Albert Ali Salah
Bruno Lepri (Eds.)

LNCS 7065

Human Behavior Understanding

Second International Workshop, HBU 2011
Amsterdam, The Netherlands, November 2011
Proceedings

Lecture Notes in Computer Science 7065

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Albert Ali Salah Bruno Lepri (Eds.)

Human Behavior Unterstanding

Second International Workshop, HBU 2011
Amsterdam, The Netherlands, November 16, 2011
Proceedings

Springer

Volume Editors

Albert Ali Salah
Bogaziçi University
Department of Computer Engineering
Bebek 34342, Istanbul, Turkey
E-mail: salah@boun.edu.tr

Bruno Lepri
FBK - Fondazione Bruno Kessler
Via Sommarive 18, 38100 Trento, Italy
E-mail: lepri@fbk.eu

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-25445-1 e-ISBN 978-3-642-25446-8
DOI 10.1007/978-3-642-25446-8
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011941084

CR Subject Classification (1998): I.5, H.5.2, I.4, I.4.8, I.2, I.2.10, H.3-4

LNCS Sublibrary: SL 6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Automatic computer analysis of human behavior is an expanding research area, with many technical challenges and many potential applications, encompassing gaming, surveillance, multimedia, ambient-assisted living, and many more. The Second International Workshop on Human Behavior Understanding (HBU) aimed to bring together researchers developing and using computer analysis tools for learning and modeling human behavior, covering both hardware or software aspects. As such, the topics link areas like pattern recognition, sensor technologies, social signal processing, and interaction design.

The International Joint Conference on Ambient Intelligence combines concepts of ubiquitous technology, intelligent systems and advanced user interface design, presenting an excellent opportunity to foster collaborations across disciplines. The first HBU Workshop had a pattern recognition focus, and was organized as a satellite to ICPR 2010. The second workshop had a focus theme on inducing behavioral change, which means moving the computer from a passive observer role to a socially active participating role and enabling it to drive some kinds of interaction, such as influencing attitudes and behaviors of people in natural or virtual environments.

This proceedings volume contains 13 papers presented at the workshop, as well as the abstracts of the keynote talks by Nuria Oliver (Telefonica Spain) and Wijnand Ijsselsteijn (Eindhoven University of Technology), and a summarizing paper. We received 32 submissions in total, and the each paper was peer-reviewed by at least two members of the Technical Program Committee.

We would like to take the opportunity to thank our Program Committee members and reviewers for their rigorous feedback, our authors and our keynote speakers for their contributions. We would also like to thank the AmI 2011 Organizing Committee, and in particular Ben Kröse, Gwenn Englebienne, and Reiner Wichert.

November 2011

Albert Ali Salah
Bruno Lepri

Organization

Conference Co-chairs

Albert Ali Salah	Boğaziçi University, Turkey
Bruno Lepri	MIT, USA, and FBK, Italy

Technical Program Committee

Hamid Aghajan	Stanford University, USA
Lale Akarun	Boğaziçi University, Turkey
Oya Aran	IDIAP, Switzerland
Mark Cavazza	Teesside University, UK
Mauro Cherubini	Telefonica Research, Spain
Jeffrey Cohn	University of Pittsburgh, USA
Theo Gevers	University of Amsterdam, The Netherlands
Jordi Gonzáles	Universidad Autónoma de Barcelona, Spain
Dirk Heylen	University of Twente, The Netherlands
Stephen Intille	Northeastern University, USA
Taemie Kim	MIT, USA
Tsvika Kuflik	University of Haifa, Israel
Bruno Lepri	MIT, USA, and FBK, Italy
Vittorio Murino	IIT, Italy
Maja Pantic	Imperial College London, UK
Eric Pauwels	CWI, The Netherlands
Alex Pentland	MIT, USA
Fabio Pianesi	University of Trento, Italy
Peter Robinson	Cambridge University, UK
Michael S. Ryoo	ETRI, South Korea
Albert Ali Salah	Boğaziçi University, Turkey
Ben Schouten	Eindhoven University of Technology, The Netherlands
Nicu Sebe	University of Trento, Italy
Alessandro Vinciarelli	University of Glasgow, UK
Massimo Zancanaro	FBK, Italy

Additional Reviewers

Hande Alemdar	Leonardo Giusti	Hatice Köse
Neşe Alyüz	Jungong Han	Ilkka Kosunen
Shlomo Berkovsky	Kyriaki Kalimeri	Oswald Lanz
Hamdi Dibeklioğlu	Tim van Kasteren	Wei Pan
Eyal Dim	Cem Keskin	

Table of Contents

Social Interactions

Human Behavior Understanding for Inducing Behavioral Change: Application Perspectives

Albert Ali Salah[1], Bruno Lepri[2,3], Fabio Pianesi[2], and Alex Sandy Pentland[3]

[1] Boğaziçi University, Department of Computer Engineering, Istanbul, Turkey
salah@boun.edu.tr
[2] FBK, via Sommarive 18, Povo, Trento, Italy
{lepri,pianesi}@fbk.eu
[3] MIT Media Lab, 20 Ames Street, 02-139 Cambridge, MA, USA
pentland@mit.edu

Abstract. Pervasive sensing and human behavior understanding can help us in implementing or improving systems that can induce behavioral change. In this introductory paper of the 2nd International Workshop on Human Behavior Understanding (HBU'11), which has a special focus theme of "Inducing Behavioral Change", we provide a taxonomy to describe where and how HBU technology can be harnessed to this end, and supply a short survey of the area from an application perspective. We also consider how social signals and settings relate to this concept.

1 Introduction

In recent years, the automatic analysis of human behavior has been attracting an increasing amount of attention from researchers because of its important potential applications and its intrinsic scientific challenges. In many technological fields (pervasive and ubiquitous computing, multimodal interaction, ambient assisted living and assisted cognition, computer supported collaborative work, user modeling, automatic visual surveillance, etc.) the awareness is emerging that a system can provide better and more appropriate services to people only if it can understand much more about users' attitudes, preferences, personality, social relationships etc., as well as about what people are doing, the activities they have been engaged in the past, their routines and life-styles, etc.

At the same time, several attempts have been made to build *persuasive technologies*. Most of the research on this topic is often comprised under the umbrella of the term 'captology', which generally refers to the study of machines designed to influence people's attitudes and behaviors. The challenge in captology is to design an engaging and stimulating environment (or technology) that in time would steer the user's behavior towards a desired behavior. In [15], Fogg stresses the distinction between a technology's side effects and its planned effects, where

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 1–15, 2011.

the latter is relevant from a design perspective. For instance, exposure to violent video games may increase aggressive thoughts, feelings, and behaviors, and decrease helping behavior, as unplanned side effects [2]. Although a better understanding of the mechanisms underlying side effects would make it possible to compensate for them, it is the planned effects themselves that have been attracting most of the attention.

Current efforts towards persuasive technologies have rarely taken into account the real-time understanding of individual traits or the social dynamics the users engage in. Technologies for human behavior understanding (HBU), however, can be gainfully employed to make these persuasive systems more context-aware, interactive, adaptive, immersive and even anthropomorphic. The goal of this paper is to give an application perspective on how to reach these goals. Since persuasion is not always an explicit goal of such systems (as we will show later via examples), the systems we describe here span a broader area than "classical" persuasive technologies.

This paper is structured as follows. Section 2 describes taxonomies for employing HBU in a persuasive environment. Then, Section 3 reports recent research focus in different pervasive sensing modalities. Section 4 gives application examples for inducing behavior change, selected from four different domains. Finally, Section 5 concludes the paper. In this work, we are not going to discuss the theoretical and social aspects of behavioral change; these are tackled in a follow-up paper explicitly dealing with these issues [34]. In the present volume, [31] gives a good overview of theories on behavior change.

2 Taxonomies

In this section we discuss where and how HBU can be employed to induce behavioral change. We should note here that in computer science, the term "behavior" usually refers to a relatively short, measurable pattern of activity, whereas in psychology, it incorporates a broad range of details, pertaining to ability, intentions, and sustainability. The construction of relations between different time scales is one of the challenges in this area. By Human Behavior Understanding (HBU), we mean here pattern recognition and modeling techniques to automatically interpret complex behavioral patterns generated when humans interact with machines or with other humans [52]. These patterns encompass actions and activities, attitudes, affective states, social signals, semantic descriptions, and contextual properties [53]. Since persuasion is a detailed framework for discussing how to induce behavior change, we will adopt it as our main guideline, and deviate from it only occasionally.

Technology can achieve persuasion by being a tool (improving abilities, providing customized information, guiding people through a process, etc.), by being the channel of persuasion (particularly relevant for ambient intelligence (AmI) scenarios, where the environment uses feedback and visualization to provide behavior changing experiences) or by being a social actor to persuade through social communication channels [15].

Adapting the terminology of the attitude change model of Petty et al. (1997) [47], consider a **source** that sends a persuasive **message** (here a computer system with some output) to a **recipient** of the message (a human). HBU technologies can play different roles in this processes:

- **Positioning:** The source uses HBU to position the recipient, and selects appropriate messages (e.g. identifying whether the source has an attitude with a cognitive or affective base to select appropriate arguments [47]).
- **Feedback:** The source uses correct feedback towards sustaining behavior (e.g. monitoring facial expressions to judge the effect of provided messages).
- **Message:** The result of HBU is a part of the message to the recipient (e.g. measuring activity levels and visualizing them for a fitness application).
- **Evaluation:** HBU measures the progress of the recipient (e.g. a sign-language tutoring tool evaluating the correctness of replicated signs).
- **Prediction:** HBU is used to predict future behavior, to allow the system to adapt and preempt (e.g. predicting when the user will engage in harmful behavior and providing a timely feedback to prevent it).
- **Social guide:** HBU is used to increase the credibility of the source (e.g. an embodied conversational agent observing social norms and responding coherently to the social signals of the receiver). Correct display and interpretation of social signals play a great role in the persuasiveness of a technology.
- **Immersion:** HBU facilitates the construction of an immersive system, where the target behavior is encapsulated in the interaction with the system (e.g. a body movement and gesture based fitness game).

Fig. 1 shows some of the potential contributions of HBU in inducing behavioral change. Initially we consider the traditional case of a source, a message, a channel and a recipient. Broadly, HBU can be used for the analysis of social and affective cues to provide a better model of the context for the message. It can also be used to scale up assessment by eliminating a resource bottleneck: Most persuasive technologies are assessed by questionnaires, and other manually performed assessment tools. While these technologies are not widely adopted yet, real-time assessment is conceivable for applications where the effects of induction is not otherwise easy to observe directly. To give a simple example, we can monitor cigarette sales, but HBU can give us the number of actual instances where a subject lights a cigarette.

The most important contribution of HBU seems to be in improving **the message**. Learning recipient behavior can tell the system something about the response patterns of the recipient, and help message selection. It can also help message timing via prediction of certain behaviors. If the system can tell when the driver is going to speed, it can provide a timely message to negatively reinforce this behavior. The second contribution is to observe the recipient and make behavior-related cues part of the message. For instance, your energy consumption can be visualized for you in green or red light, prompting you to reconsider your consumption habits [24]. Thirdly, HBU can help transform the singular message into a communicative exchange. A successful embodied conversational agent (ECA) is an engaging conversation partner, and such tools can effectively

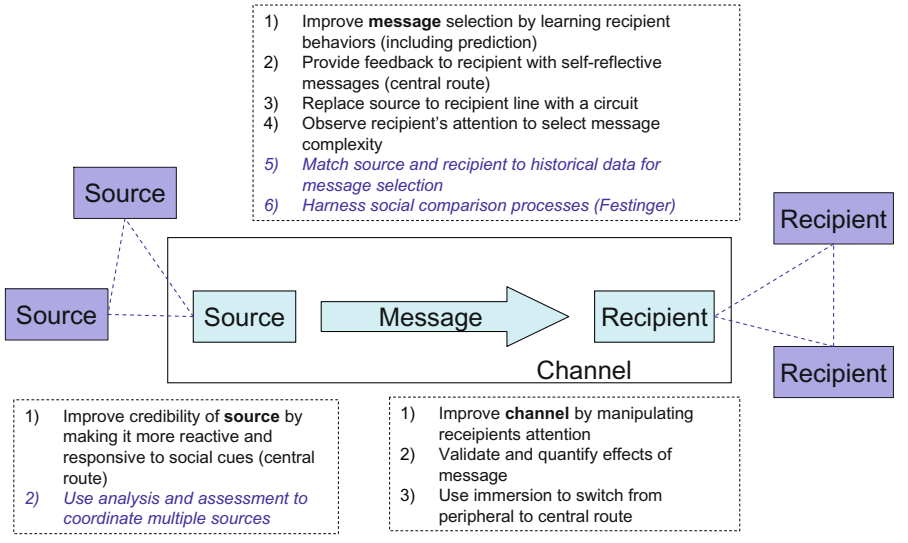


Fig. 1. Points of contribution for HBU in inducing behavioral change. Potential contributions to source, message and channel are listed. The case of multiple sources/recipients is indicated with blue italic lines.

engage more cognitive resources (i.e. from the perspective of the elaboration likelihood model [46], they engage the *central route*, as opposed to the *peripheral route*[1]). Through HBU, social interactions with systems can be augmented to pay attention to dimensions like empathy and intimacy, and thereby more natural interfaces can be obtained [22].

If the system has several messages, HBU can help in selecting the most appropriate of these. In a multiple source/recipient setting (i.e. a social setting), previous message exchanges and their observed results can serve as templates, and help in message selection (e.g. recipient A resembles recipient B, who responded well to message X, so the system follows the same strategy). Finally, observed behavior of multiple recipients can be used to mutually improve all messages sent to these recipients by harnessing social comparison processes.

The channel can benefit from HBU as well. The archetypical examples come from the ambient intelligence setting. It is possible that the environment observes the recipient's current focus of attention (e.g. by gaze tracking or speech analysis) and improves message delivery. Such an environment can potentially track

[1] According to the elaboration likelihood model (ELM) of Petty and Cacioppo, there are two different routes for changing the attitudes and the behaviors of a person, namely, a central route and a peripheral route, respectively [46]. The central route assumes attention to and elaboration of arguments, and involves coherence, logic, and clarity of arguments. The peripheral route, on the other hand, uses secondary attributes like the attractiveness of a source, its familiarity, credibility, etc.

the 'progress' of the recipient in an unobtrusive way, and quantify the effects of the message. [26] is an excellent discussion on persuasion in ambient intelligence settings, where the authors discuss many principles of human persuasion that may have repercussions for computer persuasion systems. The paper also introduces the *ambient persuasion model*, where a distinction is made between a horizontal axis of change (to model messages from a source to a recipient), and a vertical axis of change to model the temporal aspects of change, where short (the initial attitude), medium (behavior) and long-term (sustained behavior) effects are considered. Such a vertical axis of induced behavior duration is also proposed by Fogg, where the shortest change is a one-time performance of a behavior, and the longest change is a change of habit [16].

The channel can also offer an immersive experience and thus effect message processing. A conversation with an ECA where the recipient is manipulated to argue for a certain message can be effective, since through immersion more central, cognitive processing of the message can be ensured.

HBU can improve **the source** by increasing its credibility. This can be achieved for instance by making the source more responsive to social and affective cues observed in the recipient. An application example would be an ECA that acts as a source. Analysis of the recipient can make the ECA more plausible, as it conforms to social signals, observes backchannel rules, etc. In the multiple source case, HBU can be used to coordinate messages from different sources. For instance, if the messages are given in different modalities, or at different locations, the user behavior can be monitored to coordinate the messages.

In general, the use of HBU and the consequent modeling and understanding of social dynamics, individual traits, internal states, attitudes, and routines are able to contribute to a more ambitious goal, namely the design and construction of machines that act as social actors and that are able to purposely influence the attitudes and the behaviors of people in their everyday natural environments.

3 Pervasive Sensing

What kinds of behaviors can we analyze with HBU technologies? The answer to this question partly rests on the sensory modalities that we are willing to deploy, and the amount of computational resources we are willing to devote to observe and model humans in particular environments and contexts. To summarize, we can sense:

- **Specific behaviors:** Visual and auditory actions, activities and complex behaviors.
- **Psychophysical signals:** Bodily signals, physiological measurements.
- **Activities:** Amount of performed behavior per time unit, including body, head and facial movement, amount of speaking, and any other virtual and real behaviors.
- **Engagement:** Visual attention and gaze patterns, postures and gestures indicating engagement, speaker turn-taking behavior, indications of interest and motivation.

- **Empathy:** Mirroring and mimicry of speech and gestures, signals of agreement and disagreement, physiological signals of affect and empathy.
- **Other social signals:** Social role indicators, voluntary and involuntary social signals.

It is possible to group the research activity on HBU around sensory modalities.

3.1 Vision and Audio

The visual modality has traditionally been the primary channel for deriving information about human behavior at different scales. The whole gamut of spatio-temporal scales of behavior are explored with visual modalities. In the present volume, Hadid gives a broad overview of facial behavior analysis, which can occur over very short time frames, and can involve barely noticeable muscle movements [19]. On the other end of spectrum, cameras and camera networks can be used for monitoring multiple subjects. [10] uses vision for tracking multiple subjects in an indoor scenario, and to detect dyadic interactions. At an even larger scale, [49] describes an approach for detecting abnormal behavior in a crowd. In between, we find for instance analysis of hand gestures, which can have a rich vocabulary and thus pose a challenging classification task. In the present volume, [28] shows that a clever pre-processing stage can greatly improve the recognition of hand gestures.

Recent methods for gesture recognition involve the use of RGB-D cameras (e.g. the Microsoft Kinect), if the distance requirements of the RGB-D camera is met in the application [29]. The 1^{st} IEEE Workshop on Consumer Depth Cameras for Computer Vision (CDC4DV)[2], organized as a satellite to ICCV'11, received over 60 submissions, which is a clear acknowledgment of the great potential of this modality. Among the contributions of CDC4CV, we note a novel color-depth video database for human daily activity recognition [41], and new pose detection approaches [21,9]. Also for robotics applications, RGB-D cameras have quickly become the standard for estimating the pose of interacting humans, as we have witnessed in the RoboCup@Home competition in 2011[3]. In [33], such a camera is used to combine pose and motion cues to detect change during human-robot interaction. The system allows the robot to imitate the gestures of the interacting human, which is particularly useful in turn-taking for cases where the semantics of the gesture performed by the human is not precisely understood by the robot.

Apart from new modalities, progress in image retrieval techniques have also resulted in improved camera-based identification of actions and activities, as well as recognition of events [57]. The shift is towards learning the temporal aspects of activities, and generalizing from limited amount of training data [44]. In the present volume, Baccouche et al. propose a two-stage model for action recognition, where convolutional networks are extended to 3D case for learning

[2] http://www.vision.ee.ethz.ch/CDC4CV

[3] http://www.robocupathome.org

spatio-temporal features, followed by a recurrent neural network that classifies sequences into action classes [4].

HBU research in auditory modality focuses on the classification of paralinguistic information, including speaker states like affect, intimacy, deception, sleepiness, stress, etc., as well as speaker traits like gender and personality. Additional vocal behavior like yawns, laughter, and other vocalizations are automatically classified as well. Schuller provides an overview of this area in [56].

3.2 Mobile Phones as Social Sensors

Recent developments in mobile technologies and the advent of the smartphones have opened the way to a new and very useful tool for social sciences, particularly sociology, social psychology, urban studies, and network science. Smartphones allow for unobtrusive and cost-effective access to previously inaccessible sources of data related to daily social behavior [48,32]. The most important feature of a smartphone is its sensing ability. Nowadays, these devices are able to sense different kind of behavioral data: (i) location, (ii) other devices in physical proximity (e.g., through Bluetooth scanning); (iii) communication data, including both the metadata (logs of who, when, and duration) of phone calls and text messages (SMS) as well as their actual contents (e.g., recording of voice and text of SMS); (iv) scheduled events, (v) the devices status (e.g. network coverage, alarm clock, charger, status, and so on), (vi) movement patterns, and (vii) the user interaction with the mobile phone (e.g., the user is downloading some application; he/she is engaged in a call, is surfing the web and/or browsing a specific page; he/she is playing games, etc.). Additional sensors can provide researchers with futher useful information: detailed locations using a GPS, actions and activities by means of accelerometers, physiological variables (e.g., heart rate, galvanic skin response).

One of the first approaches for modeling human behaviors from mobile sensor data was the Reality Mining study [14]. In this work, the researchers followed 94 subjects using mobile phones, recording data about (i) call logs, (ii) Bluetooth devices in proximity of approximately five meters, (iii) cell tower IDs, (iv) application usage, and (v) phone status. Subjects were observed using these measurements over the course of nine months. The researchers additionally collected self-reports about relational data from each individual, where subjects were asked about their proximity to, and friendship with, others. Subjects were also asked about their satisfaction with their work group. This study compared observational data from mobile phones with standard self-report survey data, finding that the information from these two data sources is overlapping but distinct. For example, self-reports of physical proximity deviate from mobile phone records depending on the time difference and salience of the interactions.

Mobile sensor data have been used to understand a broad spectrum of sensing and modeling questions. Some examples include automatically inferring of co-location and conversational networks [60], linking social diversity and economic progress [14], automatic activity and event classification for mass market phones [39], and the adoption and diffusion of applications [45]. In particular,

three recent studies exploited mobile phones to model behavioral and attitudinal changes in individuals [35,37,36]. In the first study [35], proximity and communication data are used to improve the understanding of the link between behaviors and health symptoms at the individual level. In particular, this study demonstrated the existence of characteristic behavioral changes in subjects suffering from common diseases like colds, flu and stress.

In the second study [37], the authors described the use of mobile phones to model and understand the link between exposure to peers and weight gain among students during the course of a semester. This study demonstrates that the change in an individuals Body Max Index (BMI) can be explained by face-to-face exposure to contacts who themselves gained weight.

In the third study, mobile phones were used to measure the spread of political opinions (republicans vs. democrats) during the 2008 US presidential election campaign between Obama and McCain [36]. Mobile features in term of proximity can be used to estimate unique individual exposure to different opinions. Furthermore, the authors proposed a method to understand the link between specific behaviors and change in political opinions, for both democrats and republicans. In particular, they used the Latent Dirichlet Allocation (LDA) topic model [6] to contrast the activities of participants who changed opinions with those who did not.

3.3 Wearables, Brain-Computer Interfaces, and Other Sensing Devices

Physiological reactions can provide insight into subjects' affective states. These are typically galvanic skin response, heart rate, palmar sweat, pupillary dilation and constriction, and such. These need to be worn on the body, which means they are intrusive to different degrees. This, however, may not be a problem in certain settings, especially in working environments where special dress and equipment needs to be used. In [40], nurses and physicians in a stroke unit of a hospital were equipped with wearable electrocardiography (ECG) and acceleration sensors to measure the amount of stress they experience during their everyday work. The identification of and feedback about stressful situations can lead to avoidance behavior for these situations.

Nijholt et al. provide a survey of approaches that use brain-computer interfaces for innovative applications, among which we find many instances of inducing behavior change [42]. They have elsewhere demonstrated how one can control a character in the popular video game World of Warcraft, with the brain. They analyze the brain activity for alpha waves (in the frequency band of 8-12Hz), which relates to relaxed alertness, and allow the game character to change into a bear when the actual character is under stress. An earlier example is the Brainball game, where two gamers have to control a ball on the table by remaining calm [20]. In both cases, waves of the brain are sensed by a portable EEG sensor.

In [12] the authors use an ultrasonic device directed to a person to read micro-Doppler signatures of actions by transmitting an ultrasonic wave and measuring the reflection. These signatures are used with a k-means classifier to classify a

small number of activities with good accuracy. For activity recognition, inertial sensors can also provide detailed information [3].

The *sociometric badge* is a multi-sensor device to collect data during social interactions [30]. It can gather proximity data via sensing other badges, measure acceleration in 3D, and record speech. These badges were used to analyze dominance effects in group interactions in [25].

3.4 Ambient Settings and Immersive Platforms

RFID tags attached to objects can reveal usage patterns very easily, and these have been used in ambient settings [58]. Similarly, location occupation information can be obtained with non-intrusive passive infrared sensors (PIR) from an indoor environment [59]. The data obtained with such sensors can be mined for patterns, or triggers can be implemented if certain behaviors are expected.

3.5 Virtual Settings

Virtual behavior of people can be analyzed in similar ways to real behaviors. Here, more traditional sensors that measure physical phenomena are replaced by virtual sensors that provide data. Typical examples are social networks, from which interaction and activity patterns can be extracted [5], and mobile phones, which can reveal usage patterns and locations [13]. The existence of underlying social connection structures in most of these applications brings a wealth of additional information that can be used in prediction or profiling. Yet these media, unless enhanced specifically, have little to offer in terms of detailed face-to-face social interactions.

4 Applications

In this section we provide application examples from four domains, and discuss the mechanisms of inducing change, as well as the HBU technology used in the process.

4.1 Healthcare and Wellbeing

HBU can provide self-monitoring applications with quantitative data, and play an important role for healthcare and wellbeing applications. For instance, the *Houston* mobile application encourages individuals to take more steps each day by providing self-monitoring and social data sharing over mobile phones [11]. For the proposed application, pedometers were used to determine activity levels. Similarly, [17] encourage eating more fruits and vegetables via self-monitoring, but the setting is much more difficult to automatize data collection for target attainment.

The behavior to be influenced sometimes cannot be accurately assessed by the subject, and a computer system can be better positioned to provide assessment. These applications illustrate how the domain extends beyond persuasion.

In [23], a wearable sensor system is described to help patients after a hip replacement operation. The doctor provides the system with thresholds and rules for admissible postures and movements, and the system monitors the patient at all times, to sound an alarm when dangerous behavior is engaged. In the IS-ACTIVE project[4], the aim is to develop a person-centric solution to induce activity in people with chronic conditions, like the chronic obstructive pulmonary disease (COPD). A typical problem with these patients is that the fear of damaging themselves (or in the case of COPD of having a breathing problem) drives the patient to a mostly static lifestyle, which aggravates the condition. Real-time analysis allows self monitoring, whereby the patient can perform exercises while receiving feedback about how far they are to a critical overload condition. Alemdar and Ersoy provide an extensive survey of the usage of wireless sensor networks in health-care [1], and include applications like monitoring a patient for medication intake, status monitoring, activities of daily living and location tracking. Avci et al. survey the usage of inertial sensors for activity analysis in healthcare, sports and wellbeing applications [3].

4.2 Serious Gaming

Serious gaming is an area where games are designed with the purpose of teaching, skill acquisition, training, attitude and behavioral change. Since games serve entertainment, the primary mediator of behavioral change is the entertainment feedback, but other motivational factors like challenge, fantasy, curiosity, control, as well as social factors are also considered [8,38]. Bogost coined the term *persuasive games* to describe games designed to change behaviors and attitudes [7]. In the present volume, Rozendaal et al. describe the "Intelligent Play Environments" program, which deals with the design of playful interactive systems that stimulate physical and social activities [51]. We note here that only in rare cases is HBU integrated into gaming applications, primarily because the technology is deemed less than adequate. With new developments in sensors (for instance RGB-D cameras that facilitate real-time gesture recognition and portable EEG sets), the gaming industry sees more applications in this area [55].

4.3 Marketing

Marketing applications have been dominant particularly in virtual settings. Here, the aim is to influence buying behavior of as many people as possible. One way of achieving this is to rely on the way ideas are spread in social networks, and to seek to influence a proper subset of the population by analysing their buying behavior, letting social dynamics take care of the rest [27]. A very hot application area is **interactive marketing**, where HBU technology can be used to drive interaction. The interaction is typically an audio-visual experience, which is somehow related to the advertised product, and the viewers become a part of the whole setup. We give the example of the University of Amsterdam spinoff

[4] `http://www.is-active.eu/`

ThirdSight, which installed such a system on the biggest advertising screen in Amsterdam (on Rembrandtplein), where computer vision was used to allow people to interact and play with virtual balloons[5]. Another use of HBU in marketing is to determine location and head pose of a shop-viewer to measure presence and attention. Reitberger et al. describe a shop mannequin that turns toward and gazes at the customer, providing a more engaging interaction [50].

4.4 Energy Saving and Sustainability

In her keynote talk at HBU'11, Nuria Oliver discusses the emerging area of urban computing and smart cities in general and improving the quality of life of an urban environment by understanding the city dynamics through the data provided by ubiquitous technologies in particular [43]. This information can be used by city planners to improve infrastructure, as well as to create appropriate relief plans. The interest is also rising in systems that reduce energy consumption at home or work [24]. In [18], a system was proposed to augment thermostats by processing information from location-aware mobile phones. Even a simple context sensing approach can be useful for reducing individual energy costs.

5 Conclusions

HBU is our collective term for the toolbox of methods used in computer analysis of human behavior, and as we have shown, such analysis has many uses for persuasion, as well as for inducing behavior change beyond persuasion. Human behavior is complex; a certain behavior will be prompted by habits or intentions, modified based on skill, affect and attitude, influenced by physical and contextual conditions, and the results will be available to the computer via sensors, which capture a limited portion of reality with some noise. The models of HBU can thus be directed to make inferences about any of these influences on the outcome: What is the intention of the subject? Is the subject an expert? How does the subject feel? Does the subject have a positive attitude? What is the context of this behavior? We can certainly ask many more questions, and devise new ways of harnessing HBU for inducing behavior change.

Model-based approaches can go very deep in a particular domain. Starting from the seminal work of Schank and Abelson on scripts and plans [54], we have seen that by charting out detailed descriptions of the possibilities for actions and activities in a semantically limited domain (ordering food in a restaurant in the archetypical example) it is possible to associate meaning to sensed behavior. Yet, there are almost no limits to the domains of behavior, and one can imagine for the future vast repositories of behavior ontologies accessible over the Internet as a potential solution for substituting common knowledge. In recent years, we have witnessed the power of brute force computation in diverse domains. Also in HBU, researchers are building models sufficiently rich for particular application

[5] http://www.thirdsight.co/technologies/interactiveadvertising/

domains, which means that all relevant behaviors that can be conceived of as being salient within the context of the domain can be fruitfully distinguished. It is imperative to understand the capabilities and limits of existing HBU approaches to tailor them for practical solutions.

Amid all this scaling up of approaches to previously inconceivable computational resource expenditure, the data sources available for analysis are also vastly enriched. Beyond novel sensory modalities like RGB-D cameras and real-time data streaming smartphones, the social aspects of human behavior became the subject matter of computer analysis. Beyond doubt, this brings new challenges and new opportunities to the table.

Acknowledgments. Bruno Lepris research was funded by the Marie Curie COFUND 7th Frame- work PERSI project.

References

1. Alemdar, H., Ersoy, C.: Wireless sensor networks for healthcare: A survey. Computer Networks 54(15), 2688–2710 (2010)
2. Anderson, C.A., Bushman, B.J.: Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature. Psychological Science 12(5), 353–359 (2001)
3. Avci, A., Bosch, S., Marin-Perianu, M., Marin-Perianu, R., Havinga, P.: Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In: Proc. 23rd Int. Conf. on Architecture of Computing Systems (ARCS), pp. 167–176. VDE Verlag (2010)
4. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Sequential Deep Learning for Human Action Recognition. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 29–39. Springer, Heidelberg (2011)
5. Benevenuto, F., Rodrigues, T., Cha, M., Almeida, V.: Characterizing user behavior in online social networks. In: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference, pp. 49–62. ACM (2009)
6. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
7. Bogost, I.: Persuasive games: The expressive power of videogames. The MIT Press (2007)
8. Boyle, E., Connolly, T.M., Hainey, T.: The role of psychology in understanding the impact of computer games. Entertainment Computing 2(2), 69–74 (2011)
9. Charles, J., Everingham, M.: Learning shape models for monocular human pose estimation from the Microsoft Xbox Kinect. In: Proc. IEEE Workshop on Consumer Depth Cameras for Computer Vision (2011)
10. Chen, C.-W., Aztiria, A., Ben Allouchinst, S., Aghajan, H.: Understanding the Influence of Social Interactions on Individual's Behavior Pattern in a Work Environment. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 149–160. Springer, Heidelberg (2011)
11. Consolvo, S., Everitt, K., Smith, I., Landay, J.: Design requirements for technologies that encourage physical activity. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 457–466. ACM (2006)

12. Dura-Bernal, S., Garreau, G., Andreou, C., Andreou, A., Georgiou, J., Wennekers, T., Denham, S.: Human Action Categorization Using Ultrasound Micro-Doppler Signatures. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 18–28. Springer, Heidelberg (2011)
13. Eagle, N., Pentland, A.: Reality mining: sensing complex social systems. Personal and Ubiquitous Computing 10(4), 255–268 (2006)
14. Eagle, N., Pentland, A.S., Lazer, D.: Inferring friendship network structure by using mobile phone data. Proceedings of the National Academy of Sciences 106(36), 15274–15278 (2009)
15. Fogg, B.J.: Persuasive technologies. Communications of the ACM 42(5), 27–29 (1999)
16. Fogg, B.J.: The behavior grid: 35 ways behavior can change. In: Proceedings of the 4th International Conference on Persuasive Technology, pp. 42–46. ACM (2009)
17. Gasser, R., Brodbeck, D., Degen, M., Luthiger, J., Wyss, R., Reichlin, S.: Persuasiveness of a Mobile Lifestyle Coaching Application Using Social Facilitation. In: IJsselsteijn, W.A., de Kort, Y.A.W., Midden, C., Eggen, B., van den Hoven, E. (eds.) PERSUASIVE 2006. LNCS, vol. 3962, pp. 27–38. Springer, Heidelberg (2006)
18. Gupta, M., Intille, S., Larson, K.: Adding GPS-control to traditional thermostats: An exploration of potential energy savings and design challenges. Pervasive Computing, 95–114 (2009)
19. Hadid, A.: Analyzing Facial Behavioral Features from Videos. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 52–61. Springer, Heidelberg (2011)
20. Hjelm, S.I., Browall, C.: Brainball–using brain activity for cool competition. In: Proceedings of NordiCHI, pp. 177–188 (2000)
21. Holt, B., Bowden, R.: Putting the pieces together: Connected poselets for human pose estimation. In: Proc. IEEE Workshop on Consumer Depth Cameras for Computer Vision (2011)
22. Ijsselsteijn, W.: Augmenting Social Interactions: Experiments in Socio-Emotional Computing. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, p. 83. Springer, Heidelberg (2011)
23. Iso-Ketola, P., Karinsalo, T., Vanhala, J.: Hipguard: A wearable measurement system for patients recovering from a hip operation. In: Second International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth 2008, pp. 196–199. IEEE (2008)
24. Jentsch, M., Jahn, M., Pramudianto, F., Simon, J., Al-Akkad, A.: An Energy-Saving Support System for Office Environments. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 84–93. Springer, Heidelberg (2011)
25. Kalimeri, K., Lepri, B., Kim, T., Pianesi, F., Pentland, A.: Automatic Modeling of Dominance Effects Using Granger Causality. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 127–136. Springer, Heidelberg (2011)
26. Kaptein, M.C., Markopoulos, P., de Ruyter, B., Aarts, E.: Persuasion in ambient intelligence. Journal of Ambient Intelligence and Humanized Computing 1(1), 43–56 (2010)
27. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137–146. ACM (2003)
28. Keskin, C., Cemgil, A.T., Akarun, L.: DTW Based Clustering to Improve Hand Gesture Recognition. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 72–82. Springer, Heidelberg (2011)

29. Keskin, C., Kıraç, F., Kara, Y.E., Akarun, L.: Real time hand pose estimation using depth sensors. In: Proc. IEEE Workshop on Consumer Depth Cameras for Computer Vision (2011)
30. Kim, T., Olguín, D.O., Waber, B.N., Pentland, A.: Sensor-based feedback systems in organizational computing. In: International Conference on Computational Science and Engineering, CSE 2009, vol. 4, pp. 966–969. IEEE (2009)
31. Klein, M., Mogles, N., van Wissen, A.: Why Won't You Do What's Good For You? Using Intelligent Support for Behavior Change. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 105–116. Springer, Heidelberg (2011)
32. Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T.: A survey of mobile phone sensing. Comm. Mag. 48, 140–150 (2010)
33. Lee, J., Chao, C., Thomaz, A., Bobick, A.: Adaptive Integration of Multiple Cues for Contingency Detection. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 62–71. Springer, Heidelberg (2011)
34. Lepri, B., Salah, A.A., Pianesi, F., Pentland, A.: Human Behavior Understanding for Inducing Behavioral Change: Social and Theoretical Aspects. In: Wichert, R., Van Laerhoven, K., Gelissen, J. (eds.) Constructing Ambient Intelligence: AmI 2011 Workshops (2011)
35. Madan, A., Cebrian, M., Lazer, D., Pentland, A.: Social sensing for epidemiological behavior change. In: Proceedings of the 12th ACM International Conference on Ubiquitous Computing, Ubicomp 2010, pp. 291–300. ACM, New York (2010)
36. Madan, A., Farrahi, K., Gatica-Perez, D., Pentland, A.: Pervasive Sensing to Model Political Opinions in Face-to-Face Networks. In: Lyons, K., Hightower, J., Huang, E.M. (eds.) Pervasive 2011. LNCS, vol. 6696, pp. 214–231. Springer, Heidelberg (2011)
37. Madan, A., Moturu, S.T., Lazer, D., Pentland, A.S.: Social sensing: obesity, unhealthy eating and exercise in face-to-face networks. In: Wireless Health 2010, pp. 104–110. ACM, New York (2010)
38. Malone, T.W., Lepper, M.R., Handelsman, M.M., Briggs, W.L., Sullivan, N., Towler, A., Bryan-Kinns, N., Healey, P.G.T., Leach, J.: Making learning fun: A taxonomy of intrinsic motivations for learning. Journal of Educational Research 98(3) (2005)
39. Miluzzo, E., Lane, N.D., Fodor, K., Peterson, R., Lu, H., Musolesi, M., Eisenman, S.B., Zheng, X., Campbell, A.T.: Sensing meets mobile social networks: the design, implementation and evaluation of the CenceMe application. In: Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems, SenSys 2008, pp. 337–350. ACM, New York (2008)
40. Müller, L., Rivera-Pelayo, V., Kunzmann, C., Schmidt, A.: From Stress Awareness to Coping Strategies of Medical Staff: Supporting Reflection on Physiological Data. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 94–104. Springer, Heidelberg (2011)
41. Ni, B., Wang, G., Moulin, P.: RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In: Proc. IEEE Workshop on Consumer Depth Cameras for Computer Vision (2011)
42. Nijholt, A., Plass-Oude Bos, D., Reuderink, B.: Turning shortcomings into challenges: Brain-computer interfaces for games. Entertainment Computing 1(2), 85–94 (2009)
43. Oliver, N.: Urban Computing and Smart Cities: Opportunities and Challenges in Modelling Large-Scale Aggregated Human Behavior. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 16–17. Springer, Heidelberg (2011)

44. Orrite, C., Rodríguez, M., Montañés, M.: One-Sequence Learning of Human Actions. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 40–51. Springer, Heidelberg (2011)
45. Pan, W., Aharony, N., Pentland, A.: Composite social network for predicting mobile apps installation. In: Proc. AAAI (2011)
46. Petty, R., Cacioppo, J.: The elaboration likelihood model of persuasion. Advances in Experimental Social Psychology 19(1), 123–205 (1986)
47. Petty, R.E., Wegener, D.T., Fabrigar, L.R.: Attitudes and attitude change. Annual Review of Psychology 48(1), 609–647 (1997)
48. Raento, M., Oulasvirta, A., Eagle, N.: Smartphones. Sociological Methods & Research 37(3), 426–454 (2009)
49. Raghavendra, R., Del Bue, A., Cristani, M., Murino, V.: Abnormal Crowd Behavior Detection by Social Force Optimization. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 137–148. Springer, Heidelberg (2011)
50. Reitberger, W., Meschtscherjakov, A., Mirlacher, T., Scherndl, T., Huber, H., Tscheligi, M.: A persuasive interactive mannequin for shop windows. In: Proceedings of the 4th International Conference on Persuasive Technology. ACM (2009)
51. Rozendaal, M., Vermeeren, A., Bekker, T., De Ridder, H.: A Research Framework for Playful Persuasion Based on Psychological Needs and Bodily Interaction. In: Salah, A.A., Lepri, B. (eds.) HBU 2011. LNCS, vol. 7065, pp. 117–126. Springer, Heidelberg (2011)
52. Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A.: Challenges of human behavior understanding. In: HBU [53], pp. 1–12 (2010)
53. Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A. (eds.): HBU 2010. LNCS, vol. 6219. Springer, Heidelberg (2010)
54. Schank, R.C., Abelson, R.P.: Scripts, plans, goals and understanding: An inquiry into human knowledge structures. Lawrence Erlbaum Associates (1977)
55. Schouten, B., Tieben, R., van de Ven, A., Schouten, D.: Human Behavior Analysis in Ambient Gaming and Playful Interaction. In: Salah, A., Gevers, T. (eds.) Computer Analysis of Human Behavior. Springer, Heidelberg (2011)
56. Schuller, B.: Voice and speech analysis in search of states and traits. In: Salah, A.A., Gevers, T. (eds.) Computer Analysis of Human Behavior, pp. 227–253. Springer, Heidelberg (2011)
57. Shapovalova, N., Fernández, C., Roca, F.X., González, J.: Semantics of human behavior in image sequences. In: Salah, A.A., Gevers, T. (eds.) Computer Analysis of Human Behavior, pp. 151–182. Springer, Heidelberg (2011)
58. van Kasteren, T., Noulas, A., Englebienne, G., Kröse, B.: Accurate activity recognition in a home setting. In: Proc. 10th Int. Conf. on Ubiquitous Computing, pp. 1–9. ACM (2008)
59. Wren, C., Ivanov, Y., Leigh, D., Westhues, J.: The MERL motion detector dataset. In: Proc. 2007 Workshop on Massive Datasets, pp. 10–14. ACM (2007)
60. Wyatt, D., Choudhury, T., Bilmes, J., Kitts, J.A.: Inferring colocation and conversation networks from privacy-sensitive audio with implications for computational social science. ACM Trans. Intell. Syst. Technol. 2, 7:1–7:41 (2011)

Urban Computing and Smart Cities: Opportunities and Challenges in Modelling Large-Scale Aggregated Human Behavior

Nuria Oliver

Telefonica R&D, Via Augusta, 177,
Barcelona, Spain
nuriao@tid.es

Abstract. City-wide urban infrastructures are increasingly reliant on networked technology to improve and expand their services. As a side effect of this digitalization, large amounts of data –digital footprints– can be sensed and analyzed to uncover patterns of human urban behavior and to augment the city experience of its citizens. In my talk, I will introduce the main concepts, opportunities and challenges in the emerging area of urban computing/smart cities, which focuses on improving the quality of life of an urban environment by understanding the city dynamics through the data provided by ubiquitous technologies. This is a human-centric, data-rich area that spans multiple disciplines, including sociology, computer science and urban planning. From a computer science perspective, there are challenges in a variety of domains, including data visualization, storage, security, privacy, machine learning, data mining and pattern recognition. Some of the applications of smart cities include traffic forecasting, modeling of the spread of biological viruses, urban and transportation design and location-based services.

1 Extended Abstract

Traditionally, the study of urban environments has used data obtained from surveys to characterize specific geographical areas or the behavior of groups of individuals. However, new data sources (including GPS, bluetooth, WiFi hotspots, geo-tagged resources, etc.) are becoming more relevant as traditional techniques face important limitations, mainly:

1. the complexity and cost of capturing survey data;
2. the lack of granularity of the data given that is typically of aggregated nature;
3. the data are static and represent a snapshot of the situation in a specific moment in time; and
4. the increasing unwillingness of individuals to provide (what they perceive to be) personal information.

I will describe recent work analyzing the digital footprints from two sensors of the urban infrastructure: shared bicycling and mobile phone stations. I will show

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 16–17, 2011.

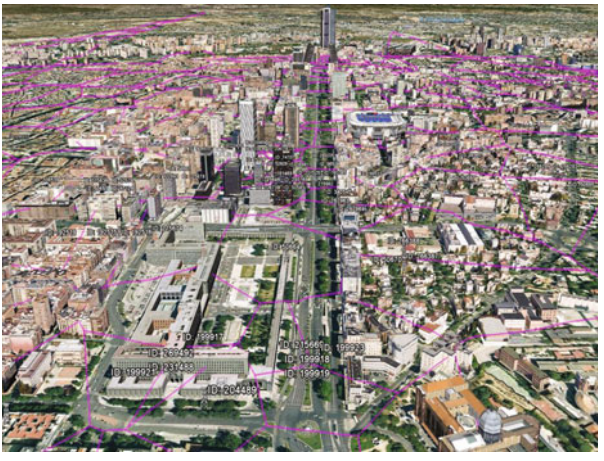


Fig. 1. Tesellation of a city into sections based on behavior similarity

how these digital footprints can be used to infer cultural and geographic aspects of the city and predict aspects of the city's behavior, such as automatically identifying hotspots of activity or segmenting the city by its behavior.

The identification of areas with high levels of people and/or activity (hotspots) is of paramount importance for e.g. urban and transport planners or emergency relief and public health officials. Urban planners can use this information to improve the public transport system by identifying dense areas that are not well covered by the current infrastructure, and determine at which specific times the service is more needed. In addition, public health officials can use the information to identify the geographical areas in which epidemics can spread faster and thus prioritize preventive and relief plans accordingly.

The spatial layout of a city has an obvious influence on the movement patterns and social behaviors found therein. Most western cities have a mixture of residential, commercial, and recreational areas connected via narrow streets, one-way avenues and a multitude of public transportation options and topographic features. Each of these areas has its own patterns of behavior which to date have only been elucidated by means of surveys and questionnaires. We have developed clustering algorithms to automatically segment the city in areas with similar behavior from two data sources: shared bicycling stations and anonymized and aggregated cell-phone records. Using our approach, only sections of the city with a given minimum similarity in their behavior will be labeled (See Figure 1). This technique could also be applied to data obtained from other ubiquitous data sources, like geo-localized tweets, Flickr or the logs of any service that includes geo-localization.

