi.org/10.1145/3474085.3478329

Fang, H.-S., Shuqin, X., Tai, Y.-W., Xie, S., Lu, C., and contributors (2021). *AlphaPose* [Python]. [Accessed September 2021] https://github.com/MVIG-SJTU/AlphaPose (Original work published 2018)

Feichtenhofer, C. (2020, June). X3D: Expanding Architectures for Efficient Video Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 200–210). IEEE. https://doi.org/10.1109/CVPR42600.2020.00028

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior*. Addison-Wesley, 1975.

Freedson, P., Pober, D., & Janz, K. F. (2005). Calibration of accelerometer output for children. *Medicine and Science in Sports and Exercise*, *37*(11), S523. https://doi.org/10.1249/01.mss.0000185658.28284.ba

French, S. A., Story, M., & Jeffery, R. W. (2001). Environmental influences on eating and physical activity. *Annual Review of Public Health*, *22*, 309–335. https://doi.org/10.1146/annurev.publhealth.22.1.309

Global Advocacy for Physical Activity (GAPA) the Advocacy Council of the International Society for Physical Activity and Health (ISPAH) (2012). NCD Prevention: Investments that Work for Physical Activity. *British Journal of Sports Medicine*, *46*(10), 709–712. https://doi.org/10.1136/bjsm.2012.091485

Gourlan, M., Bernard, P., Bortolon, C., Romain, A. J., Lareyre, O., Carayol, M., Ninot, G., & Boiché, J. (2016). Efficacy of theory-based interventions to promote physical activity. A meta-analysis of randomised controlled trials. *Health Psychology Review*, *10*(1), 50–66. https://doi.org/10.1080/17437199.2014.981777

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377. https://doi.org/10.1016/j.patcog.2017.10.013

Guthold, R., Stevens, G. A., Riley, L. M., & Bull, F. C. (2018). Worldwide trends in insufficient physical activity from 2001 to 2016: A pooled analysis of 358 population-based surveys with 1.9 million participants. *The Lancet Global Health*, *6*(10), e1077–e1086. https://doi.org/10.1016/S2214-109X(18)30357-7

Harrell, J. S., Mcmurray, R. G., Baggett, C. D., Pennell, M. L., Pearce, P. F., & Bangdiwala, S. I. (2005). Energy costs of physical activities in children and adolescents. *Medicine & Science in Sports & Exercise*, *37*(2), 329–336. https://doi.org/10.1249/01.MSS.0000153115.33762.3F

Hidalgo, C. A., Orghian, D., Canals, J. A., Almeida, F. D., & Martin, N. (2021). *How Humans Judge Machines*. MIT Press.

Hildebrand, M., Van Hees, V. T., Hansen, B. H., & Ekelund, U. (2014). Age group comparability of raw accelerometer output from wrist- and hip-worn monitors. *Medicine & Science in Sports & Exercise*, *46*(9), 1816–1824. https://doi.org/10.1249/MSS.0000000000000289

Honas, J. J., Washburn, R. A., Smith, B. K., Greene, J. L., Cook-Wiens, G., & Donnelly, J. E. (2008). The System for Observing Fitness Instruction Time (SOFIT) as a measure of energy expenditure during classroom-based physical activity. *Pediatric Exercise Science*, *20*(4), 439–445.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Ivakhnenko, A. G. (1971). Polynomial Theory of Complex Systems. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-1*(4), 364–378. https://doi.org/10.1109/TSMC.1971.4308320

Janssen, I., & LeBlanc, A. G. (2010). Systematic review of the health benefits of physical activity and fitness in school-aged children and youth. *International Journal of Behavioral Nutrition and Physical Activity*, *7*(1), 40. https://doi.org/10.1186/1479-5868-7-40

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 221–231. https://doi.org/10.1109/TPAMI.2012.59

Jones, L. K., Jennings, B. M., Goelz, R. M., Haythorn, K. W., Zivot, J. B., & de Waal, F. B. M. (2016). An ethogram to quantify operating room behavior. *Annals of Behavioral Medicine*, *50*(4), 487–496. https://doi.org/10.1007/s12160-016-9773-0

Jones, M., Defever, E., Letsinger, A., Steele, J., & Mackintosh, K. A. (2020). A mixed-studies systematic review and meta-analysis of school-based interventions to promote physical activity and/or reduce sedentary time in children. *Journal of Sport and Health Science*, *9*(1), 3–17. https://doi.org/10.1016/j.jshs.2019.06.009

Kaas, J. H., & Balaram, P. (2014). Current research on the organization and function of the visual system in primates. *Eye and Brain*, *6*(Suppl 1), 1–4. https://doi.org/10.2147/EB.S64016

Katzmarzyk, P. T., Friedenreich, C., Shiroma, E. J., & Lee, I.-M. (2022). Physical inactivity and non-communicable disease burden in low-income, middle-income and high-income countries. *British Journal of Sports Medicine*, *56*(2), 101–106. https://doi.org/10.1136/bjsports-2020-103640

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*. https://doi.org/10.48550/arXiv.1705.06950

Kemmerer, D. (2021). What modulates the mirror neuron system during action observation?: Multiple factors involving the action, the actor, the observer, the relationship between actor and observer, and the context. *Progress in Neurobiology*, *205*, 102128. https://doi.org/10.1016/j.pneurobio.2021.102128

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys*, *54*(10s), 200:1–200:41. https://doi.org/10.1145/3505244

Knuth, D. E. (1974). Structured programming with go to statements. *ACM Computing Surveys*, *6*(4), 261–301. https://doi.org/10.1145/356635.356640

Kohl, H. W., Craig, C. L., Lambert, E. V., Inoue, S., Alkandari, J. R., Leetongin, G., & Kahlmeier, S. (2012). The pandemic of physical inactivity: Global action for public health. *The Lancet*, *380*(9838), 294–305. https://doi.org/10.1016/S0140-6736(12)60898-8

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*. https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Lai, S. K., Costigan, S. A., Morgan, P. J., Lubans, D. R., Stodden, D. F., Salmon, J., & Barnett, L. M. (2014). Do school-based interventions focusing on physical activity, fitness, or fundamental movement skill competency produce a sustained impact in these outcomes in children and adolescents? A systematic review of follow-up studies. *Sports Medicine*, *44*(1), 67–79. https://doi.org/10.1007/s40279-013-0099-9

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in Neural Information Processing Systems*, *2*. https://proceedings.neurips.cc/paper/1989/hash/53c3bce66e43be4f209556518c2fcb54-Abstract.html

Lee, H. H., Emerson, J. A., & Williams, D. M. (2016). The exercise–affect–adherence pathway: An evolutionary perspective. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01285

Lee, I.-M. (2007). Dose-response relation between physical activity and fitness: Even a little is good; more is better. *JAMA*, *297*(19), 2137–2139. https://doi.org/10.1001/jama.297.19.2137

Leibold, J. (2020). Surveillance in China's Xinjiang region: ethnic sorting, coercion, and inducement. *Journal of Contemporary China*, *29*(121), 46–60. https://doi.org/10.1080/10670564.2019.1621529