I believe that the work carried out in this research area has the potential to revolutionize our understanding of human urban behavior, as it is the first time in human history that we have access and the ability to analyze such multi-dimensional, pervasive and large-scale datasets.

Human Action Categorization Using Ultrasound Micro-Doppler Signatures

Salvador Dura-Bernal[1], Guillaume Garreau[2], Charalambos Andreou[2], Andreas Andreou[2], Julius Georgiou[2], Thomas Wennekers[1], and Susan Denham[1]

[1] Centre for Robotics and Neural Systems, University of Plymouth, PL4 8AA Plymouth, United Kingdom
salvador.durabernal@plymouth.ac.uk
[2] Holistic Electronics Research Lab, University of Cyprus, Kallipoleos 75, 1678 Nicosia, Cyprus
ggarreau@ucy.ac.cy

Abstract. The spectrotemporal representation of an ultrasonar wave reflected by an object contains frequency shifts corresponding to the velocity of the object's moving parts, also known as the micro-Doppler signature. The present study describes how the micro-Doppler signature of human subjects, collected in two experiments, can be used to categorize the action performed by the subject. The proposed method segments the spectrogram into temporal events, learns prototypes and categorizes the events using a Nearest Neighbour approach. Results show an average accuracy above 95%, with some categories reaching 100%, and a strong robustness to variations in the model parameters. The low computational cost of the system, together with its high accuracy, even for short length inputs, make it appropriate for a real-time implementation with applications to intelligent surveillance, monitoring and related disciplines.

Keywords: ultrasound, micro-Doppler signature, action categorization.

1 Introduction

The velocity of a moving object relative to an observer causes a frequency shift of a wave radiated or scattered by the object. This is known as the Doppler effect. If the object is composed of smaller moving parts, each of them will produce an additional modulation of the base Doppler frequency shift, known as the micro-Doppler effect. The micro-Doppler signature enables to determine properties of the motion of an object. Current research targets on a number of applications, e.g., presence detection [11,10], gait characterization [13], gender classification [10,5] and individual identification [4,14,5] of people walking; hand gesture recognition [7], face recognition [1,6,9], fall detection [8] and classification of the mode of transport [3]. Alternative conventional technologies –primarily visible or infrared video cameras– have a number of drawbacks, such as being expensive, requiring a lot of memory, computing power or communication bandwidth to

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 18–28, 2011.

process and transmit the images, being bulky and relatively immobile, and raising privacy concerns when deploying the system in public.

The microsonar system we propose is fast, portable, has low hardware and computational cost, doesn't invade personal privacy and can be used in situations where no visual information is available, for example, at night or in smoke-filled rooms. This technology has been biologically inspired by the bat sonar system [1] and similar echolocation systems in blind people [12].

There are three main wave transmission modes for micro-Doppler systems: continuous wave at 40kHz [3,5,6,7], continuous wave RF at 10.5 GHz [4,10] and Pulse Doppler RF [13]. Similarly, existing systems employ a variety of data analysis methods, usually commencing with some kind of time-frequency representation, for example using the Fast Fourier Transform (FFT), followed by one of a broad range of categorization methods like linear classifiers [3,10], neural networks [1], Gaussian Mixture models [5,6,7], maximum correlation coefficient (MCC) [9], Support Vector Machine (SVM) [8], and k-Nearest Neighbour (kNN) classifier [8].

The system we propose employs a continuous wave at 40kHz and the data is processed using a time-corrected instantaneous frequency reassigned (IFR) spectrogram [2], which is then segmented into short temporal events, followed by k-means clustering and a k-Nearest Neighbour classifier. Two different datasets have been collected and tested using this approach, both of them yielding an extraordinarily high action categorization accuracy (>95%), which outperforms previous published studies. Additional results support a continuous time implementation of the system given the low computational cost and high accuracy, even for a small number of incoming events.

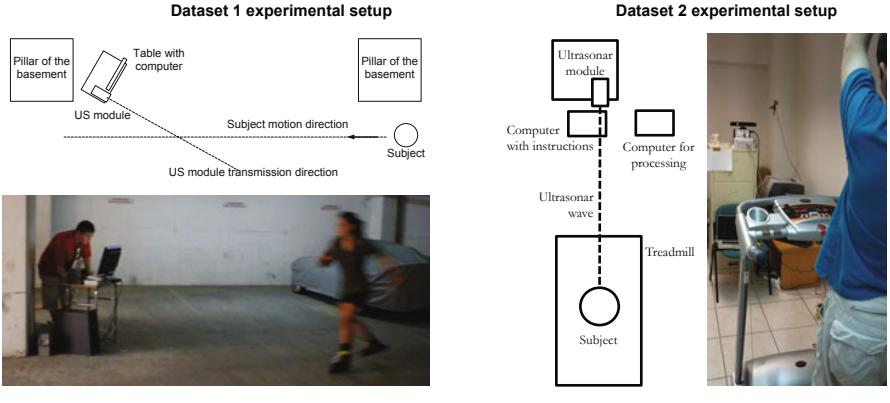


Fig. 1. Experimental setup for the collection of dataset 1 (top) and dataset 2 (bottom)

2 Methods

2.1 Acquisition Hardware

The ultrasonic device enables us to obtain the micro-Doppler signature of actions by transmitting an ultrasonic wave and measuring the reflected wave. The ultrasonic frequency determines the distance range and the measurement resolution of the device, where lower frequencies propagate over larger distances but have lower resolution. The chosen operating frequency is 40 kHz, which provides a maximum detection range of roughly five metres. The reflected sound is converted from mechanical energy, i.e., the acoustic wave, to electrical energy, i.e., voltage, and then sampled with an analog-to-digital converter.

The proposed custom hardware is composed of two printed circuit boards (PCBs), a micro-controller board and an analog board. An important factor that can increase the range of operation is the ability of the receiver to discriminate useful signal from noise. For this reason, the proposed device was designed to minimize the receiver noise. Further details of the hardware can be found in [3].

2.2 Data Collection

The results shown in this paper correspond to two different datasets. The main difference between them is that in the first one subjects were moving towards the ultrasonar device, whereas in the second dataset subjects performed the actions on top of a treadmill, thus the distance from the ultrasonar device was kept constant. Furthermore, the first dataset contains exclusively actions related to modes of transport, whereas the second dataset also contains some common human actions. The specific details for each dataset are described below:

Dataset 1: 5-category human transport modes. The ultrasonar sonar device was placed on a table at a height of 0.8m and subjects were required to move in a diagonal direction towards the device, as shown in the setup of Figure 1 (left). Three subjects (all males) performed 20 trials of five different actions: 1) walking, 2) running, 3) inline skating, 4) slow cycling (approximately 11 km/h) and 5) fast cycling (approximately 22 km/h). For inline skating, the three subjects were replaced by a different set (one male and two females) who could perform this action. Additionally, five of the inline skating trials were not performed adequately and were therefore discarded. This leads to a total of 295 recorded action trials: 60 for each action category, except for inline skating, which has only 55. The recording time for each action trial varied between 5 to 10 seconds, but only the 2 seconds centred around the peak value were used, as these represented the region of interest of the signal, i.e., the period when the action was captured by the ultrasonar device.

Dataset 2: 7-category human actions. The ultrasonar was placed at approximately 2.5 metres from the subject and at approximately 1.5 metres above

the floor in order to minimize the occluding surface of the treadmill. All actions were performed on top of the treadmill with the subject at the exact same distance from the recording device. The experimental setup is shown in Figure 1 (right). Ten subjects (eight males and two females) performed five trials of seven different actions: 1) walking slowly (3 km/h), 2) walking fast (6 km/h), 3) running slowly (9 km/h), 4) running fast (12 km/h), 5) clapping hands, 6) calling "help me" while moving their arms up in the air and 7) calling "come here" while beckoning with their right arm. The total number of action trials was therefore 350: 50 for each action category. Each trial recording lasted for ten seconds, started after the subject had already began performing the action and stopped before the subject had finished, such that the whole ten seconds contain relevant information related to the action.

2.3 Data Processing

The model used to process and analyze the data consists of two stages: the segmentation stage, which obtains a set of features or events from each micro-Doppler signature; and the categorization stage, which calculates representative prototypes from these events and uses them to decide what action is being performed given a set of incoming test events. The different steps are represented schematically in Figure 2.

The first step of the segmentation stage is to compute a time-corrected instantaneous frequency reassigned (IFR) spectrogram representation of each recorded sound wave, using the method of Nelson described in [2]. The frequency range of the recorded ultrasound waves is centred around 40 kHz, the emitter or carrier frequency. Thus prior to calculating the IFR spectrogram, the signal is undersampled such that the emitter frequency is transformed to 2.5 kHz.

The IFR spectrogram provides the micro-Doppler signature for each recorded action. Note that the frequency range of the IFR spectrogram is set differently for both datasets in order to capture the relevant frequency shifts. In dataset 1 the subject bodies where moving towards the ultrasonar device, so the region of interest are frequencies above the emitter frequency (2.55 and 3.6 kHz), whereas in the dataset 2 the subjects were at a constant distance from the device, so the region of interest are frequencies around the emitter frequency (1.85 and 3.25 kHz), i.e. those that capture the moving body parts. Other parameters of the IFR spectrogram function are the number of channels, which determines the frequency resolution and was set to 50 for both datasets; and the bandwidth, which determines the number of samples or temporal resolution and was set to 95 for both datasets.

The next step is to segment the IFR spectrogram into temporal events of fixed length and interval. For dataset 1, approximately 16 events are obtained for each 2 second recording, whereas for dataset 2, approximately 80 events are obtained for each 10 second recording. We then obtain a set of event output vectors by summing the IFR spectrogram output within the temporal window of each segmented event. Thus, each event output vector consists of N values, where N is the number of frequency channels.

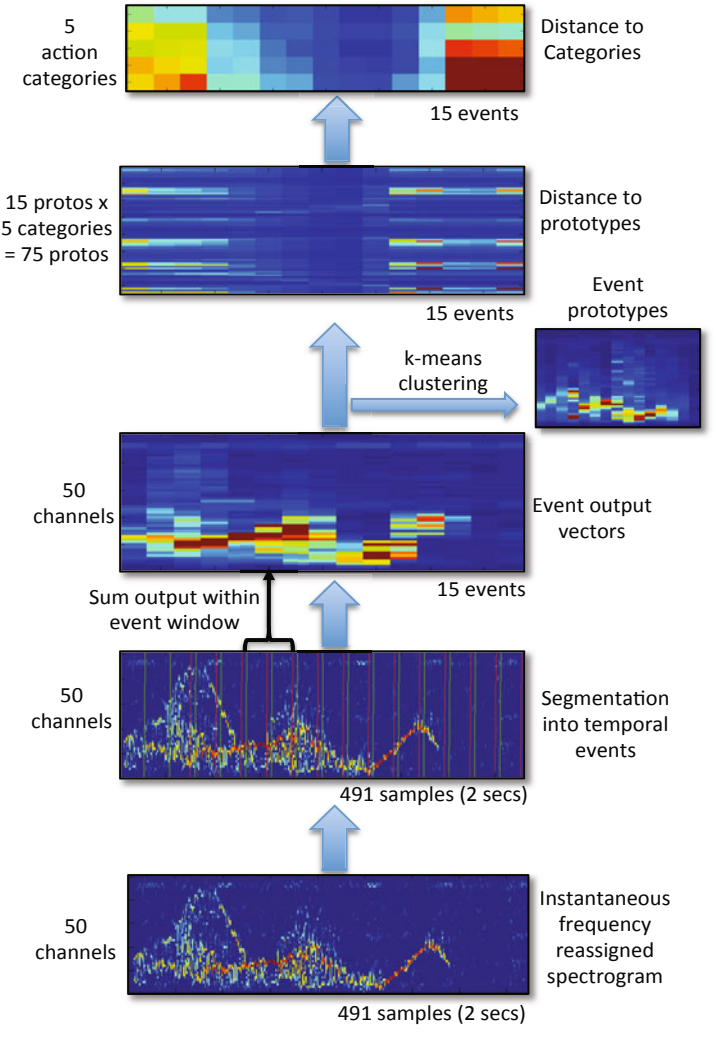


Fig. 2. Sequence of steps performed to process and categorize the microsonar data

In the categorization stage, a set of prototypes are learned from the event output vectors using k-means clustering. Note that k-means is applied to the events of each category independently and not to all events of all categories. This ensures that there are a set of prototypes strongly representative of each category and that, within each category, the prototypes are relatively uncorrelated.

During testing, the nearest neighbour approach is used to determine the winner category. Each data file or sample is composed of several test events. The Euclidean distance between each of these test events and all the prototypes of all categories is calculated. The winner category (action) for each data file is given by the average distance between its events and the action prototypes. Comparing all the test events to all the action prototypes has the advantage of showing some invariance with respect to the temporal position of the test event.

Half of the data files are randomly selected for training and the other half for testing. This procedure is repeated 30 times (cross-validation), and results averaged, in order to avoid any bias during the random file selection. This approach, which enables trials from the same subject to be present in both the training and testing datasets, was chosen instead of one-subject-out cross validation due to the high inter-trial variability of the data. In order to maximize the number of trials used in the training and testing datasets, the use of a validation dataset was discarded; instead the test results are presented directly as a function of the model parameters.

3 Results

An example of the prototypes obtained for each action category is shown in Figure 3 and action categorization results for datasets 1 and 2 as a function of the number of k-means clusters (prototypes) per category are shown in Figure 4. The optimum number of clusters for dataset 1 is 17, reaching 95.5% accuracy; and for dataset 2, 55 clusters with 95.3% accuracy.

The action categorization results for datasets 1 and 2 as a function of the number of test events used per action are shown in Figures 5. An accuracy of approximately 90% or above was achieved with only 4 out of 17 incoming events for dataset 1, and only 20 out of 70 events for dataset 2.

4 Discussion

The present study demonstrates the excellent performance of the proposed ultrasonar device and categorization model, which have achieved an accuracy above 95% for two different datasets composed of micro-Doppler signatures of human actions. Thus, our results outperform previously published results, which reported a categorization accuracy of 80% for dataset 1 [3], as well as similar action categorization studies, such as [7], which reported an 88.4% for an 8-category dataset. Importantly, all results are obtained using a 50% training ratio, which means the model is able to generalize to new data.

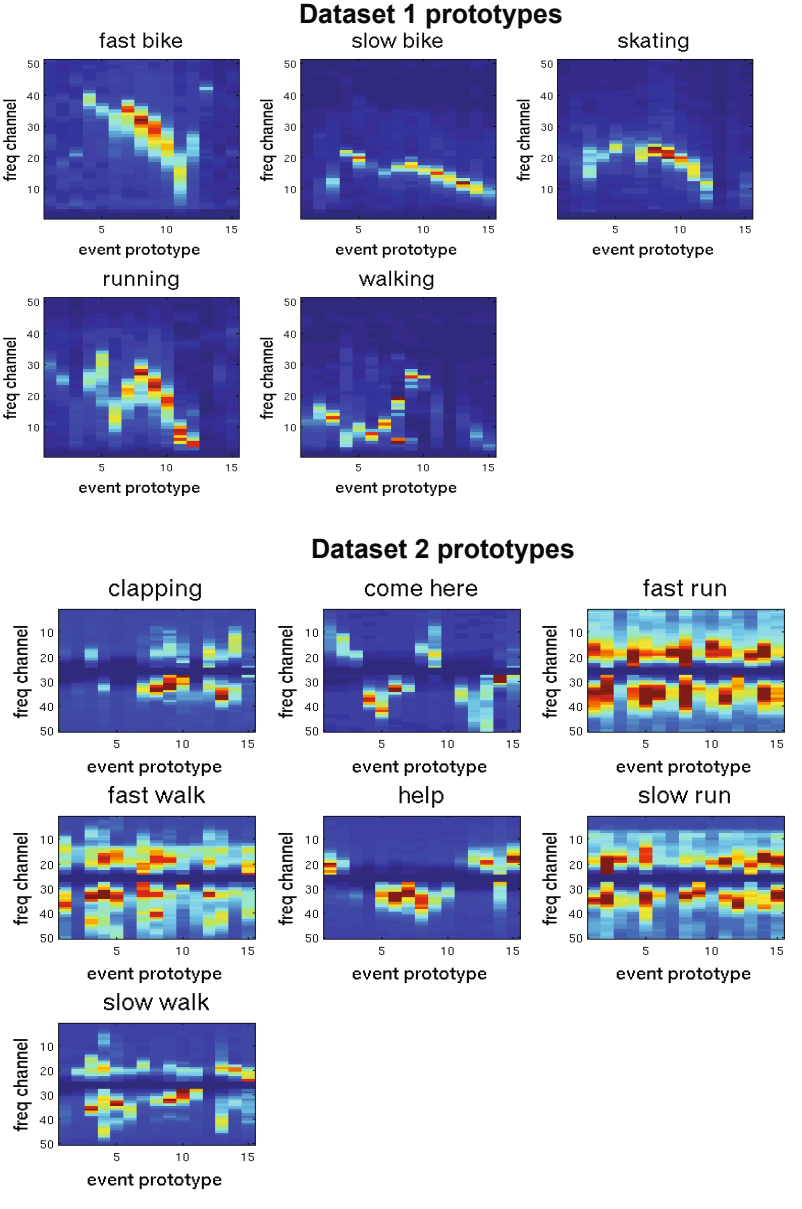


Fig. 3. Some action prototypes for dataset 1 (top) and dataset 2 (bottom)

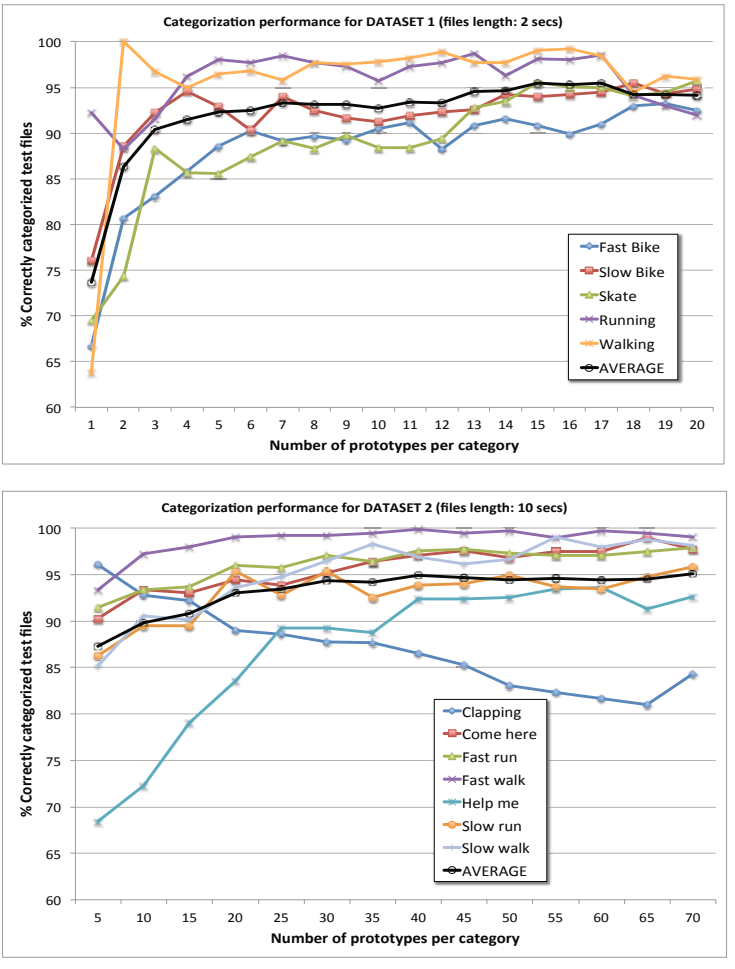


Fig. 4. Action categorization for dataset 1 (top) and dataset 2 (bottom) as a function of the number of prototypes per category

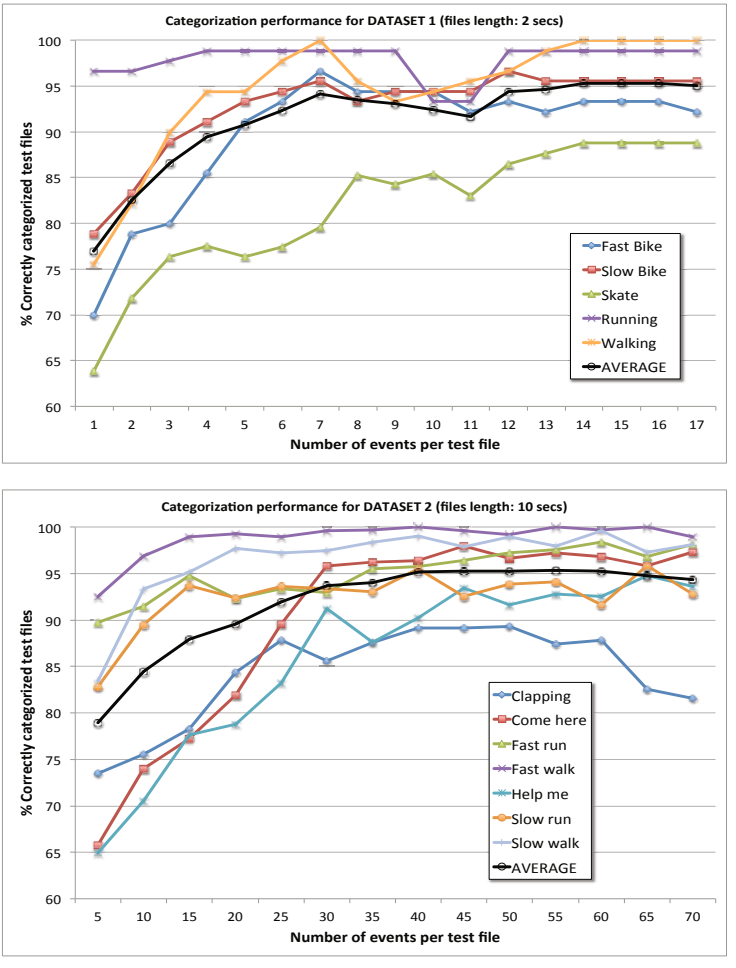


Fig. 5. Action categorization for dataset 1 (top) and dataset 2 (bottom) as a function of the number of test events used per action. Longer actions improve recognition.

In dataset 2 the accuracy of all categories was close to 100% except for 'clapping' at only 85%. This can be explained, first because the ultrasonar reflected wave was least affected by the relatively small clapping motion, and secondly, because it had the highest inter-trial variability - people clapped at many different frequencies and with different hand positions. Taking this into account, the 85%-result seems to be in consonance with the high performance of the model.

Another interesting result is that the number of prototypes can be significantly reduced compared to the number of events per action. This keeps categorization performance high ($> 90\%$ with only 3 prototypes in dataset 1) and suggests that the individual periodic components of each action, like steps in the walking condition, are captured by the prototypes. Therefore only a small number, corresponding to single cycles of an action, are required. The length of the prototypes, approximately 200 ms, is consistent with this interpretation.

In line with the previous result, reducing the number of test events per action-file also does not have a significant effect on the performance (above 90% accuracy with less than 25% of the test events). This again suggests only a small number of events, corresponding to one cycle, are required to characterize each action. The two previous results argue for a potential continuous time implementation of the model where incoming events gradually contribute to form a belief of the current action, by reinforcing or weakening the previous hypothesis.

As indicated above, the model is very robust to changes in the number of prototypes and incoming events. Further results not shown here for space limitations also demonstrate the robustness of the model to the number of IFR spectrogram channels ($< 5\%$ variation for values between 25 and 200) and bandwidth ($< 3\%$ variation for values between 25 and 200). The optimum number of channels seems to provide a compromise between the model's selectivity and invariance in the frequency dimension.

Similarly, the model also showed a strong robustness to variations of the event length ($< 3\%$ variation for values between 140 ms and 500 ms) and event interval ($< 3\%$ variation for values between 140 ms and 500 ms). The optimum value of the event length and interval provides a compromise between the model's selectivity and invariance in the temporal domain.

Finally, an important aspect to highlight is that these results are achieved using a compact hardware system characterized by low processing power and low cost. Correspondingly, the software model is computationally efficient and can be run in real-time once the system has been trained. The instantaneous frequency spectrogram provides a less noisy representation than standard spectrograms and requires less temporal samples. The segmentation method facilitates the learning of representative prototypes and permits the continuous categorization of temporally short incoming events. Overall, the system may have a wide range of applications in the context of security, surveillance and human behavior monitoring.

Future lines of research are intended to explore the applicability of the sensor to real-life scenarios. In this sense, future experiments will be developed to evaluate aspects such as the distance limits of the system, specially in outdoor

conditions, and the effects on accuracy of the angle deviation from the training condition. Furthermore, other classication approaches, based on support vector machines and hierarchical architectures, are being developed to improve recognition performance and robustness in future, more demanding real-life scenarios.

References

1. Dror, I., Florer, F., Rios, D., Zagaeski, M.: Using artificial bat sonar neural networks for complex pattern recognition: Recognizing faces and the speed of a moving target. Biol. Cyber. 74, 331–338 (1996)
2. Fulopa, S., Fitz, K.: Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. J. Acoust. Soc. Am. 119, 360–371 (2006)
3. Garreau, G., Nicolaou, N., Andreou, C., D'Urbal, C., Stuarts, G., Georgiou, J.: Computationally efficient classification of human computationally efficient classification of human transport mode using micro-Doppler signatures. In: 45th Annual Conference on Information Sciences and Systems, CISS (2011), doi:10.1109/CISS.2011.5766136
4. Geisheimer, J., Marshall, W., Greneker, E.: A continuous-wave radar for gait analysis. In: 35th Asilomar Conference on Signals, Systems and Computers, vol. 1, pp. 834–838 (2001)
5. Kalgaonkar, K., Raj, B.: Acoustic Doppler sonar for gait recognition. In: IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 27–32 (2007)
6. Kalgaonkar, K., Raj, B.: Recognizing talking faces from acoustic Doppler reflections. In: 8th IEEE International Conference on Automatic Face Gesture Recognition (FG), pp. 1–6 (2008)
7. Kalgaonkar, K., Raj, B.: One-handed gesture recognition using ultrasonic Doppler sonar. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1889–1892 (2009)
8. Liang, L., Mihail, P., Marylin, R., Marjorie, S., Paul, C., Tarik, Y.: Automatic fall detection based on Doppler radar motion signature. In: 5th International Conference on Pervasive Computing Technologies for Healthcare (2011)
9. Miao, Z., Ji, W., Xu, Y., Yang, J.: A novel ultrasonic sensing based human face recognition. In: IEEE Ultrasonics Symposium (IUS), pp. 1873–1876 (2008)
10. Otero, M.: Application of a continuous wave radar for human gait recognition. In: Proc. SPIE 5809, vol. 538 (2005), doi:10.1117/12.607176
11. Sabatini, A., Colla, V.: A method for sonar based recognition of walking people. Robotics and Autonomous Systems 25(1-2), 117–126 (1998)
12. Thaler, L., Arnott, S., Goodale, M.: Neural correlates of natural human echolocation in early and late blind echolocation experts. PLoS One 6(5), e20162 (2011)
13. Yardibi, T., Cuddihy, P., Genc, S., Bufi, C., Skubic, M., Rantz, M., Liu, L., Phillips, C.: Gait characterization via pulse-Doppler radar. In: IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), pp. 662–667 (2011)
14. Zhang, Z., Andreou, A.: Human identification experiments using acoustic micro-Doppler signatures. In: Argentine School of Micro-Nanoelectronics, Technology and Applications (EAMTA), pp. 81–86 (2008)

Sequential Deep Learning for Human Action Recognition

Moez Baccouche[1,2], Franck Mamalet[1],
Christian Wolf[2], Christophe Garcia[2], and Atilla Baskurt[2]

[1] Orange Labs, 4 rue du Clos Courtel, 35510 Cesson-Sévigné, France
firstname.surname@orange-ftgroup.com
[2] LIRIS, UMR 5205 CNRS, INSA-Lyon, F-69621, France
firstname.surname@liris.cnrs.fr

Abstract. We propose in this paper a fully automated deep model, which learns to classify human actions without using any prior knowledge. The first step of our scheme, based on the extension of *Convolutional Neural Networks* to 3D, automatically learns spatio-temporal features. A *Recurrent Neural Network* is then trained to classify each sequence considering the temporal evolution of the learned features for each timestep. Experimental results on the KTH dataset show that the proposed approach outperforms existing deep models, and gives comparable results with the best related works.

Keywords: Human action recognition, deep models, 3D convolutional neural networks, long short-term memory, KTH human actions dataset.

1 Introduction and Related Work

Automatic understanding of human behaviour and its interaction with his environment have been an active research area in the last years due to its potential application in a variety of domains. To achieve such a challenging task, several research fields focus on modeling human behaviour under its multiple facets (emotions, relational attitudes, actions, etc.). In this context, recognizing the behaviour of a person appears to be crucial when interpreting complex actions. Thus, a great interest has been granted to human action recognition, especially in real-world environments.

Among the most popular state-of-the-art methods for human action recognition, we can mention those proposed by Laptev et al. [13], Dollar et al. [3] and others [12,17,2,4], which all use engineered motion and texture descriptors calculated around spatio-temporal interest points, which are manually engineered. The *Harris-3D* detector [13] and the *Cuboid* detector [3] are likely the most used space-time salient points detectors in the literature. Nevertheless, even if their extraction process is fully automated, these so-called *hand-crafted* features are especially designed to be optimal for a specific task. Thus, despite their high performances, these approaches main drawback is that they are highly problem dependent.

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 29–39, 2011.

In last years, there has been a growing interest in approaches, so-called *deep models*, that can learn multiple layers of feature hierarchies and automatically build high-level representations of the raw input. They are thereby more generic since the feature construction process is fully automated. One of the most used deep models is the *Convolutional Neural Network* architecture [14,15], hereafter ConvNets, which is a bioinspired hierarchical multilayered neural network able to learn visual patterns directly from the image pixels without any pre-processing step. If ConvNets were shown to yield very competitive performances in many image processing tasks, their extension to the video case is still an open issue, and, so far, the few attempts either make no use of the motion information [20], or operate on hand-crafted inputs (spatio-temporal outer boundaries volume in [11] or hand-wired combination of multiple input channels in [10]). In addition, since these models take as input a small number of consecutive frames (typically less than 15), they are trained to assign a vector of features (and a label) to short sub-sequences and not to the entire sequence. Thus, even if the learned features, taken individually, contains temporal information, their evolution over time is completely ignored. Though, we have shown in our previous work [1] that such information does help discriminating between actions, and is particularly usable by a category of learning machines, adapted to sequential data, namely *Long Short-Term Memory* recurrent neural networks (LSTM) [6].

In this paper, we propose a two-steps neural-based deep model for human action recognition. The first part of the model, based on the extension of ConvNets to 3D case, automatically learns spatio-temporal features. Then, the second step consists in using these learned features to train a recurrent neural network model in order to classify the entire sequence. We evaluate the performances on the KTH dataset [24], taking particular care to follow the evaluation protocol recommendations discussed in [4]. We show that, without using the LSTM classifier, we obtain comparable results with other deep models based approaches [9,26,10]. We also demonstrate that the introduction of the LSTM classification leads to significant performance improvement, reaching average accuracies among the best related results.

The rest of the paper is organized as follows. Section 2 outlines some ConvNets fundamentals and the feature learning process. We present in Section 3 the recurrent neural scheme for entire sequence labelling. Finally, experimental results, carried out on the KTH dataset, will be presented in Section 4.

2 Deep Learning of Spatio-Temporal Features

In this section, we describe the first part of our neural recognition scheme. We first present some fundamentals of 2D-ConvNets, and then discuss their extension in 3D and describe the proposed architecture.

2.1 Convolutional Neural Networks (ConvNets)

Despite their generic nature, deep models were not used in many applications until the late nineties because of their inability to treat "real world" data.

Indeed, early deep architectures dealt only with 1-D data or small 2D-patches. The main problem was that the input was "fully connected" to the model, and thus the number of free parameters was directly related to the input dimension, making these approaches inappropriate to handle "pictoral" inputs (natural images, videos...).

Therefore, the convolutional architecture was introduced by LeCun et al. [14,15] to alleviate this problem. ConvNets are the adaptation of multilayered neural deep architectures to deal with real world data. This is done by the use of local *receptive fields* whose parameters are forced to be identical for all its possible locations, a principle called *weight sharing*. Schematically, LeCun's ConvNet architecture [14,15] is a succession of layers alternating 2D-convolutions (to capture salient information) and sub-samplings (to reduce dimension), both with trainable weights. Jarret et al. [8] have recommended the use of rectification layers (which simply apply absolute value to its input) after each convolution, which was shown to significantly improve performances, when input data is normalized.

In the next sub-section, we examine the adaptation of ConvNets to video processing, and describe the 3D-ConvNets architecture that we used in our experiments on the KTH dataset.

2.2 Automated Space-Time Feature Construction with 3D-ConvNets

The extension from 2D to 3D in terms of architecture is straightforward since 2D convolutions are simply replaced by 3D ones, to handle video inputs. Our proposed architecture, illustrated in Figure 1, also uses 3D convolutions, but is different from [11] and [10] in the fact that it uses only raw inputs.

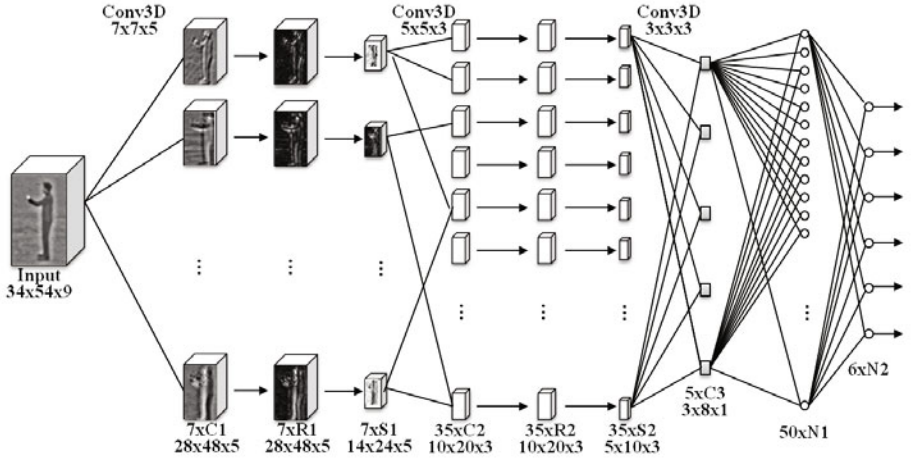


Fig. 1. Our 3D-ConvNet architecture for spatio-temporal features construction

This architecture consists of 10 layers including the input. There are two alternating convolutional, rectification and sub-sampling layers C1, R1, S1 and C2, R2, S2 followed by a third convolution layer C3 and two neuron layers N1 and N2. The size of the 3D input layer is $34 \times 54 \times 9$, corresponding to 9 successive frames of 34×54 pixels each. Layer C1 is composed of 7 feature maps of size $28 \times 48 \times 5$ pixels. Each unit in each feature map is connected to a 3D $7 \times 7 \times 5$ neighborhood into the input retina. Layer R1 is composed of 7 feature maps, each connected to one feature map in C1, and simply applies absolute value to its input. Layer S1 is composed of 7 feature maps of size $14 \times 24 \times 5$, each connected to one feature map in R1. S1 performs sub-sampling at a factor of 2 in spatial domain, aiming to build robustness to small spatial distortions. The connection scheme between layers S1 and C2 follows the same principle described in [5], so that, C2 layer has 35 feature maps performing $5 \times 5 \times 3$ convolutions. Layers R2 and S2 follow the same principle described above for R1 and S1. Finally, layer C3 consists of 5 feature maps fully-connected to S2 and performing $3 \times 3 \times 3$ convolutions. At this stage, each C3 feature map contains $3 \times 8 \times 1$ values, and thus, the input information is encoded in a vector of size 120. This vector can be interpreted as a descriptor of the salient spatio-temporal information extracted from the input. Finally, layers N1 and N2 contain a classical multilayer perceptron with one neuron per action in the output layer. This architecture corresponds to a total of 17,169 trainable parameters (which is about 15 times less than the architecture used in [10]). To train this model, we used the algorithm proposed in [14], which is the standard *online Backpropagation with momentum* algorithm, adapted *to weight sharing.*

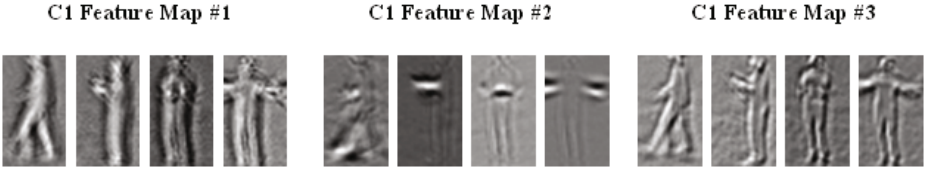


Fig. 2. A subset of 3 automatically constructed C1 feature maps (of 7 total), each one corresponding, from left to right, to the actions walking, boxing, hand-claping and hand-waving from the KTH dataset

Once the 3D-ConvNet is trained on KTH actions, and since the spatio-temporal feature construction process is fully automated, it's interesting to examine if the learned features are visually interpretable. We report in Figure 2 a subset of learned C1 feature maps, corresponding each to some actions from the KTH dataset. Even if finding a direct link with engineered features is not straightforward (and not necessarily required) the learned feature maps seem to capture visually relevant information (person/background segmentation, limbs involved during the action, edge information...).

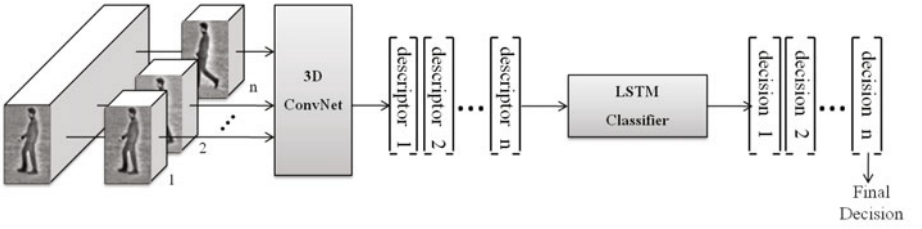


Fig. 3. An overview of our two-steps neural recognition scheme

In the next section, we describe how these features are used to feed a recurrent neural network classifier, which is trained to recognize the actions based on the temporal evolution of features.

3 Sequence Labelling Considering the Temporal Evolution of Learned Features

Once the features are automatically constructed with the 3D-ConvNet architecture as described in Section 2, we propose to learn to label the entire sequence based on the accumulation of several individual decisions corresponding each to a small temporal neighbourhood which was involved during the 3D-ConvNets learning process (see Figure 3). This allows to take advantage of the temporal evolution of the features, in comparison with the majority voting process on the individual decisions.

Among state of the art learning machines, Recurrent Neural Networks (RNN) are one of the most used for temporal analysis of data, because of their ability to take into account the context using recurrent connections in the hidden layers. It has been demonstrated in [6] that if RNN are able to learn tasks which involve short time lags between inputs and corresponding teacher signals, this *short-term* memory becomes insufficient when dealing with "real world" sequence processing, e.g video sequences. In order to alleviate this problem, Schmidhuber et al. [6] have proposed a specific recurrent architecture, namely *Long Short-Term Memory* (LSTM). These networks use a special node called *Constant Error Carousel* (CEC), that allows for constant error signal propagation through time. The second key idea in LSTM is the use of multiplicative gates to control the access to the CEC. We have shown in our previous work [1] that LSTM are efficient to label sequences of descriptors corresponding to hand-crafted features.

In order to classify the action sequences, we propose to use a *Recurrent Neural Network* architecture with one hidden layer of LSTM cells. The input layer of this RNN consists in 120 C3 output values per time step. LSTM cells are fully connected to these inputs and have also recurrent connexions with all the LSTM cells. Output layer consists in neurons connected to LSTM outputs at each time step. We have tested several network configuration, varying the number of hidden LSTM. A configuration of 50 LSTM was found to be a good compromise for

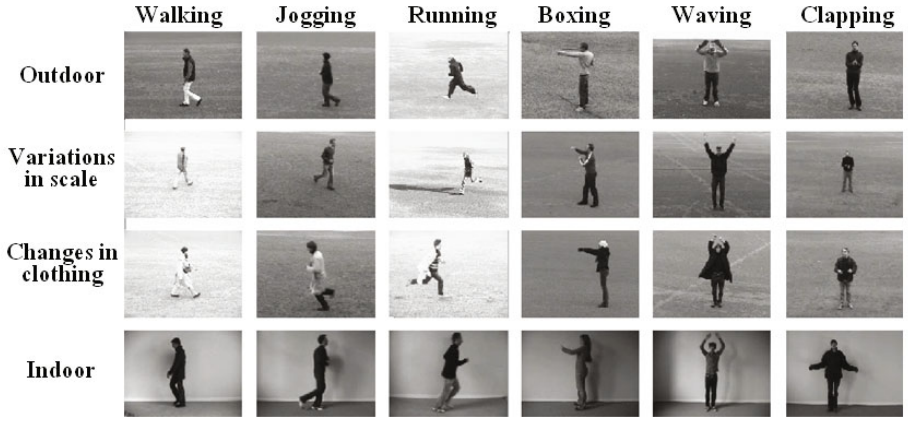


Fig. 4. A sample of actions/scenarios from the KTH dataset [24]

this classification task. This architecture corresponds to about 25,000 trainable parameters. The network was trained with *online backpropagation through time with momentum* [6].

4 Experiments on KTH Dataset

The KTH dataset was provided by Schuldt et al. [24] in 2004 and is the most commonly used public human actions dataset. It contains 6 types of actions (walking, jogging, running,boxing, hand-waving and hand-clapping) performed by 25 subjects in 4 different scenarios including indoor, outdoor, changes in clothing and variations in scale (see Figure 4). The image size is of 160×120 pixels, and temporal resolution is of 25 frames per second. There are considerable variations in duration and viewpoint. All sequences were taken over homogeneous backgrounds, but hard shadows are present.

As in [4], we rename the KTH dataset in two ways: the first one (the original one) where each person performs the same action 3 or 4 times in the same video, is named KTH1 and contains 599 long sequences (with a length between 8 and 59 seconds) with several "empty" frames between action iterations. The second, named KTH2, is obtained by splitting videos in smaller ones where a person does an action only one time, and contains 2391 sequences (with a length between 1 and 14 seconds).

4.1 Evaluation Protocol

In [4], Gao et al. presented a comprehensive study on the influence of the evaluation protocol on the final results. It was shown that the use of different experimental configurations can lead to performance differences up to 9%.

Furthermore, authors demonstrated that the same method, when evaluated on KTH1 or KTH2 can have over 5.85% performance deviations. Action recognition methods are usually directly compared although they use different testing protocols or/and datasets (KTH1 or KTH2), which distorts the conclusions. In this paper, we choose to evaluate our method using cross-validation, in which 16 randomly-selected persons are used for training, and the other 9 for testing. Recognition performance corresponds to the average across 5 trials. Evaluations are performed on both KTH1 and KTH2.

4.2 Experimental Results

The two-steps model was trained as described above. Original videos underwent the following steps: spatial down-sampling by a factor of 2 horizontally and vertically to reduce the memory requirement, extracting the person-centred bounding box as in [9,10], and applying 3D *Local Contrast Normalization* on a $7 \times 7 \times 7$ neighbourhood, as recommended in [8]. Note that we do not use any complex pre-processing (optical flow, gradients, motion history...). We also generated vertically flipped and mirrored versions of each training sample to increase the number of examples. In our experiments, we observed that, both for 3D-ConvNets and LSTM, no overtraining is observed without any validation sequence and stopping when performances on training set no longer rise. Obtained results, corresponding to 5 randomly selected training/test configurations are reported on Table 1.

Table 1. Summary of experimental results using 5 randomly selected configurations from KTH1 and KTH2

		Config.1	Config.2	Config.3	Config.4	Config.5	Average
KTH1	3D-ConvNet + Voting	90.79	90.24	91.42	91.17	91.62	91.04
	3D-ConvNet + LSTM	92.69	96.55	94.25	93.55	94.93	**94.39**
KTH2	3D-ConvNet + Voting	89.14	88.55	89.89	89.45	89.97	89.40
	3D-ConvNet + LSTM	91.50	94.64	90.47	91.31	92.97	**92.17**
	Harris-3D [13] + LSTM	84.87	90.64	88.32	90.12	84.95	87.78

The 3D-ConvNet, combined to majority voting on short sub-sequences, gives comparable results (91.04%) to other deep model based approaches [9,10,26]. We especially note that results with this simple non-sequential approach are almost the same than those obtained in [10], with a 15 times smaller 3D-ConvNet model, and without using neither gradients nor optical flow as input. We also notice that the first step of our model gives relatively stable results on the 5 configurations, compared to the fluctuations generally observed for the other methods [4]. The LSTM contribution is quite important, increasing performances of about 3%. KTH1 improvement ($+3,35\%$) is higher than KTH2, which confirms that LSTM are more suited for long sequences.

In order to point out the benefit of using automatically learned features, we also evaluated the combination of the LSTM classifier with common engineered space-time salient points. This was done by applying the *Harris-3D* [13] detector to each video sequence, and calculating the *HOF* descriptor (as recommended in [27] for KTH) around each detected point. We used the original implementation available on-line[1] and standard parameter settings. A LSTM classifier was then trained taking as input a temporally-ordered succession of descriptors. Obtained results, reported on Table 1, show that our learned 3D-ConvNet features, in addition to their generic nature, perform better on KTH2 than hand-crafted ones, with performances improvement of 4.39%.

To conclude, our two-steps sequence labelling scheme achieves an overall accuracy of 94.39% on KTH1 and 92.17% on KTH2. These results, and others among the best performing of related work on KTH dataset, are reported on Table 2.

Table 2. Obtained results and comparison with state-of-the-art on KTH dataset: methods reported in bold corresponds to deep models approaches, and the others to those using hand-crafted features

Dataset	Evaluation Protocol	Method	Accuracy
KTH1	Cross validation with 5 runs	**Our method**	**94.39**
		Jhuang et al. [9]	**91.70**
		Gao et al. [4]	95.04
		Schindler and Gool [23]	92.70
	Leave-one-out	Gao et al. [4]	96.33
		Chen and Hauptmann [2]	95.83
		Liu and Shah [17]	94.20
		Sun et al. [25]	94.0
		Niebles et al. [19]	81.50
KTH2	Cross validation with 5 runs	**Our method**	**92.17**
		Ji et al. [10]	**90.20**
		Gao et al. [4]	93.57
	Other protocols	**Taylor et al.** [26]	**90.00**
		Kim et al. [12]	95.33
		Ikizler et al. [7]	94.00
		Laptev et al. [13]	91.80
		Dollar et al. [3]	81.20

Table 2 shows that our approach outperforms all related deep model works [9,10,26], both on KTH1 and KTH2. One can notice that our recognition scheme outperforms the HMAX model, proposed by Jhaung et al. [9] although it is of hybrid nature, since low and mid level features are engineered and learned ones are constructed automatically at the very last stage.

[1] Available at `http://www.irisa.fr/vista/Equipe/People/Laptev/download.html`

For each dataset, Table 2 is divided into two groups: the first group consists of the methods which can be directly compared with ours, i.e those using the same evaluation protocol (which is cross validation with 5 randomly selected splits of the dataset into training and test). The second one includes the methods that use different protocols, and therefore those for whom the comparison is only indicative. Among the methods of the first group, to our knowledge, our method obtained the second best accuracy, both on KTH1 and KTH2, the best score being obtained by Gao et al. [4]. Note that the results in [4] corresponds to the average on the 5 best runs over 30 total, and that these classification rates decreases to 90.93% for KTH1 and 88.49% for KTH2 if averaging on the 5 worst ones.

More generally, our method gives comparable results with the best related work on KTH dataset, even with methods relying on engineered features, and those evaluated using protocols which was shown to outstandingly increase performances (e.g leave-one-out). This is a very promising result considering the fact that all the steps of our scheme are based on automatic learning, without the use of any prior knowledge.

5 Conclusion and Discussion

In this paper, we have presented a neural-based deep model to classify sequences of human actions, without a priori modeling, but only relying on automatic learning from training examples. Our two-steps scheme automatically learns spatio-temporal features and uses them to classify the entire sequences. Despite its fully automated nature, experimental results on the KTH dataset show that the proposed model gives competitive results, among the best of related work, both on KTH1 (94.39%) and KTH2 (92.17%).

As future work, we will investigate the possibility of using a single-step model, in which the 3D-ConvNet architecture described in this paper is directly connected to the LSTM sequence classifier. This could considerably reduce computation time, since the complete model is trained once. The main difficulty will be the adaptation of the training algorithm, especially when calculating the retro-propagated error.

Furthermore, even if KTH remains the most widely used dataset for human action recognition, recent works are increasingly interested by other more challenging datasets, which contains complex actions and realistic scenarios. Therefore, we plan to verify the genericity of our approach by testing it on recent challenging datasets, e.g Hollywood-2 dataset [18], UCF sports action dataset [21], YouTube action dataset [16], UT-Interaction dataset [22] or LIRIS human activities dataset[2]. This will allow us to confirm the benefit of the learning-based feature extraction process, since we expect to obtain stable performances on these datasets despite their high diversity, which is not the case of the approaches based on hand-crafted features.

[2] Available at `http://liris.cnrs.fr/voir/activities-dataset/`

References

1. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Action Classification in Soccer Videos with Long Short-Term Memory Recurrent Neural Networks. In: Diamantaras, K., Duch, W., Iliadis, L.S. (eds.) ICANN 2010. LNCS, vol. 6353, pp. 154–159. Springer, Heidelberg (2010)
2. Chen, M.y., Hauptmann, A.: MoSIFT: Recognizing human actions in. surveillance videos. Tech. Rep. CMU-CS-09-161, Carnegie Mellon University (2009)
3. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72 (2005)
4. Gao, Z., Chen, M.-y., Hauptmann, A.G., Cai, A.: Comparing Evaluation Protocols on the KTH Dataset. In: Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A. (eds.) HBU 2010. LNCS, vol. 6219, pp. 88–100. Springer, Heidelberg (2010)
5. Garcia, C., Delakis, M.: Convolutional face finder: a neural architecture for fast and robust face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(11), 1408–1423 (2004)
6. Gers, F.A., Schraudolph, N.N., Schmidhuber, J.: Learning precise timing with LSTM recurrent networks. Journal of Machine Learning Research 3, 115–143 (2003)
7. Ikizler, N., Cinbis, R., Duygulu, P.: Human action recognition with line and flow histograms. In: International Conference on Pattern Recognition, pp. 1–4 (2008)
8. Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y.: What is the best multi-stage architecture for object recognition? In: International Conference on Computer Vision, pp. 2146–2153 (2009)
9. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: International Conference on Computer Vision, pp. 1–8 (2007)
10. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. In: International Conference on Machine Learning, pp. 495–502 (2010)
11. Kim, H.J., Lee, J., Yang, H.S.: Human Action Recognition Using a Modified Convolutional Neural Network. In: Liu, D., Fei, S., Hou, Z., Zhang, H., Sun, C. (eds.) ISNN 2007. LNCS, vol. 4492, pp. 715–723. Springer, Heidelberg (2007)
12. Kim, T.K., Wong, S.F., Cipolla, R.: Tensor canonical correlation analysis for action classification. In: International Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
13. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: International Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
14. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
15. LeCun, Y., Kavukcuoglu, K., Farabet, C.: Convolutional networks and applications in vision. In: IEEE International Symposium on Circuits and Systems, pp. 253–256 (2010)
16. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos in the wild. In: International Conference on Computer Vision and Pattern Recognition, pp. 1996–2003 (2009)
17. Liu, J., Shah, M.: Learning human actions via information maximization. In: International Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
18. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: International Conference on Computer Vision and Pattern Recognition, pp. 2929–2936 (2009)

19. Niebles, J., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision 79, 299–318 (2008)
20. Ning, F., Delhomme, D., LeCun, Y., Piano, F., Bottou, L., Barbano, P.E.: Toward automatic phenotyping of developing embryos from videos. IEEE Transactions on Image Processing 14(9), 1360–1371 (2005)
21. Rodriguez, M.D., Ahmed, J., Shah, M.: Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In: Computer Vision and Pattern Recognition, pp. 1–8 (2008)
22. Ryoo, M., Aggarwal, J.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: International Conference on Computer Vision, pp. 1593–1600 (2009)
23. Schindler, K., van Gool, L.: Action snippets: How many frames does human action recognition require? In: International Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
24. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: International Conference on Pattern Recognition, vol. 3, pp. 32–36 (2004)
25. Sun, X., Chen, M., Hauptmann, A.: Action recognition via local descriptors and holistic features. In: International Conference on Computer Vision and Pattern Recognition Workshops, pp. 58–65 (2009)
26. Taylor, G.W., Fergus, R., LeCun, Y., Bregler, C.: Convolutional Learning of Spatio-temporal Features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6316, pp. 140–153. Springer, Heidelberg (2010)
27. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference (2009)

One-Sequence Learning of Human Actions

Carlos Orrite, Mario Rodríguez, and Miguel Montañés*

I3A, University of Zaragoza, Spain

Abstract. In this paper we address the problem of human action recognition from a single training sequence per class using a modified version of the Hidden Markov Model. Inspired by codebook approaches in object and scene categorization, we first construct a codebook of possible discrete observations by applying a clustering algorithm to all samples from all classes. The number of clusters defines the size of the codebook. Given a new observation, we assign to it a probability to belong to every cluster, i.e., to correspond to a discrete value of the codebook. In this sense, we change the 'winner takes all' rule in the discrete-observation HMM for a distributed probability of membership. It implies the modification of the Baum-Welch algorithm for training discrete HMM to be able to deal with fuzzy observations. We compare our approach with other models such as, dynamic time warping (DTW), continuous-observation HMM, Conditional Random Fields (CRF) and Hidden Conditional Random Fields (HCRF) for human action recognition.

1 Introduction

Human activity recognition is a challenging research field with a lot of applications such as surveillance or sport analysis. In video-surveillance the identification of unusual actions is one of the most relevant goals for security concerns and threat prevention. It is clear that in these situations, not so many training examples are available, so the necessity of learning having only few sequences become more relevant.

Markov models are a supervised approach to deal with temporal series. When the number of labelled sequences is low, the training process suffers from incomplete data. This problem becomes more significant when working with continuous observations since it usually involves the computation of a Mixture of Gaussians. For example, when training a HMM we find that the log-likelihood decreases during EM, meaning that the mixture components are not getting enough data.

In speech recognition one phoneme can be used more than once in a word so, using all the information data from the word can be useful for training a single state. The same can be applied to different words where they can share

* The authors would like to acknowledge the following institutions for their support: "Departamento de Ciencia, Tecnología y Universidad del Gobierno de Aragón", "Fondo Social Europeo" and "Ministerio de Ciencia e Innovación (TIN2010-20177)". Mario Rodríguez has got a FPI grant from the Ministerio de Ciencia e Innovación.

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 40–51, 2011.

several phonemes in common. When the number of data for training is low, the inclusion of this words can help to train a word-specific HMM, even when we are using features from other words. This idea can be applied to other fields such as human action recognition, where several actions can share the same body pose.

Inspired by codebook approaches in object and scene categorization, we first construct a codebook of possible discrete observations by applying a clustering algorithm to all samples from different classes. The number of clusters defines the size of the codebook. Given a new observation, we assign a probability to belong to any cluster, i.e., to correspond to a discrete value of the codebook. In this sense we change the 'winner takes all rule' in the discrete HMM for a distributed probability of membership. It implies the modification of the Baum-Welch algorithm for training HMM to be able to deal with fuzzy observation variables. This new algorithm is named Fuzzy Observation HMM (FO-HMM).

Fig. 1 shows an overview of the steps involved in our approach for human action recognition. Temporal templates consist in several consecutive Motion History Images (MHIs) per sequence. These templates are projected onto a reduced subspace by PCA. Afterwards, a clustering algorithm identifies the main clusters and set the codebook. The probability for every observation o_t to belong to all clusters is obtained. Then, a FO-HMM per class is trained. Once the model parameters are estimated, the maximum log-likelihood (ML) is used to find which class ω^* a test sequence belongs to.

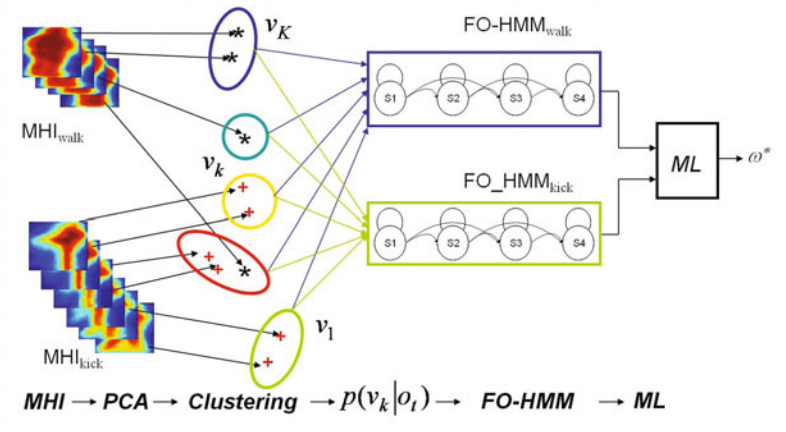


Fig. 1. Fuzzy observation HMM framework for action recognition

1.1 Previous Work

We assume a task where the goal is to predict a label ω^* from an input O. Each ω^* is a member of a set of possible labels and each vector O is a vector of local observations $O = \{o_1, o_2, \ldots, o_T\}$. The training set consists of labelled examples (O_i, ω_i) for $i = 1 \ldots M$, where each $\omega_i \in \Omega$, and each $O_i = \{o_{i,1}, o_{i,2}, \ldots, o_{i,T}\}$.

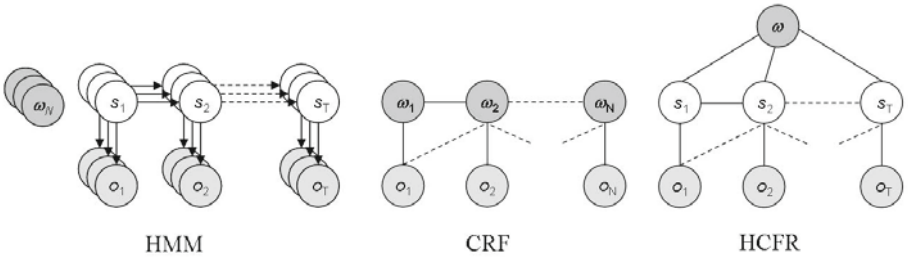


Fig. 2. Different models used for learning temporal series. ω represents the class label, o the observation, and s the hidden state labels. (Left) a 'stack of HMMs' where a separate HMM is trained for each class. (middle and right) CRF and HCRF models are depicted, in this case, only one model is trained for all classes.

This kind of task can be achieved by training a HMM model for each class respectively, as depicted in Fig. 2(left). Thus, for a given unknown sequence, the class label is assigned via Maximum Likelihood (ML).

This approach entails a HMM training process traditionally solved by using the expectation-maximization (EM) algorithm Baum-Welch. In [1] the authors tackle the problem of local maximum convergence in EM algorithm by combining the EM with an evolutionary algorithm. But, as they mention, this implies even more parameters to be estimated. So, there is another problem with EM approach when the number of states and observations grows, because the representations of the state and observation spaces require a great number of parameters to be estimated and therefore a large set of training data.

HMMs are generative models which assign a joint probability to paired observation and label sequences. The argument against the generative models is that observations are assumed to be independent given the values of the hidden variables [2]. This restriction makes it difficult, or not practical, to deal with long-range dependencies among observations or multiple interacting features of the observations. Conditional Random Fields (CRF), introduced in [3], are undirected graphical models that generalize the HMM by putting feature functions conditioned on the global observation in the place of transition probabilities. In a CRF, unlike in the HMM, abstraction feature selection does not have to be limited to the current observation, but it can also consider any combination of past and future observations. CRFs assign a label for each observation, and they neither capture hidden states nor directly provide a way to estimate the conditional probability of a class label for an entire sequence, as depicted in Fig. 2(center). In this figure we notice that the arrows used to link states to observations in HMM have been changed to some dashed lines meaning that one label can take into account combination of different temporal observations.

Hidden Conditional Random Fields (HCRF) was first introduced by Gunawardana et al. [4] for phone-conversation/speech classification and has later been applied to gesture and object recognition [2]. The authors argued that CRFs

are limited in the sense that they cannot capture intermediate structures using hidden-sate variables. HCFRs use intermediate hidden variables to model the latent structure of the input domain. Recently, Zhang and Gong [5] have presented a method for action categorization with a modified HCRF.

Learning of CRF and HCRF parameters can be achieved by using optimization methods such as quasi-Newton gradient descent. However, this optimization method takes a longer time than the training of HMMs, becoming the major disadvantage of CRF and HCRF in relation to HMMs.

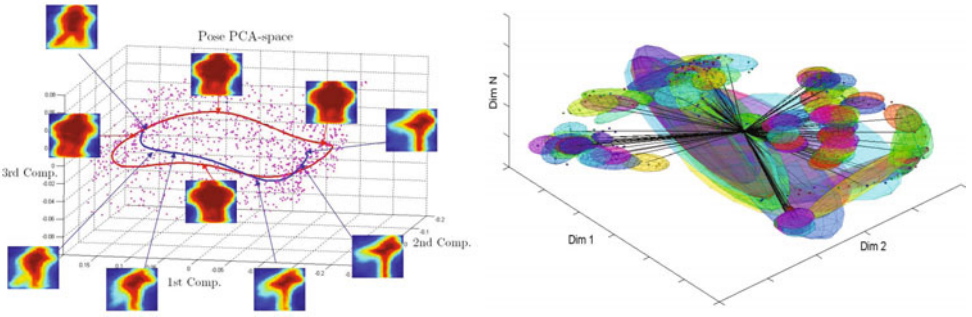
In previous work [6] the authors demonstrated that useful aspects of a new object category may be learned from a single training example per class (or just from a few). The key insight is that rather than learning from scratch one can take advantage of knowledge coming from previously learned categories, no matter how different these categories might be. In machine learning, this is known as transfer learning. The idea is to leverage the acquired knowledge from related tasks to aid in learning a new task. So this paper addresses the problem of how to use temporal observations coming from different categories to learn a new model, instead of using the only one-sequence (or just a few).

In [7] the authors carried out an attempt for human action recognition using only a single clip per action. They used a patch based motion descriptor and a method for learning a transferable distance function for these patches. The experimental results were quite impressive, but it seems difficult to generalize this approach to be used in other computer vision fields, since it does not provide any new model to deal with temporal sequences, as presented in this paper. More recent works deal with the same problematic scenario of one example per action. In [8] the authors use a novel feature representation and a detector of similar actions. Our experiments are comparable because of the usage of the same dataset. In [9] the training process uses a single video to create semi-synthetic videos manipulating an original human activity video and then the recognition system includes these videos in each training step.

Discretization and codebook construction is a common task in discrete HMM based algorithms. In [10] we find a framework composed by a posture labelling which feeds a combination of a Variable-Length Markov Model (VLMM) and a HMM. The posture labelling step consists of a codebook of posture templates creation. In this work, a human posture is represented by a silhouette image, and a shape matching process is used to assess the difference between two shapes. Moreover, the usage of spatio-temporal features to construct a vocabulary and then feed a recognition method is also common, like [11] [12] [13] do.

2 Fuzzy Observation Variables (FO)

Human silhouette extraction from videos is a well understood problem for current vision techniques when dealing with fixed cameras, uncluttered scenes and stable lighting conditions. Although there is more research needed to deal with more extensive real-world situations, most researchers in action recognition assume that the calculation of silhouettes is solved separately. So, the method presented



(a) Two sequences in PCA space **(b)** Clustering and distance in PCA space

Fig. 3. PCA space: PCA space for two human pose sequences (3a) and Euclidean distances from a temporal observation to the centres of the clusters (3b)

here directly relies on moving silhouettes. The input features used in this paper are based on Motion History Images (MHI) as introduced in [14]. MHIs capture motion information in images by encoding where motion occurred and the history of motion occurrences in the image. Given a video sequence corresponding to a specific action, several consecutive overlapping windows generate a stream of MHIs, every MHI calculated in a particular temporal window.

The feature extraction process provides a vast amount of data, making real time implementation challenging. An alternative to alleviate the dimensionality space consists of mapping MHIs to low dimensional manifolds, as for example by means of PCA.

Let us assume that we have M different sequences in PCA space denoted as $O_i = \{o_{i,1}, \cdots, o_{i,T_i}\}$, for $i = 1, \cdots, M$, corresponding to L different labels. Inspired by codebook approaches in object and scene categorization we first construct a codebook of possible discrete observations. Our approach executes this process by applying a clustering algorithm to all samples from all sequences $\mathcal{O} = \{O_1, \cdots, O_M\}$. The number of clusters defines the size of the codebook. This process is not a temporal clustering so, some samples corresponding to different time and label may belong to the same cluster. Fig. 3a represents a couple of sequences corresponding to two different human actions in a PCA space: "turn around" and "kick". Different temporal templates can be neighbours at the PCA space. Therefore, even if we have a low number of samples for action "kick", we are able to find a cluster for same temporal features thanks to those samples corresponding to a completely different action.

Once the codebook has been created it is possible to assign a belonging probability from a sample to every cluster as shown in Fig. 3b. Given a test sequence $O = \{o_1, \ldots o_T\}$, and a codebook $V = \{v_1, v_2, ..., v_K\}$ for any sample o_t we calculate the probability $p(v_k|o_t)$ for all $v_k, k = 1...K$ as:

$$p(v_k|o_t) = \frac{\mathcal{N}_k(o_t; \mu_k, \Sigma_k)}{\sum_{i=1}^{K} \mathcal{N}_i(o_t; \mu_i, \Sigma_i)} \tag{1}$$

where the denominator is a normalization factor. These probabilities are the observations we use in our algorithm, moreover we select as observation the winner word v_k from the codebook to assess a discrete HMM.

The number of clusters K defines the size of the codebook (i.e., the number of observation symbols). If K is too small it might not explain some important differences that should be detected in the temporal features. Conversely, if K is too big the differences could be too small. In the limit, when the number of training samples is too small, every sample constitutes a cluster v_k and calculate the probability of a new observation o_t to every sample (cluster) as:

$$p(v_k|o_t) = \frac{e^{-\|v_k, o_t\|^2}}{\sum_{i=1}^{K} e^{-\|v_i, o_t\|^2}} \tag{2}$$

where $\|\cdot\|$ denotes Euclidean distance.

3 Fuzzy Observation HMM (FO-HMM)

To deal with FOs in a discrete HMM framework some modifications have to be done for training and evaluation. Formally, the parameters of the Markov model are $\lambda = \{N, A, B, \pi\}$. N is the number of states, i.e., $S = \{S_1, \dots S_N\}$. $A = a_{ij}$ is the state transition matrix where the transition probability a_{ij} represents the frequency of transiting from state i to state j. $B = b_j(v_k)$ is the observation probability distribution where $b_j(v_k) = Pr(o_t = v_k \mid q_t = S_j)$ and o_t the observation at time t. Finally, $\pi = \{\pi_i\}$ is the initial state probability distribution where $\pi_i = Pr(q_1 = S_i)$, $1 \leq i \leq N$ being q_t the state at time t. The set of observation symbols is $V = \{v_1, v_2, ..., v_K\}$, where K is the number of observation symbols per state. The problem of estimating the parameters λ of a HMM given a sequence of observations $O = \{o_1, \dots o_T\}$ can be approach as a maximum likelihood (ML) problem. To solve this we have the Baum-Welch algorithm which is an EM procedure, estimating the parameters of a HMM in an iterative procedure [15]. The observation probability is calculated as:

$$\hat{b}_j(v_k) = \frac{\sum_{t=1}^{T} \gamma_t(j) \cdot \delta_{o_t, v_k}}{\sum_{t=1}^{T} \gamma_t(j)} \tag{3}$$

$$0 \leq \hat{b}_j(v_k); \text{ and } \sum_{k=1}^{K} \hat{b}_j(v_k) = 1 \tag{4}$$

where $\gamma_t(i) = Pr(q_t = S_i|O, \lambda)$ is an auxiliary probability. The quantity given in (3) is the expected number of times the output observations have been equal to v_k while in state j relative to the expected total number of times in state j. So, the term δ_{o_t, v_k} in that equation is equal to 1 when being in state j, the discrete observation o_t is exactly v_k and 0 otherwise.

If the number of temporal sequences to train the HMM is low in relation to the size of the observation symbols set V, then most of these symbols will never be reached. This constitutes a serious problem when evaluating a local observation

corresponding to one of these symbols, since the learned probability distribution is null. To avoid this problem we regularize the calculated probability values to a minimum threshold, so every value under the threshold acquires the threshold value, normalizing afterwards the probabilities. Nevertheless, due to the small training set, many symbols are not trained and that deteriorates the results. This problem becomes even worse when dealing with continuous observation since it implies the computation of Gaussian mixtures and we can get that the log-likelihood decreases during EM, meaning that the mixture components are not getting enough data.

In order to cope with this limitation we introduce a proposal to calculate the observation symbol probability density function (3) by the inclusion of the auxiliary probability function $p(v_k|o_t)$ which assigns observation probabilities to different elements of the codebook. It avoids the previous problem of non-trained probability distributions since all of the symbols in the codebook have been assigned during training. To work with fuzzy observations it is necessary to modify the Baum-Welch algorithm in a similar way as done before in [16] but in this paper we introduce a simpler and more intuitive way to accomplish it.

For the sake of clarity we briefly describe the modifications carried out to deal with fuzzy observations. Baum-Welch algorithm is a special case of the EM algorithm.

E-step: This step implies the calculation of functions $\xi_t(i,j)$ and $\gamma_t(i)$.

Let $\gamma_t(i)$ be a probability of being in state i at time t, given O and λ.

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{5}$$

Let $\xi_t(i,j)$ be a probability of being in state i at time t and at state j at time $t+1$, given O and λ.

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\Sigma_{i=1}^{N} \Sigma_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)} \tag{6}$$

where $\alpha_t(i) = Pr(O, q_t = S_i|\lambda)$ and $\beta_t(j) = Pr(O|\lambda, q_t = S_j)$ are auxiliary probabilities calculated by the forward and backward algorithms.

In (6) we introduce the first modification by computing:

$$b_j(o_{t+1}) = \sum_{k=1}^{K} p(v_k|o_{t+1}) b_j(v_k) \tag{7}$$

This implies the modification of the forward algorithm as follows:

Initialization

$$\alpha_1(i) = \pi_i \sum_{k=1}^{K} p(v_k|o_1) b_j(v_k) \tag{8}$$

Iteration

$$\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i) a_{ij} \cdot \Big(\sum_{k=1}^{K} p(v_k|o_t) b_j(v_k) \Big), t = 2, \ldots, T \tag{9}$$

where N is the number of states.

The same can be applied to the backward algorithm.

M-step: This is an iterative process involving the calculation of $\hat{\pi}_i$, $\hat{a}_{ij}$ and $\hat{b}_j(v_k)$. In this step we introduce the second modification to the Baum-Welch algorithm calculating the expected observation matrix as:

$$\hat{b}_j(v_k) = \frac{\sum_{t=1}^{T} \gamma_t(j) \cdot p(v_k|o_t)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{10}$$

In this sense, we change the winner takes all rule, given by δ_{o_t,v_k} in (3), for a distributed probability of membership, given by $p(v_k|o_t)$ in (10).

The convergence of the Baum-Welch algorithm to a local minimum is guaranteed as long as (4) is fulfilled, so:

$$\sum_{k=1}^{K} p(v_k|o_t) = 1 \tag{11}$$

This requirement is accomplished thanks to the normalization introduced in (1).

In the recognition of a new sequence, the fuzzy observation is computed as done in the learning process. These FOs, in conjunction with model parameters λ, are used in the modified forward-backward algorithm to give a log likelihood.

4 Experiments

We conducted our experiments on a standard action recognition dataset, [17]. This data set is composed by 93 low-resolution (180 x 144, 50 fps) video sequences showing nine different people, each performing 10 natural actions: "bend", "jumping-jack", "jump-forward-on-two-legs", "jump-in-place-on-two-legs", "run", "gallop-side-ways", "skip", "walk", "wave-one-hand", "wave-two-hands". The authors provide the mask for all these actions obtained by background subtraction. The input features used in this paper for all methods are based on MHIs, as it is introduced in [14]. Temporal patterns are obtained generating consecutive MHIs of five frames. To reduce the image dimensionality we project these patterns onto a subspace via PCA.

The models we compared include:

DTW: Given two time series, $Q = \{q_1, \ldots, q_n\}$ and $R = \{r_1, \ldots, r_m\}$, DTW aligns the two series so that their difference is minimized. DTW provides a way to align both temporal series by obtaining a warping path that has the minimum distance between both series. At the same time DTW provides a way to quantify the goodness of the matching by means of an accumulative cost along the warping

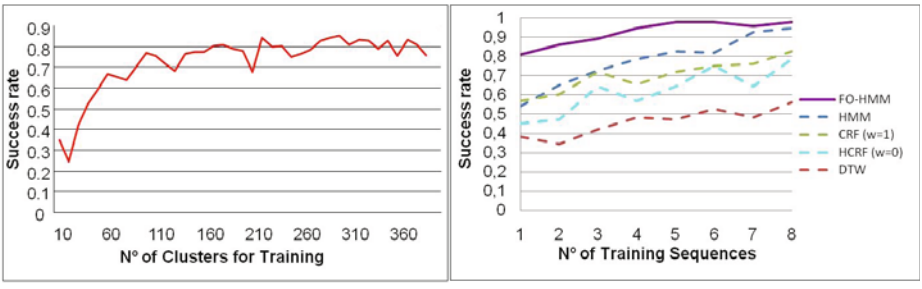


Fig. 4. (Left) Average success rate (over 93 test sequences of the Weizmann Database) provided by FO-HMM approach with a single training sequence per class for different number K of clusters. (Rigth) Average success rate for all methods obtained by increasing the number of training sequences.

path. The label for an unknown sequence O is estimated by assigning the label of the training sequence with the lowest accumulative cost.

HMM: We trained a discrete HMM model for each class respectively, denoted as HMM_ω. We regularized the calculated probability values of the HMM by including a threshold related to the number of words in the codebook by $th = 1/(100K)$ where K is the number of clusters. Once we have estimated the model parameters λ_ω for every class the label for an unknown sequence O is assigned via $\omega^* = argmax_\omega p(\text{HMM}_\omega \mid O; \lambda_\omega)$.

CRF: We trained a single CRF model where every action class has a corresponding state. In this case, the CRF predicts labels for each frame in a sequence, not the entire sequence. Therefore, during evaluation the Viterbi path is found, and assigned the sequence label based on the most frequently occurring action label per frame. As it can be seen in Table 1, different long range dependencies, denoted by w are tested. This parameter w accounts for the amount of past and future history to be used when predicting the state at time t.

HCRF: We trained a single HCFR for all actions. Once we have estimated the model parameters Θ, the test of a new sequence is carried out by maximizing the posterior probability of the learnt model, i.e., $\omega^* = argmax_\omega p(\omega \mid O; \Theta)$. We also conducted experiments that incorporated different long range dependencies in the same way as described before.

FO-HMM: We trained a FO-HMM model for each class respectively denoted as FO-HMM_ω. Once we have estimated the model parameters λ_ω, for all class, the label for an unknown sequence O is assigned via $\omega^* = argmax_\omega p(\text{FO-HMM}_\omega \mid O; \lambda_\omega)$.

We ran several experiments with different number of clusters, starting from a single training sequence per class. The number of temporal templates in different trials were around 400, so the maximum number K of clusters is limited by this amount. The results were obtained applying the *leave-one-out* rule. The methodology was tested over the 93 sequences, each time one sequence was

selected to test and n training sequences per class were selected randomly from the 92 available. n went from 1 to 8 (maximum number of available sequences for some actions). Figure 4 (left) shows the average success rate over 93 testing sequences using a single training sequence per class. It seems that when the number of clusters is higher so is the average of success rate. The maximum recognition rate was 84.3% for $K = 290$ which outperform the results on [8] for one training sequence where they obtained around 75%.

In order to introduce a FO-HMM proposal free of any parameter and independent of any cluster process we carried out some new experiments where the number of clusters K was exactly the number of samples. For these experiments the fuzzy observation was assigned by (2). Figure 4 (right) shows the average success rate over 93 testing sequence of our method in relation to continuous HMM, DTW, CRF and HCRF, for different number of training sequences (from 1 to 8). Our method converges to an average accuracy maximum value of 98% with only 5 training sequences. In a recent work using all the sequences from the same dataset and a modified HCRF [5], the authors reached an accuracy of 89.3%. As it can be noticed in Figure 4 (right), no other method outperforms FO-HMM using up to 8 sequences. It is worth noticing that HMM exhibits the best improvement when the number of sequence increases, so it might reach a similar performance as FO-HMM with more sequences.

Focusing our experiments on using one training sequence per class, Table 1 shows the average accuracy for all methods, where the CRF and HCRF have been trained with different values for the long rage dependencies parameter w. As mentioned before, this parameter takes into account the amount of past and future samples used when predicting the state at a particular time. Clearly our approach provides the best result with just one training sequence per class, i.e., 81.11%, followed by CRF (w=1) with 56.99%.

Table 1. Comparisons of recognition performance (average percentage accuracy) for action recognition using only 1 sequence for training

Models	DTW	HMM	CRF (w=0)	CRF (w=1)	CRF (w=2)	HCRF w=0)	HCRF (w=1)	HCRF (w=2)	FO-HMM
Avg.(%)	38.4	54.2	54.8	57.0	55.9	45.1	40.8	34.4	81.1

5 Conclusions

This paper has addressed the problem of supervised learning of human actions with a lack of annotated training data. The main contribution of this paper has been the novel proposal for data observation called fuzzy observation (FO). It has implied the modification of the Baum-Welch algorithm to work with this kind of observation, given as a result the FO-HMM. This approach has allowed us to take into account the whole set of possible discrete observations with some likelihood assigned to any of them. This raises another relevant area for

further research, the possibility to deal with unfeasible observations, those that are doubtful due to noise or a bad working feature extraction.

Our model has provided some encouraging results when applied to human activity recognition by using only one sequence per class for training. The performance of the recognition is slightly higher to other state-of-the-art results with just one sequence per class in the training step and comparing to other works that uses all the sequences available is slightly lower. Using all the sequences available to feed our method, the performance is then comparable to the state-of-the-art results. However, some questions are still open and they require a further research. For instance, the number of clusters used to model the data space is an important parameter which has not been solved properly. On the other hand, the step previous to clustering has consisted on a PCA projection for the temporal templates (MHIs). However, there are other approaches, such as LDA, kernel PCA or support vector machines (SVMs), to categorize the input features that should be investigated.

Finally, other research line under study is the possibility of using fuzzy observations in a CRF or HCRF model since in the previous examples we worked with continuous observation values corresponding to the raw data.

References

1. Huda, S., Yearwood, J., Togneri, R.: A Constraint-Based Evolutionary Learning Approach to the Expectation Maximization for Optimal Estimation of the Hidden Markov Model for Speech Signal Modelling. IEEE Trans. Syst., Man, Cyber. B, Cybern. 39(1), 182–197 (2009)
2. Quattoni, A., Wang, S.B., Morency, L.-P., Collins, M., Darrell, T.: Hidden Conditional Random Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(10), 1848–1852 (2007)
3. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings 18th International Conf. on Machine Learning, pp. 282–289 (2001)
4. Gunawardana, A., Mahajan, M., Acero, A., Platt, J.C.: Hidden conditional random fields for phone classification. In: International Conference on Speech Communication and Technology, pp. 1117–1120 (2005)
5. Zhang, J., Gong, S.: Action categorization with modified hidden conditional random field. Pattern Recognition 43(1), 197–203 (2010)
6. Fei-Fei, L., Fergus, R., Perona, P.: One-Shot Learning of Object Categories. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 594–611 (2006)
7. Yang, W., Wang, Y., Mori, G.: Human Action Recognition from a Single Clip per Action. In: 2nd International Workshop on Machine Learning for Vision-based Motion Analysis (2009)
8. Seo, H.J., Milanfar, P.: Action Recognition from One Example. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(5), 867–882 (2011)
9. Ryoo, M.S., Yu, W.: One Video is Sufficient? Human Activity Recognition Using Active Video Composition. In: WACV 2011 (2011)
10. Liang, Y.-M., Shih, S.-W., Shih, A.C.-C., Liao, H.-Y.M., Lin, C.-C.: Learning Atomic Human Actions Using Variable-Length Markov Models. IEEE Trans. Syst., Man, Cyber. B, Cybern. 39(1), 268–280 (2009)

11. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. IJCV 79(3) (September 2008)
12. Schuldt, C., Laptev, I., Caputo, B.: Recognizing Human Actions: a Local SVM Approach. In: ICPR 2004 (2004)
13. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior Recognition via Sparse Spatio- Temporal Features. In: IEEE Workshop on VS-PETS 2005 (2005)
14. Bobick, A.F., Davis, J.W.: The Recognition of Human Movement Using Temporal Templates. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(3), 257–267 (2001)
15. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 257–286 (1989)
16. Uguz, H., Ozturk, A., Saracoglu, R., Arslan, A.: A biomedical system based on fuzzy discrete hidden Markov model for the diagnosis of the brain diseases. Expert Systems with Applications 35, 1104–1114 (2008)
17. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as Space-Time Shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(12), 2247–2253 (2007)

Analyzing Facial Behavioral Features from Videos

Abdenour Hadid

Machine Vision Group, P.O. Box 4500, FI-90014, University of Oulu, Finland

Abstract. Face analysis from videos can be approached using two different strategies, depending on whether the temporal information is used or not. The most straightforward strategy applies still image based techniques to some selected (or all) frames and then fuses the results over the sequence. In contrast, an emerging strategy consists of encoding both facial structure and dynamics through spatiotemporal representations. To gain insight into the usefulness of facial dynamics, this paper considers two baseline systems and compares static versus spatiotemporal approaches to face analysis from videos. The first approach is based only on static images and uses spatial Local Binary Pattern features as inputs to SVM classifiers, while the second baseline system combines facial appearance and motion through a spatiotemporal representation using Volume LBP features as inputs to SVM classifiers. Preliminary experiments on classifying face patterns into different categories based on gender, identity, age, and ethnicity point out very interesting findings on the role of facial dynamics in face analysis from videos.

Keywords: Facial Dynamics, Spatiotemporal Analysis, Face Recognition, Local Binary Patterns.

1 Introduction

Automatic classification of face patterns into different categories based e.g. on gender, identity, age, ethnicity, facial expression or even eye/hair color, wear of glasses and other face characteristics (usually called soft biometrics traits in literature), are essential elements of vision systems and intelligent robots in smart environments. This is useful in several applications, like law enforcement, surveillance, more affective human-computer interaction (HCI), video-conferencing, and content-based image and video retrieval.

Though there has been a great deal of progress in face analysis in the last few years, many problems remain unsolved. For instance, research on face recognition must confront with many challenging problems, especially when dealing with illumination changes, aging, pose variations, low-image quality, and occlusion. The design of algorithms that are effective in discriminating between males and females, classifying faces into different age categories, or recognizing facial emotions in complex environments is still a major area of research.

Importantly, nowadays video cameras are omnipresent and, very often, real-world applications (such as in video surveillance and HCI) have to deal with video sequences as input data. So, the question which arises then is how to efficiently represent faces and perform automatic classification and categorization from videos? Basically, the problem

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 52–61, 2011.

can be approached in two different ways. The most common approach is to apply methods developed for still images to some selected frames and then fusing the results at decision or score levels. Obviously, such an approach is easy to implement but it only exploits the abundance of frames in the videos and thus ignores the temporal correlation between the face images (i.e. facial dynamics). The second approach which is less obvious and thus more challenging consists of using spatiotemporal analysis for combining facial structure and facial dynamics. This is an emerging direction in video-based face analysis and is inspired from psychological and neural studies (e.g. [17,10,8,2]) which indicate that when people talk their changing facial expressions and head movements provide a useful dynamic cue for face analysis.

This paper studies static versus spatiotemporal approaches to face analysis from videos by comparing two baseline systems. The first one is based only on static images and uses spatial LBP (Local Binary Pattern) features [15,16] as inputs to SVM (support vector machines) classifiers, while the second baseline system combines facial appearance and motion through a spatiotemporal representation using Volume LBP features [20] as inputs to SVM classifiers. Extensive experiments on classifying face patterns into different categories based on gender, identity, age, and ethnicity are conducted. The obtained results point out very interesting findings on the role of facial dynamics in face analysis from videos. These preliminary results will hopefully advance the ongoing research and open a debate on new opportunities and challenges in face processing from videos.

2 Psychophysics of Face Perception from Videos

Psychological and neural studies [17,10,8,2] indicate that when people talk their changing facial expressions and head movements provide a dynamic cue for face and gender analysis. Therefore, both fixed facial features and dynamic personal characteristics are used in the human visual system to recognize and analyze faces. Among the main findings related to the importance of facial dynamics in the human visual system and which have direct relevance to research on automatic face analysis are: **(i)** both static and dynamic facial information are useful for face recognition; **(ii)** people rely primarily on static information because facial dynamics provide less accurate identification information than static facial structure; **(iii)** dynamic information contributes more to recognition under a variety of degraded viewing conditions (such as poor illumination, low image resolution, recognition from distance etc.); **(iv)** facial motion is learned more slowly than static facial structure; **(v)** recognition of familiar faces is better when they are shown as an animated sequence than as a set of multiple frames without animation. However, for unfamiliar faces, the moving sequence does not provide more useful information than multiple static images; **(vi)** facial movement (i.e., dynamics) helps the discrimination between men and women; **(vii)** facial movement is fundamental to the recognition of facial expressions as analyzing an animated sequence produces more accurate results than what a collection of static images may result.

How can we interpret and exploit these findings to enhance the performance of automatic face analysis systems? A possible indication from the statements in (i), (iii), (vi) and (vii) is that motion is a useful cue to enhance the performance of static image

based systems. Importantly, the usefulness of the motion cue increases as the viewing conditions deteriorate (statement (iii)). Such an environment is often encountered in surveillance and access control applications. Thus, an automatic recognition system should exploit both dynamic and static information. From the evidence in (iv), one can interpret that facial motion is more challenging to learn and use than the face structure. Thus, a laborious training might be necessary when exploiting the facial motion.

3 Use of Facial Dynamics in Computer Vision

Despite these evidences from psychophysics and neuroscience which indicate that facial movements can provide valuable information to gender classification and face recognition, only recently have researchers started to pay an important attention to the use of facial dynamics in automatic face analysis (e.g. [12,11,21,22,13]). Facial dynamics is commonly encoded and exploited through spatiotemporal analysis.

Unsurprisingly, the facial expression recognition problem has attained the most attention and efforts in combining facial structure and motion. This is due to the fact that facial expressions (happiness, sadness, fear, disgust, surprise and anger) are generated by contractions of facial muscles which result in temporally deformed facial features such as eye lids, lips and skin texture. Hidden Markov Models and optical flow algorithms are commonly used to determine the facial expression by modeling the dynamics of facial actions caused by skin and facial feature deformation. Complete surveys on the large number of works on facial expression recognition can be found in [19]. Since the role of facial dynamics in facial expression recognition is quite obvious and well studied, we devote this article to the use of facial dynamics in less obvious problems such as face recognition, gender classification, age categorization and ethnicity classification.

There have been many attempts to exploit the facial dynamics. For instance, to recognize faces in video sequences, Li and Chellappa used the trajectories of tracked features [11]. The features are extracted using Gabor attributes on a regular 2D grid. Using a small database of 19 individuals, the authors reported performance enhancement over the frame to frame matching scheme. In another work, Zhou and Chellappa proposed a generic framework to track and recognize faces simultaneously by adding an identification variable to the state vector in the sequential important sampling method [21]. In [12], another approach exploiting spatiotemporal information for face recognition is presented. It is based on modeling face dynamics using identity surfaces. Face recognition is performed by matching the face trajectory that is constructed from the discriminating features and pose information of the face with a set of model trajectories constructed on identity surfaces. Experimental results using 12 training sequences and the testing sequences of three subjects were reported with a recognition rate of 93.9%. An alternative to model the temporal structures is the use of the condensation algorithm. This algorithm has been successfully applied for tracking and recognizing multiple spatiotemporal features. Recently, it was extended to video based face recognition problems [22,21]. Perhaps, the most popular approach to model temporal and spatial information is based on the Hidden Markov models (HMM) which have also been applied to face recognition from videos [13]. The principle of using HMMs for dynamic face recognition is quite simple: during the training phase, an HMM is created to learn

both the statistics and temporal dynamics of each individual. During the recognition process, the temporal characteristic of the face sequence is analyzed over time by the HMM corresponding to each subject. The likelihood scores provided by the HMMs are compared. The highest score provides the identity of a face in the video sequence. Recent developments also showed that volumetric features (such as volume LBP [20]) can be use to encode both facial motion and appearance [7]. The idea is to consider a face video sequence as a rectangular prism from which volumetric primitives can be collected into a histogram representing the appearance and motion of the face in the video sequence.

Very recently, Hadid and Pietikäinen proposed a spatiotemporal approach to combine facial appearance and dynamics to gender recognition from videos, yielding interesting preliminary results [6]. The experiments showed that the combination of motion and appearance was only useful for gender analysis of familiar faces while, for unfamiliar faces, motion seemed to not provide discriminative information. In [14], Matta et al. have also explored the use of head and mouth motions in combination with facial appearance for gender recognition. Experiments on a relatively small database containing 208 video sequences of 13 different persons, showed some performance enhancement when integrating the motion information compared to the use of only facial appearance.

In contrast to other facial analysis tasks, automatic age and ethnicity classification problems have received relatively far less attention despite the vast potential applications. A recent survey on different methods for age estimation can be found in [5]. Unfortunately, due to its challenging nature and lack of clear psychophysical evidences, no work has yet clearly addressed the use of facial dynamics in age and ethnicity classification.

4 Spatiotemporal *versus* Static Image Analysis

To gain insight into the use of facial dynamics, two baseline approaches based on local binary pattern (LBP) features [16] and support vector machines (SVM) are implemented and discussed in this section. The first approach is using only static images and thus ignoring the facial dynamics while the second approach uses spatiotemporal representation thus combining facial structure and dynamics. The aim of the experiments is to evaluate the benefit of incorporating the facial dynamics. The choice of adopting LBP approach [16] is motivated by the recent success of using it for combining appearance and motion for face and facial expression recognition [20,7] and also for dynamic texture recognition [20]. We describe below the two experimental approaches and then report the experimental results on face recognition, gender classification, age estimation and ethnicity classification.