Lewis, B. A., Napolitano, M. A., Buman, M. P., Williams, D. M., & Nigg, C. R. (2017). Future directions in physical activity intervention research: Expanding our focus to sedentary behaviors, technology, and dissemination. *Journal of Behavioral Medicine*, *40*(1), 112–126. https://doi.org/10.1007/s10865-016-9797-8

Li, N. P., van Vugt, M., & Colarelli, S. M. (2018). The evolutionary mismatch hypothesis: Implications for psychological science. *Current Directions in Psychological Science*, *27*(1), 38–44. https://doi.org/10.1177/0963721417731378

Lieberman, D. E. (2015). Is exercise really medicine? An evolutionary perspective. *Current Sports Medicine Reports*, *14*(4), 313–319. https://doi.org/10.1249/JSR.0000000000000168

Lightfoot, J. T., De Geus, E. J. C., Booth, F. W., Bray, M. S., den Hoed, M., Kaprio, J., Kelly, S. A., Pomp, D., Saul, M. C., Thomis, M. A., Garland, T., & Bouchard, C. (2018). Biological/genetic regulation of physical activity level: Consensus from GenBioPAC. *Medicine and Science in Sports and Exercise*, *50*(4), 863–873. https://doi.org/10.1249/MSS.0000000000001499

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: common objects in context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 740–755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48

Lindström, B., Bellander, M., Schultner, D. T., Chang, A., Tobler, P. N., & Amodio, D. M. (2021). A computational reward learning account of social media engagement. *Nature Communications*, *12*(1), 1311. https://doi.org/10.1038/s41467-020-19607-x

Linnainmaa, S. (1976). Taylor expansion of the accumulated rounding error. *BIT Numerical Mathematics*, *16*(2), 146–160. https://doi.org/10.1007/BF01931367

Liu, W., Bao, Q., Sun, Y., & Mei, T. (2022). Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective. *ACM Computing Surveys*, *55*(4), 80:1–80:41. https://doi.org/10.1145/3524497

Love, R., Adams, J., & Sluijs, van E. M. F. (2019). Are school-based physical activity interventions effective and equitable? A meta-analysis of cluster randomized controlled trials with accelerometer-assessed activity. *Obesity Reviews*, *20*(6), 859–870. https://doi.org/10.1111/obr.12823

Luepker, R. V., Perry, C. L., McKinlay, S. M., Nader, P. R., Parcel, G. S., Stone, E. J., Webber, L. S., Elder, J. P., Feldman, H. A., Johnson, C. C., Kelder, S. H., Wu, M., Nader, P., Elder, J., McKenzie, T., Bachman, K., Broyles, S., Busch, E., Danna, S., … Verter, J. (1996). Outcomes of a field trial to improve children's dietary patterns and physical activity: The child and adolescent trial for cardiovascular health (CATCH). *JAMA*, *275*(10), 768–776. https://doi.org/10.1001/jama.1996.03530340032026

Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., & Kim, T.-K. (2021). Multiple object tracking: A literature review. *Artificial Intelligence*, *293*, 103448. https://doi.org/10.1016/j.artint.2020.103448

Lustig, R. H., Schmidt, L. A., & Brindis, C. D. (2012). The toxic truth about sugar. *Nature*, *482*, 27–29. https://doi.org/10.1038/482027a

Malinowski, B. (2017). *Argonauts of the western Pacific: An account of native enterprise and adventure in the archipelagoes of Melanesian New Guinea*. Routledge, London, 1922; Project Gutenberg. Retrieved November 24, 2022, from https://www.gutenberg.org/files/55822/55822-h/55822-h.htm. (Original work published 1922)

Maltagliati, S., Sarrazin, P., Fessler, L., Lebreton, M., & Cheval, B. (2022). Why people should run after positive affective experiences instead of health benefits. *Journal of Sport and Health Science*, S2095–2546. https://doi.org/10.1016/j.jshs.2022.10.005

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*, 370–396. https://doi.org/10.1037/h0054346

McCarney, R., Warner, J., Iliffe, S., van Haselen, R., Griffin, M., & Fisher, P. (2007). The Hawthorne Effect: a randomised, controlled trial. *BMC Medical Research Methodology*, *7*, 30. https://doi.org/10.1186/1471-2288-7-30

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

Mckenzie, T. (2015). *SOFIT (System for Observing Fitness Instruction Time) Description and Procedures Manual (Generic Version)*. https://doi.org/10.13140/RG.2.2.20282.70087

McKenzie, T. L. (2002). Use of direct observation to assess physical activity. In G. J. Welk (Ed.) *Physical Activity Assessments for Health-Related Research*, (pp. 179–195). Human Kinetics.

McKenzie, T. L., Nader, P. R., Strikmiller, P. K., Yang, M., Stone, E. J., Perry, C. L., Taylor, W. C., Epping, J. N., Feldman, H. A., Luepker, R. V., & Kelder, S. H. (1996). School physical education: Effect of the child and adolescent trial for cardiovascular health. *Preventive Medicine*, *25*(4), 423–431. https://doi.org/10.1006/pmed.1996.0074

McKenzie, T. L., Sallis, J. F., & Nader, P. R. (1992). SOFIT: System for observing fitness instruction time. *Journal of Teaching in Physical Education*, *11*(2), 195–205. https://doi.org/10.1123/jtpe.11.2.195

McKenzie, T. L., & van der Mars, H. (2015). Top 10 research questions related to assessing physical activity and its contexts using systematic observation. *Research Quarterly for Exercise and Sport*, *86*(1), 13–29. https://doi.org/10.1080/02701367.2015.991264

McMullen, J. M., Kallio, J., & Tammelin, T. H. (2022). Physical activity opportunities for secondary school students: International best practices for whole-of-school physical activity programs. *European Physical Education Review*, *28*(4), 890–905. https://doi.org/10.1177/1356336X221092281

McMurray, R. G., Butte, N. F., Crouter, S. E., Trost, S. G., Pfeiffer, K. A., Bassett, D. R., Puyau, M. R., Berrigan, D., Watson, K. B., Fulton, J. E., for the CDC/NCI/NCCOR Research Group on Energy Expenditure in Children (2015). Exploring metrics to express energy expenditure of physical activity in youth. *PloS one*, *10*(6), e0130869. https://doi.org/10.1371/journal.pone.0130869

Migueles, J. H., Cadenas-Sanchez, C., Ekelund, U., Delisle Nyström, C., Mora-Gonzalez, J., Löf, M., Labayen, I., Ruiz, J. R., & Ortega, F. B. (2017). Accelerometer data collection and processing criteria to assess physical activity and other outcomes: A systematic review and practical considerations. *Sports Medicine*, *47*(9), 1821–1845. https://doi.org/10.1007/s40279-017-0716-0

Migueles, J. H., Cadenas-Sanchez, C., Tudor-Locke, C., Löf, M., Esteban-Cornejo, I., Molina-Garcia, P., Mora-Gonzalez, J., Rodriguez-Ayllon, M., Garcia-Marmol, E., Ekelund, U., & Ortega, F. B. (2019). Comparability of published cut-points for the assessment of physical activity: Implications for data harmonization. *Scandinavian Journal of Medicine & Science in Sports*, *29*(4), 566–574. https://doi.org/10.1111/sms.13356