4.1 Static Image Based Approach

The LBP texture analysis operator, introduced by Ojala *et al.* [15,16], is defined as a gray-scale invariant texture measure, derived from a general definition of texture in a local neighborhood. It is a powerful means of texture description and among its properties in real-world applications are its discriminative power, computational simplicity

and tolerance against monotonic gray-scale changes. LBP is shown to be efficient in representing and analyzing faces images [1,7,20].

The original LBP operator forms labels for the image pixels by thresholding the 3×3 neighborhood of each pixel with the center value and considering the result as a binary number. The histogram of these $2^8 = 256$ different labels can then be used as a texture descriptor. Each bin (LBP code) can be regarded as a micro-texton. Local primitives which are codified by these bins include different types of curved edges, spots, flat areas etc. The calculation of the LBP codes can be easily done in a single scan through the image. The value of the LBP code of a pixel (x_c, y_c) is given by:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \tag{1}$$

where g_c corresponds to the gray value of the center pixel (x_c, y_c), g_p refers to gray values of P equally spaced pixels on a circle of radius R, and s defines a thresholding function as follows:

$$s(x) = \begin{cases} 1, if\ x \geq 0; \\ 0, otherwise. \end{cases} \tag{2}$$

The occurrences of the LBP codes in the image are collected into a histogram. The classification is then performed by computing histogram similarities. For an efficient representation, facial images are first divided into several local regions from which LBP histograms are extracted and concatenated into an enhanced feature histogram. In such a description, the face is represented in three different levels of locality: the LBP labels for the histogram contain information about the patterns on a pixel-level, the labels are summed over a small region to produce information on a regional level and the regional histograms are concatenated to build a global description of the face. This locality property, in addition to the computational simplicity and tolerance against illumination changes, are behind the success of LBP approach for facial image analysis [1].

Given a target face video sequence, a straightforward approach to perform recognition or classification is to analyze each frame and then combine the results through majority voting which consists of determining the gender (or age or identity) in every frame and then fusing the results. Therefore, we divide each facial image (frame) into several local regions from which LBP histograms are extracted and concatenated into an enhanced feature histogram. Then, we present the results to an SVM classifier for classification. Finally, we combine the recognition results over the face sequence using majority voting. In such an approach, only static information is used while the facial dynamics are ignored.

4.2 Spatiotemporal Based Approach

For spatiotemporal analysis, volume LBP operator has been introduced in [20] and successfully used for combining appearance and motion for face and facial expression recognition [20,7] and also for dynamic texture recognition [20]. The idea behind

volume LBP (VLBP) consists of looking at a face sequence as a rectangular prism (or volume) and defining the neighborhood of each pixel in three dimensional space (X,Y,T) where X and Y denote the spatial coordinates and T denotes the frame index (time). Then, similarly to LBP in spatial domain, volume textons can be defined and extracted into histograms. Therefore, VLBP combines structure and motion together to describe the moving faces.

Once the neighborhood function is defined, we divide each face sequence into several overlapping rectangular prisms of different sizes, from which we extract local histograms of VLBP code occurrences. Then, instead of simply concatenating the local histograms into a single histogram, we use AdaBoost learning algorithm [4] for automatically determining the optimal size and locations of the local rectangular prisms, and more importantly for selecting the most discriminative VLBP patterns for classification while discarding the features which may hinder the classification process. So, to combine facial structure and dynamics, we first extract VLBP features from the face sequences and do feature selection using AdaBoost. The result is then fed to an SVM classifier for classification. In such an approach, both static facial information and facial dynamics are used.

4.3 Experiments on Face Recognition

We first applied the two approaches to the face recognition problem. In order to experiment with a large amount of facial dynamics, resulted for example from the movements of the facial features when the individuals are talking, we considered CRIM video database [3]. This is large set of 591 face sequences showing 20 persons reading broadcast news for a total of about 5 hours. The database is originally collected for audio-visual recognition. There are between 23 and 47 video sequences for each individual. The size of the extracted face images is 130×150 pixels. We randomly selected half of the face sequences of each subject for training while the other half was used for testing. We report the average recognition rates of 100 random permutations.

The performances of both static image based and spatiotemporal based approaches on CRIM video database are shown in Table 1. From the results, we can notice that the spatiotemporal based method (*i.e.* combination of face structure and dynamics) significantly outperforms the static image based method (*i.e.* using only facial structure). The better performance of the spatiotemporal method is in agreement with the neuropsychological evidence [17] stating that facial dynamics are useful for face recognition.

Table 1. Average face recognition rates using static image based and spatiotemporal based approaches on CRIM video database

Method	Face Recognition Rate
Static image based approach	93.3%
Spatiotemporal based approach	98.1%

4.4 Experiments on Gender Recognition

Determining whether the person whose face is in the given video is a man or a woman is useful for many applications such as more affective Human-Machine Interaction, restricting access to certain areas based on gender, collecting demographic information in public places, counting the number of women entering a retail store and so on. Similarly to the face recognition experiments, we adapted our methodology and applied the static image based and spatiotemporal based approaches to the problem of gender recognition. We considered three different publicly available video face databases (namely CRIM [3], VidTIMIT [18] and Cohn-Kanade [9]) containing a balanced number of male's and female's sequences and including several subjects moving their facial features by uttering phrases, reading broadcast news or expressing emotions. We randomly segmented the datasets and extracted over 4 000 video shots of 15 to 300 frames each. From each shot or sequence, we automatically detected the eye positions from the first frame. The determined eye positions are then used to crop the facial area in the whole sequence. Finally we scaled the resulted images into 40×40 pixels.

For evaluation, we adopted a 5-fold cross validation test scheme by dividing the 4 000 sequences into five groups and using the data from four groups for training and the left group for testing. We repeated this process five times and we report the average classification rates. When dividing the data into training and test sets, we explicitly considered two scenarios. In the first one, a same person may appear in both training and test sets with face sequences completely different in the two sets due to facial expression, lighting, facial pose etc. The goal of this scenario is to analyze the performance of the methods in determining the gender of familiar persons seen under different conditions. In the second scenario, the test set consists only of persons who are not included in the training sets. This is equivalent to train the system on one or more databases and then do evaluation on other (different) databases. The goal of this scenario is to test the generalization ability of the methods to determine the gender of unseen persons.

The gender classification results using the two approaches (static image based and spatiotemporal based) in both scenarios (familiar and unfamiliar) are summarized in Table 2. We can notice that both methods gave better results with familiar faces than unfamiliar ones. This is not surprising and can be explained by the fact that perhaps the methods did not rely only on gender features for classification but may also exploited information about face identity. For familiar faces, the combination of facial structure and dynamics yielded in perfect classification rate of 100%. This proves that the system succeeded in learning and recognizing the facial behaviors of the subjects even under different conditions of facial expression, lighting and facial pose. For unfamiliar faces, the combination of facial structure and dynamics yielded in classification rate of about 83% which is still encouraging although the best result for unfamiliar faces is obtained using the static image based approach. This may indicate that incorporating motion information with facial appearance was useful for only familiar faces.

4.5 Experiments on Age Estimation

Automatic age estimation (or classification) aims at determining the age range of a target face. This is a very challenging problem but also a very useful application. To study

Table 2. Gender classification results on test videos of familiar and unfamiliar subjects using static image based and spatiotemporal based methods

Method	Gender Classification Rate	
	Subjects Seen during Training	Subjects Unseen during Training
Static image based	94.4%	90.6%
Spatiotemporal based	**100%**	**82.9%**

whether facial dynamics may enhance the automatic age estimation performance, we performed a set of experiments using the static image based and spatiotemporal based approaches. We considered five age classes as follows: child= 0 to 9 years old; youth= 10 to 19; adult= 20 to 39; middle-age= 40 to 59 and elderly = above 60. Then, we built a novel classification scheme based on a tree of four SVM classifiers. The first SVM classifier is trained to learn the discrimination between `child` class and the rest. If the target face is assigned into the `child` category, then the classification is completed. Otherwise, the second SVM classifier is examined to decide whether the face belongs to the `Youth` category or not. If not, the third SVM is examined and so on.

We considered the proposed tree classification scheme and applied the static image based and spatiotemporal based approaches to age estimation from videos. For evaluation, we collected from Internet a set of video sequences mainly showing many celebrities giving speeches in TV programs and News. For the videos of unknown individuals (especially children), we manually labeled them using our (human) perception of age. Then, we randomly segmented the videos and extracted about 2000 video shots of about 300 frames each. In the experiments, we adopted a 10-fold cross validation test scheme by dividing the 2000 sequences into 10 groups and using the data from 9 groups for training and the left group for testing. We repeated this process 10 times and we report the average classification rates. The performances of both static image based and spatiotemporal based approaches are shown in Table 3. From the results, we can notice that both methods did not perform very well and this somehow confirms the difficulty of the age estimation problem. Interestingly, the static information based method significantly outperformed the spatiotemporal based method. This might be an indication that facial dynamics is not useful for age estimation. However, due to the challenging nature of the age estimation problem, it is perhaps too early to make such a conclusion and hence more investigations are needed to study the integration of facial dynamics and facial structure for age estimation from videos.

Table 3. Average age classification rates

Method	Age Classification Rate
Static image based approach	77.4%
Spatiotemporal based approach	69.2%

4.6 Experiments on Ethnicity Classification

Similarly to the previous experiments on gender recognition, the two experimental approaches are also applied to ethnicity classification from videos. Because of lack of

ground truth data for training, only two ethnic classes (namely Asian and non-Asian) are considered. The same set of 2000 video shots previously used in the experiments on age estimation is also considered here for ethnicity classification tests. A manual labeling yielded in 81% of non-Asian and 19% of Asian data samples (i.e. video shots). For evaluation, we also adopted a 5-fold cross validation test scheme.

Table 4. Average ethnicity classification rates using static image based and spatiotemporal based approaches

Method	Ethnicity Class. Rate
Static image based approach	97.0%
Spatiotemporal based approach	99.2%

The performances of both static image based and spatiotemporal based approaches are shown in Table 4. From the results, we can notice that both approaches perform quite very well but the spatiotemporal based method (i.e. combination of face structure and dynamics) slightly outperforms the static image based method (using only facial structure). This is somehow surprising because one may not expect better results using spatiotemporal methods for ethnicity classification.

5 Discussion and Conclusions

In this work, we discussed the psychological and neural findings about the importance of facial dynamics in the human visual system and reviewed the major attempts to combine facial structure and motion for automatic face analysis. To gain insight into the use facial dynamics, we considered two approaches to face analysis from videos using LBP features and SVMs, and reported preliminary experimental results on several problems including face recognition, gender classification, age estimation and ethnicity determination. The experiments results on face recognition showed that the spatiotemporal based method significantly outperforms the static image based method. This is somehow in agreement with the neuropsychological evidences stating that facial dynamics are useful for face recognition. The experiments on age estimation pointed out that combining face structure and dynamics does not enhance the performance of static image based automatic systems. Our experiments also showed that incorporating motion information with facial appearance for gender classification might be only useful for familiar faces but not with unfamiliar ones (while the psychological and neural studies indicated that facial movements do contribute to gender classification in the HVS). Finally, our experiments on the ethnicity classification problem yielded in quite surprising results indicating some relative usefulness of facial dynamics in ethnicity classification.

Acknowledgment. The financial support from the EU FP7 project TABULA RASA (grant agreement #257289), and the Academy of Finland is gratefully acknowledged.

References

1. Ahonen, T., Hadid, A., Pietikäinen, M.: Face description with local binary patterns: Application to face recognition. IEEE TPAMI 28(12), 2037–2041 (2006)
2. Bassili, J.: Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. Journal of Personality and Social Psychology 37, 2049–2059 (1979)
3. CRIM, http://www.crim.ca/
4. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997)
5. Fu, Y., Guo, G., Huang, T.: Age synthesis and estimation via faces: A survey. IEEE TPAMI 32(11), 1955 (2010)
6. Hadid, A., Pietikäinen, M.: Combining appearance and motion for face and gender recognition from videos. Pattern Recognition 42(11), 2818–2827 (2009)
7. Hadid, A., Pietikäinen, M., Li, S.Z.: Learning Personal Specific Facial Dynamics for Face Recognition from Videos. In: Zhou, S.K., Zhao, W., Tang, X., Gong, S. (eds.) AMFG 2007. LNCS, vol. 4778, pp. 1–15. Springer, Heidelberg (2007) (in conjunction with ICCV 2007)
8. Hill, H., Johnston, A.: Categorizing sex and identity from the biological motion of faces. Current Biology 11(11), 880–885 (2001)
9. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 46–53 (2000)
10. Knight, B., Johnston, A.: The role of movement in face recognition. Visual Cognition 4, 265–274 (1997)
11. Li, B., Chellappa, R.: Face verification through tracking facial features. Journal of the Optical Society of America 18, 2969–2981 (2001)
12. Li, Y.: Dynamic Face Models: Construction and Applications. Ph.D. thesis, Queen Mary, University of London (2001)
13. Liu, X., Chen, T.: Video-based face recognition using adaptive hidden markov models. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp. 340–345 (June 2003)
14. Matta, F., Saeed, U., Mallauran, C., Dugelay, J.L.: Facial gender recognition using multiple sources of visual information. In: 10th IEEE International Workshop on MultiMedia Signal Processing, MMSP 2008, Cairns, Queensland, Australia, October 8-10 (2008)
15. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on feature distributions. Pattern Recognition 29, 51–59 (1996)
16. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE TPAMI 24, 971–987 (2002)
17. O'Toole, A.J., Roark, D.A., Abdi, H.: Recognizing moving faces: A psychological and neural synthesis. Trends in Cognitive Science 6, 261–266 (2002)
18. Sanderson, C., Paliwal, K.K.: Noise compensation in a person verification system using face and multiple speech feature. Pattern Recognition 36(2), 293–302 (2003)
19. Zeng, Z.H., Pantic, M., Roisman, G.I., Huang, T.: A survey of affect recognition methods:audio, visual, and spontaneous expressions. PAMI 31(1), 39–58 (2009)
20. Zhao, G., Pietikäinen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE TPAMI 29(6), 915–928 (2007)
21. Zhou, S., Chellappa, R.: Probabilistic Human Recognition from Video. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2352, pp. 681–697. Springer, Heidelberg (2002)
22. Zhou, S., Krueger, V., Chellappa, R.: Face recognition from video: A condensation approach. In: IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 221–228 (May 2002)

Adaptive Integration of Multiple Cues for Contingency Detection

Jinhan Lee, Crystal Chao, Andrea L. Thomaz, and Aaron F. Bobick

School of Interactive Computing,
Georgia Institute of Technology,
Atlanta GA 30332, USA
{jinhlee,chao,athomaz,afb}@cc.gatech.edu

Abstract. Critical to natural human-robot interaction is the capability of robots to detect the contingent reactions by humans. In various interaction scenarios, a robot can recognize a human's intention by detecting the presence or absence of a human response to its interactive signal. In our prior work [1], we addressed the problem of detecting visible reactions by developing a method of detecting changes in human behavior resulting from a robot signal. We extend the previous behavior change detector by integrating multiple cues using a mechanism that operates at two levels of information integration and then adaptively applying these cues based on their reliability. We propose a new method for evaluating reliability of cues online during interaction. We perform a data collection experiment with help of the Wizard-of-Oz methodology in a turn-taking scenario in which a humanoid robot plays the turn-taking imitation game "Simon says" with human partners. Using this dataset, which includes motion and body pose cues from a depth and color image, we evaluate our contingency detection module with the proposed integration mechanisms and show the importance of selecting the appropriate level of cue integration.

Keywords: Contingency Detection, Human Response Detection, Cue Integration.

1 Introduction

In a variety of scenarios, a social robot can leverage the presence or absence of a contingent response on the part of a human to determine his or her intention. Such determination is often necessary for a robot to act appropriately. One such scenario is identifying willing interaction partners and subsequently initiating interactions with them such as a shop robot's attempting to engage customers who might need help. To do this, the robot would generate a signal intended to elicit a behavioral response and then look to see if any such response occurs. The presence of a change in the behavior of a human at the right time is a good indication that he or she is willing to follow-up with an interaction. Another scenario is that which involves reciprocal turn-taking situations. In such contexts, contingency can be used as a signal that helps a robot determine when a human is ready to take a turn and when it is appropriate for robot to relinquish the floor to the human.

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 62–71, 2011.

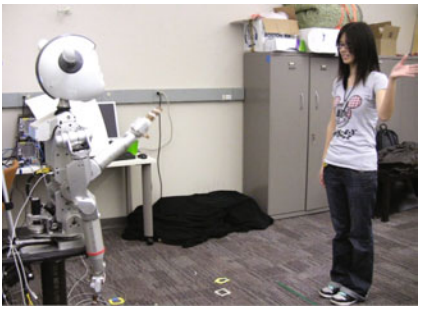


Fig. 1. A human being contingent to the robot in a turn-taking scenario based on the game "Simon says." The robot sends both motion and speech signals to the human subject by simultaneously waving and saying hello. The subject waves back to the robot in response.

When a robot understands the semantics of a human's activity and has a specific expectation of the human's next action, then checking for a contingent response might simply entail matching the human's actual human action with the expected action. This strategy makes the sometimes problematic assumptions that the set of appropriate or meaningful responses can be enumerated in advance, and that specific recognition methods can be constructed for each. For example, when a robot waves to a person across a room to see if the person is interested in engaging the robot, the human might wave back, change their gaze, shift their body orientation or - in this case of having no interest - may make no behavioral change at all. In these situations, a robot can look for a more general change in its sensory data that is indicative of a behavioral change. In this paper, we assume that behavioral change *occurring at a time that would be a contingent response* to a robot probe signal is indeed a meaningful response.

A contingent behavioral change by a human can occur in one or multiple communication channels. Here, we consider the problem of human contingency detection with multimodal sensor data as input when forming our computational model. In our prior work [1], we presented a contingency detection framework that uses visual information based upon both a depth and color image. However, we provided no principled mechanisms for combining multiple channels. In this paper we extended our work by considering a variety of approaches for integrating multiple channels. Furthermore we show how the relative significance of each channel can be assessed and then incorporated into the contingent response decision function.

We make the following contributions. We present a contingency detection framework that integrates data from multiple cues using a naïve Bayes model and propose two different levels of cue integration: the module level and the decision level. We show that for change-based detection, integration of visual cues at the module level outperform integration at the decision level. We also provide a probabilistic method for measuring the reliability of visual cues and adaptively integrating these cues based on their reliability. We evaluate our proposed contingency detection framework using multimodal data and demonstrate that multi-cue contingency detection is a necessary component for interactions with humans.

2 Related Work

The problem of response detection using visual cues contrasts with the problem of *action recognition* [3] in that responses cannot be formulated as actions. For example, a human could respond to a robotics greeting by approaching to the robot, speaking to the robot, stopping a previous action and looking at the robot, or by performing a number of other actions. In action recognition, the actions to be recognized are known a priori or trained in real time. In either case, explicit modeling of a fixed vocabulary of possible responses takes place. Modeling all possible responses would be difficult due to the large number of possible responses. Some research on the detection of abnormal or unusual events has taken an unsupervised approach [4][5]. Zhong et al. [4] splits a video into segments and clusters them using a spectral graph co-clustering method. A video segment is considered unusual if it differs from a majority of the other segments. In contrast to this approach, we solve the response detection problem by only comparing observations before a robot's interaction signal to those after that signal. Boiman and Irani [5] built a rich database of imagery for regular events to determine the irregularity of an input video, but when applied to the problem of response detection, their framework is susceptible to false negatives when responses are similar to events in the database.

For contingent response detection, other work has focused on processing other individual channels, independently demonstrating the significance of gaze shift [7][8], agent trajectory [9][10], audio cues [11], or synchrony [12] as contingent reactions. To analyze of human response for the engagement detection scenario in our prior work, we proposed an approach to detecting such a visual change. This approach focused on change detection using single cues, but it could not be generalized to integrate multiple cues in a systematic way.

3 Approach

Contingent response detection consists of two sub-problems: response detection and timing interpretation. Figure 2 shows the causal relationship between a robot signal and the corresponding human response and time windows for detecting such a response. A robot generates some interactive signal to a human by gesturing to, speaking to, or approaching her. After and even before the robot finishes conveying its signal, the human may initiate a response that could last for a certain amount of time. The point of time at which the robot starts looking for a response is important. We define t_S as the time robot's signal to the human is initiated and t_R as the initiation time of the human's response. t_S is known and available to the robot, but of course t_R must be determined. We define t_{Ref} be the time at which the robot begins to look for a human response. This parameter will need to be learned from observation of human interacting with robots. Similarly we define the maximum response delay (MRD), as the maximum time for the human's response delay after t_{Ref}. We evaluate only sensor data within the time window between t_{Ref} and $t_{Ref} + MRD$ to detect a response.

This paper focuses on the response detection problem assuming that the expected timing of a response is given. To detect a human response, we detect the human's

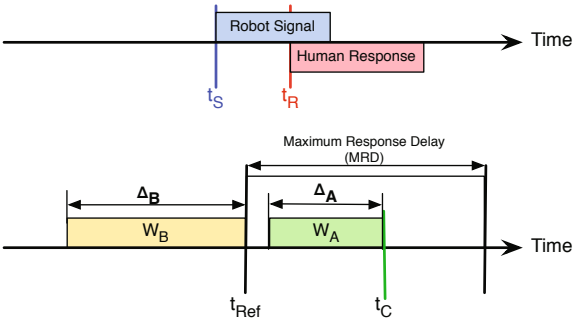


Fig. 2. Causal relationship between a robot's interactive signal and a human's response. Top: t_S is the initiation of the robot signal, and t_R is the initiation of the human response. Bottom: Two time windows, W_B and W_A, defined with respect to the reference point of the robot signal t_{Ref} and the current time t_C, are used to model the human behavior before and after the signal. The time window starting at t_{Ref} and valid over MRD is examined for the contingent human response. Note that t_{Ref} may not align with t_S.

behavior change, which we measure by comparing observations between the human behavior before and after the robot's reference signal. As shown in Figure 2, we define W_B and W_A as time windows in which data are used to model the human behavior before and after the robot's reference signal, t_{Ref}, respectively. When the human does not respond to the robot, both W_B and W_A describe the same aspect of the human's behavior. However, in the contingent cases the human changes her behavior to make a contingent response, and thus W_B and W_A model different behaviors. To detect such changes, we measure how likely that the sensor data in W_A reflects the same behavior observed in the sensor data in W_B.

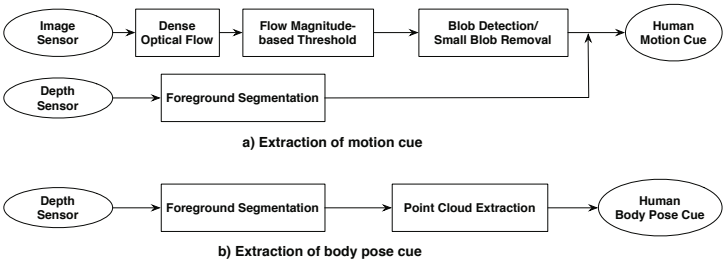


Fig. 3. Extracting cues from image and depth sensors: a) motion cue, and b) body pose cue

3.1 Cue Representation

To model a given aspect of human behavior, we derive information from observations from a single or multiple sensors, hereafter referred to as a *cue*. The choice of cues should be determined by the nature of the interaction. When a robot engages in a face-to-face interaction with a human, a shift of the human's eye gaze is often enough to determine the presence of a response. In situations in which a gaze cue is less indicative or is less reliable to perceive, other cues should be leveraged for response detection. Here, we are interested in modeling human behavior using 1) a *motion cue*, the pattern of observed human motion, and 2) a *body pose cue*, the observed human body configuration.

The motion cue models the motion patterns of the human of interest in the image coordinate system. This cue is derived from observations from image and depth sensors. To observe only motions generated by a human subject in a scene, we segment the region in which a human subject appears. To do so, we use the OpenNI API [13][1]. The process for generating the motion cue is illustrated in Figure 3(a). First, the motion in an image is estimated using a dense optical flow calculation [6]. After grouping motion regions using a connected components-based blob detection method, groups with small motions and groups that do not coincide with the location of the human of interest are filtered out. The motion cue comprises the remaining motion regions.

The body pose cue models the body configuration of the human of interest. The human body configuration is estimated from a 3D point cloud extracted from the human of interest. The process for generating the body pose cue is illustrated in Figure 3(b). As for the motion cue, we segment the region of the human from a depth scene using the OpenNI API. Then, we extract a set of 3D points by sparse sampling from the human region in a depth image and reconstructing 3D points from those depth samples.

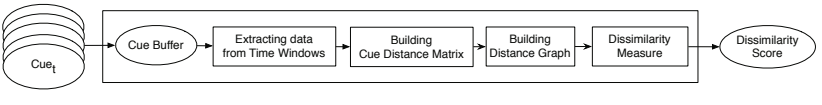


Fig. 4. Change detection framework [1]

4 Change Detection

To measure the degree of behavioral difference in the response, we proposed a change detection framework [1], as shown in Figure 4. The change detector accumulates cue data inside a buffer over time and when the buffer is full of data for W_B and W_A, a cue distance matrix is calculated between the cue data in W_B and W_A using a cue-specific distance metric. The distance matrix is converted into a distance graph by applying specific node connectivity; nodes representing cue data from W_B are fully connected to

[1] This API provides the functionalities of detecting and tracking multiple humans using depth images.

one another, nodes from W_A are never connected to one another, and nodes from W_A are connected to only the K nearest nodes from W_B. Then, we calculate the dissimilarity score by measuring a statistical difference between the graph nodes representing data from W_B and W_A. By using a learned threshold value on this dissimilarity score, the change detector determines whether or not a contingent response occurs. (Refer to [1] for technical details.)

We learned a threshold value on the dissimilarity score from the training data and use that to classify the score as being contingent or not. This simple evaluation method cannot be used in our probabilistic model for multi-cue integration because it does not have confidence on the decision made and it does not take into account how informative a used cue is.

4.1 Evaluating the Dissimilarity Score

One of contributions of this paper is the proposal of a new evaluation method that resolves two problems described above. To determine that an observed change (i.e. a dissimilarity score) actually resulted from a human response and not from changes that occurs naturally in the human's background behavior, we should evaluate the change not only under the contingency condition, but also under the non-contingency condition. To this end, we model two conditional probability distributions: a probability distribution of the dissimilarity score S under the contingency condition C, $P(S|C)$ and a probability distribution of S under the non-contingency condition, $P(S|\overline{C})$. Assuming that a human changes her behavior when responding, these two distributions need to differ for the cue to be considered informative.

We learn the distribution $P(S|C)$ off-line from the training data in which a human subject is being contingent to a robot's action. We estimate the distribution $P(S|\overline{C})$, the null hypothesis, on the fly during an interaction from observations of the human's behavior before the robot triggers a signal. It is important to note that a null hypothesis is estimated with on-line data, particularly the data gathered immediately before the robot's signal is triggered. As shown in Figure 5, this distribution is estimated from dissimilarity score samples, each of which is obtained as if the robot's signal were triggered and enough data were accumulated at each point in time. We refer to this method as *multiple evaluation-based null hypothesis building* (MENHB).

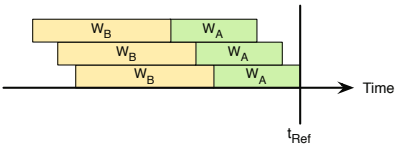


Fig. 5. Building the null hypothesis (MENHB). Dissimilarity score samples are obtained by evaluating them over time

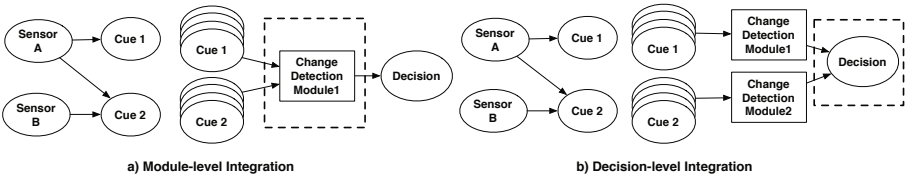


Fig. 6. Cue integration at different levels: a) the module level, and b) the decision level

4.2 Multi-cue Integration

The extracted cues from sensors should be integrated in such a way that the response detection module reduces uncertainty and increases accuracy in its decision-making capabilities. We propose two different levels of cue integration: 1) the module levels, at which cues are integrated within one change detection module; and 2) the decision level, at which outputs from multiple single-cue change detection modules merge. These levels are shown in Figure 6. Intuitively, the principal difference is whether the cues are combined into a single signal whose variation during the contingency window is evaluated for a behavioral change, or whether each cue is considered independently and the two decisions are fused to provide a final answer.

4.3 Cue Integration at the Module Level

At the module level of integration, the cue data are integrated when calculating a distance matrix. From the accumulated cue data for both the motion and body pose cues, two distance matrices are calculated independently and merged into a new distance matrix. Since the motion and body pose cues are represented in different coordinate spaces, an image space for motion and a 3D world space for body pose, the distance matrices need to be normalized. We denote the distance matrices for motion and body pose cues as DM_M and DM_D, respectively. The merged distance matrix DM_N is calculated in the following way:

$$DM_N = \frac{1}{2}\left(\frac{DM_M}{\|DM_M\|_F} + \frac{DM_D}{\|DM_D\|_F}\right), \tag{1}$$

where $\|DM_X\|_F$ is the Frobenius norm of a matrix DM_X. After building DM_N, we use this distance matrix to calculate the dissimilarity measure as explained in Section 4.

4.4 Cue Integration at the Decision Level

At the decision level of integration, we use the naïve Bayes probabilistic model to integrate dissimilarity scores obtained from cues. We chose this model because cues that are integrated at this level assumed to be conditionally independent of each given the contingency value; otherwise, they should be integrated at the module level. Let's assume that change detectors for Cue_X and Cue_Y obtain dissimilarity scores, S_X and

S_Y, respectively. The overall contingency C in terms of S_X and S_Y is estimated by standard Bayesian mechanics of Equation 2:

$$\frac{P(C|S_X, S_Y)}{P(\overline{C}|S_X, S_Y)} = \frac{P(S_X|C)}{P(S_X|\overline{C})} \frac{P(S_Y|C)}{P(S_Y|\overline{C})} \frac{P(C)}{P(\overline{C})} \tag{2}$$

If this ratio > 1, we declare that the human behavior has changed; otherwise, any changes are not detected.

5 Data Collection

We validate multi-cue contingency detection within a turn-taking scenario in which the robot plays the turn-taking imitation game "Simon says" with a human partner. This imitation game was not designed for evaluating a contingency detector, but for generating natural interactions with human-like timings. To collect naturalistic data, the robot's behavior was controlled via teleoperation with randomly generated timing variations. This imitation game was based on the traditional children's game "Simon says." The interaction setup is shown in Figure 1. In this game, one participant plays the leader and the other plays the follower. The game has two different types of a robot's signal to a human: game phase and negotiation phase. In game phase, the robot plays the leader and asks a human to imitate an action using a mixture of speech, motion, and gaze. In negotiation phase, the leader and the follower can switch roles. The experimental evaluation is performed with a multi-modal data set of 11 human subjects. About four minutes for data was collected from each subject. (Refer to [2] for data collection details.)

6 Model Building and Evaluation

To build and evaluate a computational model of social contingency detection for the "Simon says" interaction game, we used a supervised learning approach. We employed a leave-one-subject-out cross validation procedure; a single subject was iteratively left out of training data and used as testing data. First we split the recorded data sessions into shorter video segments, each of which started at the t_{Ref} of one robot signal and ended at the t_{Ref} of the following one. We used t_S as a referent event, t_{Ref}. Depending on the presence or absence of a human response, video segments were partitioned into two sets, contingent and non-contingent. Video segments were classified by two authors. If the human made noticeable body movements or vocal utterances as a response to a robot's signal, then corresponding video segments were classified as being contingent. We collected 246 video segments: 101 (97 contingent cases) and 163 (120 contingent cases) from negotiation phases and from game phases, respectively. During negotiation phases, the human responded through an audio channel, which was not detectable using our observed cues. Thus, we evaluated only on interaction cases in game phases in which humans responded through the visual channel.

We set the time window W_B and W_A to four seconds and two seconds, respectively. We chose these values because during interaction human responses usually last less than two seconds, and from our empirical observation W_B should be at least two times

longer than W_A to make a reliable dissimilarity evaluation. We learned the model of $P(S|C)$ from dissimilarity scores, which were obtained by processing the contingent video segments in the training data set. Instead of manually collecting dissimilarity scores in which a response occurred, we extracted one second of dissimilarity scores around the maximum. Note that a probability model of $P(S|\overline{C})$, the null hypothesis, is not learned from the training set since the null hypothesis should represent the amount of behavioral change that occurs naturally during interaction – not indicative of a contingent response. Therefore, it is learned on the fly by evaluating a human's normal behaviors before the referent event. In building a null hypothesis, our proposed method, MENHB, requires both W_B+W_A and an extra α, used to generate dissimilarity samples. We set α to two seconds. Overall, the method requires eight seconds of data.

7 Experiment

Our experiments test the effects of cue integration on the overall performance of contingency detection in our experimental scenario. We set the evaluation window, MRD, such that the evaluation terminates after the following interaction begins or after it lasts for eight seconds, whichever comes first.

Table 1. Accuracy by cue combination

Cues	Accuracy	Precision	Recall
Motion only	0.785	0.829	0.892
Body Pose only	0.802	0.857	0.800
Motion/Body Pose (Decision)	0.810	0.845	0.908
Motion/Body Pose (Module)	**0.859**	**0.870**	**0.950**

We run experiments to test how indicative single or combined cues are for detecting the presence of responses. We test four different cue combinations: motion only, body pose only, motion and body pose cues merged at the decision level, and motion and body pose cues merged at the module level. Table 1 shows the classification accuracy for each combination. The best accuracy, 0.859, is obtained when the motion and body pose cues are integrated at the module level. As shown, the accuracy of the classifier built with visual cues merged at the decision level (81%) is only 1% greater than that achieved by the best single visual cue of body pose. However, when integrated at the module level, the combination of pose and motion is 5% higher than pose alone. We argue that because these visual cues model the different but related aspects of the same perceptual signal, the cues are strongly related, and thus not conditionally independent. By combining at the module level this dependence is explicitly modeled and thus the merged cues generate more discriminative features. This is supported by our result that the classifiers built with a signal cue or built with cues integrated at the decision level produce more false negative cases than the classifiers built with visual cues merged at the module level.

8 Conclusion

In this paper, we proposed a contingency detection framework that integrates data from multiple cues at two different levels: the module level and the decision level. We introduced a new evaluation method that takes into account how informative cues are to detect a response. We collected multimodal sensor data from a turn-taking human-robot interaction scenario based on the imitation game "Simon says." We implemented our multiple-cue approach and evaluated it with collected data using two visual cues, motion and body pose cues. Our result shows that integrating multiple cues at the appropriate level offers an improvement over individual cues for contingency detection. We believe that our contingency detection module improves a social robot's ability to engage in multimodal interactions with humans when the semantics of the human's behavior are not known to the robot.

References

1. Lee, J.H., Keiser, J.F., Bobick, A.F., Thomaz, A.L.: Vision-based Contingency Detection. In: ACM/IEEE International Conference on Human-Robot Interaction, HRI (2011)
2. Chao, C., Lee, J.H., Begum, M., Thomaz, A.L.: Simon plays Simon says: The timing of turn-taking in an imitation game. In: IEEE International Symposium on Robot and Human Interactive Communication, ROMAN (2011)
3. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing (2010)
4. Zhong, H., Shi, J., Visontai, M.: Detecting Unusual Activity in Video. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR (2004)
5. Boiman, O., Irani, M.: Detecting irregularities in images and in video. In: International Conference on Computer Vision, ICCV (2005)
6. Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., Bishof, H.: Anisotropic Huber-L1 Optical Flow. In: Proceedings of the British Machine Vision Conference, BMVC (2009)
7. Mutlu, B., Shiwa, T., Ishiguro, T.K.H., Hagita, N.: Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In: ACM/IEEE International Conference on Human-Robot Interaction, HRI (2009)
8. Rich, C., Ponsler, B., Holroyd, A., Sidner, C.L.: Recognizing engagement in human-robot interaction. In: ACM International Conference on Human-Robot Interaction, HRI (2010)
9. Michalowski, M.P., Sabanovic, S., Simmons, R.: A spatial model of engagement for a social robot. In: International Workshop on Advanced Motion Control, AMC (2006)
10. Muller, S., Hellbach, S., Schaffernicht, E., Ober, A., Scheidig, A., Gross, H.M.: Whom to talk to? Estimating user interest from movement trajectories. In: IEEE International Symposium on Robot and Human Interactive Communication, ROMAN (2008)
11. Butko, N.J., Movellan, J.R.: Detecting Contingencies: An Infomax Approach. IEEE Transactions on Neural Networks (2010)
12. Shen, Q., Saunders, J., Kose-Bagci, H., Dautenhahn, K.: Acting and Interacting Like me? A Method for Identifying Similarity and Synchronous Behavior between a Human and a Robot. In: IEEE IROS Workshop on From Motor to Interaction Learning in Robots (2008)
13. The OpenNI API, `http://www.openni.org`

DTW Based Clustering to Improve Hand Gesture Recognition

Cem Keskin, Ali Taylan Cemgil, and Lale Akarun

Bogazici University, Computer Engineering Department
keskinc@cmpe.boun.edu.tr, {taylan.cemgil,akarun}@boun.edu.tr

Abstract. Vision based hand gesture recognition systems track the hands and extract their spatial trajectory and shape information, which are then classified with machine learning methods. In this work, we propose a dynamic time warping (DTW) based pre-clustering technique to significantly improve hand gesture recognition accuracy of various graphical models used in the human computer interaction (HCI) literature. A dataset of 1200 samples consisting of the ten digits written in the air by 12 people is used to show the efficiency of the method. Hidden Markov model (HMM), input-output HMM (IOHMM), hidden conditional random field (HCRF) and explicit duration model (EDM), which is a type of hidden semi Markov model (HSMM) are trained on the raw dataset and the clustered dataset. Optimal model complexities and recognition accuracies of each model for both cases are compared. Experiments show that the recognition rates undergo substantial improvement, reaching perfect accuracy for most of the models, and the optimal model complexities are significantly reduced.[1]

Keywords: Dynamic time warping, DTW, hand gesture recognition, HMM, IOHMM, HSMM, HCRF, preprocessing, time series clustering.

1 Introduction

Vision based hand gesture recognition has been used in the last decade as a natural interface for a variety of applications, such as games, virtual reality and modeling tools. However, using hand gestures as an input to HCI systems is challenging due to the inherent sensor noise. The impact of illumination conditions on the image, the difficulty of segmenting the hand from a cluttered background, and the cumbersome procedures for calibration of multiple cameras and other sensors, have limited the spread of vision based hand gesture interaction.

The recent release of infrared equipped depth sensors such as Kinect has accelerated the use of hand gestures for HCI, since such depth sensors can be used to segment the hand from cluttered backgrounds. Moreover, Kinect works by emitting and sensing infrared structured light, and does not depend on illumination conditions. Hence, hands can be easily detected, segmented and tracked in real time.

[1] This work has been supported by research grants Tubitak 108E161 and BU-BAP 09M101.

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 72–81, 2011.

In HCI applications, sensors detect the gesture signals by retrieving images of the hand while a gesture is performed. Features describing the shape and motion of the hand are extracted from each frame, forming a vector–valued time series, which we call a gesture sample. Gesture recognition is performed through classification of these gesture samples in real time. From a machine learning point of view, hand gestures can be considered as the output of partially observable stochastic processes [10]. Hence, the majority of related studies use graphical models such as the HMM for this task. These models have been traditionally compared in terms of their gesture classification accuracies. However, their classification speeds are also important, as the target applications are almost always meant to run in real–time.

1.1 Graphical Models for Hand Gesture Recognition

A hand gesture is generated by the hand as it assumes certain shapes while moving on a predefined trajectory. Sensors supply partial observations from this process. Both generative and discriminative graphical models have been employed for hand gesture recognition based on these observations. Generative models learn the joint distribution of their latent variables and the observations, and thus, they can produce new samples belonging to a gesture class by sampling from this distribution. On the other hand, discriminative models condition their hidden states on a suitable function of observations, and learn to distinguish between different gesture classes.

The most basic generative graphical model is the HMM [11]. The ability of generative models to generate samples is not required for classification. Instead, Markov random fields can be used to attack the problem by directly modeling the probability of model parameters conditioned on observation features. The simplest type of Markov random field is the conditional random field (CRF), which is the discriminative counterpart of HMM [4].

CRFs are not suitable for sequence classification tasks, since they associate a class label with each frame instead of the entire sequence. For such tasks, a CRF variant called hidden CRF (HCRF) is used, that incorporates a single class label with a sequence [13]. This is achieved by adding a new variable for the class label that is connected to all of the hidden state variables of the graph.

The input–output HMM (IOHMM) is an HMM variant, which conditions model parameters on an external input sequence [1]. This sequence is used to estimate the HMM parameters at each time frame. The sequence can contain any information that is known to be correlated with observations and state transitions, i.e., regime changes in the data. HMM parameter estimation is done through common regression methods such as artificial neural networks or radial basis functions.

HSMM is a natural extension to HMM, where each state produces a sequence of observations instead of a single item. These segments can be generated using a variety of methods, such as using counters to keep track of the number of symbols or employing local HMMs that produce subsequences. Explicit duration model (EDM) is a type of HSMM, where state visit durations and sojourn times are

explicitly modeled [14]. Hence, EDM can be interpreted as a special case of HMMs, where the state variable is augmented by a counter variable that keeps track of the time that the process has spent in a given state. Likewise, an HMM is an EDM with durations set to a single frame.

Some gestures are subject to spatio–temporal variability, i.e., the exact starting point, speed or scale of the gesture do not change their meaning. Speed, scale and sampling rate have a direct effect on the gesture sample lengths. Graphical models need to take the variance of sample length into account, usually by modeling durations at each hidden state. Nevertheless, as long as there are no alternative trajectories or hand shapes for a gesture class, the model can have a *left–right architecture*, which has considerably lower complexity than an unrestricted model and a lower evaluation time complexity. A common example is the left–right HMM [5], which is also extensively used for speech recognition.

To analyze and compare different graphical models, we use a challenging dataset in the sense that it does not conform to the assumptions of a left–right architecture. This dataset is created from the ten digits written in the air by several users, and captured by Kinect. There are no universally accepted trajectories for drawing digits, and different gesturers are likely to follow different paths; e.g.,the digit zero can be drawn clockwise or counter–clockwise. Likewise, the changes in speed along the path do not change the meaning of a digit. Due to these additional challenges, a left–right architecture cannot be directly assumed.

In this work, we compare HMMs, HCRFs, IOHMMs, and EDMs on the basis of recognition accuracy and speed using the digit dataset. We show that the models need to be more complex (i.e., need more hidden states) and are not restricted to have a left–right architecture, due to the complexity of the dataset. We propose a a preprocessing method, which eliminates the need for more complex models by transforming the dataset. This new dataset is formed by rescaling, resampling and clustering of gesture samples.

1.2 Clustering Time Series

Clustering of time series has been shown to be effective in many application domains [6]. The goal of clustering is to identify sets of samples that form homogeneous groups, in the sense that a certain distance measure, such as Euclidean distance for static data, is minimized among the samples in the formed clusters. Thus, a direct benefit of clustering the dataset is that modeling clusters is easier than modeling the original classes. For instance, such clusters can be modeled using simple graphical models with left–right architecture in a hand gesture recognition framework.

There are two main approaches to time series clustering. In the first approach, a distance measure that is applicable to time series is used to calculate a distance matrix from pairwise distances of samples. A common measure is the DTW cost, which is the cost of aligning one sample to the other. Likewise, pairwise distances can be trivially transformed to similarities, forming a similarity matrix instead. Some clustering methods, such as hierarchical and spectral clustering use these similarity or distance matrices as input to cluster the data [6]. In the second

approach, static features are extracted from each sample, essentially converting the time series data to static data. Common static data clustering methods such as k–means can then be used to cluster the features.

While DTW can esimate the similarity of two samples, graphical models such as HMM can measure similarity of a sample respective to a set of samples. This can be used to formulate a k–means type of clustering approach, where the HMMs play the role of cluster means. Hence, each sample is assigned to the *closest* HMM, and each HMM is re–estimated using their own set of sequences. For instance, Oates et al. use DTW to hierarchically cluster the data to form the initial clusters [9]. Hu et al. use DTW iteratively to form initial clusters and for model selection [2]. Ma et al. recursively model the dataset with HMM, calculate a feature called weighted transition occurring matrix and use normalized cut algorithm to divide the set into two clusters [7].

In this work, we first train HMM, IOHMM, HCRF and EDM on the digit dataset and optimize the model parameters. Next, we apply DTW to calculate pairwise distances of samples belonging to the same gesture class and form a distance matrix, and a corresponding similarity matrix. We use these matrices to apply spectral and hierarchical clustering to the digit dataset. Then, we train each graphical model on the resulting clusters and optimize model parameters for the new dataset. We compare the results and show that after clustering, model complexities are significantly reduced and the recognition accuracies reach nearly perfect scores.

The rest of the paper is organized as follows: We explain the time series clustering method in more detail in Section 2. Section 3 introduces the models used for hand gesture modeling. In Section 4, we explain the experiment setup and present the results. Finally, we conclude and discuss future work in Section 5.

2 Time Series Clustering Methodology

Hand gesture samples are time series consisting of concatenated observation vectors corresponding to each time frame, where the observations are features describing the shape and motion of the hand. Thus, any measure of similarity or distance is based on these observations and their sequence. The efficiency of clustering methods directly depends on the selection of these features.