Moore, S. C., Lee, I.-M., Weiderpass, E., Campbell, P. T., Sampson, J. N., Kitahara, C. M., Keadle, S. K., Arem, H., Berrington de Gonzalez, A., Hartge, P., Adami, H.-O., Blair, C. K., Borch, K. B., Boyd, E., Check, D. P., Fournier, A., Freedman, N. D., Gunter, M., Johannson, M., … Patel, A. V. (2016). Association of leisure-time physical activity with risk of 26 types of cancer in 1.44 million adults. *JAMA Internal Medicine*, *176*(6), 816–825. https://doi.org/10.1001/jamainternmed.2016.1548

Mooses, K., Vihalemm, T., Uibu, M., Mägi, K., Korp, L., Kalma, M., Mäestu, E., & Kull, M. (2021). Developing a comprehensive school-based physical activity program with flexible design – from pilot to national program. *BMC Public Health*, *21*(1), 92. https://doi.org/10.1186/s12889-020-10111-x

Neil-Sztramko, S. E., Caldwell, H., & Dobbins, M. (2021). School-based physical activity programs for promoting physical activity and fitness in children and adolescents aged 6 to 18. *Cochrane Database of Systematic Reviews*, *9*. https://doi.org/10.1002/14651858.CD007651.pub3

Neishabouri, A., Nguyen, J., Samuelsson, J., Guthrie, T., Biggs, M., Wyatt, J., Cross, D., Karas, M., Migueles, J. H., Khan, S., & Guo, C. C. (2022). Quantification of acceleration as activity counts in ActiGraph wearable. *Scientific Reports*, *12*(1), 11958. https://doi.org/10.1038/s41598-022-16003-x

Nissenbaum, H. (2004). Privacy as Contextual Integrity. *Washington Law Review*, *79*(1), 119–158.

Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Ogidi, F. (2022). *A TensorFlow implementation of X3D* [Python]. [Accessed February 2022] https://github.com/fcogidi/X3D-tf

Oja, P., Kelly, P., Pedisic, Z., Titze, S., Bauman, A., Foster, C., Hamer, M., Hillsdon, M., & Stamatakis, E. (2017). Associations of specific types of sports and exercise with all-cause and cardiovascular-disease mortality: A cohort study of 80 306 British adults. *British Journal of Sports Medicine*, *51*(10), 812–817. https://doi.org/10.1136/bjsports-2016-096822

Parish, R. (1995). Health promotion Rhetoric and reality. In R. Bunton, R. Burrows, S. Nettleton (Eds.). *The Sociology of Health Promotion*. (pp. 11–21). Routledge.

Pate, R. R., Davis, M. G., Robinson, T. N., Stone, E. J., McKenzie, T. L., & Young, J. C. (2006). Promoting physical activity in children and youth: a leadership role for schools: a scientific statement from the American Heart Association Council on Nutrition, Physical Activity, and Metabolism (Physical Activity Committee) in collaboration with the Councils on Cardiovascular Disease in the Young and Cardiovascular Nursing. *Circulation*, *114*(11), 1214–1224. https://doi.org/10.1161/CIRCULATIONAHA.106.177052

Pate, R. R., Pratt, M., Blair, S. N., Haskell, W. L., Macera, C. A., Bouchard, C., Buchner, D., Ettinger, W., Heath, G. W., King, A. C., Kriska, A., Leon, A. S., Marcus, B. H., Morris, J., Paffenbarger, R. S., Jr, Patrick, K., Pollock, M. L., Rippe, J. M., Sallis, J., & Wilmore, J. H. (1995). Physical activity and public health: A recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. *JAMA*, *273*(5), 402–407. https://doi.org/10.1001/jama.1995.03520290054029

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pfattheicher, S., & Keller, J. (2015). The watching eyes phenomenon: The role of a sense of being seen and public self-awareness. *European Journal of Social Psychology*, *45*(5), 560–566. https://doi.org/10.1002/ejsp.2122

Physical Activity Guidelines Advisory Committee *(2008). Physical Activity Guidelines Advisory Committee Report, 2008.* Washington, DC: U.S. Department of Health and Human Services, 2008.

Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences*, *25*(2), 100–110. https://doi.org/10.1016/j.tics.2020.11.006

Platt, J. (1983). The development of the "participant observation" method in sociology: Origin myth and history. *Journal of the History of the Behavioral Sciences*, *19*(4), 379–393. https://doi.org/10.1002/1520-6696(198310)19:4<379::AID-JHBS2300190407>3.0.CO;2-5

Plotnikoff, R. C., Costigan, S. A., Karunamuni, N., & Lubans, D. R. (2013). Social cognitive theories used to explain physical activity behavior in adolescents: A systematic review and meta-analysis. *Preventive Medicine*, *56*(5), 245–253. https://doi.org/10.1016/j.ypmed.2013.01.013

Pope, R. P., Coleman, K. J., Gonzalez, E. C., Barron, F., & Heath, E. M. (2002). Validity of a revised system for observing fitness instruction time (SOFIT). *Pediatric Exercise Science*, *14*(2), 135–146. https://doi.org/10.1123/pes.14.2.135

Pulling Kuhn, A., Stoepker, P., Dauenhauer, B., & Carson, R. L. (2021). A systematic review of multi-component comprehensive school physical activity program (CSPAP) interventions. *American Journal of Health Promotion: AJHP*, *35*(8), 1129–1149. https://doi.org/10.1177/08901171211013281

Raiber, L., Christensen, R. A. G., Jamnik, V. K., & Kuk, J. L. (2017). Accelerometer thresholds: Accounting for body mass reduces discrepancies between measures of physical activity for individuals with overweight and obesity. *Applied Physiology, Nutrition, and Metabolism*, *42*(1), 53–58. https://doi.org/10.1139/apnm-2016-0303

Raiber, L., Christensen, R. A. G., Randhawa, A. K., Jamnik, V. K., & Kuk, J. L. (2019). Do moderate- to vigorous-intensity accelerometer count thresholds correspond to relative moderate- to vigorous-intensity physical activity? *Applied Physiology, Nutrition, and Metabolism*, *44*(4), 407–413. https://doi.org/10.1139/apnm-2017-0643

Ramirez, E., Kulinna, P. H., & Cothran, D. (2012). Constructs of physical activity behaviour in children: The usefulness of social cognitive theory. *Psychology of Sport and Exercise*, *13*(3), 303–310. https://doi.org/10.1016/j.psychsport.2011.11.007

Rebar, A. L., Dimmock, J. A., Jackson, B., Rhodes, R. E., Kates, A., Starling, J., & Vandelanotte, C. (2016). A systematic review of the effects of non-conscious regulatory processes in physical activity. *Health Psychology Review*, *10*(4), 395–407. https://doi.org/10.1080/17437199.2016.1183505

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788). https://doi.org/10.1109/CVPR.2016.91

Regulation 2016/679/EC. *General Data Protection Regulation.* European Parliament and Council. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Reis, R. S., Salvo, D., Ogilvie, D., Lambert, E. V., Goenka, S., & Brownson, R. C. (2016). Scaling up physical activity interventions worldwide: Stepping up to larger and smarter approaches to get people moving. *The Lancet*, *388*(10051), 1337–1348. https://doi.org/10.1016/S0140-6736(16)30728-0