2.1 Feature Selection

The digit dataset used in this work consists of motion–only gestures, i.e., the hand shape is not important. On the other hand, the shape of the trajectory contains most of the information for digits. Yet, clustering only according to the shape of the trajectory will not produce homogeneous clusters that can be modeled with simple graphical models, as the distribution of the hand speed over the trajectory might be different for two samples, even though they share the same path. Such datasets are not suitable for left–right architecture, and should be further clustered. To ensure production of homogeneous clusters in this sense, we use both location and velocity based features.

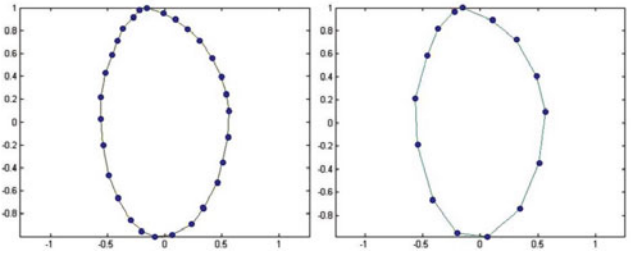


Fig. 1. Effect of resampling a signal. The digit zero is resampled to 15 samples.

First, we normalize the hand coordinates for each sample by tightly mapping the digit in the vertical interval $[-1, 1]$. Then, we resample each gesture sample using cubic interpolation, so that every sequence is of the same length. These steps ensure that some of the spatio–temporal variability is handled manually. Finally, we use these normalized locations, and the differences between consecutive frames as features. Since the digit dataset is essentially in 2D, each real valued observation vector consists of four numbers: two for location, two for velocity.

The effect of resampling a gesture signal can be seen in Figure 1. Here, a sample of digit zero is normalized and resampled to length 15.

2.2 Clustering Methods

As mentioned in Section 1.2, a distance measure for gesture pairs is needed for most clustering algorithms. We use DTW to estimate the cost of aligning two rescaled and resampled sequences. The cost of aligning individual frames of two sequences is taken as the Euclidean distance between the feature vectors corresponding to each frame. Therefore, if the normalized locations are close and the velocities in these locations are similar, this cost is small. Thus, the overall DTW cost is low if the shapes of the paths as well as the velocity distribution over the trajectories are similar.

We applied both spectral and hierarchical clustering methods to the digit dataset. Spectral clustering methods are based on the Min–Cut algorithm, which partitions graph nodes by minimizing a certain cost associated with each edge in the graph [12]. This is a binary clustering method, which can be used to hierarchically cluster data into multiple clusters. A related algorithm has been proposed by Meila and Shi [8], which can estimate multiple clusters. We first form the distance matrix D using pairwise DTW costs, and convert it into the similarity matrix S by taking the reciprocals of each element. Then we normalize each column using the row sums, to obtain the matrix P as follows:

$$D_{i,j} = DTW(G_i, G_j) \tag{1}$$

$$S_{i,j} = 1/\left(D_{i,j} + \epsilon\right) \tag{2}$$

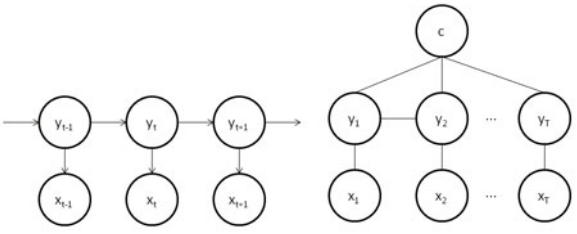


Fig. 2. Graphical models of HMM (left), HCRF (right)

$$R_{i,i} = \sum_j S_{i,j} \tag{3}$$

$$P = SR^{-1} \tag{4}$$

where $DTW(G_i, G_j)$ is the cost of aligning gesture samples G_i and G_j. This cost is symmetric due to resampling of the data. Finally, we take the eigenvectors corresponding to the k largest eigenvalues of the matrix P. We cluster these eigenvectors using the conventional k-means method.

Hierarchical clustering method creates a cluster tree using the distance matrix D [6]. Both bottom–up and top–down strategies can then be followed to merge or divide clusters according to certain criteria. We followed the bottom–up strategy called agglomerative hierarchical clustering. Initially, the algorithm regards each sample as a separate cluster and forms a tree. Then, starting from the leaves, the method merges clusters that have the minimum distance, until a termination criterion is satisfied. We force the algorithm to terminate if the number of clusters reaches a predefined number.

We cluster the digit dataset using both of these methods. Both of the methods manage to detect all the meaningful clusters in the dataset.

3 Hand Gesture Modeling

In this section, we briefly introduce the graphical models mentioned in Section 1.1. The graphical models are depicted in Figures 2 and 3. Here, c is the class label, y_t is the state variable, x_t is the observation, s_t is the input sequence and T_t is the counter value at time t.

3.1 Hidden Markov Models

HMM is one of the simplest graphical models, consisting of discrete states producing observations conditioned on the state and a state transition network with fixed probabilities. Each hidden state of an HMM represents a section of the sequence. Since HMMs are generative models, we train a separate model for each gesture class or cluster. The complexity of HMMs is directly based on the number of hidden states and the allowed transitions.

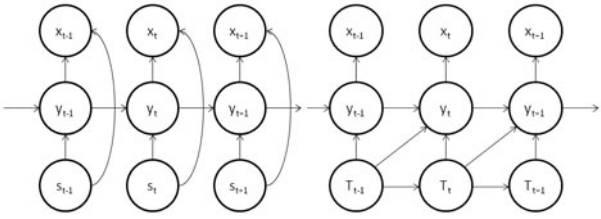


Fig. 3. Graphical model of IOHMM (left) and HSMM (right)

3.2 Hidden Conditional Random Fields

CRFs are the discriminative counterparts of HMMs [4]. CRFs do not have the naive Bayes assumption; each state is conditioned on features extracted from an overlapping set of observations. However, CRFs do not model intra-class dynamics, i.e., each gesture is represented by a single latent variable. The model needs to determine the class label at each time frame based on the current observations. Therefore, CRFs are not suitable for modeling time series. To extend the modeling capabilities of CRFs, Hidden Conditional Random Fields (HCRFs) [13] have been introduced. HCRFs relate a single class variable to the entire sequence. As HCRFs are discriminative models, we train a single model that learns to differentiate between every class or cluster pair.

3.3 Input-Output HMMs

Input-Output HMMs (IOHMM), as hybrids of generative and discriminative models, have shown considerable success in hand gesture recognition [3]. These models condition the state transition and emission probability distributions on an input sequence, which is a function of the observations [1]. The transition and observation probabilities are estimated via local models using the input sequence. In the literature, it is common to use multi–layer perceptrons (MLP) or radial basis function as local function approximators. Consequently, IOHMMs require careful design by an expert and are harder to train. In this study, we use MLPs as local models and train a separate IOHMM for each gesture.

3.4 Hidden Semi Markov Models

The special HSMM called EDM allows explicit modeling of state durations. As in the case of HMMs, the observations x_t are conditioned on the states y_t. Each hidden state is augmented by a positive counter variable τ_t that is initialized and deterministically decreased until it becomes zero. If the counter reaches zero, both the state y_t is allowed to make a transition, and the counter variable τ is reinitialized.

HSMMs are generative models, and a separate model is trained for each gesture.

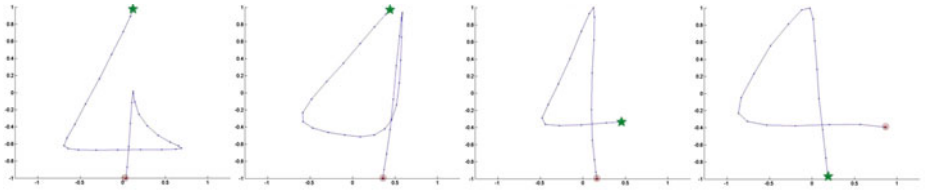


Fig. 4. Some common choices of trajectories by different users for the digit 4. Start and end points are swapped for the third and fourth trajectories.

4 Experiments

To justify the claim that pre–clustering is useful in terms of both speed and accuracy, we conduct several experiments on the digit dataset. First, the digit classes in the dataset are modeled with HMM, EDM, IOHMM and HCRF without pre–clustering. Then, the dataset is clustered using both spectral and hierarchical clustering, and the same graphical models are trained on the pre–clustered datasets. Finally, the accuracies and speeds of the models on these datasets are compared.

4.1 Gesture Dataset

The gestures were performed by 12 subjects, in 10 repetitions, yielding 120 exemplars for each class and a total of 1200 samples. Subjects were not instructed to follow a specific path. As a result, subjects used a wide variety of trajectories, yielding a difficult dataset with high variance. An example is given for the digit 4 in Figure 4.

4.2 Training Methods

HMMs are trained using the Baum-Welch algorithm, EDMs are trained using a generalized version of Baum-Welch algorithm extended for explicit durations [14], HCRFs are trained using the Broyden Fletcher Goldfarb Shanno [13] method, and IOHMMs are trained using generalized expectation maximization method [3]. Training IOHMMs and HCRFs take significantly more time than training HMMs and EDMs. To reduce training times of EDMs and IOHMMs, we initially constructed the models using a priori information obtained from trained HMMs. This reduced training times and increased the accuracy.

4.3 Parameter Optimization

We applied grid search and 5x2 cross validation for parameter optimization over all possible parameters. For HMMs and EDMs, the optimized parameter is the number of states. For HCRFs, both the number of states and the window size is considered. For IOHMMs, MLPs are used as local models. Therefore, IOHMMs have both two parameters: The number of hidden states, and the number of

hidden nodes. The optimum parameters for the models on both datasets are shown in Table 1. Here, the number of hidden states is depicted as N_S. N_H is the number of hidden neurons, L is the maximum duration for EDMs, and w is the window size for HCRF.

4.4 Results

The recognition results for the models are given in Table 1. Here, D_O is the original dataset and D_C is the clustered dataset. The models trained on the clustered dataset are constrained to have a left–right architecture. HMM, EDM and IOHMM reach 100% accuracy on the clustered dataset, and HCRF has a recognition rate of 98.95%. This shows that clustering is significantly effective for this problem.

For the original dataset, the number of hidden states that maximize the recognition rates are 16 for the HMM, 19 for the EDM, 8 for the IOHMM and 7 for the HCRF. IOHMM uses 5 hidden neurons, EDM states have a maximum duration 15, and HCRF has a window size of 3 in this case. On the clustered dataset, a left–right HMM with 3 states is capable of achieving perfect accuracy. As HMMs are special cases of EDMs and IOHMMs, these too need only 3 states. HCRFs, however, need more states to be able to distinguish between the increased number of class labels.

Table 1. Recognition rates and optimum model parameters on the original dataset D_O and on the clustered dataset D_C. N is the number of states, H is the number of hidden neurons, w is the window size and L is the maximum duration allowed for EDMs.

	Accuracy on D_O	N_S	N_H	w	L	Accuracy on D_C	N_S	N_H	w	L
HMM	89.7%	16				100%	3			
EDM	91.17%	19			15	100%	3			5
IOHMM	94.33%	8	5			100%	3	2		
HCRF	95.17%	7		3		98.95%	13		3	

Furthermore, the models trained on the clustered dataset are faster in comparison to their original counterparts, both due to their lower complexities and due to their left–right architectures, which are N_S times faster in general.

5 Conclusion

In this study, we proposed unsupervised clustering of gesture samples belonging to gesture classes to improve gesture recognition accuracy of commonly used graphical models. To justify our claims, we collected a challenging digit dataset and trained several graphical models on this dataset. Then we applied a DTW based clustering method to the original dataset and formed a clustered dataset.

We trained the same models on this dataset and achieved perfect accuracy for even very simple models.

This study shows that, rather than solving the isolated gesture recognition task by increasing the complexity of models, one can decrease the complexity of the gesture classes through preprocessing and clustering. Then, fast and simple models are able to attain good accuracies.

References

1. Bengio, Y., Frasconi, P.: Input-output HMM's for sequence processing. IEEE Transactions on Neural Networks 7(5), 1231–1249 (1996)
2. Hu, J., Ray, B., Han, L.: An interweaved hmm/dtw approach to robust time series clustering. In: 18th International Conference on Pattern Recognition, ICPR 2006, vol. 3, pp. 145–148 (August 2006)
3. Keskin, C., Akarun, L.: Stars: Sign tracking and recognition system using input-output hmms. Pattern Recogn. Lett. 30, 1086–1095 (2009)
4. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289. Morgan Kaufmann Publishers Inc. (2001)
5. Lee, H.-K., Kim, J.-H.: Gesture spotting from continuous hand motion. Pattern Recognition Letters 19(5-6), 513–520 (1998)
6. Liao, T.W.: Clustering of time series data - a survey. Pattern Recognition, 1857–1874 (2005)
7. Ma, G., Lin, X.: Typical Sequences Extraction and Recognition. In: Sebe, N., Lew, M., Huang, T.S. (eds.) ECCV/HCI 2004. LNCS, vol. 3058, pp. 60–71. Springer, Heidelberg (2004)
8. Meila, M., Shi, J.: A random walks view of spectral segmentation (2001)
9. Oates, T., Firoiu, L., Cohen, P.: Using Dynamic Time Warping to Bootstrap Hmm-Based Clustering of Time Series. In: Sun, R., Giles, C.L. (eds.) Sequence Learning. LNCS (LNAI), vol. 1828, pp. 35–52. Springer, Heidelberg (2001)
10. Pavlovic, V., Sharma, R., Huang, T.S.: Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Tran. on Patt. Anal. and Machine Intel. 19(7), 677–695 (1997)
11. Rabiner, L., Juang, B.: An introduction to hidden markov models. In: IEEE Acoustic Speech Signal Processing Magazine, pp. 3–4 (1986)
12. Shi, J., Malik, J.: Normalized cuts and image segmentation. In: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR 1997), p. 731. IEEE Computer Society, Washington, DC (1997)
13. Wang, S.B., Quattoni, A., Morency, L.-P., Demirdjian, D.: Hidden conditional random fields for gesture recognition. In: CVPR 2006: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1521–1527. IEEE Computer Society, Washington, DC (2006)
14. Yu, S.-Z., Kobayashi, H.: Practical implementation of an efficient forward-backward algorithm for an explicit-duration hidden markov model. IEEE Transactions on Signal Processing 54(5), 1947–1951 (2006)

Augmenting Social Interactions: Experiments in Socio-emotional Computing

Wijnand IJsselsteijn

Eindhoven University of Technology,
Eindhoven, The Netherlands
w.a.ijsselsteijn@tue.nl

Abstract. In recent decades, research on affective computing, social signal processing, and mediated communication has flourished. Combining these diverse fields leads to the new, multidisciplinary area of socio-emotional computing, where computing technologies are applied to transform and enrich communication between people, either mediated or face-to-face. As a research field, socio-emotional computing serves a number of goals. First, it aims to inform the design of communication media through identifying, implementing and validating those socio-emotional elements that enable or augment awareness, mutual understanding, empathy, and intimacy between people. Augmented social interactions can be beneficial to many application areas, including mental healthcare, training and coaching, behavior change, negotiation, and intimate social interactions. Secondly, research in socio-emotional computing allows us to obtain a more fundamental understanding of the impact of mediated communication on human intimacy and social connectedness. Finally, media tools developed to augment social interactions can, at the same time, serve as research tools to extend and improve research on the fundamental emotional and interpersonal processes underlying intimate communication. In this presentation, I will highlight some of the exciting research opportunities that emerge in this multidisciplinary field, and will present a number of experiments that exemplify socio-emotional computing as applied to intimacy, empathy, and persuasion.

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, p. 82, 2011.

An Energy-Saving Support System for Office Environments

Marc Jentsch, Marco Jahn, Ferry Pramudianto, Jonathan Simon, and Amro Al-Akkad

Fraunhofer FIT, Schloss Birlinghoven, 53754 Sankt Augustin, Germany
{marc.jentsch,marco.jahn,ferry.pramudianto,jonathan.simon,
amro.al-akkad}@fit.fraunhofer.de

Abstract. We present a system that helps office workers to save energy at work. The system features two concepts which are differing from current smart metering systems. It takes the special characteristics of office environments into account where saving energy has lower priority than the actual working processes. Firstly, the system uses unobtrusive technology in order not to interrupt the normal working processes of office workers. Secondly, the system minimizes the effort of workers to deal with the topic of saving energy so that it can be done en passant. In an explorative user study, we examine if the system is considered useful by users.

Keywords: Energy saving support, sustainability, ambient display, visualization, situation detection, mobile application.

1 Introduction

Raising energy efficiency is a current social topic in order to limit ecological damage like global warming and exhaustion of natural resources. At the same time, a more efficient usage of energy can help private households and companies to save costs. Equipment automation can help to save energy. For example, light barrier controlled lamps make sure that the light is switched off when nobody needs it. However, some energy-saving potential cannot be tapped by automation. For instance, when leaving a room, only the user herself knows whether the heater shall be left on the current level, because she will get back to the room soon, or the heater may be turned down to save energy. So, by involving the user, additional energy can be saved beyond automation.

To support users in saving energy, many related works focus on enhancing users' energy awareness by providing feedback on current energy consumption [1]. But knowing how much energy a particular device consumes does not always imply the knowledge how to change behavior in order to save energy. For example, although a user is told that her heater consumes 1000W she may not be aware that bleeding the radiator increases its efficiency. Hence, current approaches to energy-saving support offer potential for improvement.

Since many people spend much time of the day at the office, this environment provides huge potential for saving energy. At the office, working processes are usually higher prioritized than saving energy [8]. Working processes may not be interrupted; saving energy must happen alongside.

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 83–92, 2011.

In this paper, we present a system that supports users in actively participating in saving energy while it takes the special characteristics of office environments into account. The focus is on new types of visualization and presentation of energy information. More specifically, we use unobtrusive technology and minimize the effort which users have to take to save energy. We implement these concepts by using an ambient display to minimize intrusiveness of information and a mobile application to allow the user to take action when she wants. Furthermore, we give concrete recommendations how to save energy which can be put into practice by using remote control functionality.

The remainder of the paper is structured as follows. In the succeeding Section 2, we discuss related work. After that, we present our system in Section 3 followed by a preliminary user evaluation in Section 4. In the final Section 5, we conclude and give an outlook on future work.

2 Related Work

Early research on motivating energy efficient behavior in private households has been conducted by behavioral and environmental psychologists since the 1970s [1]. Most current research is based on the assumption that continuous real-time feedback is the main driver for creating awareness of energy consumption [4].

While most research on motivating sustainable behavior focuses on private households [7, 11], only little attention is paid to office environments. When talking about energy efficiency in office buildings, we mainly deal with two concepts: Structural changes (i.e. exchanging old devices with new energy efficient hardware) and automation (i.e. building management systems (BMS)). While both are necessary, neither of them takes into account the personal energy consumption behavior of the office workers.

Taherian et al. [15] tried to find out the potential of user-driven energy saving at work. Twenty-two employees at the Cambridge Computer Laboratory were told to switch off their devices (i.e. computers and lights) during their outside office hours for one week. The achieved energy savings of 15% for power sockets and 10% for lights show that behavioral changes can lead to significant energy savings.

Siero et al. [14] investigated the effects of comparative feedback to influence energy consumption behavior in work environments. The experiments were conducted in a metallurgical company, where saving energy has a similar low priority as in office environments. Employees who were given comparative feedback saved more energy than the ones who only received feedback on their own energy consumption. This shows that energy efficient behavior change in work environments can be facilitated by software systems.

In a recent study, Schwartz et al. [13] researched the opportunities and risks posed by using smart metering technology to support sustainable energy practices at work and to make energy saving potentials in work environments more transparent. They conducted a series of participatory design studies and an online survey, to explore the role of workers in energy consumption. They draw the conclusion that smart metering technology has the potential to make energy consumption transparent in office environments if conflicting interests of different stakeholders (e.g. privacy, data ownership) are taken seriously.

Prototypes of ambient displays to convey information on energy consumption have been developed [3, 6, 2] and applied mainly in the domain of private households. Rogers et al. [12] examine the question if ambient displays can, besides creating awareness, influence peoples' behavior at the point of decision making. They installed a display of twinkly lights in the floor to nudge people to use the stairs instead of the elevator. An interesting result was that the number of people using the stairs increased significantly but only few people said it was because of the lights. A possible explanation for this phenomenon is that the decision of taking the stairs has been made unconsciously, influenced by the ambient display.

Maan et al. [9] conducted a lab study where participants had the task to use as little energy as possible. At the same time they had to perform an ongoing task as cognitive load. The current energy consumption was visualized by LED light which could have red and green color for one group while the other group saw the current amount of Watt as numerical feedback. The visualization of energy consumption by ambient light led to 21% lower energy consumption than numerical feedback. The results show that ambient light can be used to influence energy saving behavior, which needs to be confirmed in practice.

3 System

3.1 Concept

In office environments, energy-saving has a low priority because companies prioritize working processes higher [8]. At the same time, while people save their own money by economizing on energy at home, they only save the money of their company when saving energy at work. The personal benefit of saving the own money by economizing energy at home is not given in the office. So, we cannot transfer the systems and concepts from private households to office environments without adaptation. Hence, we designed our system with a focus on two characteristics. Firstly, we use unobtrusive technology which is working in the background in order not to interrupt the normal working processes. Secondly, we minimize the effort which users have to make to save energy. The concepts were identified by Jahn et al. [8] in a series of reflection workshops about energy practices of office workers. The aim of the Business Ethnography process was to look for opportunities to enhance the energy practices.

Unobtrusive Technology. As first part of visualization, we implemented an ambient light concept which gives energy-saving hints that do not distract users from their activities. Ambient displays aim at conveying information to users in an unobtrusive way. They don't pop up requesting the user's full attention but stay in the background, addressing the peripheral vision. "Ambient Displays present information within a space through subtle changes in light, sound, and movement, which can be processed in the background of awareness." [16] Thus, ambient displays are not suited for providing complex or highly important information; they are suited to "[...] support monitoring of non-critical information." [10]

We believe that ambient displays are a promising approach to convey information about energy consumption in office environments for several reasons:

- The information is of low priority.
- It is not distracting.
- It does not interfere with work processes.
- It can influence users' behavior [12], [9].

The ambient light is mounted at the rear of a desk so that the light illuminates the adjacent wall. In a status without energy-saving potential, the light is green. Parts of the light turn red when the system detects potential for saving energy. Besides indicating that there is energy-saving potential, the light also gives rough hints where the cause is. For instance, if the heater, which is situated above the central part of the desk, is unnecessarily switched on, the ambient light turns red on the central part of the desk (cf. Figure 1). There is no gradual change of the color, is it always either red or green so that the status is obvious.

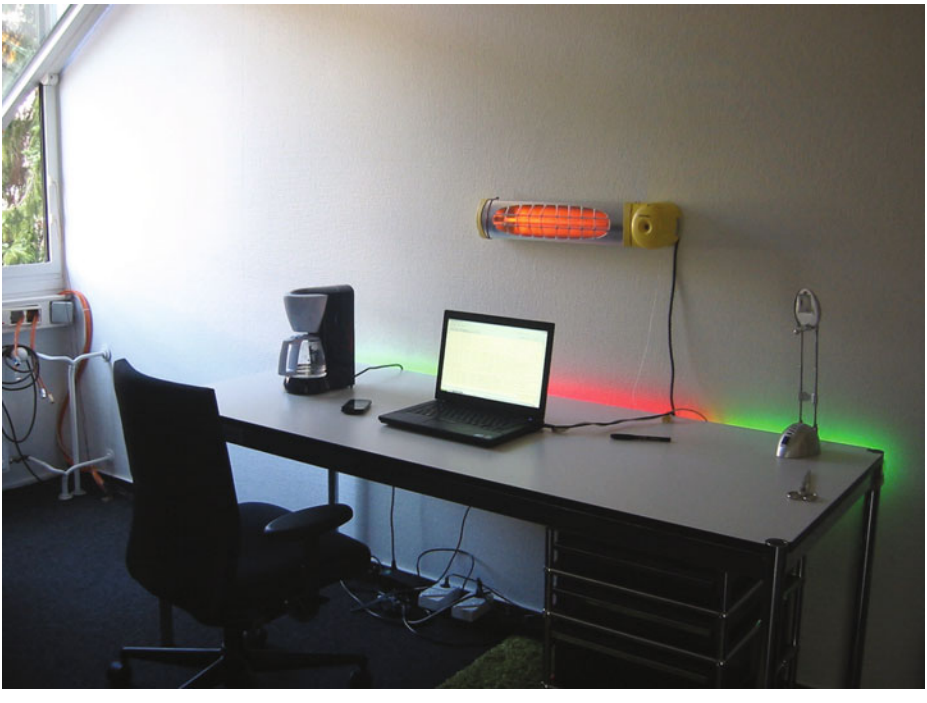


Fig. 1. Center of the desk is augmented with red light because the heater is switched on while the window is open. Left and right part are augmented with green light since the coffee machine and desk lamp are switched off (best viewed in color).

The ambient light is intended to push hazy hints to the users. If users wish to get more detailed information, they can utilize a second means of visualization. This second part is a mobile phone application that provides an overview of the office environment. Each object is represented by a small image which is supplemented by a traffic light icon

(cf. Figure 2). The traffic light visualizes energy-saving potential at a more detailed level for each object. Green color still represents no potential. We differentiate between low and high energy-saving potential. The first is indicated by yellow traffic light, the latter by red light.

Fig. 2. Overview of the office environment. The traffic lights indicate which object has energy-saving potential (best viewed in color).

This mobile application is implemented under the premise not to interrupt users during their working processes. That is why it is designed as pull service which gives users control when to request personal energy information.

Minimizing Effort. Instead of only visualizing energy consumption values, we provide direct recommendations what can be done to save energy. When the user touches an object image, the mobile application explains why there is energy-saving potential and what can be done to optimize efficiency. An example text is: ”The heater is turned on while the window is opened. Please turn off the heater while airing.” As an advantage, users do not have to spend effort on interpreting consumption values and drawing conclusions what to do. Users only have to decide if the recommended action would imply unacceptable consequences for them. They can also quickly compare the effort to the level of energy-saving potential which is visualized by the traffic lights.

As another feature for minimizing effort, we implemented a remote control functionality for devices that can be turned off by draining power. This feature is integrated in the recommendation window. For example, if the lamp is unnecessarily switched on, the recommendation window provides the text ”Please switch off the light” and a button for turning it off remotely.

3.2 Implementation

In our main application, we implemented a situation manager that subscribes to events which are generated by different sensors. The situation manager fuses these events to derive the current situation and its recommendations. It then calculates how much energy-saving potential the corresponding situation has. Finally, this module generates events to notify the user interfaces to display hints to the users. As a prototype, we modeled three situations that can be considered as wasting energy:

- Window is tilted or open and at the same time the heater is on (High energy-saving potential because the heater consumes 586W)
- Light is left on when the user leaves the room (Low energy-saving potential because the lamp consumes 42W)
- Coffee maker is left on when the user leaves the room (High energy-saving potential because the coffee maker consumes 637W).

We utilize several sensors to measure the current situation in the rooms. Each window is equipped with two contact sensors that are placed on the top and bottom of the opening side. These sensors are connected to an Arduino[1] board. When the window is tilted, the top sensor generates an event, and when the window is opened widely, both of the contact sensors generate events.

For measuring the power consumption, we use Plogg[2], a plug-socket bridge that has a power meter and a ZigBee radio inside. Ploggs send the real time power consumption wirelessly to the main application which raises an event if the application logic detects a state change e.g.: on, off, standby. Ploggs can also be remote controlled for switching power on or off.

Since we conducted our study on a trade fair, there was no office room available. Instead, we placed a rug in front of the desk and hid a dance mat game controller beneath it. We detect that a user is near the desk from pressure on the rug and defined this as the user being present in the office. However, as the data acquisition process is decoupled from the situation recognition process, these sensors can be flexibly exchanged, e.g. one can use an IR motion detector to sense the user's presence.

The ambient light is produced by a RGB LED stripe that allows each LED to be controlled independently. An Arduino board drives the stripe. The Arduino board is controlled by the main application via serial port. It indicates which parts of the stripe must change to which color. The introduced setup of LEDs and Arduinos consumes about 2W. The processing can be done on mobile phones and PCs that are switched on anyways so that there is no significant increase in energy consumption.

The mobile phone application is implemented for Android and communicates to the main application via wireless LAN. The situation events are relayed via MQTT[3].

4 Study

As the visualization concept is new for energy-saving support systems, we conducted a preliminary explorative user study to find out if people can imagine to use the system.

[1] `http://www.arduino.cc/`
[2] `http://www.plogginternational.com/`
[3] `http://mqtt.org/`

We investigated which part of the system users especially like and which features they miss. Questionnaires were favored over personal interviews because we aimed at receiving a representative number of answers. The results are the starting point for more detailed interviews and user tests.

4.1 Design

The study was conducted in conjunction with an appearance on an IT trade-fair. We set up a showcase with the described system (cf. Figure 1). On the left side of the desk, a coffee machine was placed. On the right side, there was a desk lamp. Above the center of the desk, an electronic heater was installed. Left to the desk, there was also a window that could be opened, tilted and closed. All electronic devices were fully functional.

For each participant, we first explained the concept and the functionality of the system. Afterwards, the participants could freely try it out as long as they liked to get a feeling for the system in use. Finally, after having everything tested to their satisfaction, we asked the participants to fill in a questionnaire covering their professional background and three questions concerning the system. The questionnaire was filled in anonymously to ensure that nobody feels forced to vote for green approaches.

4.2 Participants

We interviewed 31 volunteers, aged between 25 and 65 years, who are working in office jobs. Four participants work in micro enterprises, six in small enterprises, seven in medium-sized enterprises (definition according to [5]). Seven test persons work in corporations with less than 1000 employees and another seven in companies with more than 1000 employees. Ten participants work in research related jobs, nine work in IT jobs, the remaining participants are spread among several industrial sectors. Eleven test subjects held leading positions like management or head of department, 20 were employees. Six participants work in single, eleven in two-person offices and eight in offices with three or four people. Five participants work in open-plan offices with five or more people and one does not have a permanent office.

4.3 Results and Discussion

24 participants (7 managers and 17 employees) answered “Yes” to the multiple-choice question, *Can you imagine using a system like the one presented here at work?* Only one answered “No”, the remaining six checked “I don’t know”. So, our system seems to be deemed useful by the majority of the test persons. Chi-square test proved there was neither correlation of the employment positions (p=0.25) nor business sectors (p=0.42) on the acceptance of the system.

To find out which part of the system is considered particularly useful, we asked, *Which features do you find particularly useful to help you save energy?* The pre-defined options were: *Ambient light*, *Overview of devices on the phone*, *Energy-saving tips on the phone*, *Remote control of devices from the phone* and *Miscellaneous*. Multiple answers as well as no answer were allowed. Figure 3 summarizes the number of answers.

Only 1 user checked no answer, the user who checked *Miscellaneous* wrote she likes in general that energy consumption is visualized by the system. Overall, all features are considered useful by a large part of participants.

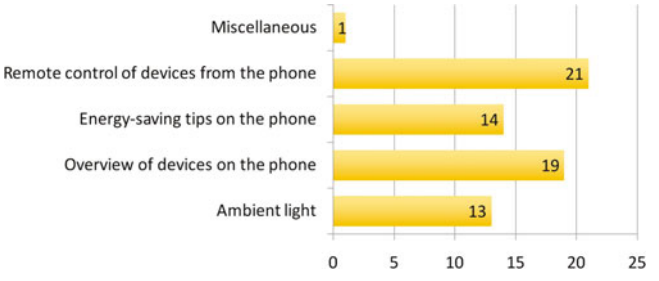


Fig. 3. Number of answers to the question: Which features do you find particularly useful to help you save energy?

To find out what is missing from the user's perspective, we finally asked *What additional features would you like to use?* without giving pre-defined options. Five participants wanted more detailed information like "information about the currently cheapest energy provider" or a "statistical analysis of energy consumption". Two test persons asked for different technology, namely "use of Bluetooth" and an "iPhone app". One participant liked to have something similar to the traffic light visualization "directly on the physical objects". One participant wanted the features to be "customized" to her own needs. She did not specify further in which sense this customization was meant.

The frequently mentioned features represent small changes which can be added in future work. Nevertheless, the few requested features of customization and extended traffic light visualization represent different concepts, which need further investigation.

5 Conclusion and Outlook

We introduced an energy-saving support system for office environments. It is using unobtrusive technology in form of ambient light and pull services in order not to distract workers from their activities. As second concept, the system is minimizing the effort that users have to make to save energy by presenting recommendations of what users can do to save energy and providing remote control functionality of devices. This shall help workers to save energy at the office en passant. The results of our user study indicate that every part of our system is considered useful by the users.

As the next step, we will conduct a user observation in which we let participants use the system over several weeks to find out if the system is still accepted in practice, especially after a long period of usage. Doing so, we will be able to find out if users really perceive the system as unobtrusive and helpful for changing behavior. This will also allow us to investigate practice-oriented questions: For example, if an energy-inefficient state is necessary for a work purpose for an extended period, is a red glow going to be

tolerated by users? By conducting interviews, we will deeper investigate user requests, which we already broached in the current questionnaire. Such qualitative approaches will make clear what the next steps are for refining the system.

Furthermore, we need to find out if the introduction of the system in an office can save more energy than it is consuming. Hence, we will solve the issues that currently prevent the prototype from being applicable in practice. We will substitute the dance mat, which we used as presence detector for the showcase, through more sophisticated presence sensors. We will also come up with an intuitive configuration concept for the position of devices. At the moment, the position of lamp, heater and coffee machine is fixed due to the implementation of the ambient light. Finally, we will investigate the suitability of the ambient light concept if two devices are situated close to each other.

Acknowledgements. This research was supported by the European Commission within the SEEMPubS project (project No. 260139) and the SEAM4US project (project No. 285408).

References

[1] Abrahamse, W., Steg, L., Vlek, C., Rothengatter, T.: A review of intervention studies aimed at household energy conservation. Journal of Environmental Psychology 25(3), 273–291 (2005)

[2] Backlund, S., Gyllensward, M., Gustafsson, A., Hjelm, I., Maze, R., Redstrom, J.: Static! the aesthetics of energy in everyday things. In: Proceedings of Design Research Society Wonderground International Conference 2006 (2007)

[3] Bartram, L., Rodgers, J., Muise, K.: Chasing the Negawatt: Visualization for Sustainable Living. IEEE Computer Graphics and Applications 30(3), 8–14 (2010)

[4] Darby, S.: The effectiveness of feedback on energy consumption. A Review for DEFRA of the Literature on Metering, Billing and direct Displays (April 2006)

[5] European Commission: SME Definition - Small and medium-sized enterprises (SMEs), `http://ec.europa.eu/enterprise/policies/sme/facts-figures-analysis/sme-definition/index_en.htm` (last accessed March 21, 2011)

[6] Gustafsson, A., Gyllensward, M.: The power-aware cord: energy awareness through ambient information display. In: CHI 2005 Extended Abstracts on Human Factors in Computing Systems, pp. 1423–1426 (2005)

[7] Jahn, M., Jentsch, M., Prause, C., Pramudianto, F., Al-Akkad, A., Reiners, R.: The energy aware smart home. In: 5th International Conference on Future Information Technology, FutureTech 2010, pp. 1–8 (May 2010)

[8] Jahn, M., Schwartz, T., Simon, J., Jentsch, M.: Energypulse: Tracking sustainable behavior in office environments. In: Proceedings of 2nd International Conference on Energy-Efficient Computing and Networking 2011 (2011)

[9] Maan, S., Merkus, B., Ham, J., Midden, C.: Making it not too obvious: the effect of ambient light feedback on space heating energy consumption. Energy Efficiency, 1–9 (2011)

[10] Mankoff, J., Dey, A.K., Hsieh, G., Kientz, J., Lederer, S., Ames, M.: Heuristic evaluation of ambient displays. In: Proceedings of the Conference on Human Factors in Computing Systems - CHI 2003, vol. (5), p. 169. ACM Press, New York (2003)

[11] Mattern, F., Staake, T., Weiss, M.: ICT for green. In: Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking - e-Energy 2010. ACM Press, Passau (2010)

[12] Rogers, Y., Hazlewood, W., Marshall, P., Dalto, N.: Ambient influence: can twinkly lights lure and abstract representations trigger behavioral change? In: Proceedings of Ubicomp 2010, pp. 261–270 (2010)

[13] Schwartz, T., Betz, M., Ramirez, L., Stevens, G.: Sustainable energy practices at work: understanding the role of workers in energy conservation. In: Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, pp. 452–462. ACM (2010)

[14] Siero, F.W., Bakker, A.B., Dekker, G.B., van Den Burg, M.T.: Changing Organizational Energy Consumption Behaviour Through Comparative Feedback. Journal of Environmental Psychology 16, 235–246 (1996)

[15] Taherian, S., Pias, M., Coulouris, G., Crowcroft, J.: Profiling energy use in households and office spaces. In: Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking - e-Energy 2010, p. 21. ACM Press, New York (2010), http://portal.acm.org/citation.cfm?doid=1791314.1791318

[16] Wisneski, C., Ishii, H., Dahley, A., Gorbet, M., Brave, S., Ullmer, B., Yarin, P.: Ambient Displays: Turning Architectural Space into an Interface between People and Digital Information. In: Yuan, F., Konomi, S., Burkhardt, H.-J. (eds.) CoBuild 1998. LNCS, vol. 1370, pp. 22–32. Springer, Heidelberg (1998)

From Stress Awareness to Coping Strategies of Medical Staff: Supporting Reflection on Physiological Data

Lars Müller, Verónica Rivera-Pelayo,
Christine Kunzmann, and Andreas Schmidt

FZI Research Center for Information Technologies, Karlsruhe, Germany

Abstract. Nurses and physicians on a stroke unit constantly face pressure and emotional stress. Physiological sensors can create awareness of one's own stress and persuade medical staff to reflect on their own behavior and coping strategies. In this study, eight nurses and physicians of a stroke unit were equipped with a wearable electrocardiography (ECG) and acceleration sensor during their everyday work in order to (a) make them aware of stress and (b) support the re-calling of experiences to identify stressors. In an interview one week later, the participants were asked to recollect stress related events through the examination of the sensor data. Although high activity levels diminished the expressiveness of the data, physicians and nurses could recall stressful events and were interested in their physiological signals. However, existing coping strategies turned out as barriers to the adoption of new tools. Future persuasive applications should focus on integration with existing coping strategies to scaffold the reflection process.

Keywords: Reflective learning, physiological sensor, user study, healthcare.

1 Introduction

According to the Health and Safety Executive [17] Stress has consistently been the second most commonly reported type of work-related illness in the UK and is responsible for 55% of workdays lost due to illnesses. Besides, employees in health and social work have the highest rate of illnesses across all occupations and industries. Employees are not aware or deliberately ignore their stress level until they develop depression, anxiety or burnout.

Today, wearable sensors [2,14] are becoming commercially available that measure stress indicators and can be unobtrusively worn during the whole day. These sensors provide a base for new persuasive computing [9] applications that create awareness of one's own stress level and provide assistance in avoiding stress or appropriating coping strategies.

However, there is no one-fits-all solution to cope with stress, because stressors largely depend on the specific workplace. While there are some techniques like relaxation exercises and breathing techniques, the causes of stress need a careful

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 93–103, 2011.

evaluation. Reflective practice is seen as a particularly promising approach in the healthcare profession[10], improving quality of care and supporting personal and organizational competence development. At its core are reflective learning processes, which can be understood as the re-evaluation of past experiences by attending to its various aspects (including feelings and emotions) and thereby producing outcomes [6]. More precisely, reflective learning at work means returning to and evaluating past work performances and personal experiences in order to promote continuous learning and improve future experiences. Sensors can provide the necessary data and moreover support the selection of relevant time spans for reflection.

In this paper, we analyze the potential of physiological sensors to (a) make the employee aware of stress and (b) support the re-calling of experiences to identify stressors and ultimately change their behavior. Towards that end, we used physiological sensors and conducted an ethnographically informed study based on a method similar to [4] in a German hospital, which will be described in the following section. Section 3 shows a concrete example how stressful events were examined from physiological data. The following Section 4 summarizes our results regarding data quality, usability and the subjective potential to recognize stress from the captured data. In Section 5, we discuss our results with special focus on existing coping strategies and conclude in Section 6.

2 Study Design

We have combined a sensor-based study with an ethnographically informed study, which has been extended with a contextualized interview that was based on preliminary findings of the observations.

For the first part, we selected the Movisens sensor [2] (as described in detail in Section 2.2). The sensors were worn by the study participants during at least two consecutive shifts of approximately 8 hours. After each shift, the participants were asked to state their experienced stress level on a 5 point scale for each hour of their shift. The sensors captured the daily activity and the physiological reactions of nurses and physicians.

The second part was based on the adapted rapid ethnographic method [12], which has been further developed in the context of the MATURE project [4]. Ethnographically informed methods are becoming increasingly popular in design-based research approaches, and their key characteristic is active participation in social settings to understand why things happen [8,11]. In contrast to field observation which describes what happens, ethnography focuses also on the why and how things happen. While traditional ethnography is based on long-term studies, the adapted method compensates the much shorter time frames with (a) a more focused observation scheme and (b) an interview at the end of the study that is used to clarify issues that arise from a preliminary analysis of the data. In this study, we had four days of observation, followed by an interview one week later.

In observation and interviews, special attention was given to existing coping strategies. For this purpose, it was essential to capture information about the employees interaction with their colleagues and patients. This provided insights into the employee's mindset, possible stressors and their reaction to stress during their work. In the interviews, we followed up on our insights and talked with the participants about their coping strategies, having as starting points the recall of the experiences made during the observation period.

The next sections describe the used sensor equipment, the tasks of the observer, and the structure of the concluding interviews.

2.1 Target Context

For the study, we have selected a stroke unit in a German hospital, the Neurological Clinic Bad Neustadt. A stroke unit is a specialized entity in hospitals that deals with acute cases of strokes. The slogan: "time is brain" shows the time pressure at the stroke unit. There are two main types of strokes. If these two types are confused the wrong treatment will aggravate the situation of the patient.

The time pressure and the daily work with emergencies and their results are a burden for all employees on a stroke unit. Some patients die, other will have to cope with disabilities for the rest of their lives. Currently the number of younger patients is increasing, which are in the same age group as the employees. Therefore, it is easier for nurses and physicians to relate to the individual patient and the emotional stress increases. Reflection about current practices and the knowledge of one's own physiological reaction might support the employees in their daily work.

Four physicians and four nurses took part in the study. The participants included all age groups at the stroke unit (22-44), men and women (3:5) and different levels of experience (1.5-25 years). The first part of the study, wearing sensors and observing employees, took place during four consecutive days. The interview was scheduled on two days one week after the study.

2.2 Sensor Equipement

Stress and cognitive load can be measured by monitoring the activity of the heart (electrocardiography - ECG) or the electrodermal activity (EDA) of the skin [7]. While EDA is more closely linked to the sympathetic activity of the autonomous nervous system [16], there are only a few appropriate positions to measure EDA, including fingers, palms and under the feet. In hospitals hands have to remain free and even wrist watches are forbidden. Hence, commercially available EDA sensors cannot be used in a critical environment like a hospital. The activity of the heart, especially the heart rate, is easier to capture by wearable sensors at the chest.

The ambulatory measurement system from Movisens [2] was selected to capture the activity level and physiological reactions, because of its simplicity for the user and the quality of the data. Commercial heart rate monitors for sports

[3] do not provide the necessary data quality and use wet electrodes that depend on the sweat of the user. Standard electrodes for ECG measurements use gel electrodes. They provide accurate measurement results, but are inconvenient. The ambulatory measurement system from Movisens uses dry electrodes that do not need special preparation before usage. Thus, test persons can use the system after a short introduction.

The ambulatory measurement system from Movisens, a shown in Figure 1, consists of a breast belt and a small sensor that captures a single channel ECG, the acceleration of the sensor in 3 dimensions, temperature and air pressure. The ECG monitors the physiological reaction of the user's heart. The acceleration sensors at the breast capture the main movements of the upper body and can be used to measure the physical activity of the user.

Fig. 1. The Movisens sensor and the sensor belt: On the inside of the sensor belt one of the two dry electrodes is visible. The sensor's battery supports 24 hours of recording ECG and acceleration data.

2.3 Ethnographically Informed Study

A subset of the participants (3 nurses and 2 physicians, 1 male/ 4 female) was followed by one observer who took the role of an ethnographer during their shift who collects additional data about the work practices and environment for later qualitative analysis as well as benchmarking of the sensor data.

The ethnographers (in total 3) had mixed professional background and experience healthcare to avoid bias in this respect. Each of them was in charge of following a participant during a whole shift.

The tasks of the ethnographers included being close to the participant and annotating (a) time, (b) place, (c) activity of the participant (d) and people that interact with the participant or influence his/her behavior and activities. The ethnographers followed their assigned participants during the whole shift, including work time and breaks. The annotations were made in a traditional notebook, which facilitates the skill to take notes anywhere and anytime, and have a level of detail of about 1 minute.

2.4 Reflection Interview

One week later, there were interviews of about one hour with each participant of the study, where the interviewer corresponded to the ethnographer. This seems to be a plausible scenario for reflection because time is limited and prevents daily reflection. The interview was structured into two parts.

In the first part, the sensor data of the participant was shown and chronologically analyzed with the UnisensViewer software [1], inviting the participant to remember what could have happened in specific timestamps where the curves indicated a special event. The UnisensViewer allows the visualization of different sensor data as line charts in a single window with flexible zooming. Where the data pool allowed it, the selected sensor data was taken from a quiet and a stressful day, in order to have the possibility to compare them. The participants were also given a printed report with the aggregation of the sensor data (heart rate histogram, Poincar plot of heart rate variability, heart rate per hour and activity in steps per hour).

In the second part, the participant was asked about the support that the data offers him/her to remember stress related experiences and which representation of the data is more suitable for her.