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *28*. Available at: https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html

Rhodes, R. E., & Dickau, L. (2012). Experimental evidence for the intention–behavior relationship in the physical activity domain: a meta-analysis. *Health Psychology*, *31*(6), 724–727. https://doi.org/10.1037/a0027290

Rhodes, R. E., McEwan, D., & Rebar, A. L. (2019). Theories of physical activity behaviour change: A history and synthesis of approaches. *Psychology of Sport and Exercise*, *42*, 100–109. https://doi.org/10.1016/j.psychsport.2018.11.010

Ricciardelli, P., Baylis, G., & Driver, J. (2000). The positive and negative of human expertise in gaze perception. *Cognition*, *77*(1), B1–B14. https://doi.org/10.1016/S0010-0277(00)00092-5

Richer, J. (2017). Direct observation: impediments and approaches. *Human Ethology Bulletin*, *32*(4), 6–14. https://doi.org/10.22330/heb/324/006-014

Ridgers, N. D., Timperio, A., Cerin, E., & Salmon, J. (2014). Compensation of physical activity and sedentary time in primary school children. *Medicine and Science in Sports and Exercise*, *46*(8), 1564–1569. https://doi.org/10.1249/mss.0000000000000275

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408. https://doi.org/10.1037/h0042519

Rowe, P. J., Schuldheisz, J. M., & Mars, H. van der. (1997). Validation of SOFIT for measuring physical activity of first- to eighth-grade students. *Pediatric Exercise Science*, *9*(2), 136–149. https://doi.org/10.1123/pes.9.2.136

Rowe, P., Mars, H. van der, Schuldheisz, J., & Fox, S. (2004). Measuring students' physical activity levels: Validating SOFIT for use with high-school students. *Journal of Teaching in Physical Education*, *23*(3), 235–251. https://doi.org/10.1123/jtpe.23.3.235

Rule, J. B. (2019). Contextual integrity and its discontents: A critique of Helen Nissenbaum's normative arguments. *Policy & Internet*, *11*(3), 260–279. https://doi.org/10.1002/poi3.215

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), Article 6088. https://doi.org/10.1038/323533a0

Russell, S. J., Norvig, P., & Davis, E. (2010). *Artificial intelligence: a modern approach* (3rd ed). Prentice Hall.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Sahoo, K., Sahoo, B., Choudhury, A. K., Sofi, N. Y., Kumar, R., & Bhadoria, A. S. (2015). Childhood obesity: causes and consequences. *Journal of Family Medicine and Primary Care*, *4*(2), 187–192. https://doi.org/10.4103/2249-4863.154628

Saint-Maurice, P. F., Kim, Y., Welk, G. J., & Gaesser, G. A. (2016). Kids are not little adults: What MET threshold captures sedentary behavior in children? *European Journal of Applied Physiology*, *116*(1), 29–38. https://doi.org/10.1007/s00421-015-3238-1

Sallis, J. F., Cervero, R. B., Ascher, W., Henderson, K. A., Kraft, M. K., & Kerr, J. (2006). An ecological approach to creating active living communities. *Annual Review of Public Health*, *27*, 297–322. https://doi.org/10.1146/annurev.publhealth.27.021405.102100

Sallis, J. F., & Owen, N. (2015). Ecological models of health behavior. In *Health behavior: Theory, research, and practice, 5th ed* (pp. 43–64). Jossey-Bass/Wiley.

Sawyer, S. M., Afifi, R. A., Bearinger, L. H., Blakemore, S.-J., Dick, B., Ezeh, A. C., & Patton, G. C. (2012). Adolescence: A foundation for future health. *The Lancet*, *379* (9826), 1630–1640. https://doi.org/10.1016/S0140-6736(12)60072-5

Schaar, P. (2010). Privacy by design. *Identity in the Information Society*, *3*(2), 267–274. https://doi.org/10.1007/s12394-010-0055-x

Schofield, W. N. (1985). Predicting basal metabolic rate, new standards and review of previous work. *Human Nutrition Clinical Nutrition*, *39 Suppl 1*, 5–41.

Sekachev, B., Manovich, N., Zhiltsov, M., Zhavoronkov, A., Kalinin, D., Hoff, B., TOsmanov, Kruchinin, D., Zankevich, A., DmitriySidnev, Markelov, M., Johannes222, Chenuet, M., a-andre, telenachos, Melnikov, A., Kim, J., Ilouz, L., Glazov, N., … Truong, T. (2020). *opencv/cvat: V1.1.0*. Zenodo. https://doi.org/10.5281/zenodo.4009388

Shephard, R. J. (2001). Absolute versus relative intensity of physical activity in a dose-response context. *Medicine and Science in Sports and Exercise*, *33*(6 Suppl), S400-S418. https://doi.org/10.1097/00005768-200106001-00008

Silva, P., Santiago, C., Reis, L. P., Sousa, A., Mota, J., & Welk, G. (2015). Assessing physical activity intensity by video analysis. *Physiological Measurement*, *36*(5), 1037. https://doi.org/10.1088/0967-3334/36/5/1037

Simon, H. A. (1988). The science of design: creating the artificial. *Design Issues*, *4*(1/2), 67–82. https://doi.org/10.2307/1511391

Smith, N. J., McKenzie, T. L., & Hammons, A. J. (2018). International studies of physical education using SOFIT: a review. *Advances in Physical Education*, *9*(1), 53–74. https://doi.org/10.4236/ape.2019.91005

Spotswood, F., Vihalemm, T., Uibu, M., & Korp, L. (2021). Understanding whole school physical activity transition from a practice theory perspective. *Health Education*, *121*(5), 523–539. https://doi.org/10.1108/HE-04-2021-0066

Spotswood, F., Wiltshire, G., Spear, S., Morey, Y., & Harris, J. (2021). A practice theory approach to primary school physical activity: Opportunities and challenges for intervention. *Critical Public Health*, *31*(4), 392–403. https://doi.org/10.1080/09581596.2019.1695746

Staiano, A. E., Abraham, A. A., & Calvert, S. L. (2013). Adolescent exergame play for weight loss and psychosocial improvement: A controlled physical activity intervention. *Obesity*, *21*(3), 598–601. https://doi.org/10.1002/oby.20282

Strobach, T., Englert, C., Jekauc, D., & Pfeffer, I. (2020). Predicting adoption and maintenance of physical activity in the context of dual-process theories. *Performance Enhancement & Health*, *8*(1), 100162. https://doi.org/10.1016/j.peh.2020.100162

Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep High-Resolution Representation Learning for Human Pose Estimation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5686–5696. https://doi.org/10.1109/CVPR.2019.00584

Tang, M. Y., Smith, D. M., Mc Sharry, J., Hann, M., & French, D. P. (2019). Behavior change techniques associated with changes in postintervention and maintained changes in self-efficacy for physical activity: a systematic review with meta-analysis. *Annals of Behavioral Medicine*, *53*(9), 801–815. https://doi.org/10.1093/abm/kay090

Teixeira, P. J., Carraça, E. V., Markland, D., Silva, M. N., & Ryan, R. M. (2012). Exercise, physical activity, and self-determination theory: A systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, *9*(1), 78. https://doi.org/10.1186/1479-5868-9-78

Tokson, M. (2010). Automation and the Fourth Amendment. *Iowa Law Review*, *96*(2), 581–648.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 4489–4497). https://doi.org/10.1109/ICCV.2015.510

U.S. Department of Health and Human Services (2008). *2008 Physical Activity Guidelines for Americans: Be Active, Healthy, and Happy!* Available at: http://www.health.gov/Paguidelines/Guidelines/Default.Aspx.