3 Examining Stressful Events

We used unprocessed data of whole shifts and asked users to recollect stress related events. In this section we present one example from our interviews. Figure 2 shows the overview of the data of an eight hour shift of a nurse. Figure 3 shows the details of a specific event during this shift. This event is clearly visible in

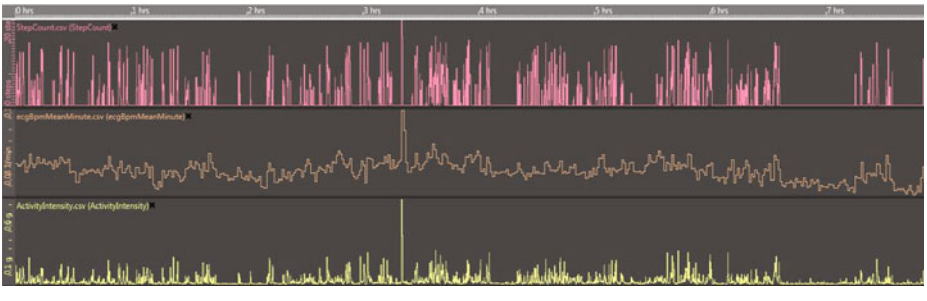


Fig. 2. Screenshot of the captured data as displayed in the UnisensViewer. The first row shows the number of steps, the second row the heart rate, and the bottom row the general activity. Clearly visible is a sudden peak of the heart rate after 3 hours and 20 minutes.

the data shown in Figure 2. At 3 hours and 20 minutes after the start of the recording a peak in heart rate and activity is visible. The participant immediately remembered this event and requested to see more details of this event. The

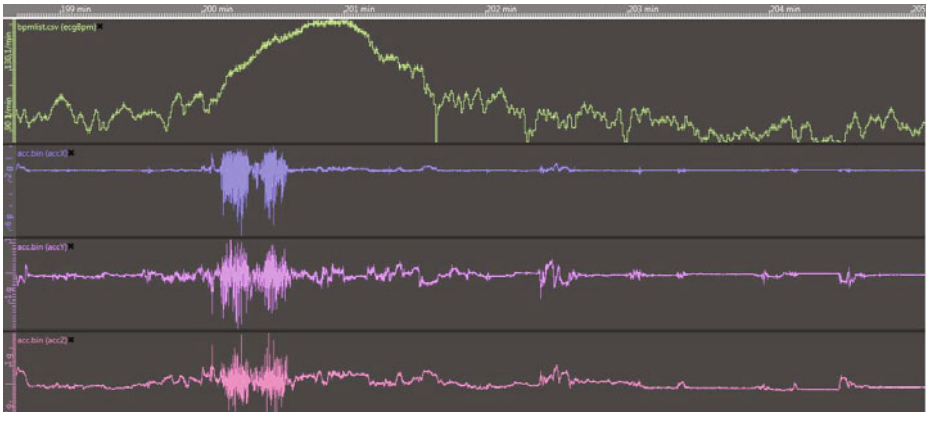


Fig. 3. Details of the reaction to a sudden emergency: the heart rate shown in the first row rises from 90 beats per minute to 155. This is mainly caused by the intense physical activity shown in the lower 3 rows. Close examination of the three activity curves shows the two sprints and a very short stop in between.

observation protocol notes that heavy muscle spasms of a patient surprised a novice nurse. She sprinted to alarm a physician and returned immediately to the bed. Back at the bed, she was about to inject the required drug without waiting for the physician. Other nurses calmed her down and the incident was resolved after 3 minutes.

The example in Figure 3 shows the stress reaction of the heart rate and the physical activity. Although, both effects overlap in the heart rate, these events can be clearly distinguished from normal activity, if the heart rate jumps to 155bpm. Minor stressful events, e.g. a heated discussion while walking, are difficult to capture by using the heart rate of a physical active person.

4 Results

152 hours of sensor data were captured and 49 of them were annotated with observed physical and estimated cognitive activity. 7 clearly and 15 probably stressful events were identified from the observation data.

In the interview, participants confirmed that dealing with stress is an important reason to use sensors. Measuring the own physiological data at work was interesting for all of them and the participants expressed their interest about recalling how were their work days and what had happened. Most of them stated that this interest is much higher when they had stressful days and that they would like to compare how the measures look like on different days.

> *'N1: Yes, it would interest me, especially when I had stress or emergencies.'*
> *'D1: How often I would use it... I can't tell you... If I had a 24 hours shift with 10 admissions with reanimation... '*

In the remainder of this section, we will outline the results from the analysis of the sensor data, the observation and the concluding interviews, regarding the captured data, usability aspects and the subjective potential to recognize stress.

4.1 ECG and Activity Data

The participants expressed that we monitored a set of rather quiet days.

> *'D1: It would have been more interesting for me if it hadn't been so quietly. I was waiting for an emergency to come, but nothing happened.'*

The daily questionnaire on the experienced stress level confirms these statements. The average over all participants and days was rated as 2.38 on the 5 point scale where 5 indicates extreme stress and 1 means a very calm day. Only three hours during the four days of the study were rated as stressful (4) by a single participant.

We used the ECG to calculate an accurate heart rate and the variability of the heart rate. Both parameters are well known parameters from psychophysiology and correlate to the stress level of a person [7,5]. During stress the 'fight-or-flight' response increases the heart rate. However, the activity of the heart is influenced to a larger extent by physical activity. Both effects overlap and make an analysis of stress levels complicated.

The data shown to the participants during the concluding interview was dominated by high activity levels. This activity mostly comprises walking between patients and offices. Figure 4 shows how many steps the different professions walked during their shifts. This physical activity results in an increased heart rate and hides potential stress related reactions of the heart rate. The observed breaks between activities were mainly used for documentation tasks and small

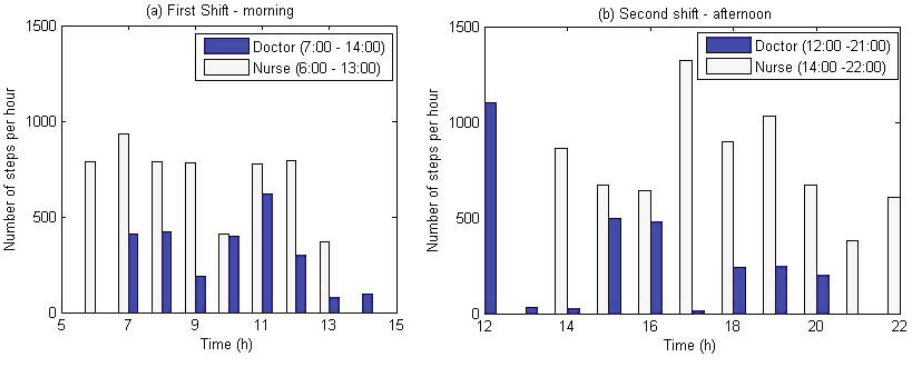


Fig. 4. Nurses are walking more than physicians and have fewer breaks for documentation. The figure shows the number of steps for a physician and a nurse during each hour of (a) first and (b) second shift. This high and varying activity levels influence the heart rate and conceal the effects of cognitive effort on the heart rate.

talk. Stressful events during inactivity could not be recorded. However, as shown in Section 3 there were stressful events that could be identified from the data even during activity. Nevertheless, these events represent only extremes that do not cover all interesting events for reflection. More information about the individual is necessary to distinguish physical and cognitive activity. This could be accomplished by longer term measurement or the recording of a baseline during several activity levels. Furthermore there are first approaches towards the calculation of the so called additional heart rate from heart rate and activity data [13].

4.2 Recognizing Stress and Identifying Stressors

Watching the curves in detail allowed the participants to compare their expectations to the measured sensor data and thus increase their awareness of their stress. In some cases, the data was surprising for them concerning e.g. the range of their heart rate or their appreciation of a specific moment.

'N2: . . . I thought I was calm. . . but now I see it wasn't like that. . . '

All participants acknowledged that the sensors had supported them to remember the course of the day. The curves helped them to structure the day and remember the overview of the day. Some of them could think about what happened in a specific moment, where the heart rate curve showed that something could have happened. Some participants could even explicitly say what had happened and why.

'N4: Yes! I can remember the two patients. They annoyed me. . . '
'N1: Yes [it helps me to remember]. I can say approximately when some things happened.'
'D1: Yes, it was interesting [the support of the sensors to remember]. It was interesting to see it graphically.'

However, reflection is a matter of time, and one participant explained how the physicians have to act quickly in specific situations, without having time to reflect.

'D1: We have to hurry up. On duty you can't do anything against it. What could I do better? You don't think. You are there, and you have to do it.'

Nevertheless, participants used the data to reflect on their behavior during the interview. They thought aloud about the experienced events and the reasons why they happened.

4.3 Application and Sensor Usability

The general interest of the participants about the use of sensors for tracking their work activities and doing a subsequent analysis was very positive. They were used to see physiological measures and such curves in their patients' monitors, but had not used them on their own before. As one of the participants stated:

'D4: . . . I don't like staying in hospitals and going to the doctor. I am not type of person keen on trying new things out. . . but it was actually interesting for me. I would mainly like to know about activity and movement.'

In general the participants accepted the belt for the study, but all of them saw room for improvement. Hence, they would not like to wear the sensor everyday but accepted it for the purpose of a study. One participant described the belt:

'D3: a badly fitting bra that is a little bit inconvenient but still wearable.'

The participants criticized that the belt was itching, that it was hard to adjust to the right size but the main point of critique was the electrodes. One participant noted that the dry electrodes tend to stick to the dry skin and that this is painful when the electrodes are moving and the electrodes are pulled off. However, a third participant said that the sweating caused the belt to become uncomfortable and proposed that the sensors could be integrated in an ordinary bra.

Concerning the type of visualization, the participants were asked to choose between the UnisensViewer application or a printed summary of the data. This summary showed the captured data aggregated on an hourly basis. All participants except one preferred the UnisensViewer, because it shows the measures in detail and they can discern the impact of specific events.

'D4: Amazing, it is easy to understand.'
'N2: I like the UnisensViewer more than the graphics. I can see everything what happened there and make a guess.'
'N1: Maybe UnisensViewer, then I can exactly see when, what time, something happened. .. With the graphics I can't see, when a seizure occurs, for example.'

However, nurses and physicians are already used to read heart rate diagrams. Users from other contexts might be confused by the amount of data.

5 Replacing Existing Coping Strategies

While our results have shown the potential of combining physiological sensors and reflection, the interviews showed that there are already coping strategies in place. If a persuasive application should replace or support these strategies, it has to provide clear benefits over the existing solutions, which are often non-technical but have proved to be successful for the individual. We gained several insights regarding existing coping strategies from our interviews:

– Developing effective coping strategies is part of gaining experience. Consistently, more experienced individuals had more developed coping strategies. They shared the characteristic of "professional distance" although they varied in their form. Some tried to suppress emotions and emotional reactions altogether in their work life, while others had more balanced views.

– Some coping strategies (especially for nurses) consisted of a clear separation between work life and private life that was linked to symbols, like changing clothes. It has been explicitly mentioned by one participant of the study that there is resistance to more reflection, as this might lead to rumination after work, which could result in negative emotional effects.

Moreover, we encountered different types of users that reacted completely different. While optimism of participants is important [15] further aspects have to be researched. Some users showed great interest, while others said that they do not want to know about this data to protect themselves. Applications that capture and present this data for reflection have to target a specific type of users, or these anxieties have to be addressed as part of scaffolding reflective processes for inexperienced employees.

6 Conclusion

This study has shown how a stress management solution can combine reflection and physiological sensors to analyze stressful events at work. We have used off-the-shelf components to create a basic system, and evaluated its impact in a real work environment.

This is only a first step towards stress management. Sensors are accepted for a limited time but are still not comfortable enough to use them every day. The captured and annotated data provides rich content for additional research. However, a straight forward analysis of the ECG data is complicated by the overlap between physical and cognitive activity. The additional measurement of the activity level provides an approach to distinguish both components.

Existing coping strategies turned out to be a barrier to the introduction of persuasive applications. Employees in hospitals already have their solutions, e.g. ignoring stress and they do not want to give them up. Especially, reflection about stress and stressors collides with their concept of "professional distance". They do not want to ruminate about stressful events. Most of our participants have been interested in stress management but accepted stress as part of their job.

In the short time of the study, participants did not change their behavior. However, the study has shown promising results to create awareness about stress and remember stressful situations to identify stressors. The system was positively accepted by the participants and in most cases the sensor data supported the recall of their personal work experiences. Further investigation is needed with respect to the acceptance by different types of personalities and the role of experience.

With these promising results, we are planning to develop tools that facilitate the reviewing of sensor data with more possibilities of aggregation, benchmarking with other people, and highlighting of stressful time spans. These tools will be evaluated on a larger scale and with a longer time frame.

Acknowledgements. This work has been co-funded by the European Commission within the 7th Framework Programme in the MIRROR project (http://www.mirror-project.eu).

References

1. Unisens - a universal data format, `http://unisens.org/viewer.php`
2. Movisens - ECG- and activity sensor (2011), `http://www.movisens.com/`
3. Polar - listen to your body (2011), `http://www.polar.fi`
4. Barnes, S., Bimrose, J., Brown, A., Feldkamp, D., Kaschig, A., Kunzmann, C., Maier, R., Nelkner, T., Sandow, A., Thalmann, S.: Knowledge maturing at workplaces of knowledge workers: Results of an ethnographically informed study. In: Proceedings I-KNOW (2009)
5. Berntson, G., et al.: Heart rate variability: Origins methods and interpretive caveats. In: Psychophysiology, pp. 623–648. Cambridge University Press (1997)
6. Boud, D., Keogh, R., Walker, D.: Promoting Reflection in Learning: a Model. In: Reflection: Turning Experience into Learning, pp. 18–40. Routledge Falmer, New York (1985)
7. Cacioppo, J., Tassinary, L., Berntson, G.: Handbook of Psychophysiology. Cambridge University Press (2007)
8. Fetterman, D.M.: Ethnography Step by Step. Sage Publications Ltd., London (1999)
9. Fogg, B.: Persuasive technology: using computers to change what we think and do. Morgan Kaufmann Publishers, US (2003)
10. Jasper, M.: Beginning Reflective Practice (Foundations in Nursing & Health Care). Nelson Thomas Ltd. (2003)
11. Jordan, B.: Ethnographic workplace studies and cscw. Human Factors in Information Technology 12, 17–42 (1996)
12. Millen, D.R.: Rapid ethnography: time deepening strategies for hci field research. In: 3rd Conference on Designing Interactive Systems, New York (2000)
13. Myrtek, M.: Heart and emotion: Ambulatory monitoring studies in everyday life. Hogrefe & Huber Publishers (2004)
14. Poh, M., Swenson, N., Picard, R.: A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. IEEE Transactions on Biomedical Engineering 57(5), 1243–1252 (2010)
15. Scheier, M.F., Weintraub, J.K., Carver, C.S.: Coping with stress: Divergent strategies of optimists and pessimists. Journal of Personality and Social Psychology 51(6) (1986)
16. Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., Ehlert, U.: Discriminating stress from cognitive load using a wearable eda device. IEEE Transactions on Information Technology in Biomedicine, 410–417 (2010)
17. The Health and Safety Executive: Statistics 2009/10. A National Statistics Publication

Why Won't You Do What's Good for You? Using Intelligent Support for Behavior Change

Michel Klein, Nataliya Mogles, and Arlette van Wissen

Department of Artificial Intelligence,
VU University Amsterdam
{michel.klein,n.m.mogles,a.van.wissen}@vu.nl

Abstract. Human health depends to a large extent on their behavior. Adopting a healthy lifestyle often requires behavior change. This paper presents a computational model of behavior change that describes formal relations between the determinants of behavior change, based on existing psychological theories. This model is developed to function as the core of a reasoning mechanism of an intelligent support system that is able to create theory-based intervention messages. The system first tries to determine the reason of the occurrence of the unwanted behavior by asking short questions via a mobile phone application and by gathering information from an online lifestyle diary. The system then attempts to influence the user using tailored information and persuasive motivational messages.

1 Introduction

A good health requires a healthy lifestyle. However, it is not easy to find (and keep) the optimal balance between work, a social life and, for example, a healthy diet or medicine schedule. Moreover, people with a chronic disease have extra barriers to overcome, such as physical discomforts and side-effects of medicine intake. In short, people have lots of reasons not to do what's good for them. As a consequence, the amount of people that have obesity or a chronic disease such as diabetes type 2 has increased considerably over the past years [28].

It has been shown that patient engagement and empowerment could improve patient therapy adherence and consequently their health condition [18]. This engagement and empowerment is often referred to as *self-management*: the individual's ability to monitor one's condition (symptoms, treatment) and to effect the cognitive, behavioral and emotional responses necessary to maintain a satisfactory quality of life [4]. But how can we get patients to increase their self-management? The use of computers to support people with their self-management has proven to be an effective approach [33,17]. These systems are able to provide personalized (*tailored*) interventions at low costs [8] and at home [27]. Interventions that are closely tailored to the individual's convictions and motivations have shown to be more likely to be read and remembered [29].

Although intelligent persuasive assistants are increasing popularity for the use of behavior interventions, those assistants are rarely based on formal models of behavior change. In their 2008 article, Michie, Johnston, Francis, Hardeman and

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 104–115, 2011.

Eccles stated that: "Ideally, researchers designing interventions would choose a small number of the theoretical frameworks based on empirical evidence of their predictive and intervention value, i.e., there should be evidence that the theory can predict the behaviour and that interventions which change these determinants achieve change in behavior."[20] In literature however, very few works can be found that provide a model based on formal theories. One notable exception is the the iChange model [34], which describes the factors that influence behavior change, but fails to explicate how these factors interact. Yet in order to design an effective support system, it is necessary to take a closer look at the underlying mechanisms of behavior change and how the they can be influenced to establish the desired behavior. The current paper addresses this and presents a computational model based on theoretical frameworks of behavior change. It is used by an intelligent support system to understand human behavior and to detect the cause of unhealthy behavior, which enables the system to provide users both tailored information and persuasive motivational messages on how to improve their behavior.

An overview of approaches for intelligent support systems is provided in Section 2. In Section 3 a model that formalizes the interaction between the different determinants of behavior change is presented. Section 4 demonstrates how this model can function as the basis of an intelligent system that is able to provide support for individuals with a health condition (such as diabetes, HIV or obesity) by stimulating their self-management. Section 5 concludes the paper and gives some implications for further research.

2 Approaches for Intelligent Coaching and Mobile Persuasion

In order to point out the differences between the proposed system en other approaches, this section provides a brief overview of existing approaches for intelligent coaching. The main component of many contemporary approaches is the mobile phone, as they are easily available to the user and support both user and system initiated interactions. Also, information provided by the mobile phone can be personalized and can even be designed to persuade or manipulate [10]. Because of these capabilities, the mobile phone is an ideal platform to provide us with the power to induce behavior change [10]. Mobile phones and web-based interfaces have proven to be very effective in similar approaches [10].

The simplest of the approaches to induce behavior change are 'reminder systems', which do not use complex persuasive techniques but instead use simple messages to remind the patient of the desired behavior (e.g., [1,7]). For example, CARDS (Computerized Automated Reminder Diabetes System) [12] sends diabetic patients SMS messages and e-mails with reminders about blood monitoring, without further medical advice from a healthcare team. Under this category fall also the popular mobile phone and web-based applications that help patients keep track of data such as calorie intake, blood monitoring and exercise by means of an online mobile dairy (e.g., [15,30]).

More complex systems are able to provide tailored feedback based on user data that is gathered by sensors (such as an accelerometer or GPS) or user input (such as a dairy function). Most systems use some kind of human coaching to supplement their system (e.g., [5,23]). The ODA (Online Digital Assistance) system [32], for example, is developed to support self-management of patients with chronic migraine by training behavioral attack prevention. ODA combines a mobile electronic diary with direct human online coaching, based on the diary entries. Persuasive systems that do not rely on human coaches (that is, while the system is active; healthcare professionals can still be part of the design process), are less common. Recently, this area has been given more attention. The system developed by D. Preuveneers and Y. Berbers in [24] assists diabetic patients to keep track of their food intake, blood glucose levels and insulin dosage. It uses relevant user context and activities (provided by user input and GPS) to learn trends and give tailored advice to the user. As another example, the persuasive computer assistant implemented by Blanson Henkemans et al. uses an online lifestyle dairy to improve exercise behavior of people who are overweight [13]. This assistant follows the principles of motivational interviewing and offers support by monitoring the dairy and providing tailored feedback.

All mentioned studies stress the potential of mobile and online support for patient self-management. The system presented in this work differs from previous approaches in that it does not only target the user's behavior, but also the underlying mechanisms causing that behavior. Because the system uses a computational model based on the theoretical frameworks of behavior and behavior change, it is able to provide tailored feedback that is not just focused on displayed behavior, but on the underlying individual cause of non-adherence. Furthermore, the system uses validated persuasion techniques without having to rely on a human coach, and combines support on three lifestyle domains: medicine, diet and exercise.

3 Modelling Behavior Change

3.1 Theories on Behavior Change

For health interventions to be effective, they need to incorporate existing theories on behavior change and persuasive design. The model of behavior change designed in this work is based on several existing models from psychology literature that describe determinants for behavior change. This section will describe their key constructs and how they are combined.

The Transtheoretical Model (TM) [25] forms the basis for the proposed model of behavior change. This model was successfully applied in many programs aiming at the elimination of addictive behavior, improving mental health, exercise, and dietary change [2,26]. It assumes that behavior change is a five-stage process with the stages of *precontemplation*, *contemplation*, *preparation*, *action*, and *maintenance*. Depending on the awareness, motivation and commitment of an individual, he or she progresses through the stages. In the precontemplation stage individuals have no intention to change their behavior and will likely be

unaware of their problems. In the contemplation stage individuals are aware that a problem exists and are seriously thinking of changing their behavior in the next six months, but they do not have any concrete plans of change. Individuals are defined as precontemplative when they are intending to take action in the next month but have not or not successfully taken action in the past year. During the action stage individuals modify their experiences and environment in order to overcome their problems and actively changing their behavior. Those who have engaged in a new behavior for more than six months are classified as being in the maintenance stage. Although a person advances through the stages in sequential order, relapse to a previous stage is possible. For an elaborate description of the separate stages, see [25].

According to the **Social Cognitive Theory (SCT)** of Bandura [3] behavior is executed if one perceives (i) control over the outcome, (ii) few external barriers and (iii) confidence in ones own ability. Bandura introduces a new concept that relates to the expectancies concerning the outcome: *self-efficacy*, defined as confidence in one's own ability to carry out a particular behavior. The concept of self-efficacy has shown to be a good predictor of behavior, related to coping with stress and recovery from illness [2].

Self-Regulation Theories (SRT) regard an individual as an active problem solver whose behavior reflects an attempt to close the gap between his current status and a goal. Levental's self-regulation model of illness identifies 3 stages of variables regulating the adaptive behavior: cognitive representation, action plan, *coping* and appraisal stage [22]. Important aspect of this approach is the possible influence of emotions, or *mood*, on behavior.

The Theory of Planned Behavior (TPB) is a revised version of the **Theory of Reasoned Action (TRA)** that was proposed by Fishbein and Ajzen [9]. The Theory of Reasoned Action is based on the assumption that *intention* is an immediate determinant of behaviour, and that intention, in turn, is predicted from *attitude* (which is a function of the *beliefs* held about the specific behaviour, as well as the evaluation (value) of the likely outcomes) and (subjective) *social normative* factors. In a more recent version of the theory, the Theory of Planned Behavior, one more component was added: *perceived behavioral control*, which has a motivational effect on intentions. This version was an attempt to account for behaviour under 'incomplete' volitional control. There is substantial overlap between the concept of self-efficacy in Bandura's Social Cognitive Theory and the concept of behavioral control in the theory of Ajzen and Fishbein.

The Theory of Reasoned Action does not describe explicitly what the determinants of attitude formation are. The **Attitude Formation (AF)** theory defines *attitude* as an important aspect of behavior, influenced by the *beliefs* about an object (in this case, behavior), emotional connotations associated with the object, and social norms concerning this object in this case [31].

The Health Belief Model (HBM)[16] includes six determinants of behavior related to perception: *susceptibility*, *severity*, *benefits*, *barriers*, *motivation* and *cues* for action. According to this theory, a combination of perceived

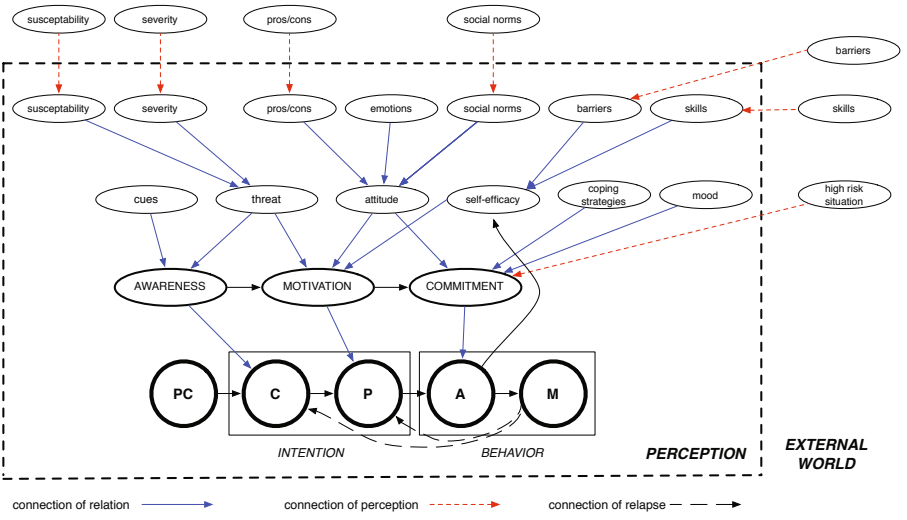


Fig. 1. The integrated model of behavior change COMBI

susceptibility with perceived severity produce perceived *threat*, and the combination of perceived benefits with perceived barriers produce evaluation of the course of action taken.

Marlatt and Gordon's [19] **Relapse Prevention Model (RPM)** describes the influence of environmental factors along with the cognitive determinants, such as *self-efficacy* and *coping*. The emphasis lies on *high risk situations* and the ability of *coping* with them. The theory provides an explanation of relapse from the acquired behavior stage to the stage of the previously performed behavior in the terms of the Transtheoretical Model.

3.2 Integrated Model: COMBI

It is evident that there is a lot of overlap between the existing theories of behavior change, and many of the theories use similar constructs with sometimes different names. The COMBI model –which stands for Computerized Behavior Intervention– is an attempt to integrate these theories (see Figure 1) into a formal representation.

The description of all factors and the theories they originate from can be found in Table 1. The model differentiates between the internal and external determinants of behavior. External factors are depicted beyond the dotted line, these are *susceptability*, *severity, pros/cons, social norms, barriers, skills* and *high risk situation*. *Susceptability* and *severity* represent how one perceives the severity of the consequences of the performed behavior and the likeliness of being affected by them, *pros/cons* correspond to the beliefs about the importance of healthy lifestyle. *Social norms* reflect the influence of culture and environment

Table 1. The concepts of the model and the related theories

concept	description	related theory
susceptibility	likeliness of being affected by behavior's consequences	HBM
severity	severity of the consequences of the behavior	HBM
pros/cons	beliefs about the importance of healthy lifestyle	TPB,AF,HBM
emotions	feelings concerning the behavior change	SRT
social norms	the influence of culture and environment of a person	TPB
barriers	practical obstacles that prevent behavior change	HBM
skills	experience and capabilities to overcome the barriers	TPB,SCT
cues	environmental or physical stimuli	HBM
threat	perceived (health) risk of continuing to perform behavior	HBM
attitude	mental state involving beliefs, emotions and dispositions	TPB,AF
self-efficacy	perceived behavioral control	SCT,TPB,RPM
coping strategies	the ability to deal with tempting situations and cues	SRT,RPM
mood	temporary state of mind defined by feelings and dispositions	SRT
high-risk situations	contexts/environments that influence a person's behavior	RPM
awareness	conscious knowledge of one's health condition, the health threat and the influence of current behavior	TM
motivation	incentives to perform goal-directed actions	HBM,TM
commitment	(intellectual or emotional) binding to a course of action	TM

of a person, and *barriers* correspond to real obstacles that prevent a person from adopting a healthy lifestyle. *Skills* determine how much experience and capabilities one has in order to overcome these barriers. *High risk situation* reflects the possibility of certain contexts to influence person's behavior. Examples of *high risk situations* are negative emotions as a result of an interaction with others, experienced pressure and some cues in the environment that lead to a particular behavior.

The stages of change from the Transtheoretical Model are represented as five circles with the initial letters of the names of the stages at the bottom of Figure 1. The *contemplation* and *preparation* stages ('C' and 'P', respectively) are embedded in the 'intention' block and *action* and *maintenance* stages ('A' and 'M', respectively) are embedded in the 'behavior' block. All internal factors that determine the stage of change of an individual consist of 3 layers, showing the causal hierarchy between them. The action stage has also a feedback loop to self-efficacy, in accordance with the Self-Regulation Theory.

3.3 Formalization and Simulations

The COMBI model has been implemented in the numerical simulation environment Matlab. In this section, the formalization of the model is described and some simulation results are provided as illustration.

The arrows in Figure 1 denote causal dependencies (with the exception of the arrows between the stages of change): they represent transitions from one state to another that occur if the value of a state exceeds a certain threshold. For example, if the value of awareness, motivation or commitment is greater than 0.5, a transition to the next relevant state occurs; if the value drops to the level lower than 0.5, the person relapses to the previous state. Dependences between the concepts are expressed by weighted sums:

```
Rule 1: Calculation attitude value

If  pros/cons have value V1
and emotions have value V2
and social norms have value V3
and connection strength between pros/cons and attitude has value w1
and connection strength between emotions and attitude has value w2
and connection strength between social norms and attitude has value w3
Then attitude will have value w1 * V1 + w2 * V2 + w3 * V3
```

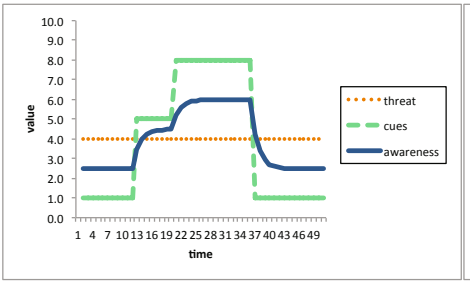
All other values in the model are calculated in a similar way. The formal model can be personalized by adjusting the links (connection strengths) between the determinants. For example, the behavior of some people is much more affected by mood or the lack of social support than that of others. The same argumentation holds for the transition from the external world to it's perceived internal representation. By increasing or decreasing the connection strengths between the determinants, these personal variations can be accounted for. In principle, the relevant connections can be updated when a discrepancy is discovered between observed patient behavior and the predicted behavior from the model.

Figure 2 shows some simulation results displaying the interplay between the different determinants of the model. These simulations show that the model can account for behavioral phenomena found in psychology and sociology.

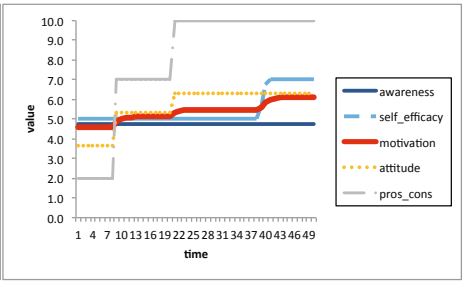
In Figure 2a it can be seen how the values of threat and cue contribute to the value of awareness. In this scenario (let's say it is about Alan), the threat Alan perceives – i.e., how likely he thinks it that he has this disease and how severe it's consequences are for him – remains constant. However, his cues (physical discomforts) increase drastically, making him much more aware of the condition he is in, until his symptoms recede again. (This is a well-known phenomenon, see e.g. [11].) Take a look at Figure 2b for another scenario, Betty's. At t=0, both awareness and motivation are low. Betty then (at t=8 and t=20) receives some information about how changing her behavior can contribute to a better health. Thus, she becomes better informed about the positive and negative consequences of her behavior. Unfortunately, Betty's attitude (and hence motivation) improve only slightly on learning this new information, as other factors –such as emotions and social norms– are stronger determinants of her attitude. Figure 2c shows how the commitment of Carol fluctuates with her mood. Although the strong improvement in social support gives her commitment a little boost, her mood is the key determinant of her commitment.

4 Implementation

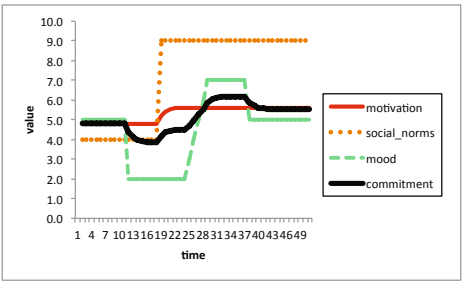
The model described in the previous section has been used as basis for an intelligent coaching system, called **eMate**. This section describes the system, explains how the model is used to reason about the state of the user, and how the system interacts with the user.



(a) Dynamics of awareness



(b) Dynamics of motivation



(c) Dynamics of commitment

Fig. 2. Model dynamics

4.1 The eMate System

The **eMate** system aims to support patients with Diabetes Mellitus type II or HIV in adhering to their therapy, which consists of lifestyle advice and/or precise instructions for medication intake. Previous research has shown that a 'cooperative assistant' – i.e., with a coaching character, able to explain and educate, and expecting high participation of the user – is more effective than a 'direct assistant' – i.e., with an instructing character with brief reporting and low expectations on participation [14]. The **eMate** system therefore operates as a coach, using both a mobile phone and a website to interact with the user. Via the website, the user can get an overview of his progress on three different domains: medication intake, physical exercise, and healthy food intake. If one of the domains is not relevant for a specific user, it will be hidden. An overview shows the extent to which the user has reached his/her goals in the past week, which is represented as a percentage and a iconic thumb. See Figure 3 for an example. A mobile phone application for the Android platform has been developed that can pose questions and send messages to a user.

4.2 Model-Based Reasoning

The model is used to analyze the state of the patient with respect to his/her behavior change goals. It does so by investigating via simple questions which of the factors that influence behavior change are probably the most problematic for

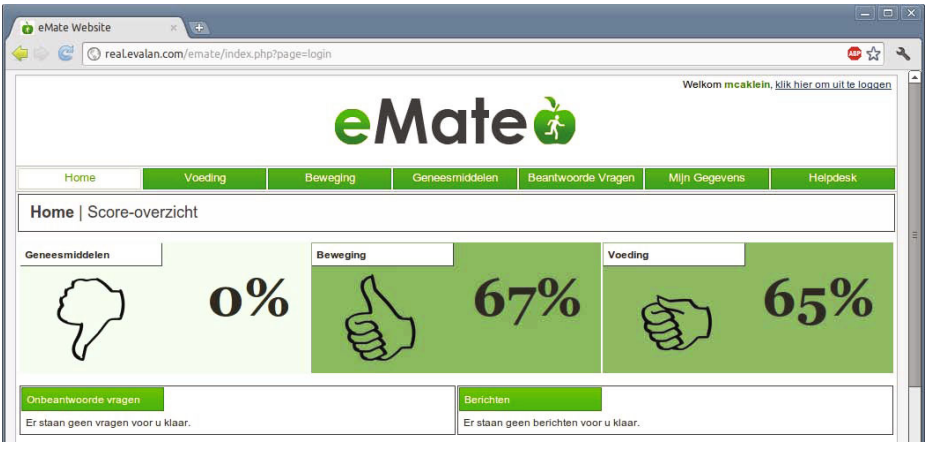


Fig. 3. Screenshot of the **eMate** website

this patient. This mechanism is called model-based diagnosis [6]. These factors are then targeted with specific messages and interventions. For this purpose the model has been translated into a rule-based representation that allows for backward reasoning over the psychological factors in the model. To achieve this, the rules relate factors in the model that have an 'influence'-relation, i.e. if there is an arrow between two factors in the model (see Figure 1), a rule specifies that a low value of factor A could be caused by a low value of an influencing factor B. For example, two rules specify: [1]

"if threat has_value < 5 & threat is_hypothesis **then** severity set_hypothesis"
"if severity is_hypothesis & has_value NULL **then** severity investigate"

The rules are implemented in a Java-based rules engine (Drools). Using these rules, the system determines for which psychological factors the value should be determined. This reasoning is performed on a regular basis and is done by posing specific questions about that concept from psychological surveys to the user. As some factors are more dynamic than others, the values will be redetermined after some time; the lifetime of the values is specified per factor. The user answers to the questions translate to values for each concept. These are stored in a database along with a timestamp of their determination. This way, the system maintains an up-to-date representation of the mental state of the user. The reasoning is performed separately for all domains that are active for a patient. However, some values, i.e. 'mood', 'cues', 'skills', 'severity', 'susceptibility' and 'threat' are considered to be equal for the different domains, and their value is automatically propagated to the other domains via the rules.

After this diagnostic phase, the system determines which factor should be targeted at to support the user in the most effective way. This is calculated by

[1] Note that the personalized parameters of the strengths between factors are ignored in this representation.

combining the 'urgency' of the value (how low the value is for a factor) with the 'changeability' of the factor, which is a parameter that represents to what extent the factor can be changed. For example, the social norms of a person are more difficult to change than the perception of the severity of the disease. Each week, the user will receive for each of the domains a summary of his behavior and a motivating message related to the factor on which the intervention should focus. The system contains several messages for each factor, so if the same factor is targeted in two subsequent weeks, the messages will still be different.

4.3 Questions and Messages

Due to the model-based reasoning, **eMate** is able to address the right problems at the right time. However, in order to persuade a user, the formulation of the queries and messages are also important. All messages are designed in such a way that the user won't be annoyed or bored by lengthy information messages (this approach is typical for tailored health messages that are commonly used in web-based solutions [17]). Furthermore, the motivational messages adhere to the principles of motivational interviewing, which have proven to be effective for purposes of coaching and therapy [21]. These principles focus on the social functioning of the user and on providing feedback by giving advice and direction. Expressing empathy, cheering and complimenting, and the support of self-efficacy and optimism, are some examples of the principles that are incorporated by the **eMate** system.

5 Discussion and Conclusions

This work presents the design and use of a computational model for behavior change. It has been shown that the model can be incorporated in a coaching system, which has a strong potential of providing support for individuals with respect to their lifestyles. The integrated model is an example of a causal modeling approach to developing complex, user-tailored interventions aimed at behavior change. **eMate** differs from other intervention approaches in that it targets the user's motivation and interests, and tailors intervention messages based on the underlying mechanisms of behavior change, thus attempting to understand the behavior.

Although developed for HIV and diabetes type II patients, it is expected that the flexible setup of the system is able to deal with other behavior change goals (such as quitting smoking or increasing the level of physical exercise for healthy persons), as the general mechanisms for these changes of behavior are similar to the ones implemented in **eMate**. Moreover, the rules and tailored messages can easily be changed to include different conditions and requirements. In the future, the model could also be used to predict the effect of an intervention, in order to let the system choose the most effective one.

Of course, the model is not able to capture every aspect of human behavior, as human behavior is the result of an interplay between different external

and internal factors, including biological, cognitive, environmental and socio-demographic factors. As the current model has been designed for intelligent health intervention applications that aim at behavior change, only the variables that are potentially amenable to change in the course of an intervention have been taken into consideration.

We intend to test and validate the model by setting up experiments with real users. A group of patients with either HIV or diabetes will be provided with the system. Behavior of this experimental group will be compared to a control group consisting of similar patients that were provided with only a website with static information about the importance of a healthy lifestyle and medication adherence. Validated pre- and post-questionnaires will be used to determine whether behavior change occurred in both groups.

Acknowledgement. This work is supported by the ZonMW programme "chronic disease management", grant number 300020005.

References

1. Anhoj, J., Nielsen, L.: Quantitative and qualitative usage data of an internet-based asthma monitoring tool. Journal of Medical Internet Resources 6(23) (2004)
2. Armitage, C.J., Conner, M.: Social cognition models and health behaviour: A structured review. Psychology and Health 15, 173–189 (2000)
3. Bandura, A.: Self-efficacy: Toward a unifying theory of behavioral change. Psychological Review 84(2), 191–215 (1977)
4. Barlow, J., Wright, C., Sheasby, J., Turner, A., Hainsworth, J.: Self-management approaches for people with chronic conditions: a review. Patient Education and Counseling 48, 177–187 (2002)
5. van den Berg, H.M., Rondaly, H.K., Peeters, A.J., van der Harst, E.M.V., Munneke, M., Breedveld, F.C., Vlieland, T.P.M.V.: Engagement and satisfaction with an internet-based physical activity intervention in patients with rheumatoid arthritis. Rheumatoloty 46(3), 545–552 (2007)
6. Davis, R.: Diagnostic reasoning based on structure and behavior. Artificial Intelligence 24(1-3), 347–410 (1984)
7. Downer, S.R., Meara, J.G., Costa, A.C.D.: Use of sms text messaging to improve outpatient attendance. Medical Journal of Australia 183(7) (2005)
8. Eysenbach, G.: What is e-health? Journal of Medical Internet Research 3(2)
9. Fishbein, M., Ajzen, I.: Belief, attitude, Intention and behaviour: An Introduction to Theory and research. Addison-Wesley, Reading (1975)
10. Fogg, B.J., Eckles, D.: Mobile Persuasion, 20 Perspectives on the Future of Behavior Change. Stanford Captology Media (2007)
11. Halm, E.A., Mora, P., Leventhal, H.: No symptoms, no asthma*. Chest 129(3), 573–580 (2006)
12. Hanauer, D.A., Wentzell, K., Laffel, N., Laffel, L.M.: Computerized automated reminder diabetes system (cards): E-mail and sms cell phone text messaging reminders to support diabetes management. Diabetes Techn. and Therapeutics 11(2) (2009)

13. Henkemans, O.A.B., van der Boog, P.J.M., Lindenberg, J., van der Mast, C.A.P.G., Neerincx, M.A., Zwetsloot-Schonk, B.J.H.M.: Technology and Health Care 17, 253–267 (2009)
14. Henkemans, O.A.B., Rogers, W.A., Fisk, A.D., Neerincx, M.A., Lindenberg, J., van der Mast, C.A.P.G.: Ussability of an adaptive computer assistant that improves self-care and health literacy of older adults. Methods of Inf. in Medicine 47, 82–88 (2007)
15. Inc., F.: Lose it! (July 2011), http://www.loseit.com/
16. Janz, N., Becker, M.: The health belief model: A decade later. Health Education Querterly, 1–47 (1984)
17. Kreuter, M., Farrell, D., Olevitch, L., Brennan, L.: Tailoring Health Messages: Customizing Communication With Computer Technology. Lawrence Erlbaum Associates, Inc. (2000)
18. Maes, S., Karoly, P.: Self-regulation assessment and intervention in physical health and illness: A review. Applied Psychology 54(2), 267–299 (2005)
19. Marlatt, G., Gordon, J.: Determinants of relapse: Implications for the maintenance of behavior change, pp. 410–452 (1980)
20. Michie, S., Johnston, M., Francis, J., Hardeman, W., Eccles, M.: Applied Psychology 57(4) (2008)
21. Miller, W., Rollnick, S.: Motivational Interviewing: Preparing People to Change Addictive Behaviour. Guilford Press, New York (1991)
22. Nerenz, H.L.D., Steele, D.: Ilness representation and coping with health threats. In: Baum, A., Singer, J. (eds.) A Handbook of Psychology and Health, pp. 219–252. Erlbaum Associates, Hillsdale (1984)
23. Philips: Directlife (July 2011), http://www.directlife.philips.com/
24. Preuveneers, D., Berbers, Y.: Mobile phones assisting with health self-care: a diabetes case study. In: MobileHCI (2008)
25. Prochaska, J., DiClemente, C.: The Transtheoretical Approach: Crossing the Traditional Boundaries of change. J. Irwin, Homewood (1984)
26. Prochaska, J., Norcross, J.: Stages of change.psychotherapy: theory. Research and Practice 38(4), 443 (2001)
27. Rogers, W.A., Mynatt, E.D.: How can technology contribute to the quality of life of older adults? In: Mitchell, M.E. (ed.) The Technology of Humanity: Can Technology Contribute to the Quality of Life?, pp. 22–30. Illinois Institute of Technology, Chicago, IL (2003)
28. Shaw, J., Sicree, R., Zimmet, P.: Global estimates of the prevalence of diabetes for 2010 and 2030. Diabetes Research and Clinical Practice 87(1), 4–14 (2010)
29. Skinner, C., Campbell, M., Rimer, B., et al.: How effective is tailored print communication? Annual Behavioral Medicine 21, 290–298 (1999)
30. SkyHealth: Glucosebuddy (July 2011), http://www.glucosebuddy.com/
31. Smith, E., Mackie, D.: Social Psychology. Psychology Press, Philadelphia (2000)
32. Sorbi, M.J., Mak, S.B., Houtveen, J.H., Kleiboer, A.M., van Doornen, L.J.: Mobile web-based monitoring and coaching: Feasibility in chronic migraine. Journal of Medical Internet Research 9(5) (2007)
33. de Vries, H., Brug, J.: Computer-tailored interventions motivating people to adopt health promoting behaviours: introduction to a new approach. Patient Educational Counseling 36, 99–105 (1999)
34. de Vries, H., Mudde, A.: Predicting stage transitions for smoking cessation applying the attitude social influence efficacy model. Psychology and Health 13, 369–385 (1998)

A Research Framework for Playful Persuasion Based on Psychological Needs and Bodily Interaction

Marco Rozendaal[1], Arnold Vermeeren[1], Tilde Bekker[2], and Huib de Ridder[1]

[1] Faculty of Industrial Design Engineering, Delft University of Technology
[2] Department of Industrial Design, Eindhoven University of Technology,
The Netherlands
{m.c.rozendaal,a.p.o.s.vermeeren,h.deridder}@tudelft.nl,
m.m.bekker@tue.nl

Abstract. This paper presents a research framework that relates interactive systems to behavioral change with psychological needs and bodily interaction as intermediating variables. The framework is being developed in a multidisciplinary research project that focuses on how to design intelligent play environments that promote physical and social activities. Here, the framework serves to generate design relevant research questions and to guide communication amongst group members.