Vahdani, E., & Tian, Y. (2022). Deep learning-based action detection in untrimmed videos: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20. https://doi.org/10.1109/TPAMI.2022.3193611

van Hees, V. T., Gorzelniak, L., León, E. C. D., Eder, M., Pias, M., Taherian, S., Ekelund, U., Renström, F., Franks, P. W., Horsch, A., & Brage, S. (2013). Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. *PloS one*, *8*(4), e61691. https://doi.org/10.1371/journal.pone.0061691

van Rompay, T. J. L., Vonk, D. J., & Fransen, M. L. (2009). The eye of the camera: effects of security cameras on prosocial behavior. *Environment and Behavior*, *41*(1), 60–74. https://doi.org/10.1177/0013916507309996

van Sluijs, E. M. F., Ekelund, U., Crochemore-Silva, I., Guthold, R., Ha, A., Lubans, D., Oyeyemi, A. L., Ding, D., & Katzmarzyk, P. T. (2021). Physical activity behaviours in adolescence: current evidence and opportunities for intervention. *The Lancet*, *398* (10298), 429–442. https://doi.org/10.1016/S0140-6736(21)01259-9

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. Available at: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2021, June). *Scaled-YOLOv4: scaling cross stage partial network*. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 13024–13033). *IEEE*. https://doi.org/10.1109/CVPR46437.2021.01283

Wang, Z., Emmerich, A., Pillon, N. J., Moore, T., Hemerich, D., Cornelis, M. C., Mazzaferro, E., Broos, S., Ahluwalia, T. S., Bartz, T. M., Bentley, A. R., Bielak, L. F., Chong, M., Chu, A. Y., Berry, D., Dorajoo, R., Dueker, N. D., Kasbohm, E., Feenstra, B., … Hoed, M. den. (2022). Genome-wide association analyses of physical activity and sedentary behavior provide insights into underlying mechanisms and roles in disease prevention. *Nature Genetics*, *54*(9), 1332–1344. https://doi.org/10.1038/s41588-022-01165-1

Warburton, D. E. R., Nicol, C. W., & Bredin, S. S. D. (2006). Health benefits of physical activity: the evidence. *CMAJ*, *174*(6), 801–809. https://doi.org/10.1503/cmaj.051351

Webster, C. A. (2022). The comprehensive school physical activity program: an invited review. *American Journal of Lifestyle Medicine*, 15598276221093544. https://doi.org/10.1177/15598276221093543

Welch, D. (2017). Behaviour change and theories of practice: Contributions, limitations and developments. *Social Business*, *7*(3–4), 241–261. https://doi.org/10.1362/204440817X15108539431488

Westbrook, A., Ghosh, A., van den Bosch, R., Määttä, J. I., Hofmans, L., & Cools, R. (2021). Striatal dopamine synthesis capacity reflects smartphone social activity. *IScience*, *24*(5), 102497. https://doi.org/10.1016/j.isci.2021.102497

Whitehead, M. E., Durden-Myers, E. J., & Pot, N. (2018). The value of fostering physical literacy. *Journal of Teaching in Physical Education*, *37*(3), 252–261. https://doi.org/10.1123/jtpe.2018-0139

Williams, S. L., & French, D. P. (2011). What are the most effective intervention techniques for changing physical activity self-efficacy and physical activity behaviour—And are they the same? *Health Education Research*, *26*(2), 308–322. https://doi.org/10.1093/her/cyr005

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. *2017 IEEE International Conference on Image Processing (ICIP)*, 3645–3649. https://doi.org/10.1109/ICIP.2017.8296962

World Health Organization. (1985). *Targets for Health for All 2000*. Available at: https://www.cabdirect.org/cabdirect/abstract/19862027131

World Health Organization. (2022). *The Global Status Report on Physical Activity 2022*. Available at: https://www.who.int/teams/health-promotion/physical-activity/global-status-report-on-physical-activity-2022

Yang, Y. (2020). *FastMOT: High-performance multiple object tracking based on Deep SORT and KLT*. Zenodo. https://doi.org/10.5281/zenodo.4294717

Zhang, S., Feng, Y., Bauer, L., Cranor, L. F., Das, A., & Sadeh, N. (2021). "Did you know this camera tracks your mood?": Understanding privacy expectations and preferences in the age of video analytics. *Proceedings on Privacy Enhancing Technologies*, *2021*(2). https://doi.org/10.2478/popets-2021-0028

Zhang, Y. (2022). *ByteTrack* [Python]. [Accessed January 2022] https://github.com/ifzhang/ByteTrack (Original work published 2021)

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., & Wang, X. (2022). ByteTrack: Multi-object tracking by associating every detection box. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Computer Vision – ECCV 2022* (pp. 1–21). Springer Nature. https://doi.org/10.1007/978-3-031-20047-2_1

Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2021). FairMOT: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, *129*(11), 3069–3087. https://doi.org/10.1007/s11263-021-01513-4

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

ANNEX I. SCENES FROM THE DATA SET

Figure 10. REC1 (top row, this is the data used for modelling the MVPA threshold and belongs to the test set), REC2 (second row left), REC3, REC4 (third row right) and REC5 (bottom row). REC 2-5 belong in the train set. Natural lighting in the Delta building atrium provides variable lighting conditions. First five sessions filmed with one camera.

Figure 11. REC6 A (top two rows) REC6 B (middle two rows) and REC 6 C. Filmed with two cameras. Train set.

Figure 12. REC7 A Twins with different body weight. Two cameras and two accelerometers each (left and right hip). Test set.

Figure 13. REC7 B. Test set.

ANNEX II. INFORMED CONSENT FORM

Lugupeetud lapsevanem!

Kutsun Teie last osalema kehalise aktiivsuse videoandmestiku filmimise seansis. Tegemist on masinõppe treeningandmestikuga, mille eesmärgiks on luua kehalise aktiivsuse nutisensor. Kavandatava algoritmiga peaks saama analüüsida kooliõpilaste liikumist kaamera vaateväljas ilma videot salvestamata. Sellise tehisintellekti eesmärgiks on mõõta kehalist aktiivsust koolis ilma lapsi tülitamata ega isikuid tuvastamata.

Andmekogumise käigus kinnitan laste puusale aktseleromeetri (kiirenduse mõõtja) ja filmin neid liikumas erineva kehalise aktiivsusega. Videole jäädvustatakse lapsi selliselt nagu nad võiksid välja näha kooli koridoris. Filmin neid vabas olekus, tegemas erinevaid võimlemisharjutusi ja mängimas mänge, et tehisintellekt saaks õppida erinevaid laste kehalise aktiivsuse väljendusi. Seanss kestab ligikaudu tund aega ning tegevused ei eelda rasket füüsilist pingutust. Mõõdan ka lapse kaalu ja pikkuse. Nime ega teisi andmeid ei salvestata, küll aga jääb kogutud videomaterjali lapse nägu ning muud silmaga nähtavad tunnused. Filmimise ajal võib kasutatav kaamera ka salvestada heli, aga selle eemaldan videomaterjalilt esimesel võimalusel ning kustutan heli sisaldavad failiversioonid. Andmestik on mõeldud tehisnärvivõrkude õpetamiseks, mitte filmitud laste uurimiseks.

Kogun videoandmeid 2021. aasta mai kuus Põltsamaa Ühisgümnaasiumi ruumides. Kutsun Teie last osalema ühes seansis, mille täpse aja ja koha kohta saate küsida täpsustusi minult kui andmete kogujalt.

Soovin kogutud andmestiku avaldada internetis avaandmetena. Sedasi oleks töömahukas andmestik kättesaadav kõigile masinõppest huvitatud tudengitele ja teadlastele üle maailma ilma, et nad peaksid ise lapsi filmima. Kuna andmestik ei ole ühekordseks kasutamiseks, vaid mõeldud muuseas ka õppevahendiks, ei saa määrata andmete säilitamise lõppkuupäeva (andmestiku eesmärgi täitumine ei ole selgelt piiritletav). Kui nõustute oma lapse osalemisega, peate arvestama, et kui teie või teie laps soovite nõusolekut tagasi võtta ja kasutada isikuandmete kaitse üldmäärusest tulenevat õigust olla unustatud, saame garanteerida teie lapse andmete kustutamise üksnes Tartu Ülikooli andmekogust.

Andmekogumist on kirjeldatud ka järgmisel veebilehel: icv.tuit.ut.ee/KEKA

Andmekogumises osaleva lapse vanema/seadusliku esindaja teadliku nõusoleku vorm

Mind (*lapsevanema/ seadusliku esindaja nimi*), ……..……..……

(*lapse nimi*) ……..... on informeeritud ülalmainitud andmekogumisest ja ma olen teadlik läbiviidava teadustöö eesmärgist ja andmekogumise metoodikast ja kinnitan oma nõusolekut selles osalemises allkirjaga.

Tean, et andmekogumises osalemine on lapsele vabatahtlik ning ta võib sellest igal hetkel loobuda.

Tean, et teadustöö käigus tekkivate küsimuste kohta saan vajalikku täiendavat informatsiooni andmete kogujalt.

Luban oma last kujutavat videomaterjali vabalt jagada ja publitseerida.

Filmitava informeerimise ja teadliku nõusoleku leht vormistatakse 2 eksemplaris, millest üks jääb osaleva lapse vanemale ja teine andmekogujale.

Küsimuste korral võtke palun ühendust teadustöö teostajatega:

Andmete peamine koguja, valdaja ja käitleja:
Hans Hõrak
Tartu Ülikool, doktorant
e-post: hans.horak@ut.ee
Tel +372 5331 6215
Narva mnt 18–3048, 51009, Tartu

Doktorandi juhendajad:
Triin Vihalemm
Tartu Ülikool, kommunikatsiooniuuringute professor
e-post: triin.vihalemm@ut.ee
Tel +372 525 7731

Gholamreza Anbarjafari
Tartu Ülikool, masinnägemise professor
e-post: gholamreza.anbarjafari@ut.ee
Tel +372 737 4855

Lapsevanemale andis infot (nimi, kuupäev, allkiri) ...Hans Hõrak 11.05.21….....

Lapsevanema/seadusliku esindaja allkiri: ……..……

Kuupäev, aasta…….……...

TRANSLATED

Dear parent,

I invite your child to participate in a physical activity video dataset filming session. It is a machine learning training dataset for developing a smart physical activity sensor. The proposed algorithm should be able to analyze the movement of school students in the camera's field of view without recording any video. The goal of this artificial intelligence is to measure physical activity at school without bothering the children or identifying any individuals.

During data collection, I will attach an accelerometer to the child's hip and film them moving with different levels of physical activity. The video aims to capture children as they might look in the school hallway. I film them in their natural state, doing various exercises, and playing games so that artificial intelligence can learn different expressions of children's physical activity. The session lasts approximately an hour, and the activities do not require strenuous physical effort. I will also measure the weight and height of the child. The name and other data are not recorded, but the face and other visual features of the child will be visible in the video. While filming, the camera can also record audio, but I will remove it from the footage as soon as possible and delete the versions with audio. The dataset is meant for training artificial neural networks, not for studying the children filmed.

I collect the video data at the premises of Põltsamaa Ühisgümnaasium in May 2021. I invite your child to participate in a single session, the exact time and place of which you can clarify with myself as the primary data collector.

I want to publish the collected data openly on the internet. In this way, a laborious dataset would be available to all students and researchers around the world interested in machine learning without having to film children themselves. Since the data set is not a one-off, but is also intended as a learning tool, the end date for storing the data cannot be determined (the fulfilment of the purpose of the data set can not be clearly defined). If you consent to your child's participation, you must consider that if you or your child wish to withdraw your consent and exercise the right to be forgotten under the General Data Protection Regulation, we can only guarantee that your child's data will be deleted from the University of Tartu database.

The data collection is also described on the following website: icv.tuit.ut.ee/KEKA

Form of informed consent of the parent/legal representative of the child participating in the data collection.

I (*name of parent/legal representative*), ……..…… (*child's name*) ……..... have been informed of the above-mentioned data collection and am aware of the purpose of the research being carried out and the methodology of the data collection and confirm my consent to this participation by signing.

I know that participating in the data collection is voluntary for the child and he or she can refuse at any moment.

I know that I will receive the necessary additional information from the data collector on the issues arising from the research.

I allow to freely share and publish video footage depicting my child.

The informed consent form shall be prepared in 2 copies, one of which shall be kept by the parent of the participating child and the other by the data collector.

If you have any questions, please contact the researchers:

The main collector, possessor, and handler of the data:
Hans Hõrak
University of Tartu, PhD student
e-post: hans.horak@ut.ee
Phone: +372 5331 6215
Narva mnt 18–3048 ,51009, Tartu

PhD supervisors:
Triin Vihalemm
University of Tartu, professor of communications studies
e-mail: triin.vihalemm@ut.ee
Phone: +372 525 7731

Gholamreza Anbarjafari
University of Tartu, professor of computer vision
e-mail: gholamreza.anbarjafari@ut.ee
Phone: +372 737 4855

Parent was informed by (name, date, signature) ...Hans Hõrak 11.05.21….....

Signature of parent/legal representative: ……

Date, year ……..…….…….