Keywords: persuasive technology, play, user experience, bodily interaction, interactive systems, design, ambient intelligence, research through design.

1 Introduction

This paper presents a research framework that relates interactive systems to behavioral change with psychological needs and bodily interaction as intermediating variables. Due to the potential of technology to help solve pressing societal problems, the design community is giving increased attention to the design of systems for behavioral and societal change. We believe that to design persuasive systems within the ambient intelligence paradigm (e.g., interactive systems encompassing products, services and environments), we need to address the full spectrum of human psychological needs as well as the rich bodily interactions people engage in while trying to fulfill them.

Our research framework is being developed in a research program entitled 'Intelligent Play Environments' (IPE). The IPE program deals with the design of playful interactive systems that stimulate physical and social activities. Such systems comprise intelligent software agents embodied in physical play objects, which can sense and react to the human players. The systems should stimulate 'open-ended' play, a form of improvisational play that emerges by providing local interaction opportunities [1]. Thus, the design challenge in the IPE project lies in designing for open-ended play while at the same guiding players towards predefined behavioral patterns.

Interactive systems, such as the ones envisioned in IPE, are made possible by novel media technologies, wireless broadband communication and embedded intelligence; also referred to as "ambient intelligence" [2], "internet of things" [3] and "ubiquitous

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 116–123, 2011.

computing" [4]. In a sense many of our on-line activities (including work, play and communication) are already realized through interactive systems, as they can be carried out on a variety of platforms concurrently, such as on smartphones, dedicated game-systems and in-build car systems. Designing interactive systems is a complex activity, aligning hardware and software components with individual, situational and societal demands.

Interactive systems designed to stimulate behavioral change are called persuasive systems [5] based on the term *persuasive technology* coined by Fogg [6]. In the field of persuasive technology several strategies are presented that can change people's behavior by taking into account human computer interaction principles and human motivation. Behavioral change can take on many forms, such as changing a person's attitude, motivation or actually influencing a person's behavioral repertoire [7]. Another view on persuasive systems has its roots in the philosophy of technology. Due to the phenomenon of *technological mediation*, new technologies either allow for or restrict certain types of behavior [8]. For example, the technology of the microscope allows us to look into a visual micro-world while at the same time visually disconnecting us from our immediate environment.

A broad perspective on psychological needs and bodily interaction is needed to design persuasive systems within the ambient intelligence paradigm. In such a perspective, people fully engage (emotional, social, sensorial, etc.) with intelligent environments and systems of products rather than in a visual-cognitive manner only, which is often associated with traditional screen-based information systems. People share universal needs that drive behavior; relating to feelings of pleasure, intrinsic motivation and wellbeing [9]. Further, only those aspects of interactive systems that affect our sensorium, our bodily interface to make sense of the environment [10], are essential when relating system features to human needs.

Our framework is being developed in a multidisciplinary research project focusing on how to design intelligent play environments that promote physical and social play. The framework serves to generate design relevant research questions and to guide communication amongst group members. This paper is set up as follows: First, each level of the framework is discussed with respect to the relevant literature and its value for the framework as a whole. Second, the issue of how to operationalize the framework in research will be discussed.

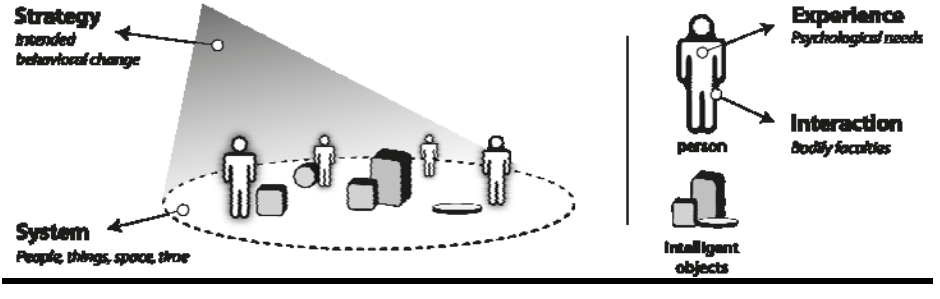
2 A Four-Leveled Framework

Our research framework relates interactive systems to behavioral change with psychological needs and bodily interaction as intermediating variables. The first level describes the behavioral change the designers intend to achieve; the second level describes the experienced psychological needs that can drive the intended behavior; the third level describes the bodily interactions that fulfill these psychological needs, and lastly, the fourth level describes features of interactive systems that afford the bodily interactions. These levels bear resemblance to the levels proposed by Ward et al. [11] aimed at connecting product attributes to human values in four intermediate steps to establish powerful 'brands' (e.g., functional attributes, functional benefits, emotional

benefits and human values). The levels of the framework are represented in Table 1 and described in more detail below.

Table 1. The four levels of the research framework ordered top-down (from Transformation to System level). The two examples show how similar transformations can be realized applying different ways of addressing psychological needs (experience level), based on different bodily interactions (interaction level) and different systems (system level). Below it, the research framework is visualized.

Level name	Focus	Aspects	Example1	Example2
Transformation	"What is the intended behavioral change?"	Attitudes, motivations, behavior, etc.	Seduce people to jump up and down for a specific period.	Seduce people to jump up and down for a specific period.
Experience	"Which psychological needs support behavioral change?"	Autonomy, stimulation, connectedness, progression, etc.	Need for self-expression	Need for discovery
Interaction	"How do bodily interactions fulfill psychological needs?"	Thinking, feeling, sensing, doing, etc.	Touching the floor with one's feet elicits musical sounds.	Jumping upwards allows one to see just one piece of a puzzle at a time.
System	"How does the interactive system afford bodily interaction?"	People, things, space, time, context, etc.	Dancing on a musical staircase.	Peaking trough a heightened show box.



To further illustrate the four levels of the framework we use the example of a car cabin. When viewing the cabin of a car as the interactive system, the system comprises chairs, a steering wheel, dashboard, air-conditioning and possibly other people. One can imagine how our bodies are involved. The chair is pressing against our backs; the steering wheel can be grasped and manipulated by our hands; the dashboard can be seen and heard; our skins feel the air-conditioning while our minds give meaning to this cabin based on previous encounters.

This bodily involvement can be experienced subjectively. The chair feels soft and supportive, the dashboard looks colourful and clean, rotating the steering wheel feels responsive and smooth, and hitting the pedal while hearing the feedback of the engine results in a feeling of power. In this example, the chair supports the need for comfort while the steering wheel supports the need for competence. Together, these different experiences combine into a unified whole affecting our behaviour. Depending on whether the need for comfort and competence is more dominant in the overall experience, the car cabin can either promote a relaxed or a sporty driving style respectively.

2.1 Transformation

The transformation-level deals with the designers' intended behavioral change. Thus, for the IPE project this entails stimulating physical and social play. Specifying a behavioral target will guide the construction of the interactive system, shaping the design activity as described by the 'design with intent' approach [12]. Setting a behavioral target involves specifying the desired behavior, the context in which it takes place and the timeframe in which the behavior is sustained [13]. For example, one can imagine that one aims to increase social interaction - captured by the amount of conversation - to occur during a play activity lasting for about ten minutes but also surpassing the play activity itself, sustained for several months in one's everyday live.

The value of the transformation-level of the framework lies in guiding interactive system design as well as in providing a means for assessment, monitoring and system adaptation. Assessments may have different forms, ranging from behavioral observations (using sensor technologies) or by interviewing people about their own attitudes and behaviors. For the IPE project this can entail monitoring physical play by measuring physical movement through low-resolution cameras for instance. Further, if behavioral targets can be translated into decisional algorithms, it would become possible to embed them in the interactive system, thus creating intelligent systems that can respond to users adequately based on sensor data. In the IPE project we intend to empirically explore the feasibility and potential of this approach.

2.2 Experience

The experience-level deals with the psychological needs that intrinsically motivate people and foster their development. Some of these well-described needs are the need for autonomy, competence, social relatedness, health and hedonic stimulation [9]. In a recent study, the first three needs of this list were found to be the most satisfying

ones [14]. Further, Laschke and Hassenzahl advocate a *gamification* approach that makes novel behaviors intrinsically rewarding by connecting them to psychological needs rather than by providing extrinsic rewards. [15]. See Korhonen et al [16] for an extensive list of playful experiences that intrinsically motivate play. For example, the playful experiences of *expression* may relate to the need for autonomy while the experiences of *competition* and *fellowship* may both relate to the need for social connectedness.

The value of the experience-level for the framework is to create behavioral change through intrinsic motivation and generate design requirements at the same time. Given the predefined behavioral target, people can be motivated differently to attain it. With respect to the IPE, one player might be stimulated into physical activity because others do it as well (need for social connectedness) while another player might feel motivated because of the individual challenge that lies in the activity (need for personal growth). Different requirements are needed when designing for each psychological need. For example, designing for social connectedness requires interconnecting each player and allowing communication and interaction between them, while designing for individual challenge entails providing feedback on individual performance over time.

2.3 Interaction

The interaction-level describes the manner in which bodily interactions are able to fulfill psychological needs. Our bodies can be viewed as our interface with the environment through our senses, feelings, thoughts and movements [10]. Bodily faculties and psychological needs are deeply intertwined: Humans are endowed with a hedonic system in the brain supporting human functioning [17, 18]. This hedonic system connects many bodily areas to cognitive processing, allowing us to experience (dis)pleasure in many different ways and directing our behavior to optimize wellbeing. A previously conducted qualitative study found that experiences related to psychological needs, involved multiple bodily faculties with a prominence of two or three specific ones [19].

The value of the interaction-level for the framework lies within guiding the design of the interactive system based upon our 'bodily interface'. This opens up the design space to allow for full-body environmental interaction. Different modes of bodily interaction uniquely shape a design. Norman distinguishes between visceral, behavioral and reflective design [20]. With IPE, the type of bodily interaction pursued in a design should fit with the psychological needs addressed to motivate the players. For example, play that is tuned to need for challenge may be more cognition-based (reflective) while play that is tuned to the need for fantasy may be more sensory-based (visceral).

2.4 System

The system-level describes the components of interactive systems. System design is a new focus in the design community influenced by the merging of products, services

and environments. These systems consist of many 'nodes' with non-collocated inputs and outputs that are socially and culturally situated hereby making such systems inherently complex and unpredictable [21]. With respect to IPE, the system contains tangible and intangible play objects, the players whom can be either virtually of physically present, the spatial configuration of players and objects, and the rules and procedures that may evolve over time.

The value of the system-level for the framework lies within the ability to focus on the components of interactive systems that have human significance, allowing designers to shape the interactive system without the need to first specify a technological paradigm. Due to the inherent complexity of these interactive systems, designing them requires an experiential approach and assessment of these systems call for qualitative and ethnographic research methods [22]. Thus, the IPE project will follow a research through design approach that will generate experiential knowledge at each design iteration, informing the framework and guiding successive iterations.

3 Discussion

Although we are too early in the process to have evaluated the framework empirically, we can highlight how we envision the four levels to be operationalized in research. For example, we plan to investigate the relationships between the four levels, (a: transformation-experience) 'which psychological needs are most influential in stimulating physical and social play?'; (b: experience-interaction) 'how can bodily interactions fulfill these psychological needs?' and (c: interaction-system) 'how can we construct interactive systems that afford the appropriate bodily interactions?' The insights gained can be assessed in relation to using the framework as an evaluation tool, in which the framework is used bottom-up (flowing from system to transformation) or top-down, as a design-generation tool (flowing from transformation to system).

It is further of importance to acknowledge in the framework differences between individuals (such as gender and age) and differences over time. Given differences in strength and levels of endurance it would be unrealistic to expect identical behavioral patterns for younger and older players. Further, for some players, the need for vitality may be dominant to feel engaged while for other players this might be the need for competence, requiring different types of bodily interaction (that may well be afforded by the same interactive system). Also, the dominant psychological need that creates engagement for an individual player may change over time since people develop skills and knowledge while playing. Additionally, depending on a user's mood and short term energy level (e.g., physical fatigue or mental fatigue after having concentrated for a long time) dominance of psychological needs may vary, affecting the type of interactions people would be willing to engage in.

The research can inform both the fields of persuasive system design and of user experience (UX), in which UX is defined as "the experience(s) derived from encountering systems" where encountering involves actual usage but also passive confrontations [23]. New insights in persuasive systems can be gained when exploring the

power of psychological needs to affect behavior in playful applications. Knowledge on UX can be strengthened as well. For example, how are psychological needs experienced emotionally and how do different bodily interactions (as afforded in a design) fulfill them?

4 Conclusion

In this paper, we have proposed a research framework that relates interactive systems to behavioral change with psychological needs and bodily interaction as intermediating variables. When used in an iterative design process, the framework guides the successive design iterations and is tested empirically at the same time. We will investigate the research framework as a tool for design generation (guiding technology development), design evaluation (translated into decisional algorithms) and multidisciplinary communication. As one of the tools for these investigations we are currently testing a preliminary self-report software tool for assessing and analyzing the results. The tool is based on elements of this framework. We envision it to be used in combination with other methods and tools, including camera observations of behavior.

References

1. Bekker, T., Sturm, J., Eggen, B.: Designing playful interactions for social interaction and physical play. Personal and Ubiquitous Computing 14(5), 385–396 (2010)
2. Weber, W., Rabaey, J.M., Aarts, E.H.L.: Ambient intelligence. Springer, Heidelberg (2005)
3. Gershenfeld, N., Krikorian, R., Cohen, D.: The Internet of things. Scientific American 291(4), 76–81 (2004)
4. Weiser, M.: The computer for the 21st century. Scientific American (1991)
5. Oinas-Kukkonen, H., Harjumaa, M.: A Systematic Framework for Designing and Evaluating Persuasive Systems. In: Oinas-Kukkonen, H., Hasle, P., Harjumaa, M., Segerståhl, K., Øhrstrøm, P. (eds.) PERSUASIVE 2008. LNCS, vol. 5033, pp. 164–176. Springer, Heidelberg (2008)
6. Fogg, B.J.: Persuasive technology: Using computers to change what we think and do. Morgan Kaufmann, New York (2010)
7. Oinas-Kukkonen, H.: Behavior Change Support Systems: A Research Model and Agenda. In: Ploug, T., Hasle, P., Oinas-Kukkonen, H. (eds.) PERSUASIVE 2010. LNCS, vol. 6137, pp. 4–14. Springer, Heidelberg (2010)
8. Verbeek, P.P.: What things do: Philosophical reflections on technology, agency, and design. Pennsylvania State University Press (2005)
9. Alkire, S.: Dimensions of Human Development. World Development 30(2), 181–205 (2002)
10. Howes, D. (ed.): Empire of the senses: The sensual culture reader. Berg Publishers, Oxford (2004)
11. Ward, S., Light, L., Goldstine, J.: What high-tech managers need to know about brands. Harvard Business Review, 85–95 (1999)

12. Lockton, D., Harrison, D., Stanton, N.: Design with Intent: Persuasive Technology in a Wider Context. In: Oinas-Kukkonen, H., Hasle, P., Harjumaa, M., Segerståhl, K., Øhrstrøm, P. (eds.) PERSUASIVE 2008. LNCS, vol. 5033, pp. 274–278. Springer, Heidelberg (2008)
13. Fogg, B.J.: The behavior grid: 35 ways behavior can change. In: Persuasive 2009, p. 42. ACM Press (2009)
14. Sheldon, K.M., Kasser, T., Elliot, A.J., Kim, Y.: What is satisfying about satisfying events? Testing 10 candidate psychological needs. Journal of Personality and Social Psychology 80(2), 325–339 (2001)
15. Laschke, M., Hassenzahl, M.: Mayor or patron? The difference between a badge and a meaningful story. In: CHI 2011 (Extended Abstracts): Conference on Computer Human Interaction. ACM Press, Vancouver (2011)
16. Korhonen, H., Montola, M., Arrasvuori, J.: Understanding playful experiences through digital games. In: Designing Pleasurable Products and Interfaces, Compiegne, France, pp. 274–285 (2009)
17. Johnston, V.S.: The origin and function of pleasure. Cognition & Emotion 17(2), 167–179 (2003)
18. Berridge, K.C.: Pleasures of the brain. Brain and Cognition 52(1), 106–128 (2003)
19. Rozendaal, M.C., Schifferstein, H.N.J.: Pleasantness in bodily experience: A phenomenological inquiry. International Journal of Design 4(2), 55–63 (2010)
20. Norman, D.A.: Emotional design: Why we love (or hate) everyday things. Basic Books, New York (2004)
21. Frens, J.W., Overbeeke, C.J.: Setting the stage for the design of highly interactive systems. In: Proceedings of International Association of Societies of Design Research, Seoul, Korea, pp. 1–10 (2009)
22. Forlizzi, J.: The product ecology: Understanding social product use and supporting design culture. International Journal of Design 2(1), 11–20 (2008)
23. Roto, V., Law, E., Vermeeren, A.P.O.S., Hoonhout, J. (eds.): User experience white paper. Bringing clarity to the concept of user experience. Result from Dagstuhl Seminar on Demarcating User Experience, September 15-18 (2010), http://www.allaboutux.org/uxwhitepaper

Automatic Modeling of Dominance Effects Using Granger Causality

Kyriaki Kalimeri[1], Bruno Lepri[2,3], Taemie Kim[3], Fabio Pianesi[2], and Alex Sandy Pentland[3]

[1] CIMeC, Corso Bettini 30, 38068, Rovereto, Italy
[2] FBK, via Sommarive 18, Povo, Trento, Italy
[3] MIT Media Lab, 20 Ames Street, 02-139 Cambridge, MA, USA

Abstract. We propose the use of Granger Causality to model the effects that dominant people induce on the other participants' behavioral patterns during small group interactions. We test the proposed approach on a dataset of brainstorming and problem solving tasks collected using the sociometric badges' accelerometers. The expectation that more dominant people have generalized higher influence is not borne out; however some more nuanced patterns emerge. In the first place, more dominant people tend to behave differently according to the nature of the task: during brainstorming they engage in complex relations where they simultaneously play the role of influencer and of influencee, whereas during problem solving they tend to be influenced by less dominant people. Moreover, dominant people adopt a complementarity stance, increasing or decreasing their body activity in an opposite manner to their influencers. On the other hand, less dominant people react (almost) as frequently with mimicry as with complementary. Finally, we can also see that the overall level of influence in a group can be associated with the group's performance, in particular for problem solving task.

Keywords: Dominance, Small Group Interactions, Granger Causality.

1 Introduction

Management, scientific research, politics and many other activities are accomplished by groups. For this reason, it is increasingly becoming important to understand the dynamics of group interactions and how they relate to group performance. Dominant behavior is a key determinant in the formation of a group's social structure, and consequently, in group dynamics [10]. Many social psychology studies have shown that individuals higher in trait dominance tend to attain more influence in face-to-face interactions [1,10,19]. Moreover, a meta-analysis of 85 years of research found that dominance can predict who emerges as the leader in group interactions more consistently than other individual characteristics, including intelligence [12]. For this reason, in the last years dominance aroused much interest in the automatic behavior analysis community. In particular, different researchers have dealt with the automatic detection of the most

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 124–133, 2011.

dominant person and/or of the least dominant person in small group interactions (e.g. meetings) using different non-verbal acoustic and visual cues [14,7,8,9].

However, to our knowledge there are not works that automatically model the causal effects that people displaying dominant non-verbal behaviors have on the non-verbal behaviors of the other participants. In order to investigate these effects and how they relate to group performance, we apply Granger causality, an approach that detects and estimates the direction of causal influence in time series analysis. To exemplify the approach, in this work we focus on people's body activity and on how it affects the body activity of other group members. In doing so, we investigate the kind of effects dominance display produces on the body activity of the influencees.

Previous studies in social psychology [20] have shown that observers can respond to dominant behaviors with mimicry or complementarity behaviors, where the former amounts to a reproduction of the behavior of the dominant person and the latter to an opposite behavior. Hence, people may respond to dominance displays by exhibiting similar (dominant) behavior and/or respond to submissive behaviors with submissive ones (mimicry). On the other hand, they could also match dominant and submissive behaviors with contrasting behaviors, with dominant displays inviting submissive responses and submissive displays soliciting dominant behaviors (complementarity). According to Chartrand and Bargh [3], mere correlational analysis are not enough to conclude that person X_1 is mimicking (or complementing) person X_2; rather, they can only inform whether X_1 and X_2 are displaying similar or contrasting behavioral patterns at the same time. Those associations, however, could be due to third, confounding, factors that are ultimately responsible for the observed behavioral patterns (e.g. a hot room causing all present to fan their face). In order to conclude for the presence of true mimicry/complementarity, a causal relationship must be proven in which Person X_1 first engages in the particular behavior and then Person X_2 mimics (or complement) that behavior. Granger causality [6] is a promising approach to this end: widely used in neuroscience to infer the existence of causal relationships among neural circuits [18], it has originated in econometrics [6] to detect and model causal relationships among temporal series. To our knowledge, it has been seldom, if ever, applied to the automatic analysis of human behavior [13] and to social behavior, in particular.

2 Twenty-Question Game Dataset

The dataset consists of 13 groups with four participants. Each participant wore a sociometric badge - a wearable electronic badge with multiple sensors collecting interaction data. By interacting with other badges, it can collect proximity data, other badges in direct line of sight, body movement data by means of accelerometers, and speech features. In this paper, we use only the accelerometer data and more specifically the variation of body movement energy, obtained by computing the amplitude of the movement vector in the *3*-dimensional space (x,y,z). An example of the participants wearing sociometric badges can be found in Figure 1.

Fig. 1. Meeting participants wearing the sociometric badges

The data encompass two co-located meeting types, brainstorming and problem solving. The task used is based on a modification of the game "Twenty-Questions", which integrates both brainstorming and problem-solving scenarios by closely replicating Wilson's experiments [21]. At the beginning of a task, each group was given a set of ten yes/no question-and-answer pairs. For the first phase of each task, groups were given eight minutes to collaboratively brainstorm as many ideas that satisfy the set of question-and-answers. Then, continuing into the second phase, groups were given 10 minutes to ask the remaining 10 questions of the Twenty-Question Game to determine the correct solution.

2.1 Dominance

In the post-task questionnaire, one of the questions asked users to rate the self-perceived level of dominance. The subjects answered using a 5-point Likert scale. Following [11], the participants with values higher than one standard deviation over the mean were considered dominant. We also asked all the participants to rate each other's dominance level and as for self-perceived dominance the participants with values higher than one standard deviation over the mean were labelled as dominant.

2.2 Performance

The performance scoring is determined by (i) the number of correct ideas in the brainstorming phase and (ii) the number of questions used to arrive at the correct answer in the problem-solving phase. As the goal of brainstorming is to generate as many ideas as possible. We use the total number of ideas generated as a measure for the performance of the brain- storming phase. In the problem solving phase, groups were asked up to 10 questions to find the correct solution. They received a higher score if they used fewer questions and a zero score if they could not get the answer correct within 10 questions. Hence, we use the number of questions each team used as a negated measure of the team's performance.

3 Our Approach

To understand the direction of the influence flow in social interactions, it is of fundamental importance to distinguish the driver from the recipient. One of the most prominent methods to estimate the direction of the causal influence in time series analysis is the Granger Causality(GC)[6]. This method is based on asymmetric prediction accuracies of one time series on the future of another. In specific, let two time series X_1 and X_2,

$$X_1(t) = \Sigma_{j=1}^{p} A_{11,j} X_1(t-j) + \Sigma_{j=1}^{p} A_{21,j} X_2(t-j) + \xi_1(t)$$

$$X_2(t) = \Sigma_{j=1}^{p} A_{21,j} X_1(t-j) + \Sigma_{j=1}^{p} A_{22,j} X_2(t-j) + \xi_2(t)$$

where A is the matrix containing the coefficients of the model and ξ_1,ξ_2 are the residuals of X_1 and X_2 respectively. A time series X_1, is said to Granger-cause X_2 if the inclusion of past observations of X_1 reduces the prediction error of X_2 in a linear regression model of X_2 and X_1, as compared to a model including only the previous observations of X_2. An important aspect of GC is its generalizability to the multivariate case in which the GC of X_1 on X_2 is tested in the context of multiple additional variables (in our scenario the other two meeting participants W and Z). In this case, X_1 is said to Granger-cause X_2 if knowing X_1 reduces the variance in X_2's prediction error when all the other variables are also included in the model [5]. In our case, the time seriesX_1, X_2, X_3, X_4 of the system $\mathbf{X}$ are reffering to the body movement of each of our subjects as described above. To remove every linear trend from the data, all series have been detrended and their temporal mean has been removed as an initial preprocessing step. We estimate the best order of the multivariate autoregressive model (MVAR) using the Bayesian Information Criterion (BIC)[15]. The estimated model was further checked both (i) to control whether it accounted for a sufficient amount of variance in the data and (ii) using the Durbin-Watson [4] test to validate whether its residuals are serially uncorrelated. Then, once the set of significant lagged values for X_2 is found, the regression is augmented with lagged levels of X_1. Having estimated the G-causality magnitudes, their statistical significance was evaluated via an F-test on the null hypothesis that the coefficients $A_{i,j}$ are zero. If the coefficients in the corresponding $A_{i,j}$ were jointly significantly different from zero, then the causal interaction was considered to be statistically significant. To correct the tests from multiple comparisons, the Bonferroni correction [2] approach was chosen thresholded at $\frac{P}{n(n-1)}$, with P=0.01.

Let our small group of participants be a small causal network of four interacting nodes. In causal networks, nodes represent variables and the directed edges represent causal interactions.A measure of the causal interactivity of a system $\boldsymbol{X}$ is the causal desity [16], which is defined as the mean of all pairwise G-causalities between system elements, conditioned on the system's statistically significant interactions.

$$cd(X) \equiv \frac{1}{n(n-1)} \Sigma_{i \neq j} F_{X_i \to X_j | X_{[ij]}}$$

where $X_{[ij]}$ is the network from which the variables X_i and X_j are omitted. For each of our nodes (i.e. each subject), we estimate the unit causal density $cd_u(i)$ which is the summed causal interactions involving a node i normalized by the number of nodes. In particular, we computed two versions (i) one weighted by the GC magnitudes, *weighted unit causal density* (WUCD) and (ii) the unweighted *unit causal density* (UCD) obtained by setting all the significant causal interactions to 1. The nodes with high values of UCD or WUCD can be considered to be the causal hubs inside the meeting. Furthermore, to identify nodes with distinctive causal effects on the network dynamics, we estimated the *causal flow* of a subject X_1, both weighted by the Granger magnitudes (WFLOW) and unweighted (FLOW). The causal flow is defined as the difference between the in-degree and the out-degree of a given node. Therefore, a subject with a high positive causal flow exerts a strong causal influence on the meeting and it can be called causal *source*. On the other side, a subject with a highly negative causal flow can be called a causal *sink*. From the GC relationships in the causal network, we are only able to determine if the body activity of subject X_1 has a causal effect on the body activity of subject X_2; however, we are not able to discriminate between mimicry and complementarity effects.

In order to assess these phenomena, we investigated the correlation between the time series of the subjects for which we found some significant causal effect. For example, once determined that subject X_1 Granger-causes X_2, we checked if the correlation among the time series X_1 and the time series X_2 is positive, revealing mimicry effects, or is negative, showing complementarity ones.

4 Experimental Results

First of all, we focus our attention on the relationships between influence,behavior and the dominance scores. Our expectation is that more dominant people have generalized higher influence, measured in terms of higher UCD and/or WUCD; positive and higher flow; higher out-flow. As a first step we compute the Spearman rank-correlation between a number of GC-related quantities (UCD, WUCD, FLOW, WFLOW, Out and In) and the dominance scores both those obtained on the basis of self-assessment (DomSelf) and those provided by the other members of the group (DomOther). In both cases, ranks are computed on a group-by-group basis. The results are reported in the Table 1. As emerges from them, the rank correlations are uniformly low and non-significant, with the exception of the correlation between the self-dominance rank and the Flow rank and of the correlation between the self-dominance rank and the WFLOW rank. In both cases, they are negative, so that more dominant people tend to have lower values of both simple and weighted causal flow in the problem solving condition. None of these results seems to support our expectations. In order to deepen our analysis, we classified our subject along two dimensions, the first (BrainS) addresses their behavior in the brainstorming session and the second (ProS) does the same for the problem solving session. The two dimensions consist each of four classes:

Table 1. Rank Correlations between GC quantities and dominance scores

	BrainS	BrainS	ProS	ProS
Rank	DomSelf Rank	DomOtherRank	DomSelf Rank	DomOtherRank
UCD	0.66	0.085	- 0.123	-0.197
WUCD	0.057	0.096	-0.119	-0.96
FLOW	-0.035	-0.077	-0.320	-0.153
WFLOW	0.014	-0.049	-0.256	-0.103
Out	0.100	0.058	-0.188	-0.138
In	0.100	0.155	0.081	0.054

- Class "0": the subject was neither a source nor a target of influence
- Class "1": the subject acted only as a source of influence
- Class "2": the subject acted only as a targets of influence from other subjects
- Class "3": the subjected acted both as a source and a target of influence.

The distribution of subjects according to the two classification schema is as in Figure 2.

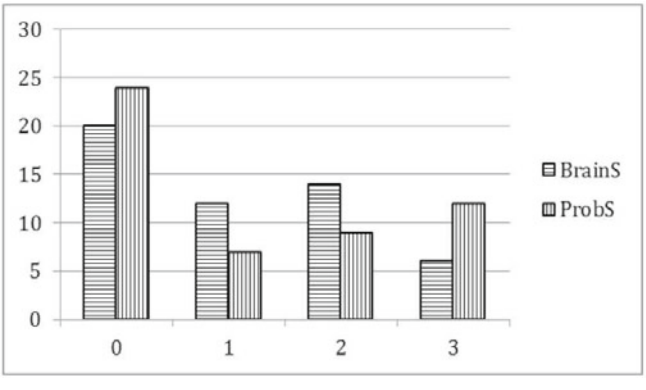


Fig. 2. The distribution of subjects according to the two classification schema

As observed, both in the brainstorming and in the problem solving conditions a large part of our subjects never took part in any influence exchange. On the other hand, the number of those who acted both as targets and as sources doubles in the ProS conditions, with a marked decrease of those playing just one of those two roles. Moreover, the number of those influence exchanges involving the same two people as both influencers and influencees increases from just one pair in the BrainS condition to four couples in the ProS one; in other words,

Table 2. Average and Median Dominance Ranks for each Behavioral Class

	DomSelf Rank	DomOtherRank	DomSelf Rank	DomOtherRank
BrainS	Average	Median	Average	Median
Class0	2.52	2.5	2.53	2.5
Class1	2.21	2.00	2.08	1.75
Class2	2.36	2.00	2.46	2.75
Class3	**3.33**	**3.50**	**3.33**	**3.5**
ProS	Average	Median	Average	Median
Class0	2.52	2.00	2.58	2.50
Class1	2.07	2.00	2.29	2.50
Class2	**3.00**	**3.00**	**2.83**	**3.00**
Class3	2.33	2.25	2.21	2.00

of the 12 people falling in class "3" in the ProS condition, 8 are part of influencer/influencees dyads. In summary, once the subjects who do not take part in any influence interaction are set apart, it seems that, in problem-solving people are more willing to get involved in complex influence interactions, whereas they stick more to a mono-directional format while brainstorming. The average and median dominance ranks for each behavior clas are reported in the Table 2.

The average and median ranks for the two dominance assessment (self vs. other) are substantially consistent. Interestingly, a trend emerges for higher dominance rankings to fall in class "3" of the BrainS classification and in class "2" of the ProS one. In other words, the subjects who act both as influencers and influences while brainstorming tend to be higher in dominance, while, in turn, the most dominant subjects seem to play the role of influencees in the problem solving condition. This latter fact explains, at least in part, the significant negative correlations between the two measures of causal flow and the dominance ranking in the ProS condition. The Figure 3 reports the correlation between the performance scores in the BrainS and ProS conditions and two measures of overall (body activity based) group-internal amount of influence. The first, AV-UCD, is the average value of the so-called Unit Causal Densities (UCD), which measures the causal density of a given person in terms of the number of incoming and outcoming influences he/she plays a role in. The second quantity, Av-WUCD, is the average of the Weighted Causal Densities (WUCD), which weights the casual density of a given person in terms of the GC values attached to the single influences he/she participates in.

Whereas the correlation values are either very close to zero for the BrainS condition, they have higher negative values in ProS. Recalling that the highest performance in the ProS condition correspond to a score of zero (zero question asked to reach the conclusion), it can be concluded that an increase of body-motion related influence during problem solving corresponds to a moderate increase in performance. No interesting trend emerges in the brainstorming condition.

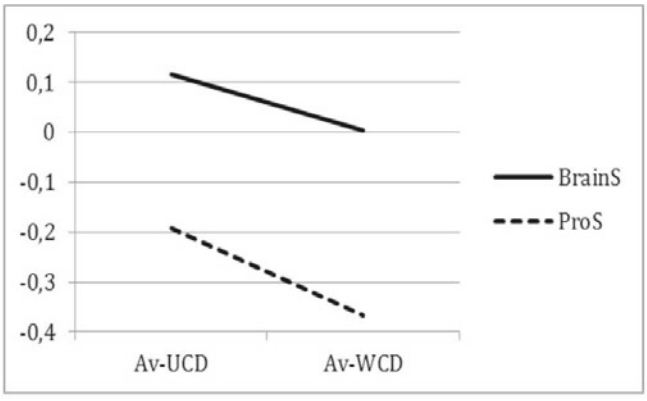


Fig. 3. Correlation between the performance scores in the BrainS and ProS conditions and the group causal densities

We conclude the analysis of the data concerning the relationships between dominance and influence by considering what happens when influence relations involve subjects of unequal dominance status - that is the relationships where one of the subject scores highest in the group and the other is lower. There are two fundamental modes described in the literature[20] in which influence can unfold: the influencee follows the behavior triggered by the influencer (mimicry) or he/she can exhibit the opposite behavior (complementarity). In our case, we can capture these differences by computing the correlation between the two corresponding signals: if there is an influence relationship (that is, the GC is significant) and the correlation is positive, then we speak in terms of mimicry, otherwise, we consider it as a case of complementarity. Given the exploratory nature of this paper, we have retained all the correlation coefficients corresponding to any significant influence relationship.

The results are promising: out of 10 cases in which a dominant person affects a non-dominant one, six were cases with mimicry (positive correlation) and four cases of complementarity; when the influence relationship was reversed and a non-dominant person affected a dominant one, only in three cases out of 14 there was mimicry.

5 Conclusion

The expectation that more dominant people have generalized higher influence (higher UCD and/or WUCD; positive and higher flow; higher out-flow) is not borne out; rather more nuanced patterns emerge. In the first place, more dominant people tend to behave differently according to the nature of the task: during brainstorming they engage in complex relations where they simultaneously play the role of influencer and of influencee, whereas during problem solving they tend to be influenced by less dominant people. However, while doing so, they

adopt a complementarity stance, increasing or decreasing their body activity in an opposite manner to their influencers. On the other hand, when less dominant people are the target of influence from more dominant ones, they react (almost) as frequently with mimicry as with complementary.

We have also seen that there are signs that the overall level of influence in a group can be associated with the group's performance, and that this seems to be the case in problem solving condition; an interesting question could be whether this is related in any way with the fact that dominant people play more often the role of influence targets in this condition, this way possibly making it possible to more focused effort to be deployed.

Before concluding, we emphasize the exploratory nature of this study and the fact that, with a few exceptions (GC values and correlation values in Table 1) none of our suggestions is supported by statistical evidence, because of the limited size of the used sample. Still, we believe that we have shown the power of the notion of Granger causality and the flexibility it allows for in the investigation of complex social phenomena.

Acknowledgements. Bruno Lepri's research was funded by the Marie Curie "COFUND" 7th Framework PERSI project.

References

1. Aries, E.J., Gold, C., Weigel, R.H.: Dispositional and situational influences on dominance behavior in small groups. Journal of Personality and Social Psychology 44, 779–786 (1983)
2. Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289–300 (1995)
3. Chartrand, T.L., Bargh, J.A.: The chameleon effect: the perception-behavior link and social interaction. Journal of Personality and Social Psychology 76, 893–910 (1999)
4. Durbin, J., Watson, G.: Testing for serial correlation in least squares regression.I. Biometrika 37, 409–428 (1950)
5. Geweke, J.: Measurement of Linear Dependence and Feedback Between Multiple Time Series. Journal of the American Statistical Association 77, 304–313 (1982)
6. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424–438 (1969)
7. Hung, H., Jayagopi, D., Yeo, C., Friedland, G., Ba, S., Odobez, J.-M., Ramchandran, K., Mirghafori, N., Gatica-Perez, D.: Using audio and video features to classify the most dominant person in a group meeting. In: Proceedings of the 15th International Conference on Multimedia, pp. 835–838. ACM Press (2007)
8. Hung, H., Jayagopi, D.B., Ba, S., Odobez, J.-M., Gatica-Perez, D.: Investigating automatic dominance estimation in groups from visual attention and speaking activity. In: Proceedings of the 10th International Conference on Multimodal Interfaces, IMCI 2008, pp. 233–236. ACM (2008)
9. Jayagopi, D.B., Hung, H., Yeo, C., Gatica-Perez, D.: Modeling Dominance in Group Conversations Using Nonverbal Activity Cues. IEEE Transactions on Audio, Speech and Language Processing 17, 501–513 (2009)

10. Judge, T.A., Bono, J.E., Ilies, R., Gerhardt, M.W.: Personality and leadership: a qualitative and quantitative review. Journal of Applied Psychology 87, 765–780 (2002)
11. Kim, T., Chang, A., Holland, L., Pentland, A.: Meeting mediator: enhancing group collaborationusing sociometric feedback. In: CSCW, pp. 457–466 (2008)
12. Lord, R.G., De Vader, C.L., Alliger, G.M.: A meta-analysis of the relation between personality traits and leadership perceptions: An application of validity generalization procedures. Journal of Applied Psychology 71, 402–410 (1986)
13. Prabhakar, K., Oh, S.M., Wang, P., Abowd, G.D., Rehg, J.M.: Temporal causality for the analysis of visual events. In: CVPR, pp. 1967–1974 (2010)
14. Rienks, R., Zhang, D., Gatica-Perez, D., Post, W.: Detection and application of influence rankings in small group meetings. In: ICMI, pp. 257–264 (2006)
15. Schwarz, G.: Estimating the Dimension of a Model. The Annals of Statistics 6, 461–464 (1978)
16. Seth, A.K.: A MATLAB toolbox for Granger causal connectivity analysis. Journal of Neuroscience Methods 186, 262–273 (2010)
17. Seth, A.K.: Causal networks in simulated neural systems. Cognitive Neurodynamics 2, 49–64 (2008)
18. Seth, A.K., Edelman, G.M.: Distinguishing causal interactions in neural populations. Neural Computation 19, 910–933 (2007)
19. Smith, J.A., Foti, R.J.: A pattern approach to the study of leader emergence. The Leadership Quarterly 9, 147–160 (1998)
20. Tiedens, L.Z., Fragale, A.R.: Power moves: complementarity in dominant and submissive nonverbal behavior (2003)
21. Wilson, D.S., Timmel, J., Miller, R.R.: Cognitive cooperation: when the going gets tough, think as a group. Human Nature 15, 225–250 (2004)

Abnormal Crowd Behavior Detection by Social Force Optimization

R. Raghavendra[1], Alessio Del Bue[1], Marco Cristani[1,2], and Vittorio Murino[1,2]

[1] Istituto Italiano di Tecnologia (IIT), Italy
[2] Dipartimento di Informatica, University of Verona, Italy

Abstract. We propose a new scheme for detecting and localizing the abnormal crowd behavior in video sequences. The proposed method starts from the assumption that the interaction force, as estimated by the Social Force Model (SFM), is a significant feature to analyze crowd behavior. We step forward this hypothesis by optimizing this force using Particle Swarm Optimization (PSO) to perform the advection of a particle population spread randomly over the image frames. The population of particles is drifted towards the areas of the main image motion, driven by the PSO fitness function aimed at minimizing the interaction force, so as to model the most diffused, normal, behavior of the crowd. In this way, anomalies can be detected by checking if some particles (forces) do not fit the estimated distribution, and this is done by a RANSAC-like method followed by a segmentation algorithm to finely localize the abnormal areas. A large set of experiments are carried out on public available datasets, and results show the consistent higher performances of the proposed method as compared to other state-of-the-art algorithms, proving the goodness of the proposed approach.

1 Introduction

Crowd behavior analysis is currently gaining more and more attention in many applicative disciplines, and recently in the surveillance context [7] as well. Nowadays, crowds are viewed as the very outliers of the social sciences [14]; such an attitude is reflected by the remarkable paucity of psychological research on crowd processes [14].

During recent years, anomaly detection for crowd dynamics are gaining more popularity. As discussed earlier, the available schemes can be divided into two types, namely, model based schemes and particle advection based schemes. In [15], anomalies are modeled as a distribution over low-level visual features which are then learned in a unsupervised way using hierarchical latent Dirichlet allocation. In [9], the optical flow is computed after dividing the whole sequence into number of cuboids; then, the activity patterns are generated with a mixture of probabilistic PCA models which is then inferred using Markov Random field to localize the anomaly in the crowd. In [1], a histogram is used to measure the

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 134–145, 2011.

probability of optical flow in local patterns of the image and then, ambiguity based threshold is selected to monitor and detect the anomalies in video. In [12], a dynamic texture model is employed to jointly model the appearance and dynamics of the crowded scene. This method explicitly addresses the detection of both temporal and spatial anomalies. In [10], spatio-temporal gradients whose statistics are modeled with a coupled HMM to detect abnormalities in densely crowded scene is proposed.

In case of particle advection schemes [13] [2], a rectangular grid of particles are placed on each frame and advected using the underlying motion. Here, each particle is assumed as an individual in the crowd and this assumption is reasonable as it overcome the limitations of tracking people in high density crowds. In [2], high density crowd analysis is carried out based on coherent structures from fluid dynamics and particle advection. This approach is further enhanced in [16] by introducing a chaotic invariant to describe the event. In [13], a crowd behavior analysis is carried out by analyzing the interaction force estimated using Social Force Model (SFM) and rectangular particle advection scheme.

In this paper, we propose a new scheme for abnormal crowd behavior detection and accurate localization from video sequences. we start from random positioned particles over each frame, still assuming that the interaction force flow is discriminant for characterizing crowd behavior. However, after the flow force estimation, we apply a minimization process aimed at optimizing the position of the particles at each frame. In this way, particles converge naturally towards the significant moving areas in the scene, and in particular towards the parts which likely show a high interaction force. After that, a RANSAC-like methods [4] and a simple segmentation algorithm (Mean-Shift [5]) are used to localize and identify the anomaly in each frame.

The rest of the paper is organized as follows. Section 2 provides the basic notions of the PSO and describes the SFM. Section 3 discusses the proposed method for local abnormality detection, and Section4 reports the results on a set of public video datasets, also compared with other methods in the literature. Finally, Section 5 draws the conclusions.

2 Fundamentals of the Computational Methods

2.1 Particle Swarm Optimization (PSO)

Particle Swarm Optimization is a stochastic, iterative, population-based optimization technique aimed at finding a solution to an optimization problem in a search space [8]. The main objective of PSO is to optimize a given criterion function called the fitness function f. The PSO algorithm is initialized with a population, namely a *swarm*, of N-dimensional particles distributed randomly over the search space (of dimension N too): each particle is so considered as a point in this N-dimensional space and the optimization process manages to move the particles according to the evaluation of the fitness function in an iterative way. More specifically, at each iteration, each particle is updated according to the best values called $pbest_i$. Such value is depending on the $i-th$ particle

and *gbest* that is independent from the specific particle i.e. *gbest* is valid for the whole swarm. The $pbest_i$ value represents the position associated with the best (i.e., minimum or maximum) fitness value of particle i obtained at each iteration. The *gbest* value represents the best position among all the particles in the swarm, i.e., the position of the particle assuming the minimum or maximum value when evaluated by the fitness function. The rate of the position change (velocity) for the particle i is called v_i, which is updated according to the following equations [8]:

$$v_i^{new} = W \cdot v_i^{old} + C_1 \cdot rand_1 \cdot (pbest_i - x_i^{old}) + C_2 \cdot rand_2 \cdot (gbest - x_i^{old}) \quad (1)$$

$$x_i^{new} = x_i^{old} + v_i^{new}, \quad (2)$$

Where, W is the inertia weight, whose value should be tuned to provide a good balance between global and local explorations. The scalars C_1 and C_2 are acceleration parameters used to drive each particle towards $pbest_i$ and *gbest*. $rand_1$ and $rand_2$ are random numbers between 0 and 1. Finally, x_i^{old} and x_i^{new} are the current and updated particle positions, respectively, and the same applies for the deviation v_i^{old} (v_i^{new}).

2.2 Social Force Model

The SFM [6] provides a mathematical formalization to describe the movement of each individual in a crowd on the basis of its interaction with the environment and other obstacles. The SFM can be written as:

$$m_i \frac{dW_i}{dt} = m_i \left(\frac{W_i^p - W_i}{\tau_i} \right) + F_{int}, \quad (3)$$

where m_i denotes the mass of the individual, W_i indicates its actual velocity which varies given the presence of obstacles in the scene, τ_i is a relaxing parameter, F_{int} indicates the interaction force experienced by the individual which is defined as the sum of attraction and repulsive forces, and W_i^p is the desired velocity of the individual. The SFM was successfully employed in different research fields like computer simulation of crowds, transportation/evacuation to analyze the pedestrians motion as a whole. The union of PSO and SFM in a single optimization framework constitute the basic elements used to develop our model for crowd behavior analysis which will be described in the following section.