SUMMARY IN ESTONIAN

Privaatsust säilitava raalnägemise meetodi arendamine kehalise aktiivsuse automaatseks jälgimiseks koolis

Tänapäeval elab suur osa inimkonnast oludes, mis võimaldavad sissetulekut teenida ja meeldivalt vaba aega veeta ilma oma keha oluliselt liigutamata. Istuva eluviisi laialdane, lausa pandeemiline levik (Kohl et al., 2012) on viinud rahvatervise kriisini, kus 7,2% globaalsest üldsuremusest tuleneb ebapiisavast kehalisest aktiivsusest (Katzmarzyk et al., 2022). Liikumisaktiivsuse edendamiseks ühiskonna tasemel on üheks tähtsaks strateegiliseks sekkumiskohaks koolisüsteem. Koolipõhised kehalise aktiivsuse sekkumised võiksid valmistada inimesi ette tervislikumaks eluks istumist soosivas keskkonnas (GAPA, 2012; Sawyer et al., 2012). Võitlemaks vähese kehalise aktiivsuse pandeemiaga on tähtis tõenduspõhiselt (Lewis et al., 2017) välja selgitada parimad praktikad kehalise aktiivsuse edendamiseks, et siis neid sekkumismeetmeid võimalikult kiiresti ja laialt rakendada (Reis et al., 2016; Ding et al., 2020). See omakorda eeldab uuringuid, kus sekkumismeetmete proovimisel mõõdetakse ka mõjutatavate kehalist aktiivsust.

Käesolevas doktoritöös arendati privaatsust säilitavat videoanalüüsipõhist kehalise aktiivsuse mõõtmise meetodit, mida saaks kasutada koolipõhiste kehalise aktiivsuse sekkumiste tõhususe hindamiseks. Kehalist aktiivsust saab mõõta sammulugejate ja aktseleromeetritega, aga sedalaadi indiviidipõhised meetodid on privaatsust riivavad ning seetõttu eeldavad laste uurimisel lapsevanema informeeritud nõusolekute omandamist. Kui raalnägemise ja masinõppe põhine automaatne videoanalüüs võimaldaks mõõta kehalist aktiivsust koolimajas sedasi, et videot ei salvestata ning inimesed neid kaadreid näha ei saa, siis saaksime koguda vaatlusandmed kehalise aktiivsuse kohta ilma vaadeldavate privaatsust rikkumata. Kuna selline tehnoloogia võimaldaks analüüsida visuaalset infot ilma seda nägemata, nimetan meetodit **pimevaatluseks**. Doktoritöö eesmärgiks oli välja selgitada ja demonstreerida, kuidas sellistele kriteeriumitele vastav videoanalüüsi meetod saavutada ning kuidas seda rakendada.

Uurimuses I anti ülevaade olemasolevatest kehalise aktiivsuse mõõtmise meetoditest ning raalnägemise valdkonna arengutest, mis viitavad automaatse, reaalaja kiirusel toimiva videoanalüüsi võimalikkusele. Seejärel esitati **uurimuses I** algne visioon kavandatavast sensortehnoloogiast ning selle arendamise protsessist koos toetava empiirilise osaga, kus uuriti puusal kantavate aktseleromeetrite signaalide korrelatsioone liikumisinfoga videos. Meetod põhineb juhendamisega masinõppel, kus videoandmetele määratakse kehalise aktiivsuse intensiivsuse märgend kasutades puusal kantavaid aktseleromeetreid. Masinõppe treeningandmestiku kogumiseks värvati mugavusvalimina 7–14-aastaseid lapsi, kelle puusale kinnitati aktseleromeeter ning filmiti statsionaarse, üle kahe meetri kõrgusel asuva kaameraga. Kiirenduse signaalid sünkroniseeriti videoga ning andmesubjektid annoteeriti ristkülikutega, et moodustada masinõppeks kasutatav andmestruktuur – kehalise aktiivsuse märgendiga teotoru (**uurimus I**, Joonis 1).

Annoteeritud andmestikus on kokku 12 tundi unikaalseid kehalise aktiivsuse väljendusi 24-lt lapselt.

Selleks, et tõlkida andmestikus kajastuvad kiirenduse signaalid üldtunnustatud kehalise aktiivsuse skaalale, viidi läbi **uurimus II**. Veebiküsitlusega paluti kehalise aktiivsuse teadusvaldkonna ekspertidel klassifitseerida kehalist aktiivsust 24-s lühikeses videoklipis – kas see laps liigub nendel sekunditel mõõduka kehalise aktiivsuse piirist (hoogsa kõnniga võrdsustatav liikumise intensiivsuse tase, kus kulub kolm korda rohkem energiat, kui sellel inimesel kulub puhkeasendis) rohkem või vähem intensiivselt. Lisaks kiirendussignaalidele, kasutati moodustunud ekspertmudelit ka uudse, videost arvutatava kehalise aktiivsuse indikaatori tõlgendamiseks. Nimelt rakendati videoandmetele kahemõõtmelise poosituvastuse mudeleid (Sun et al., 2019) ning kehalise aktiivsuse indikaatoriks arvutati tuvastatud poosi puusa nurkade muutumise määr kaadrite vahel – mida kiiremini muutub põlve ja kaela vaheline nurk tuvastatud poosi kahemõõtmelises projektsioonis, seda intensiivsemat liikumist võib eeldada. Ekspertide küsitlemisest moodustunud mõõduka kehalise aktiivsuse piiri mudel laiendati nende kahe indikaatori kaudu kogu masinõppeandmestikule.

Doktoritöö viimases faasis (**uurimus III**) arendati meetod prototüübini. Videoanalüüs toimub kahesekundilise väljundsagedusega, kus analüüsiühiku mõõtmeteks on 1280×720×20 (kümme kaadrit sekundis RGB video). Andmetöötlusjada esimeses etapis kasutatakse kahemõõtmelist konvolutsioonilist tehisnärvivõrku (C.-Y. Wang et al., 2021) inimeste tuvastamiseks sissetulevates kaadrites. Siis rakendatakse ByteTrack järgimisalgoritmi (Y. Zhang, 2021/2022), mis ühendab tuvastused järjestikustes kaadrites teotorudeks. Järgmiseks transformeeritakse teotorud vastamaks kehalise aktiivsusse klassifitseerija sisendile: järjestikused tuvastused viiakse kujule 160×160 pikslit ilma pilti moonutamata ning kui teotoru on lühem kui 20 kaadrit, sisestatakse toru algusesse või/ja lõppu vastav hulk tühje kaadreid. Moodustunud 160×160×20 teotorud klassifitseeritakse kolmemõõtmelise konvolutsioonilise tehisnärvivõrguga (Feichtenhofer, 2020), mis on treenitud kogutud andmestikult. Kehalise aktiivsuse klassifitseerimise mudel saavutab testandmestikul makrokeskmise F1 skoori 0,83 ning kogu pimevaatlussüsteemi võimekuseks antud kujul hindame ligikaudu 0,66. Andmetöötlusjada rakendati väiksele, 15W võimsusega seadmele (Nvidia Jetson Xavier NX 8GB) ning reaalaja andmetöötlusvõimekus saavutati ilma väga põhjaliku koodi optimeerimiseta. Prototüüp suudab reaalaja võimekust säilitada vähemalt kuni 15 korraga tuvastatud inimesega kaamera vaateväljas.

Doktoritööst saab järeldada, et tehnoloogiad on küpsed inimese käitumise automaatseks vaatlemiseks reaalaja kiirusel videoanalüüsiga. Kuigi kogutud masinõppeandmestik on teaduslike mõõtmiste jaoks soovitava täpsuse saavutamiseks liiga väike, pakub doktoritöö korralduslikke, tehnilisi, juriidilisi ja teadus-eetilisi suuniseid meetodi edasiseks arendamiseks ning rakendamiseks koolipõhistes kehalise aktiivsuse sekkumisuuringutes. Doktoritöö katusartiklis lahatakse vaatlusmeetodeid ka laiemalt, et mõtestada automaatsete, tehisintellektil põhinevate vaatlusmeetodite potentsiaali ja võimalikke kaasnevaid probleeme inimkäitumise uurimisel.

PUBLICATIONS

CURRICULUM VITAE

Name: Hans Hõrak
Date of Birth: 26.08.1988
Phone: +372 5331 6215
E-mail: hans.horak@ut.ee

Education

2019–2023	University of Tartu, PhD studies in sociology
2012–2015	University of Tartu, Master studies in sociology (*cum laude*)
2008–2011	University of Tartu, Bachelor studies in sociology, social work, and social policy

Work experience

2023–present	Statistics Estonia, Data Scientist
2021–2022	University of Tartu, Faculty of Social Sciences, Institute of Social Studies, Project Manager
2020–2021	University of Tartu, Faculty of Social Sciences, Institute of Social Studies, Junior Research Fellow
2017–2020	University of Tartu, Faculty of Social Sciences, Johan Skytte Institute of Political Studies, Center for Applied Social Sciences, Analyst

Teaching experiences:
Co-instructor: Big Data and Society, University of Tartu, Faculty of Social Sciences, Institute of Social Studies. Spring 2021, 2022 and 2023.

ELULOOKIRJELDUS

Nimi: Hans Hõrak
Sünniaeg: 26.08.1988
Telefon: +372 5331 6215
E-post: hans.horak@ut.ee

Haridus

2019–2023	Tartu Ülikool, sotsioloogia doktoriõpe
2012–2015	Tartu Ülikool, sotsioloogia magistriõpe (*cum laude*)
2008–2011	Tartu Ülikool, sotsioloogia, sotsiaaltöö ja sotsiaalpoliitika bakalaureuseõpe

Teenistuskäik

2023–täna	Statistikaamet, andmeteadur
2021–2022	Tartu Ülikool, Sotsiaalteaduste valdkond, ühiskonnateaduste instituut, projektijuht
2020–2021	Tartu Ülikool, Sotsiaalteaduste valdkond, ühiskonnateaduste instituut, automatiseeritud vaatluse meetodite nooremteadur
2017–2020	Tartu Ülikool, Sotsiaalteaduste valdkond, Johan Skytte poliitikauuringute instituut, sotsiaalteaduslike rakendusuuringute keskus (RAKE), analüütik

Õpetamiskogemus:
Kaaslektor ja praktikumide läbiviija: Big Data and Society, Tartu Ülikool, Sotsiaalteaduste valdkond, ühiskonnateaduste instituut. 2021, 2022 ja 2023 kevad.

DISSERTATIONES SOCIOLOGICAE UNIVERSITATIS TARTUENSIS

1. **Veronika Kalmus**. School textbooks in the field of socialisation. Tartu, 2003, 206 p.
2. **Kairi Kõlves**. Estonians' and Russian minority's suicides and suicide risk factors: studies on aggregate and individual level. Tartu, 2004, 111 p.
3. **Kairi Kasearu**. Structural changes or individual preferences? A study of unmarried cohabitation in Estonia. Tartu, 2010, 126 p.
4. **Avo Trumm**. Poverty in the context of societal transitions in Estonia. Tartu, 2011, 215 p.
5. **Kadri Koreinik**. Language ideologies in the contemporary Estonian public discourse: With a focus on South Estonian. Tartu, 2011, 128 p.
6. **Marre Karu**. Fathers and parental leave: slow steps towards dual earner/ dual carer family model in Estonia. Tartu, 2011, 125 p.
7. **Algi Samm**. The relationship between perceived poor family communication and suicidal ideation among adolescents in Estonia. Tartu, 2012, 121 p.
8. **Tatjana Kiilo**. Promoting teachers' efficacy through social constructivist language learning: challenges of accommodating structure and agency. The case of Russian-speaking teachers in Estonia. Tartu, 2013, 156 p.
9. **Ave Roots**. Occupational and income mobility during post-socialist transformation of 1991–2004 in Estonia. Tartu, 2013, 130 p.
10. **Tarmo Strenze**. Intelligence and socioeconomic success A study of correlations, causes and consequences. Tartu, 2015, 119 p.
11. **Mervi Raudsaar**. Developments of social entrepreneurship in Estonia. Tartu, 2016, 141 p.
12. **Ero Liivik**. Otsedemokraatia Eestis: õigussotsioloogilisi aspekte. Tartu, 2017, 166 p.
13. **Mai Beilmann**. Social Capital and Individualism – Collectivism at the Individual Level. Tartu, 2017, 145 p.
14. **Rainer Reile**. Self-rated health: assessment, social variance and association with mortality. Tartu, 2017, 123 p.
15. **Katri Lamesoo**. Social Construction of Sexual Harassment in the Post-Soviet Context on the Example of Estonian Nurses. Tartu, 2017, 185 p.
16. **Andu Rämmer**. Sotsiaalse tunnetuse muutused Eesti siirdeühiskonna kontekstis. Tartu, 2017, 230 p.
17. **Kadri Rootalu**. Antecedents and consequences of divorce in Estonia from longitudinal and multigenerational perspectives. Tartu, 2017, 128 p.
18. **Kairi Talves**. The dynamics of gender representations in the context of Estonian social transformations. Tartu, 2018, 129 p.
19. **Aare Kasemets**. Institutionalisation of Knowledge-Based Policy Design and Better Regulation Principles in Estonian Draft Legislation. Tartu, 2018, 252 p.

20. **Dagmar Narusson**. Personal-recovery and agency-enhancing client work in the field of mental health and social rehabilitation: Perspectives of persons with lived experience and specialists. Tartu, 2019, 139 p.
21. **Oliver Nahkur**. Measurement of Interpersonal Destructiveness: the Societal Perspective. Tartu, 2019, 164 p.
22. **Tayfun Kasapoglu**. Algorithmic Imaginaries of Syrian Refugees: Exploring Hierarchical Data Relations from the Perspective of Refugees. Tartu, 2021, 152 p.
23. **Kristjan Kikerpill**. Crime-as-communication: detecting diagnostically useful information from the content and context of social engineering attacks. Tartu, 2021, 162 p.
24. **Taavi Laanepere**. Looking at the Military Service Readiness of Estonian Reserve Soldiers through the Prism of Bourdieu's Theory of Practice. Tartu, 2021, 174 p.
25. **Tiia-Triin Truusa**. The entangled gap: the male Estonian citizen and the interconnections between civilian and military spheres in society. Tartu, 2021, 159 p.
26. **Kadri Soo**. School as a source of child subjective well-being in the framework of children's rights: Perspectives of children and young adults. Tartu, 2023, 146 p.