3 Our Approach

As described in the previous section, the SFM computes the social force of the individual given its current velocity, desired velocity and interaction force. In a video surveillance scenario, we need to compute the SFM given the image motion of each pedestrian roaming in the controlled area, as also proposed in [13].

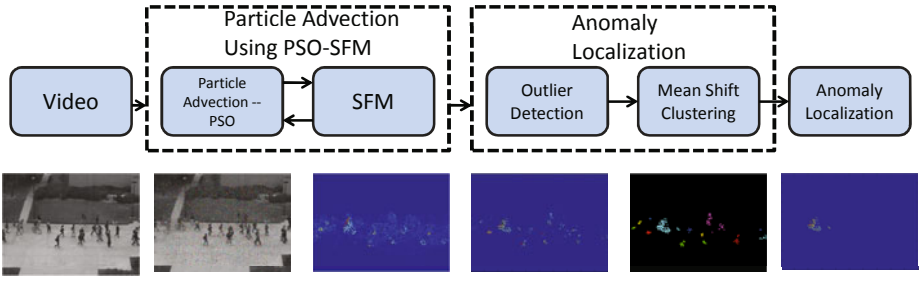


Fig. 1. Block diagram of the proposed scheme for anomaly detection and localization

The best cue for computing such interaction force is given by the visual motion of each component in the crowd. However, having a reliable tracking system in a crowded environment is not a realistic assumption because of the high degree of occlusions. An accurate tracking algorithm in crowded scenes is in fact challenging because of the following reasons: (1) overlapping between individual subjects; (2) random variations in the density of people over time; (3) low resolution videos with temporal variations of the scene background.

Thus, instead of analyzing the motion of each pedestrian, we adopt the intensity of the optical flow calculated over the video frames as the driving visual cue. The image flow is then used to calculate the interaction forces associated to every particle by means of the SFM, which are then evaluated by the fitness function which drives the displacement of the particles in the PSO optimization process. In earlier attempts [13,2], the particle advection was carried out by placing a rectangular grid of particles over the video frames. Then, the velocity for each particle is calculated using fourth-order Runge-Kutta-Fehlberg algorithm [11] along with the bilinear interpolation of the optical flow field.

In general, a drawback of this approach is that it assumes that a crowd follows a fluid-dynamical model which is too restrictive when modeling masses of people. The elements of the crowd may also move with unpredictable trajectories that will result in an unstructured flow. Moreover, the use of a rectangular grid for particles is a coarse approximation with respect to the continuous evolution of the social force. To overcome these drawbacks, we propose a novel particle advection using PSO to first localize the image areas of highest magnitude of social force. The output of the PSO is then further processed to localize abnormalities in the video sequence using an outliers detector and segmentation algorithms.

Figure 1 summarizes the proposed scheme for accurate localization of the anomaly in a crowd. First, given a video sequence, the PSO begins with a random initialization of the particles in the first frame. From such initial stage, we obtain a first guess of $pbest_i$, for each particle i, and the global $gbest$. The particles are defined by their 2-D value corresponding to the pixel coordinates in the frames. At each iteration, the $pbest_i$ value is updated only if the present position of the particle is better than the previous position according to the fitness function

evaluated on the model interaction force. Finally, the *gbest* is updated with the best position obtained from the $pbest_i$ after reaching the maximum number of iterations or if the desired fitness value is achieved. We then use the final particle positions as the initial guess in the next frame and the same iterative process is repeated until the end of the video sequence. Therefore, in the proposed approach, the movement of the particles are updated according to the fitness function which drives the particle towards the areas of minimum interaction force.

3.1 Computing the Fitness Function Using SFM and PSO

The fitness function aims at capturing the best interaction force exhibited by each motion in the crowded scene. Each particle is evaluated according to its interaction force calculated using SFM and optical flow (OF) [3]. The optical flow is actually a suitable candidate to substitute the pedestrian velocities in the SFM model.

In order to use OF in the SFM, we first define the intensity of the optical flow at a given position in the image for the particle i as:

$$W_i = O_{avg}(x_i^{new}), \tag{4}$$

where $O_{avg}(x_i^{new})$ indicates the average OF at the particle coordinates x_i^{new}. The average is computed over L previous frames. Then, the desired velocity of the particle W_i^p is defined as:

$$W_i^p = O(x_i^{new}), \tag{5}$$

where $O(x_i^{new})$ represents the OF intensity of the particle i, whose coordinates are estimated using equation (1). In fact, this OF value is an average value computed in a small spatial neighborhood to avoid numerical instabilities of the OF. Finally, we calculate the interaction force F_{int} using equation (3) as follows:

$$F_{int}(x_i^{new}) = m_i \cdot \frac{dW_i}{dt} - \frac{m_i}{\tau_i}\left(W_i^p - W_i\right), \tag{6}$$

where the velocity derivative is approximated as the difference of the OF at the current frame t and $t-1$, i.e., $\frac{dW_i}{dt} = [O(x_i^{new})|_t - O(x_i^{new})|_{t-1}]$. As observed from equation (3), the interaction force allows an individual to change its movement from the desired path to the actual one. This process is in some way mimicked by the particles which are driven by the OF towards the image areas of larger motion. In this way, the more regular the pedestrians' motion, the less the interaction force, since the people motion flow varies smoothly. So, in a normal crowded scenario the interaction force is expected to stabilize at a certain (low) value complying with the typical motion flow of the mass of people. It is then reasonable to define a fitness function aimed at minimizing the interaction force, and moving particles towards these sinks of small interaction force, thereby allowing particles to simulate a "normal" situation of the crowd.

Hence, we can write our fitness function as:

$$FitX = \min_{i=1,\ldots,K} \{F_{int}\,(x_i^{new})\} \tag{7}$$

where, x_i denotes the $i-th$ particle and K denotes the total number of particles. In our experiments, we used $K = 15,000$ particles with 100 iterations, and these values are selected experimentally and kept constant throughout all the experiments. In the figures, we map onto the image plane the magnitude of the interaction forces assigned to every particle. Figure 2(a)-(b) shows the input

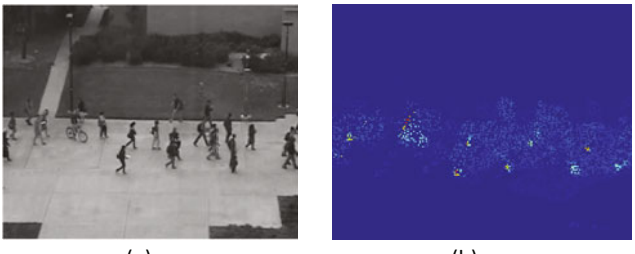


Fig. 2. (a) Input frame. (b) Interaction force

frame and the corresponding interaction force, respectively. It is interesting to observe that, higher magnitudes of the force are present in the region that moves differently from the overall image flow. Although patterns of high magnitude of the interaction force over a certain period of time can provide useful information about the presence of an anomaly, not necessarily large magnitudes of the force is a direct consequence of the presence of an anomaly. This is due to the fact that particles are not associated to a whole person, but only to person's parts, so, for instance, legs motion can lead to high interaction force which is obviously not an anomaly.

3.2 Anomaly Detection

An anomaly in a scene may be detected by finding high magnitudes of interaction forces over a certain time period. In order to detect structured interaction forces over time, we use an outliers detection scheme to eliminate isolated fluctuations of the social force at each time instant. These "outliers" effects are in general due to the approximation of the pedestrians velocities with a dense OF computation. For instance, as above observed, we noted that the leg swinging of a walking pedestrian is a cause for false positive (anomaly) detections. This occurs because the local optical flow in this small areas is noisy and may cause some disturbances in the anomaly detection.

This process is performed using a custom implementation of the well-known RANdom SAmple Consensus (RANSAC) algorithm [4]. RANSAC is an iterative method used to estimate the parameters of a mathematical model from observed data containing outliers. This algorithm basically assumes that data consists of inliers whose distribution can be explained by a known parametric model. From experimental data, we have observed that the statistics of the interaction forces associated to a crowd situation in the video datasets we considered can be reasonably well approximated by a Gaussian distribution.

So, given the interaction force magnitude of the particles at each frame we perform the following steps:

1. Randomly select 5000 particles (out of 15000 particles) and their corresponding interaction force magnitude.
2. Estimate the Gaussian distribution on the selected interaction force magnitude. Let the estimated mean and standard deviation be $\hat{\mu}$ and $\hat{\sigma}$.
3. Consider the remaining particles and evaluate those that are inliers and outliers. Inliers are detected by checking if the particle's force is within the typical $3\hat{\sigma}$ of the estimated model, particles whose force is outside this interval are considered outliers.
4. Repeat the steps $1-3$ for R number of iterations, $R = 1000$ iterations in our case.
5. Finally, choose the Gaussian model with the highest number of inliers.

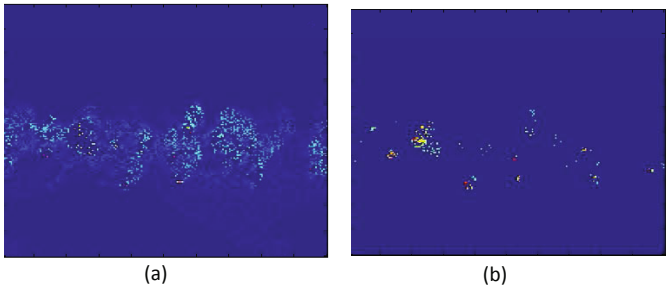


Fig. 3. Results of RANSAC-like algorithm. (a) shows the inlier particles and their corresponding social force magnitude. (b) shows the selected outlier particles.

Figure 3 shows the inliers and outliers that are obtained using the RANSAC-like algorithm. It is interesting to observe that all high magnitude interaction forces are detected as outliers. In order to achieve a better localization, we perform the spatial clustering of the detected outliers using mean-shift clustering [5] as it works independently on the assumptions regarding the shape of the distribution and the number of modes/clusters (see Figure 4). In the end, we finally select the interaction force corresponding to the clusters with a number of members larger than a certain threshold, implicitly assuming that clusters having a small number of particles do not reach a significant size, and so are discarded. This

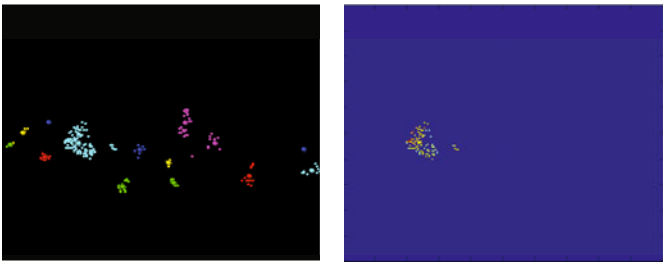


Fig. 4. Results of mean-shift clustering. (a) shows the detected clusters. (b) shows the final anomaly localization of the largest cluster with the corresponding particles force magnitude.)

threshold is fixed and kept constant in all the performed experiments. Further, assuming that the geometry of the scene is roughly known, this threshold can be set to define the minimal (abnormal) event to be detected.

4 Experimental Results and Comparisons

To evaluate the performances of the proposed method compared to previous approaches, we consider two standard datasets used for abnormal activities detection: UCSD [12] and MALL [1] datasets.

UCSD dataset: The UCSD dataset contains two different sets of surveillance videos called PED1 and PED2. The dataset has a reasonable density of people and anomalies including bikes, skaters, motor vehicles crossing the scenes. The PED1 has 34 training and 36 testing image sequence and PED2 has 16 training and 12 test image sequences. These video sequences have two evaluation protocol as presented in [12], namely: (1) frame-level anomaly detection, and (2) pixel-level anomaly detection. At frame-level, we verify if the current frame contains a labeled abnormal pixel. In such case, the frame is considered containing an abnormal event and compared with the annotated ground truth status (either normal or abnormal). At pixel-level, the detection of abnormality is compared against the ground truth on a subset of 10 test sequence. If at least 40% of the detected abnormal pixels match the ground truth pixels, it is presumed that anomaly has been localized otherwise it is treated as a false positive.

Figure 5 shows the ROC curve of our method for the frame-level anomaly detection criteria for PED1 and PED2 datasets. We then compare the performance against the state of the art approaches such as the SFM based method [13], MPPCA [9], Adam et al. [1] and Mixture of dynamic textures (MDT) [12]. Table 1 shows the quantitative results of the proposed method on frame-level anomaly detection on PED1 and PED2 datasets, and Table 2 shows the results on anomaly localization. Notice that the Equal Error Rate (EER) in Table 1

and 2 is defined as the point where false positive rate is equal to false negative rate. Remarkably, the proposed method outperforms all the previous approaches on both frame-level and pixel-level detection, reaching the best performances in the frame-level anomaly detection on the PED2 dataset.

Table 1. Equal Error Rates for frame level anomaly detection on PED1 and PED2 datasets

Approach	SF[13]	MPPCA[9]	Adam [1] et al.	MDT[12]	Proposed Method
PED1	31%	40%	38%	25%	21%
PED2	42%	30%	42%	25%	14%
Average	37%	35%	40%	25%	17%

Table 2. Anomaly localization: Rate of detection at the EER

Method	SF[13]	MPPCA[9]	Adam [1] et al.	MDT[12]	Proposed Method
Localization	21%	18%	24%	45%	52%

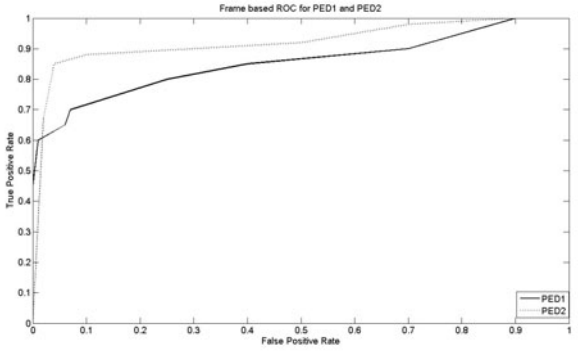


Fig. 5. ROC curves obtained on UCSD PED1 and PED2 Datasets

Figure 6 shows an example of image frames with anomaly detection and localization for PED1 and PED2 datasets. It can be observed that the proposed method is capable of detecting anomalies even from the far end of the scene (see Figure 6(a), last two frames).

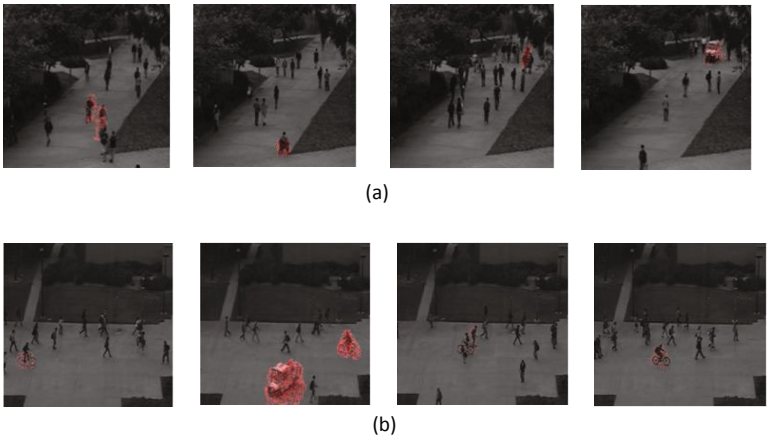


Fig. 6. Examples of anomaly frame detection and localization on PED1 (a) and PED2 (b) datasets (best viewed in color)

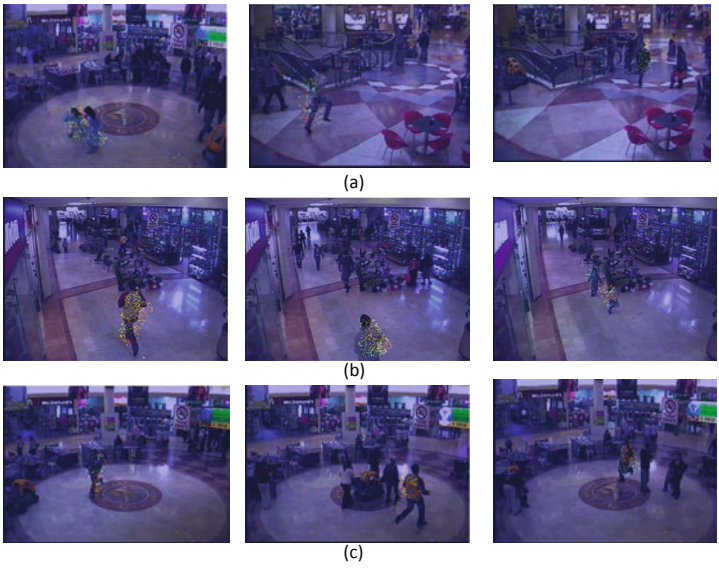


Fig. 7. Examples of anomaly detection on Mall dataset. (a) Mall Camera 1. (b) Mall Camera 2. (c) Mall Camera 3 (best viewed in color).

Mall dataset: The Mall dataset [1] consists of a set of video sequences recorded using three cameras placed in different locations of a shopping mall during working days. The annotated anomalies in such dataset are individuals running erratically in the scene. The evaluation protocol uses only the frame-level anomaly detection criteria. Figure 7 shows an example of frames from this dataset in

Table 3. Performances on the Mall dataset

Dataset	Methods	RD	FA
Mall Cam 1	Adam et al. [1]	95%(19/20)	1
	Proposed Method	100%(20/20)	2
Mall Cam 2	Adam et al. [1]	100%(17/17)	6
	Proposed Method	100%(17/17)	4
Mall Cam 3	Adam et al. [1]	95%(20/21)	4
	Proposed Method	100%(21/21)	3

which the anomaly is detected using the proposed method. Table 3 shows that the proposed method is extremely accurate in detecting all the frames with anomaly. Moreover, our approach outperforms the state-of-the-art schemes as for the best Rate of Detection (RD) and fewer False Alarms (FA).

5 Conclusions

We proposed a new algorithm for detection and localization of anomalies present in crowded videos by employing the SFM interaction force in combination with the Particle Swarm Optimization. The main contribution of this work lies in introducing the optimization of the social force and performing particle advection to obtain the optimized interaction force according to the underlying optical flow field. The main advantage of the proposed scheme is that the whole anomaly detection/localization process is carried out without any learning phase. This implies that the proposed method is very well suited for real scenarios. Further, the extensive experiments conducted on UCSD dataset show the goodness of the approach, whose results outperforms those obtained by all state-of-the-art algorithms.

References

1. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(3), 555–560 (2008)
2. Ali, S., Shah, M.: A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–6 (2007)
3. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High Accuracy Optical Flow Estimation Based on a Theory for Warping. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3024, pp. 25–36. Springer, Heidelberg (2004)
4. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24(1), 381–395 (1981)
5. Fukunaga, K., Hostetler, L.: The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory 21(1), 32–40 (1975)

6. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Physical Review E 51(4), 42–82 (1995)
7. Junior, J.C.S.J., Musse, S.R., Jung, C.R.: Crowd analysis using computer vision techniques: A survey. IEEE Signal Processing Magazine 27(5), 66–77 (2010)
8. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: IEEE International Conference on Neural Networks, Perth, Australia, pp. 1942–1948 (1995)
9. Kim, J., Grauman, K.: Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2009)
10. Kratz, L., Nishino, K.: Tracking with local spatio-temporal motion patterns in extremely crowded scenes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 693–700 (2010)
11. Lekien, F., Marsden, J.: Tricubic interpolation in three dimensions. Journal of Numerical Methods and Engineering 63(3), 455–471 (2005)
12. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1975–1981 (2010)
13. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 935–942 (2009)
14. Reicher, S.: The Psychology of Crowd Dynamics, pp. 182–208. Blackwell, Oxford (2001)
15. Wang, X., Ma, X., Grimson, W.: Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(3), 539–555 (2009)
16. Wu, S., Shah, M.: Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2010)

Understanding the Influence of Social Interactions on Individual's Behavior Pattern in a Work Environment

Chih-Wei Chen[1], Asier Aztiria[2],
Somaya Ben Allouch[3], and Hamid Aghajan[1]

[1] Stanford University, Stanford, CA 94305, USA
{louistw,aghajan}@stanford.edu
[2] Mondragon University, Mondragon, Spain
aaztiria@eps.mondragon.edu
[3] University of Twente, Enschede, The Netherlands
s.benallouch@utwente.nl

Abstract. In this work, we study social interactions in a work environment and investigate how the presence of other people changes personal behavior patterns. We design the visual processing algorithms to track multiple people in the environment and detect dyadic interactions using a discriminative classifier. The locations of the users are associated with semantic tasks based on the functions of the areas. Our learning method then deduces patterns from the trajectories of people and their interactions. We propose an algorithm to compare the patterns of a user in the presence and absence of social interactions. We evaluate our method on a video dataset collected in a real office. By detecting interactions, we gain insights in not only how often people interact, but also in how these interactions affect the usual routines of the users.

Keywords: behavioral patterns, social interaction, behavioral change.

1 Introduction

Advances in technology have seen tremendous progress in the past years, and have enabled computers to understand human behavior more thoroughly. From simple tracking, motion detection, gesture recognition to complex activity classification and frequent behavior understanding, system design is shifting to a human-centered paradigm where the awareness of users plays the central role in the development of the applications.

In this work, we attempt to understand human behavior pattern, and how the pattern changes under the influence of social interactions. In particular, the human behavior pattern we study here refers to the order, temporal duration, and conditions a person performs tasks. Understanding the behavior pattern benefits individual wellbeing and personal productivity; change toward more healthy or more efficiency habits can be made only if unhealthy or inefficient behavior patterns are detected in the first place. Computers can prompt the

A.A. Salah and B. Lepri (Eds.): HBU 2011, LNCS 7065, pp. 146–157, 2011.

users to change their behaviors based on observations and learned models, and compare how the observations deviate from previous patterns, or a desirable goal. Concrete examples include suggesting the user to take a break after working long hours continuously.

Human behavior is social and adaptive. The behavior pattern changes in the presence of other people. Understanding the behavior patterns and the influence of social interactions on them, while challenging, enables the computer to interpret human behaviors and opens up a new horizon for human behavior reasoning. Awareness of social interactions and their implications can assist a system aiming to induce behavioral change. For instance, if the system sees the user engaged in a task and, according to previous observations, the user is about to take a break, the system can then suggest other people not to disturb him until the break, especially if interruptions by other people tend to lead to work inefficiency. Similarly, change in behavior can be further motivated if it also encourages social interactions. Following our previous example, a user might be more willing to take a break from his work if the system notifies him that a group discussion is taking place.

To achieve the goal, we propose to use visual sensors, i.e. video cameras, to monitor the users. A set of visual processing algorithms is developed to extract information from the videos in real time. The locations of the users are tracked. Head poses are estimated from the video frames, and a discriminative classifier determines whether or not a pair or people is interacting. The dyadic social interaction considered in our work refers to direct interactions such greetings, eye contact, and conversations. The classifier learns from a labeled dataset, and uses relative location and head orientation as features.

The location of the user can be mapped to a semantic task based on the function of the area. For example, desk can be associated with working and dining table can be associated with eating. We use a data mining approach to learn the pattern of these location-oriented tasks. We learn patterns of a user from two sets of data collected in the same environment. One contains the user working alone, and the other one includes the user and other people interacting. We compare the two sets of patterns by considering the number of modifications required to change one to the other.

The contributions of this paper are as follows. First, we address the problem of understanding the influence of social interactions on a person's behavior pattern. Second, we present the algorithms to learn frequent behaviors, construct pattern models, and measure the differences between two patterns. In particular, a new method that defines and computes the distance between two sequences is presented. Finally, we have built the full system that monitors the users' behavior patterns using visual sensors. We evaluate our system on a challenging dataset recorded in a real office environment. Our system runs in real-time on a laptop. This implies that users can use their personal computers with a webcam to process the videos without actually saving or transmitting them. Personal recommendations can still be provided without jeopardizing privacy.

2 Related Work

The problem of constructing behavior patterns from frequent behaviors has been widely studied [5, 6]. The focus of our work is on comparing patterns of the same user in the presence of other users and social interactions, and measuring the deviations of the patterns from learned models.

Unlike the behavior pattern construction, comparing behavioral patterns has not attracted so much attention yet. Anyway, human behaviors have been analyzed in many other domains such as web navigation and activity workflow. Shifts in human behaviors have been examined in these domains [10, 2]. However, it is necessary to obtain a specific solution taking into account the special features of work environments.

Understanding social interactions has also attracted extensive research. By analyzing people's walking patterns, interactions can be detected [8]; social groups and their respective leaders can also be identified [12]. The scenario we consider is more similar to that studied in [7], where social interactions in more static work environments are detected based on the relative location and head orientation. In this paper, we acquire statistics from observed interactions to gain insights into the structure of the group, and investigate the impact of social interactions or interruptions on the productivities of workers.

3 Overview

Our proposed system consists of three main components. The *Sensor* layer tracks the location of multiple users and detects interactions among them. The second layer of the system, the *Behavior* layer, constructs behavior models for the users. Finally, the *Service* layer prompts suggestions and recommendations to the users based on the current observation and learned behavior patterns. The system overview is illustrated in Figure 1.

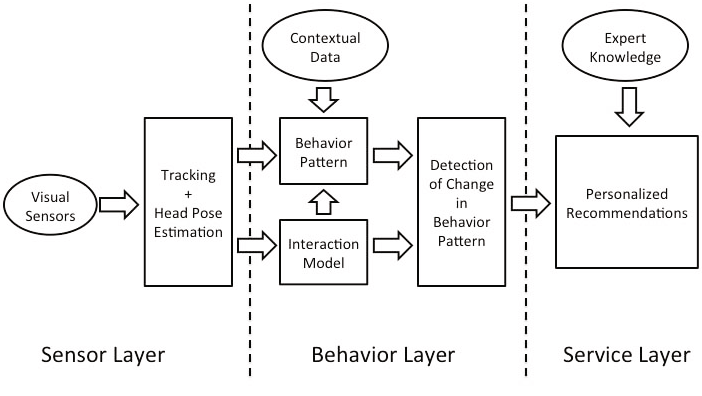


Fig. 1. System overview

4 Human Tracking and Interaction Detection

In order to construct behavior models for individual users in the environment, the system needs to know the identities of the users, track their locations, and monitor their behavior. We propose to use visual sensors, i.e. video cameras, to achieve the tasks.

4.1 Identification

To identify the users, a radio-frequency identification (RFID) system is incorporated. Each user is given a unique RFID tag. Upon entering the office, the tag is read by the RFID reader at the entrance, and the identity is associated to the person tracked by the cameras. An appearance model is initialized to track the target person throughout the day. While state-of-the-art face recognition has reaches high accuracy, the RFID system supplies reliable information and relieves the burden of the visual processing algorithms. Moreover, the RFID system provides the ground truth labels for a growing data set of human appearance that other recognition methods can be trained on and incorporated in the future.

4.2 Tracking

The system tracks the locations of multiple people in the environment. We use a tracking-by-detection approach [11], where an edge-based head detector is used. The head detector matches an Ω-shape head and shoulder silhouette against the edge map of the incoming frame. When the track is initialized with a detection, a head appearance model is also constructed. The head appearance model contains the grayscle image patch within the head detection bounding box. A new detection is the next frame is associated with the current track if the detected head appearance is close to the current model, where the distance metric is normalized correlation. The head appearance model is updated if the new detection is strong enough, i.e. the detection score is above a predefined threshold. When there is no new detection, the tracking falls back to low-level tracking, where the head appearance model finds a match that maximizes normalized correlation within a local region.

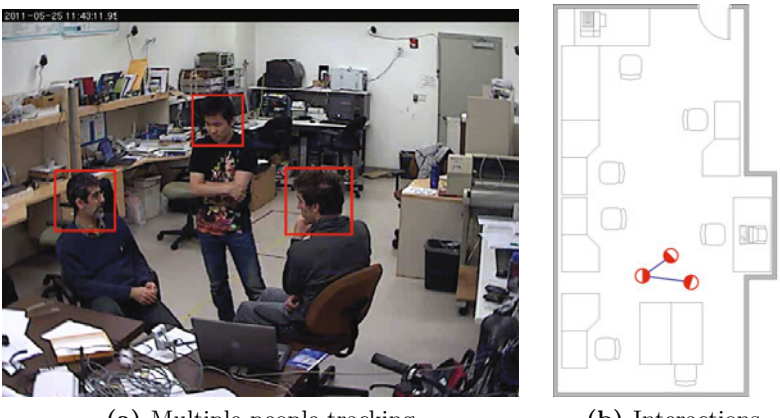
When the RFID system is triggered, an upper body appearance model is also initialized at the same time. The upper body appearance model captures the clothing of the person by building a color histogram around the upper body region, which is a rectangular box right below the head position in the video frame. The body appearance model is particularly useful when multiple people are in close proximity, and associating tracks to different people needs to be resolved. This additional cue, together with location and head appearance information, guarantees the tracks follow the same targets and not be confused. Examples of tracking results can be found in Figure 2.

4.3 Interaction Detection

A discriminative interaction classifier considers all pairs of people in the office, and estimates whether they are interacting or not. Formally, a support vector machine (SVM) classifier is trained. The features include the relative head orientation and distance between the pair of people. This is motivated by the observation that interacting people usually face each other, and tend to stay close together. A training set with ground truth label is provided for learning the classifier.

The head pose is estimated from the image patch returned by the head tracker. A manifold embedding method [7] is used to learn the mapping from the image patch to the head orientation space. Here we only consider the side-to-side orientation of the head, or the *yaw* angle. The video cameras are calibrated, so location in the video frame can be projected back to real-world coordinate. In the case of a single camera, we assume the height of the target is known to recover the location within the environment.

The interaction detector operates on single frames; for each input frame, the locations of the people are tracked by the tracker, and head orientations are estimated. The interaction detector then consider every pair of people and returns a decision for each pair. To mitigate erroneous detections, averaging is applied to the results. A running temporal window of size w is used, and an interaction detection is valid only if more than a fraction δ of frames are classified as *interacting*. In practice, we choose $w = 10$ and $\delta = 0.4$. Figure 2 shows examples of detected interactions.



(a) Multiple people tracking **(b)** Interactions

Fig. 2. (a) Example result of multiple people tracking. From the tracked head location and image patch, the real world location and head pose are estimated. **(b)** The red half-filled circles mark the locations of the tracked people, where the solid side points to the direction the person is facing. A blue line links two circles if the dyadic interaction classifier considers the two people interacting.

5 Learning Behavior Patterns

In order to understand how social interactions influence on individuals' behaviors, it is necessary to identify their behavior patterns in both cases, with and without social interactions. Behavior patterns represent users' frequent behaviors or habits in a comprehensible way. These patterns are identified using the data collected by the sensor module, so that it is totally transparent for the user.

The Learning module uses the LFPUBS [4] algorithm in order to identify behavior patterns. This algorithm is made up of four steps (see Figure 3).

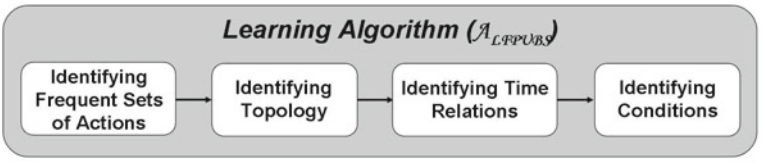


Fig. 3. Steps to be performed in order to discover frequent behaviors

1. *Identifying Frequent Sets of Actions*: The objective of this step is to discover the sets of actions that a user frequently performs together (Frequent Sets). The underlying idea of the first step is both simple and efficient. Defining a demanded minimum level (minimum confidence level), it discovers all those sets of actions that occur more times than the minimum level. For that, the Apriori algorithm [3] is used.
2. *Identifying Topology*: The step 'Identifying Frequent Sets' discovers which sets of actions frequently occur together. In order to properly model the user's behaviors defined by such sets of actions, it is necessary to define the order of such actions. For that, techniques of Workflow mining [1] have been used.
3. *Identifying Time Relations*: Topologies define a first temporal representation of the actions (qualitative representation). Qualitative relations allow one to understand the logical order of the actions. Even so, more accurate information could be provided if the relations were defined, if possible, by means of quantitative relations. The relations to study are already defined by the previous step. Thus, we applied clustering techniques in order to group data and identify quantitative time relations. This step is essential because, for example, it identifies how long a user works continuously.
4. *Identifying Conditions*: Finally, LFPUBS discovers the contextual information under what each frequent behavior occurs. By contextual information we understand either calendar information (e.g. time of day, day of week, etc.) or context information (e.g. temperature, humidity, etc.). In order to identify such conditions, classification techniques are used.

In this work, the Learning module was applied to two different datasets. On the one hand, the dataset that showed how user behaved without any social interaction ($P_{NoInteraction}$). On the other hand, the dataset that showed how

user behaved when interaction with some other people ($P_{Interaction}$). Thus, two different behavior patterns were identified, one representing user's frequent behaviours without any social interaction whereas the other one represents user's frequent behaviours with social interactions.

6 Comparing Behaviors

Once behavior patterns have been discovered, different analyses can be done. In this work, we attempt to understand how user's behavior changes under the influence of social interactions. For that, we identify the *differences* among the patterns discovered by the Learning module. By *differences* we understand:

- Insertion of an action. An action that was not frequent in $P_{NoInteraction}$ appears as frequent in $P_{Interaction}$.
- Deletion of an action. An action that was frequent in $P_{NoInteraction}$ does not appear as frequent in $P_{Interaction}$.
- Substitution of an action. An action that was frequent in $P_{NoInteraction}$ is replaced by a different action as frequent in $P_{Interaction}$.
- Swapping of two actions. The order of two action in $P_{NoInteraction}$ is reversed in $P_{Interaction}$.

The process to identify modifications is an adaptation of the Levenshtein distance [9]. Given two sequences of actions, $P_{NoInteraction}$ and $P_{Interaction}$, it calculates the set of modifications in $P_{NoInteraction}$ to get $P_{Interaction}$.

The algorithm for identifying modifications is based on the constructing of *distance matrix*. The distance matrix, $D = [d_{m,n}]_{|P_{NoInteraction}| \times |P_{Interaction}|}$, is constructed as follows:

Algorithm. Constructdistancematrix ($P_{NoInteraction}, P_{Interaction}$)

Input: $P_{NoInteraction}$ and $P_{Interaction}$ **Output:** distance matrix (D)

for $m = 0$ to $m = |P_{NoInteraction}|$
  for $n = 0$ to $n = |P_{Interaction}|$
    if $P_{NoInteraction}(m) == P_{Interaction}(n)$ then
      $d_{m,n} = d_{m-1,n-1}$ // no modification needed
    else
      $d_{m,n} = minimum($
        $d_{m-1,n} + 1$ // insertion
        $d_{m,n-1} + 1$ // deletion
        $d_{m-1,n-1} + 1$ // substitution
        $if((P_{NoInteraction}(m-1) == P_{Interaction}(n-2))\&$
        $(P_{NoInteraction}(m-2) == P_{Interaction}(n-1))$ then
          $d_{m-2,n-2} + 1$ // swap
)
return D, $d_{|P_{NoInteraction}|,|P_{Interaction}|}$

The number of modifications is given by the value of $d_{|P_{NoInteraction}|,|P_{Interaction}|}$. In addition, the construction of the distance matrix allows to identify the set of modifications. For each value, the distance matrix records what modification(s) has/have been considered (*insertion*, *deletion*, *substitution*, *swap*), so that they can be easily retrieved.

7 Experimental Results

The experiments to validate different modules are carried out in a real office environment, i.e. our own research lab. The layout, as well as the semantic task areas, is shown in Figure 4. Data were collected using an Axis network camera, capturing VGA (640 x 480 pixels) videos at 30 fps. The visual processing algorithms process the videos at 6 fps on a laptop with a 1.66 GHz Core Duo processor.

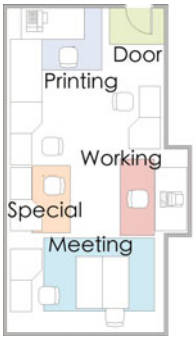
The first part of the dataset contains a user's behavior when he is by himself. During 19 days, the system recorded the morning behavior of the employee (User A), who behaved in his usual manner and performed his daily routine without interactions with other people. The second part of the data includes multiple users and various social interactions. The system recorded the behavior of the same employee during eight days, but this time he shared the office with four other officemates (User B, User C, User D and User E) and occasionally visitors. The average length of the videos per day is about 35 minutes for both parts of the dataset.

First, the *Sensor* layer tracks User A's locations and his interactions with other people. The output of this layer is essential for the discovery of user's behavior patterns, but, at the same time it allows the analysis of how he interacts with his officemates. Frequency and duration of the interactions can be summarized. Table 1 shows how User A interacts with his officemates. User A has more short interactions with User B and E. This is not surprising since they collaborate on the same projects and need to touch base very often. User C is actually User A's supervisor, and their less frequent but longer interactions are meetings. Greetings with Visitors are also recorded by our system.

Table 1. User A's interactions with his officemates

	How many times	Average Durations
User B	12	48 sec.
User C	2	261 sec.
User D	5	154 sec.
User E	13	65 sec.
Visitors	3	14 sec.

Once the user's actions were identified, we run LFPUBS to discover his patterns with and without interactions. For each case the system discovered one



(a) Office layout.

(b) Tracking results.

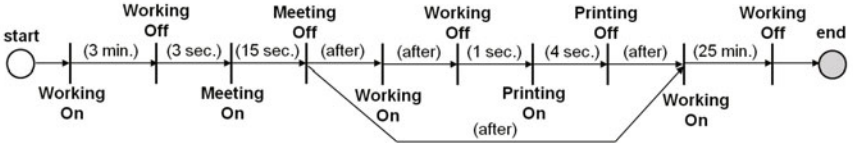
Fig. 4. (a) Office layout Users are associated with tasks based on the areas they are in. The office is divided into semantic areas: *Working, Meeting, Printing, Door* and *Special*. **(b) Tracking results** The composite of tracking results from several days of the same user in each of the different areas.

pattern. The learned patterns are shown in Figure 5. In the absence of interactions, User A starts his day by going to the working area. He then visits the meeting area momentarily. Sometimes he goes to the printer to fetch the papers he prints. He then resumes working for an average length of 25 minutes. On the other hand, when there are other people in the lab, the discovered pattern is quite different. He would work for about 20 minutes, usually interrupted by short interactions in between. He would then leave his working area. Sometimes he would interact with other people more, then return to his working area for another 14 minutes.

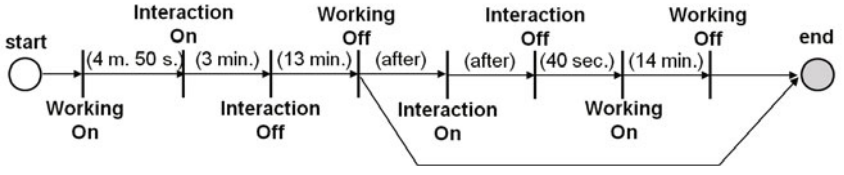
Given the two discovered patterns using LFPUBS, the next step is to compare these two patterns. Modifications are identified by constructing the distance matrix from the two patterns. The generated distance matrix is shown in Figure 6. The distance matrix not only records the distance between the sequences, but also keeps track of the types of modifications used. Theses modification records show the actions involved in the transformation from one sequence to another in the pattern and topology. In our experiment, the differences are identified to be:

- User A used to print papers regularly when he was by himself, whereas this action was less likely when he was with his officemates.
- In the presence of other people, User A interacted with his officemates. The new action substituted going to the meeting area in his pattern while he was alone.

The distance matrix shows the difference between two sequences in terms of modifications. Apart from the modifications, many analyses, quantitative or qualitative, can be performed given the two learned behavior patterns. One of the interesting observations is how User A's behavior of taking break changes depending on the interactions. In our experiment, it can be observed that without



(a) Behavior pattern without interactions ($P_{NoInteraction}$).



(b) Behavior pattern with interactions ($P_{Interaction}$).

Fig. 5. The learned behavior patterns of the same user without and with social interactions. When working alone, the user tends to work continuously without taking a break (25 minutes). In the presence of other people, the user changes his behavior, e.g. no printing anymore, but also leaves his work area more often.

	Working on	Working off	Meeting on	Meeting off	Working on	Working off	Printing on	Printing off	Working on	Working off
Working on	0	1	2	3	4	5	6	7	8	9
Interaction on	1	1	2	3	4	5	6	7	8	9
Interaction off	2	2	2	3	4	5	6	7	8	9
Working off	3	2	3	3	4	4	5	6	7	8
Interaction on	4	3	3	4	4	5	5	6	7	8
Interaction off	5	4	4	4	5	5	6	6	7	8
Working on	6	5	5	5	4	5	6	7	6	7
Working off	7	6	6	6	5	4	5	6	7	6

Fig. 6. Distance matrix generated to identify modifications. The first column shows the behavior pattern with interactions ($P_{Interaction}$), and the top row shows that without interactions ($P_{NoInteraction}$). The matrix shows the number of modifications for the subsequences of the patterns, and the last entry in the matrix is the total number of modifications for the entire sequence. The types of modifications are also recorded, so the difference between the two can be readily obtained.

any interaction, User A worked, on average, 25 minutes continuously, whereas interactions with his officemates interrupted his continuous working, breaking the time into shorter periods. Even so, it has to be pointed out that after short interactions he went back to work for another 14 minutes continuously before taking a break.

The overall length of working is similar in two scenarios for User A. It can be conjectured that the presence of his officemates encourages User A to take short breaks to interact with his officemates. This change is highly recommended by ergonomics experts, because it decreases the risk of injuries.

8 Conclusion and Future Work

We have presented a system that extracts social interactions and behavior patterns using visual sensors. The statistics of the direct interactions shed light on the social structure in an interactive, social environment. We have also demonstrated the effectiveness of our learning algorithms on data recorded in a real office environment. Two sets of patterns were learned, and our method enables the comparison between the two. Change in behavior pattern due to social interactions and interruptions by other people was identified.

The algorithms and methods we propose are general. In our setup, only one digital video camera that has a good view on the target user and the environment is used. More complex patterns can be discovered by refining the granularity of the sensors and including more activities. For instance, state-of-the-art visual processing algorithms, or sensors of other modularities, are capable of identifying the activities of the users. We plan to provide real-time service to the users in the form of recommendation or intervention, so the wellbeing of the users can be improved. Input from domain experts would have a crucial role in providing the feedbacks, such as how often one should take a break.

Our proposed method measures the topological differences between patterns. It would be useful to identify not just structural changes, but also quantitative ones. For example, one would like to understand how a worker's work cycle deviates from his previous pattern. The ability to detect these quantitative changes, including more frequent breaks and shorter work hours, is required for improving the ergonomics of the work environment, and essential for building a model for the social influence and behavior changes.

We also plan to include contextual information, such as schedules of the users, to the pattern learning algorithm to better understand the conditional relationship of the behavior pattern and change. The analysis and reasoning of social interactions will also benefit from the additional information, such as increased interactions within a team before deadlines. We are continuing our data collection. We believe more interesting results can be discovered with a richer set of data.

Acknowledgments. The authors would like to thank Amir Hossein Khalili for his contribution to the data collection. The authors also thank the reviewers for

their valuable comments and constructive suggestions. This research was partly funded by the Department of Energy under grant DE-AR0000018.

References

[1] van der Aalst, W., Weitjers, A., Maruster, L.: Workflow mining discovering process models from event logs. IEEE Transactions on Knowledge and Data Engineering 18(9), 1128–1142 (2004)
[2] Adams, M., Edmond, D., Hofstede, A.: The applications of activity theory to dynamic workflow adaptation issues. In: 7th Pacific Asia Conference on Information Systems, PACIS (2003)
[3] Agrawal, R., Srikant, R.: Mining sequential patterns. In: Proc. 11th International Conference on Data Engineering, pp. 3–14 (1995)
[4] Aztiria, A., Izaguirre, A., Basagoiti, R., Augusto, J.: Learning about preferences and common behaviours of the user in an intelligent environment. In: Behaviour Monitoring and Interpretation-BMI-Smart Environments, Ambient Intelligence and Smart Environments, pp. 289–315. IOS Press (2009)
[5] Aztiria, A., Izaguirre, A., Augusto, J.C.: Learning patterns in ambient intelligence environments: a survey. Artificial Intelligence Review 34(1), 35–51 (2010)
[6] Chen, C.W., Aztiria, A., Aghajan, H.: Learning Human Behaviour Patterns in Work Environments. In: Workshop on CVPR for Human Communicative Behavior Analysis (2011)
[7] Chen, C.W., Ugarte, R.C., Wu, C., Aghajan, H.: Discovering Social Interactions in Real Work Environments. In: IEEE Automatic Face and Gesture Recognition, Workshop on Social Behavior Analysis (2011)
[8] Habe, H., Honda, K., Kidode, M.: Human interaction analysis based on walking pattern transitions. In: ACM/IEEE International Conference on Distributed Smart Cameras. IEEE (August 2009)
[9] Levenshtein, V.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10, 707–710 (1965)
[10] Madhuri, B., Chandulal, A., Ramya, K., Phanidra, M.: Analysis of users web navigation behavior using GRPA with variable length Markov chains. International Journal of Data Mining and Knowledge Management Process 1(2), 1–20 (2011)
[11] Ren, X.: Finding people in archive films through tracking. In: CVPR, vol. 2, pp. 1–8. IEEE (June 2008)
[12] Yu, T., Lim, S.N., Patwardhan, K., Krahnstoever, N.: Monitoring, recognizing and discovering social networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1462–1469. IEEE (June 2009)

Author Index